LETTER

Action Recognition Using Visual-Neuron Feature

Ning LI[†a)], ***Student Member and*** **De XU**[†b)], ***Member***

SUMMARY This letter proposes a neurobiological approach for action recognition. In this approach, actions are represented by a visual-neuron feature (VNF) based on a quantitative model of object representation in the primate visual cortex. A supervised classification technique is then used to classify the actions. The proposed VNF is invariant to affine translation and scaling of moving objects while maintaining action specificity. Moreover, it is robust to the deformation of actors. Experiments on publicly available action datasets demonstrate the proposed approach outperforms conventional action recognition models based on computer-vision features.

key words: *neurobiological approach for action recognition (NAAR), visual-neuron template (VNT), visual-neuron feature (VNF), visual cortex*

1. Introduction

Action recognition (AR) is one of the most active research areas in computer vision due to its potential applications such as video surveillance, content based video retrieval and sports events analysis. Most of AR research has focused on studying the computer vision-based features of moving objects such as the contour, interesting points, local or global spatio-temporal volume. In these works, the affine translation, scaling and moving direction of actors can impact the performance of the systems. Humans and primates outperform the best computer vision systems for AR, so building a system that emulates AR in the primate visual cortex has always been an attractive idea.

The mechanism of AR in the primate visual cortex has progressed over the past decades [1]. AR in visual cortex is organized in two streams: a ventral stream dealing with shape information and a dorsal stream dominating motion information. Neurophysiologic experiments have shown that, in monkey and human brains, these two streams originate in the primary visual cortex (V1) and separate along two populations of cells: cells responding to spatial orientations project to extrastriate visual areas V2, V4 and inferotemporal (IT) cortex of the ventral stream; and cells responding to direction of motions project to area MT (V5) and MST in the dorsal stream. Eventually, the two streams twist and interact at higher levels [2].

Motivated by the quantitative experiments on AR in the primate visual cortex, Giese et al. [3] speculate neurons in MT and MST respond to optical flow patterns of target objects. They model the processing mechanism of these two streams separately for simplification. However, the model has only been applied to simple artificial stimuli. Jhuang et al. [4] extend the model in [3] and presume that neurons in intermediate visual areas of the dorsal stream such as MT,MST respond to spatio-temporal features of target objects. This model shows a better performance than typical computer vision-based algorithms for AR. However, its experimental results have high standard error of mean value (up to 9.9%), thus the performance of the method is not consistent.

In this letter, we propose a neurobiological approach for AR (NAAR). The approach is closely related to the feedforward template matching architecture for static object representation in the primate visual cortex [5]. The NAAR is performed in two procedures: an action described in the form of the average motion energy is firstly represented by a visual-neuron feature (VNF); the VNF is then classified to a template action category based on a supervised classification technique. The main contribution of the proposed AR approach is twofold: (1) the VNF of actions is invariant to affine translation and scaling while maintaining action specificity; (2) it is robust to the deformation of actors. Experiments on publicly available action datasets show that our approach outperforms conventional AR models based on computer-vision feature. The rest of the paper is organized as follows. Section 2 introduces the proposed NAAR. Experimental results will be analyzed in Sect. 3. Conclusive remarks are addressed at the end of this paper.

2. The Proposed NAAR

2.1 Average Motion Energy

Actions are essentially spatio-temporal variations of silhouettes which encode spatial information of postures and dynamic information of actions. To characterize an action, we represent the associated sequence of action silhouettes as the informative "average motion energy (AME)" image which implicitly captures the global motion properties of actions and has been successfully used in gait-based human identification [6]. Given a sequence of binary silhouette frames $B(x, y, t)$ containing postures, the AME is defined by Eq. (1). x and y are the coordinates of pixels in the frames, and τ is the duration of a complete action.

$$A = \frac{1}{\tau}\sum_{t=1}^{\tau} B(x, y, t) \tag{1}$$

Manuscript received July 22, 2008.
Manuscript revised October 16, 2008.

[†]The authors are with Institute of Computer Science and Engineering, Beijing Jiaotong University, Beijing China.
a) E-mail: 06112075@bjtu.edu.cn
b) E-mail: dxu@bjtu.edu.cn

DOI: 10.1587/transinf.E92.D.361

2.2 Visual-Neuron Feature for Static Objects

The extraction of visual-neuron feature (VNF) for static object is proposed by Serre et al. [5]. It comprises a 4-layer feedforward architecture (S1-C1-S2-C2). This architecture models the representation of static objects along the ventral stream in the primate visual cortex. S1 units only respond to simple bar-like stimuli. They are obtained by applying input image to a battery of Gabor filters with 8 bands (each contains 2 filter scales) and 4 orientations, see Eq. (2), and Eq. (3):

$$G(x, y) = \exp\left(-\frac{X^2 + \gamma^2 Y^2}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}X\right), s.t. \quad (2)$$

$$X = x\cos\theta + y\sin\theta, Y = -x\sin\theta + y\cos\theta. \quad (3)$$

All filter parameters are adjusted so that the S1 unit matches the response property of visual neurons in V1 area. For an input image, there are $16 * 4 = 64$ maps generated by S1 units. C1 units are more tolerant to shift and size of objects than the S1 units. In C1 units, a sliding window is applied to the input image obtained at S1 layer of all bands and orientations. In each band, C1 units select the maximal value from the two windows locating at the same position of the two S1 maps to represent the window area. Therefore, there are $8 * 4 = 32$ maps generated in this stage. S2 units are where template matching occurs. Before the matching stage, M small patches of various sizes in all the 4 orientations are extracted from random positions of C1 images of template objects. In this letter, the collection of the M patches is defined as the visual-neuron template (VNT). The S2 behaves as a Gaussian-like Radial Basis Function (RBF), where each of the patches functions as the center of the RBF. That is, for a small window within a C1 image of a particular scale, the response R of the corresponding S2 unit is given by:

$$R = \exp\left(-\|C1_{win} - P\|^2\right). \quad (4)$$

where $C1_{win}$ denotes a small window locating at every position of the C1 image in all the 4 orientations and P represents a patch extracted from template images at C1 layer. At runtime, S2 maps are computed for each of the eight scale bands. Therefore, the number of S2 maps equals to $M * 8$. Finally, at the C2 layer, the translation- and scale-invariant C2 feature is obtained by taking a global maximum across all bands and positions over the entire S2 images. Therefore, units at this layer have the largest receptive fields and respond to complex stimuli such as cars, faces and pedestrians [7]. The C2 feature is the VNF of an object. The dimension of the VNF equals to M and is independent of the size of the input image that contains static objects.

2.3 Action Recognition Using VNF

The NAAR is implemented using a 6-layer feedforward model. The key to the success of the method is that the AME images of actions are represented by simulating the process of object representation in the primate visual cortex. In the NAAR, the VNF extraction process is modified by introducing corner point information of AME images. The action classification is then completed by adding a "template action matching (TAM)" layer and a "maximal likelihood decision (MLD)" layer. Figure 1 shows the flow chart of the NAAR.

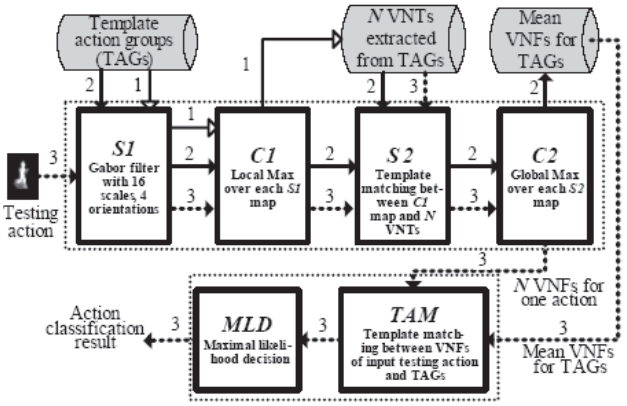

Fig. 1 The 6-layer NAAR model for action recognition. The first step indicates the process of obtaining the VNTs database. The second step indicates the extraction of mean VNFs for TAGs; the third step shows the extraction of VNF for a testing action and the classification of it.

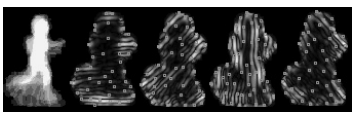

Fig. 2 Corner point images (denoted by the rectangles) for the action "run". The number of corner points is assigned by users. For the 95*120 image, about 50 corner points can be sufficient. From left to right: the AME image of "run", the corner point images of the corresponding C1 images at band 1 in the orientation $0°, 45°, 90°$ and $135°$.

In the first step, suppose there are N types of actions in a dataset, then we define N template action groups (TAGs), each consisting of m congener actions. One TAG is used to extract one VNT, thus N VNTs are obtained. The VNT extraction process is implemented along the S1 and C1 layers. We modify this process proposed in [5] by replacing the extraction of patches from random positions of template images at the C1 layer with the selective extraction of patches from corner points of these C1 images. The Harris Corner Detector is used to detect the corner points. Figure 2 shows the corner points of the C1 images for the action "run". Corner points can effectively eliminate the redundant information induced by homogeneous region, therefore the new VNT is efficient to characterize an action.

In the second step, the mean VNFs are obtained to represent TAGs. This step passes the S1, C1, S2 and C2 layers, which is the same as the process of VNF extraction for static objects proposed by [5]. Each TAG has an exclusive VNT which is used to determine the VNFs for all the template actions contained in the TAG. The mean of these VNFs is then calculated to represent their corresponding group. Conse-

quently, the "Mean VNFs for TAGs" database contains N mean VNFs, each representing one TAG.

The third step is for action classification, where a testing action is firstly represented by N VNFs through the S1, C1, S2 and C2 layer. Each VNF is determined by a VNT extracted from one TAG. And then, in the TAM layer, the similarity between the testing action and a certain TAG is calculated based on their Mahalanobis Distance, see Eq. (5).

$$M_{dist}(k) = (X(k) - Y(k))^T \sum_k^{-1} (X(k) - Y(k)), (k = 1, \ldots, N). \quad (5)$$

Where, the $X(k)$ represents the VNF of the testing action determined by the VNT extracted from the k^{th} TAG, the $Y(k)$ denotes the mean VNF of the k^{th} TAG, and the $\sum_k$ is the covariance matrix of the actions in the k^{th} TAG. The smaller the above distance measure is, the more similar the two actions are. Finally, in the MLD layer, a supervised classification algorithm is performed, where the category of the TAG that has the minimal distance to the testing action is assigned to this action, see Eq. (6).

$$A = \arg\min_k \{M_{dist}(k) \mid k = 1, \ldots, N\}. \quad (6)$$

3. Experiment Evaluation

3.1 Action Datasets

The Weizmann and Weiz.Robust action datasets are used in this letter. The Weizmann provides 90 video sequences shown by nine subjects, each performing 10 types of natural actions repeatedly. The actions are "run", "walk", "skip", "jack", "jump-forward", "jump-up-down", "gallop-sideways", "wave-one-hand", "wave-two-hands" and "bend". We split the dataset as: 6 subjects are used as templates and the remaining 3 subjects are used for testing. Thus, the size of each TAG is $m = 6$. The experiment is repeated by twenty-five random splits. The Weiz.Robust is designed for the robustness evaluation on recognition systems. It contains 10 types of walk actions: "normal walk", "moonwalk", "limp", "walk with bag, case, dog, skirt", "knees-up", "feet occluded by boxes", and "body occluded by pole". Each type of walks is performed by only one subject. For such a small dataset, the leave-one-out cross-validation rule is adopted to compute an unbiased estimate of the true recognition rate. We split the actions dataset as: 9 randomly drawn subjects are used as templates and the remaining 1 subject is used for testing. Thus, the size of each TAG is $m = 9$. One hundred random splits are repeated for the experiment. Figure 3 shows the example AME images for the datasets.

3.2 The Benchmark Approaches

For benchmarking, we use the approaches proposed by Mokhber et al. [8] and Chen et al. [9]. The [8] is designed to be invariant to the affine translation and scaling of moving objects, where the spatio-temporal volume of a silhouette sequence is represented by a 3D geometric moments. Action recognition is carried out using a nearest neighbor classifier based on Mahanalobis distance. In the [9], the time-sequential silhouettes of an action are transformed into a symbol sequence by referring to a posture codebook, and then the HMM is used to classify the symbol sequence.

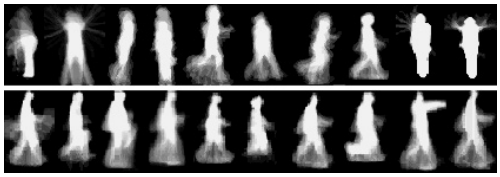

Fig. 3 First row: AME images for 10 types of actions in Weizmann. Second row: AME images for 10 types of walk actions in Weiz.Robust.

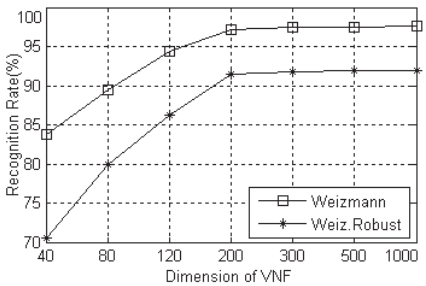

Fig. 4 Action recognition results obtained by the $NAAR_{cp}$ with VNF of variant dimensions. The y-axis represents the average recognition rate (%), and the x-axis denotes the variation of dimension.

3.3 Experimental Results

The NAAR approach with VNT extracted from random positions of C1 images is denoted as $NAAR_{rand}$, and the one with VNT extracted from corner points is denoted as $NAAR_{cp}$. The VNF of low dimension may not be sufficient to characterize target objects. On the other hand, redundant dimensions will increase the computational intensity of the NAAR. Figure 4 illustrates the recognition results obtained by the $NAAR_{cp}$ with the dimensions of VNF increasing from 40 to 1000. The average recognition rate does not increase from 200 on for all the testing datasets. Therefore, we chose the $200D$ VNF to evaluate the performance of NAAR.

The action confusion matrix in classification experiment using the $NAAR_{cp}$ tested on Weizmann dataset is shown in Table 1. The y-axis represents the ground truth, and the x-axis represents the template action groups. The numbers on the diagonal are the average correct classification rate. The numbers that are not on the diagonal are the average misclassification rate. In this table, a large confusion (16%) occurs between the action pair "skip" and "run" in the last row. In our opinion, this is because the AME image of the action "skip" is very similar to the AME image of the action "run". It is even hard for human beings to distinguish.

Table 1 Action confusion matrix obtained by the $NAAR_{cp}$ tested on Weizmann dataset. A1-A10 are action "walk", "jack", "run", "jump-forward", "wave-one-hand", "jump-up-down", "wave-two-hands", "bend", "gallop-sideways" and "skip".

	*A*1	*A*2	*A*3	*A*4	*A*5	*A*6	*A*7	*A*8	*A*9	*A*10
*A*1	**1.0**	0	0	0	0	0	0	0	0	0
*A*2	0	**1.0**	0	0	0	0	0	0	0	0
*A*3	0	0	**0.98**	0	0	0.01	0	0	0	0.01
*A*4	0	0	0	**1.0**	0	0	0	0	0	0
*A*5	0	0	0.01	0	**0.96**	0	0.02	0	0	0.01
*A*6	0	0.02	0	0	0	**0.98**	0	0	0	0
*A*7	0	0	0	0	0.02	0	**0.97**	0	0	0.01
*A*8	0	0	0	0	0	0	0	**1.0**	0	0
*A*9	0	0	0	0	0	0	0	0	**1.0**	0
*A*10	0	0	**0.16**	0.02	0	0	0	0	0	**0.82**

Table 2 Comparison between the proposed NAAR and the benchmark algorithms (denoted as [8] and [9]). The numbers are the average recognition rates (%). The standard error of the mean (s.e.m.) (%) is also indicated below the rate.

	[8]	[9]	$NAAR_{cp}$	$NAAR_{rand}$
Weizmann	95.8	94.7	**97.1**	95.3
s.e.m	±1.7	±1.9	±1.7	±3.1
Weiz.Robust	87.3	86.4	**91.4**	89.2
s.e.m	±6.1	±6.6	±3.8	±5.3

Finally, we compare the performance of our approach with benchmarks and the $NAAR_{rand}$ on Weizmann and Weiz.Robust. Table 2 shows that the $NAAR_{cp}$ generates higher recognition rate and lower standard error of the mean than the $NAAR_{rand}$, which proves that the VNT extracted from corner points is more informative to characterize actions than their counterpart in [5]. The $NAAR_{cp}$ roundly outperforms the benchmark algorithms [8] and [9], especially on the Weiz.Robust dataset, which shows our approach is more invariant to affine translation and scaling and is more robust to the deformation of actors than the action recognition models based on computer-vision features. The comparison results show that the neurobiology based feature can represent actions more successfully than computer-vision features, therefore the proposed NAAR can possibly be used as a suggestive substitute for the action recognition model using computer-vision feature.

4. Conclusion

In this letter, the visual-neuron feature (VNF) of action representation in the primate visual cortex is studied and used for action recognition. The VNF is invariant to affine translation and scaling of moving objects while maintaining action specificity. Moreover it is robust to the deformation of actors. Experiments show that our approach outperforms the benchmark approaches using computer-vision feature. In the future work, we plan to extract uniform VNFs which are independent of action datasets and apply this feature to the recognition of continuous action sequences.

Acknowledgement

This work is supported by National Nature Science Foundation of China (60803072) and National High Technology Research of China (2007AA01Z168).

References

[1] R. Blake and M. Shiffrar, "Perception of human motion," Annu. Rev. Psychol., vol.58, pp.47–73, 2007.
[2] K.S. Saleem, W. Suzuki, K. Tanaka, and T. Hashikawa, "Connections between anterior inferotemporal cortex and superior temporal sulcus regions in the macaque monkey," J. Neurosci., vol.20, pp.5083–5101, 2000.
[3] M. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements," Nat. Rev. Neurosci., vol.4, pp.179–192, 2003.
[4] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," IEEE 11th ICCV, pp.1–8, Oct. 2007.
[5] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Object recognition with cortex-like mechanisms," IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.3, pp.411–426, March 2007.
[6] J. Han and B. Bhanu, "Statistical feature fusion for gait-based human recognition," Proc. 2004 IEEE Computer Society Conf. on CVPR, vol.2, pp.842–847, June-July 2004.
[7] K. Tanaka, "Inferotemporal cortex and object vision," Annu. Rev. Neurosci., vol.19, pp.109–139, 1996.
[8] A. Mokhber, C. Achard, and M. Milgram, "Recognition of human behavior by space-time silhouette characterization," Pattern Recognit. Lett., vol.29, no.1, pp.81–89, Jan. 2008.
[9] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, "Human action recognition using star skeleton," Proc. 4th ACM Inter. workshop on Video surveillance and sensor networks, pp.171–178, Oct. 2006.