Computing and Informatics, Vol. 29, 2010, 281–301

ACTION RECOGNITION USING VISUAL-NEURON FEATURE OF MOTION-SALIENCE REGION

Ning Li, De Xu

Institute of Computer Science and Engineering
Beijing Jiaotong University
Xizhi Men Wai, Shangyuan Cun 3, 100044 Beijing, P.R. China
e-mail: {06112075, dxu}@bjtu.edu.cn

Lu Liu

School of Electronic Engineering
Beijing University of Posts and Telecommunications
Xitu Cheng street 10, 100088 Beijing, P.R. China
e-mail: luludeerdeer@gmail.com

Manuscript received 3 September 2008; revised 24 August 2009
Communicated by Horst Bischof

Abstract. This paper proposes a shape-based neurobiological approach for action recognition. Our work is motivated by the successful quantitative model for the organization of the shape pathways in primate visual cortex. In our approach the motion-salience region (MSR) is firstly extracted from the sequential silhouettes of an action. Then, the MSR is represented by simulating the static object representation in the ventral stream of primate visual cortex. Finally, a linear multi-class classifier is used to classify the action. Experiments on publicly available action datasets demonstrate the proposed approach is robust to partial occlusion and deformation of actors and has lower computational cost than the neurobiological models that simulate the motion representation in primate dorsal stream.

Keywords: Action recognition, shape-based neurobiological approach (SBNA), motion-salience region, visual-neuron template, visual-neuron feature, visual cortex

Mathematics Subject Classification 2000: 15A99, 6207, 62H30, 62H35, 68T45, 68U05, 68U10, 68U20

1 INTRODUCTION

Action recognition has become one of the most active research areas in computer vision due to its potential applications such as video surveillance, content based video retrieval, sports events analysis and virtual reality. Most of the action recognition researches have focused on studying the classical features of target objects such as the contour, interest points, local or global spatio-temporal volume, etc. In these works the shooting angle, scaling, occlusion and imperfect extraction of actors can impact the performance of the systems. Primates outperform the best computer vision systems for action recognition, so building a system that simulates action recognition in the primate visual cortex has always been an attractive idea.

The mechanism of action recognition in primate visual cortex has progressed over the past decades. The visual cortex is organized in two different streams: a dorsal stream dominating motion information [1] and a ventral stream dealing with shape information. Many electrophysiological experiments show that, in monkey and human brains, these two streams originate in the primary visual cortex (V1) and separate along two populations of cells [2]: cells responding to speed and direction of motions project to the area MT (V5) and MST in the dorsal stream [3] and cells responding to static object recognition and spatial representation project to extrastriate visual areas V2, V4 and inferotemporal (IT) cortex in the ventral stream [4]. Eventually, the two streams twist and interact at higher levels [5]. In fact, the organization of these two pathways is very similar [6]. Both of them follow the template-based feedforward hierarchy [7]. The contribution of the two streams for action recognition in the cortex is still a matter of debate [8].With the extraction of features inspired by the primate visual cortex, the action recognition model can be designed under two types of frameworks: the framework relies on a model of the dorsal stream and the framework relies on a model of the ventral stream. It would have been interesting to compare the performance of the two frameworks.

In the first framework, the motion information of actions has been used for action classification. However, the mechanism for action recognition in dorsal stream is still unclear. Giese et al. [9] speculate that 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. [10] extend the model in [9] 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 action recognition algorithms using classical features.

In the second framework, the successful quantitative model for static object representation in primate cortex has been studied by [11], which inspires us to use the shape-based feature to represent actions for the action classification. Using shape for action recognition has been done in many works. Traditional approaches, such as [12, 13, 14], model actions as space-time representations which explicitly or implicitly encode the dynamics of an action through temporal dependencies. The recent study from Weinland et al. [15] proposes motion sequences are represented with respect to a set of discriminative static key-pose exemplars and without modelling any temporal ordering.

Motivated by the recent work [16] which has shown the benefit of using shape features in addition to motion features for the recognition of actions, we propose a shape-based neurobiological approach (SBNA) for action recognition. The SBNA is performed in three procedures. First, the motion-salience region (MSR) of an actor is extracted from his/her silhouette sequence. The MSR is the spatio-temporal image of the exterior body parts that take remarkable local movement. Then the MSR is represented by a visual-neuron feature (VNF), which is implemented by simulating the hierarchical template matching architecture for static object representation in primate visual cortex [11, 17]. In this part, we improve the static object representation procedure in [11] by replacing the 2D Gabor function set by the 2D log-Gabor function set that is the standard method for representing the spatial-frequency response of primate visual neurons. Finally, a linear multi-class SVM classifier is used to classify the VNF of test actions. The main contribution of the proposed action recognition approach is twofold:

1. the VNF of the MSR image is robust to the deformation of actors, the imperfect extraction of silhouette (foreground mask), and partial occlusion;
2. it is more efficient in computation than the neurobiological models that try to simulate the motion perception mechanism in primate dorsal stream.

Experiments on publicly available action datasets show that our approach outperforms the action recognition model using classical features and the model [10] computing the mechanism of action representation in primate dorsal stream.

The rest of the paper is organized as follows. In section two we address how the MSR is extracted to represent an action. In section three, we briefly introduce the quantitative model of static object representation in primate visual cortex. Section four introduces in detail the representation of the MSR using VNF, and the action classification method. Experimental results will be discussed in section five. Finally, conclusive remarks are addressed at the end of this paper.

2 THE MOTION-SALIENCE REGION OF ACTION

The use of local image features has become a trend in action recognition [13, 18]. The sparse and informative local features provide compact and abstract representations of actions in video sequences. These features make the problem more manageable and provide increased robustness to noise and pose variation.

The body parts taking local movement varies more significantly than the parts taking global movement when actors perform different actions. Specifically, the exterior body parts (e.g., the head, limbs, and joints) take more frequent and remarkable local movement than the torso, thus these parts can be used as the local feature to effectively distinguish different types of actions. In this paper we define the spatio-temporal image of the exterior body parts as the motion-salience region (MSR) of actions. The MSR is calculated using the combination of the Poisson equation in spatial domain and the average motion energy in temporal domain.

2.1 Discrete Approximation of the Poisson Equation

In this paper, the binary silhouette (foreground mask) images of sequential postures are used for the action classification. A reliable approach to characterize the properties of the silhouette is to assign every internal point to a value that is directly proportional to the distance between that point and the contour of the silhouette. One popular approach is the distance transform [19]. Another one is to let a set of particles at an internal point move in a random walk until they hit the contour. The statistics of this random walk can be measured using the mean distance required for a particle to hit the boundaries.

We use $U(x, y)$ to denote the mean distance of the random walk and S to represent the mask of a posture in a silhouette image. Gorelick et al. [20] prove such random walk can be formulated as the solution to the Poisson equation. They proposed the discrete approximation of this solution, see Equations (1) and (2),

$$\begin{aligned} U(x,y) = c + \tfrac{1}{4}[&U(x+h,y) + U(x-h,y) \\ &+ U(x,y+h) + U(x,y-h)]. \end{aligned} \tag{1}$$

$$U(x,y) = 0, \quad (x,y) \in \partial S. \tag{2}$$

The $U(x, y)$ is computed recursively: Under the Dirichlet boundary condition $U(x, y) = 0$ at the bounding contour ∂S, for any point (x, y) inside the S, the $U(x, y)$ is equal to the average value of its immediate four neighbors plus a constant c. Therefore, high values of $U(x, y)$ are obtained in the central part of the shape, whereas the external protrusions have relatively low values. The constant c and h are equal to one, which represents the distance between immediate pixels.

The value obtained using the distance transform is determined by a single nearest contour point, while the solution to the Poisson equation takes into account many neighboring points. So the Poisson equation reflects more global properties of the silhouette.

2.2 Extraction of the Exterior Parts of Bodies

One interesting property of the $U(x, y)$ is that the its gradient keeps constant inside the mask S of a posture and increases when it falls near the boundary. The value becomes the largest at the concave regions, which normally denote the joints of a body. Figure 1 shows the $\|\nabla U(x,y)\|$ images for the silhouette images of some postures.

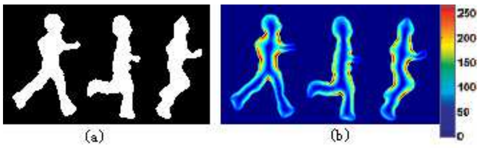


Fig. 1. The $\|\nabla U(x,y)\|$ image for the silhouette images of example postures. a) The example postures; b) The $\|\nabla U(x,y)\|$ image for the postures. The level set (from dark blue to red) is used to indicate the intensity of pixels in the images.

The exterior body parts usually consist of the regions moving with wide range and high frequency. These regions involve the head, limbs and joints, etc. In order to identify the exterior parts of a posture, we define the Equation (3) using the combination of the normalized $U(x,y)$ and $\|\nabla U(x,y)\|$:

$$V(x,y) = \log\left(\frac{1}{U(x,y)} + \|\nabla U(x,y)\|\right). \tag{3}$$

Here, we obtain a smooth version of the exterior body parts for the mask of a posture in a silhouette image. Figure 2 shows the sequential silhouettes with the exterior parts highlighted for the running actor.

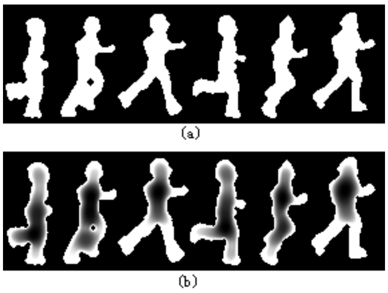


Fig. 2. The extraction of exterior body parts. a) The sequential silhouettes for the example action "run"; b) The head, limbs and joints of the body are highlighted in high intensity.

2.3 Extraction of the Motion-Salience Region

Actions are essentially spatio-temporal variations which encode spatial information of postures and dynamic information of actions. To characterize an action, we represent the sequential silhouettes (foreground masks) as an informative average motion energy image that implicitly captures the global motion properties of an action and has been successfully used in gait-based human identification [21]. The advantages of this action representation are its low computational complexity and simple implementation.

The extraction of the motion-salience region (MSR) is to obtain the spatio-temporal shape of the body parts taking remarkable local movement. These parts mainly refer to the exterior body parts. We extend the $V(x, y)$ in Equation (3) to $V(x, y, t)$, which denotes the time-sequential images with exterior body parts highlighted. Then, the average motion energy image is calculated by:

$$A = \frac{1}{\tau} \sum_{t=1}^{\tau} V(x, y, t) \tag{4}$$

where τ is the duration of a complete action, Figure 3 a) shows the average motion energy image for the action "run".

In the average motion energy image, the regions of high intensity denote the body parts undergoing frequent local movement. Because the motion mode of these parts varies significantly when objects perform different types of actions, these regions can characterize the spatio-temporal shape of actions. On the other hand, the regions of low intensity denote the body parts taking global movement which do not contribute much to distinguish different types of actions. Using an appropriate threshold, we can identify the MSR that is larger than the threshold. Figure 3 b) shows the MSR images extracted from the average motion energy image using different thresholds. The impact of the threshold on the performance of the proposed approach is illustrated in Figure 6.

The MSR image appears to be similar to the motion energy images proposed by [12]. Both methods extract the region where the action occurs. The difference is our method only extracts the body parts taking local movement (e.g. the arms in the "waving hand" action) and eliminates the central parts that only take global movement (e.g. the torso in the "walking" and "running" actions).

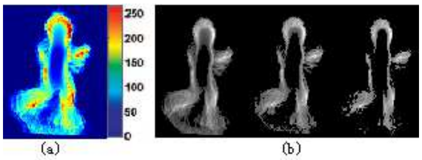


Fig. 3. The motion-salience region for the action "run". a) The average motion energy image of the posture sequence illustrated in Figure 2 b); b) The motion-salience region extracted from a) using different thresholds (from left to right: 30 %, 50 %, and 70 % of the highest pixel intensity).

3 THE VISUAL-NEURON FEATURE FOR STATIC OBJECTS

The extraction of visual-neuron feature (VNF) for static object representation is proposed by Serre et al. [11]. It comprises a 4-layer feedforward template matching architecture ($S1$-$C1$-$S2$-$C2$), which models the representation of objects along the ventral stream in primate visual cortex.

The model starts at $S1$ units that respond to simple bar-like stimuli. They are obtained by applying an input image to a battery of Gabor filters with 8 bands (each contains 2 filter scales) and 4 orientations, see Equations (5) and (6):

$$G(x,y) = \exp(-\frac{X^2+\gamma^2 Y^2}{2\sigma^2})\cos(\frac{2\pi}{\lambda}X), s.t. \tag{5}$$

$$X = x\cos\theta + y\sin\theta, Y = -x\sin\theta + y\cos\theta. \tag{6}$$

All filter parameters are adjusted so that the $S1$ units match the response property of visual neurons in $V1$ area. For an input image, there are 8 bands × 2 scales × 4 orientations = 64 images generated by $S1$ units. These images then serve as input of $C1$ units.

In the $C1$ units, a sliding window is applied to the $S1$ images in all the bands (each band contains two $S1$ images of different scales) and orientations. In each band, $C1$ units select the maximal value from the two window areas located at the same position of the two $S1$ images, and use this value to represent the window area. In this way, the feature at the $C1$ layer is more tolerant to translation and scaling of objects than the feature at the $S1$ layer [22]. The process repeats in all the 4 orientations, therefore there are 8 bands × 4 orientations = 32 $C1$ images generated in this stage.

$S2$ units are where template matching occurs. Before the matching stage, a group of template objects for all action categories is predefined in the $C1$ layer, and then M small patches of various sizes in all the 4 orientations are extracted from *random positions* of $C1$ images of the template objects. In this paper, the collection of the M patches is defined as the visual-neuron template (VNT). The $S2$ behaves as a Gaussian-like radial basis function, where each of the patches functions as the center of the function. That is, for a small window within a $C1$ image of a particular scale and orientation, the response R of the corresponding $S2$ unit is given by:

$$R = \exp(-\left\|C1_{win} - P\right\|^2), \tag{7}$$

where $C1_{win}$ denotes a small window locating at every position of the $C1$ image in all the 4 orientation and P represents a patch extracted from template images at $C1$ layer. For each $C1_{win}$, there are M patches which need to be compared, thus M responses obtained. At runtime, $S2$ images are computed for each of the eight bands. Therefore, the total number of the $S2$ images equals to $M \times 8$.

Finally, at the $C2$ layer, the translation- and scale-invariant $C2$ feature is obtained by taking a global maximum across all scales over the entire $S2$ images.

Therefore, units at this layer have the largest receptive fields and respond to complex stimuli such as cars, faces, etc [23]. In this paper, the $C2$ feature is defined as the visual-neuron feature (VNF) of an object. The dimension of the VNF is equal to the number of patches (M) and is independent of the size of the input image containing the object.

The significant successes of this model are:

1. the complexity of the feature and the size of receptive field in each layer increase simultaneously. S1 units have the smallest receptive field, thus they are low invariant to the translation and scaling. C2 units correspond to the most complex feature, thus show a greater degree of invariance;
2. more complex feature at higher layer of the model are built from simpler feature at lower layer; therefore, the complex features are tolerant to local deformations as a result of the invariance properties of the simpler features.

4 ACTION RECOGNITION USING VISUAL-NEURON FEATURE

In this part, we propose a shape-based neurobiological approach (SBNA) for action recognition, which consists of two procedures: action representation in primate visual cortex and action classification.

4.1 Representation of Actions Using Visual-Neuron Feature

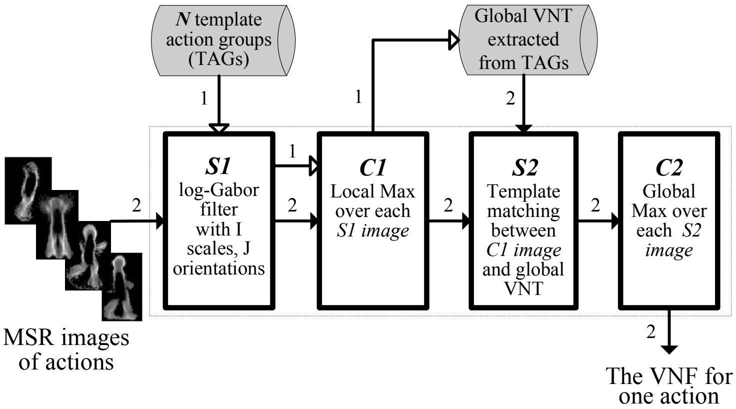
The key to the success of the proposed SBNA is that the motion-salience region (MSR) of an action is analyzed, the motion sensitive component is then represented by the visual-neuron feature (VNF) that has been used to represent static objects in primate visual cortex [11]. In the proposed approach, we make two points of improvement on the [11]:

1. replace the 2D Gabor function by the 2D log-Gabor function in the $S1$ layer;
2. use the corner point information for the VNT extraction in $C1$ layer.

Figure 4 shows the flow chart for the representation of actions using the VNF in the proposed SBNA. The basic workflow is similar to that in the [11], the improvements mainly reside in the first step, where the global visual-neuron template (VNT) is extracted.

4.1.1 2D Log-Gabor Function in $S1$ Layer

The first step is for the extraction of visual-neuron template (VNT), which is implemented along the $S1$ and $C1$ layer. The $S1$ layer agrees quantitatively with the tuning properties of simple cells in the primary visual cortex. As it is proposed in [11], the Gabor function is usually used to simulate object perception in primate simple visual neurons. However, Hawken et al. [24] suggest that the Gabor function



N template action groups (TAGs)
Global VNT extracted from TAGs
1
1
2
S1 log-Gabor filter with I scales, J orientations
C1 Local Max over each *S1 image*
S2 Template matching between *C1 image* and global VNT
C2 Global Max over each *S2 image*
2
2
2
2
2
MSR images of actions
The VNF for one action

Fig. 4. The flow chart for the representation of MSR images using the VNF in the proposed SBNA. The rectangular cases indicate different layers. The gray buckets represent the databases for "Template action groups (TAG)", and "Global VNT for a certain action dataset", respectively. The first step along the hollow arrows labeled with 1 indicates the process of obtaining the global VNT. The second step along the solid arrows labeled with 2 indicates the extraction of VNF for the MSR image of an action.

fails to capture the precise form of the spatial-frequency tuning curves in monkey cortical cells. The Gabor function can not be constructed in terms of arbitrarily wide bandwidth. It also over-represents low frequency components, which, in essence, produces a correlated and redundant response to the low frequencies.

The study of Field [25] shows that compared with the Gabor function, the frequency response of log-Gabor function can provide much broader bandwidth, thus it gives wider coverage of the spectrum; meanwhile the log-Gabor function has extended tails at high frequency, which ensures it to locate more precise local variation of natural image in the spatial domain. Another point in support of the log-Gabor function is that its frequency response has the Gaussian form when viewed on the logarithmic frequency scale, which is consistent with the measurements on primate visual systems. In these measurements, the responses of primate visual cells are symmetric on the log frequency scale.

Due to the singularity in the log function at the origin, the analytic expression for the shape of the log-Gabor function can not be constructed in the spatial domain. Equation (8) shows the 2D log-Gabor function set in the frequency domain.

$$H_i^j(f,\theta) = \exp\left\{\frac{-\left[\ln(f/f_i)\right]^2}{2\left[\ln(\sigma_f/f_i)\right]^2}\right\} \times \exp\left\{\frac{-(\theta-\theta_j)^2}{2\sigma_\theta^2}\right\}, (i=1\ldots I, j=1\ldots J) \quad (8)$$

where f_i represents the central frequency of the filter with the scaling index i; σ_f determines the bandwidth of the log-Gabor filter in the radial direction; σ_θ determines the bandwidth in the orientation direction; θ_j represents the orientation angle of the filter with the angle index j. The central frequency is obtained by Equation (9):

$$f_i = \left(\lambda_1 \times s^{i-1}\right)^{-1} \tag{9}$$

where λ_1 denotes the wavelength of the smallest filter scale, s is the scaling factor between center frequencies of successive filters. The orientation angle is fixed by the number of filter orientations which is predefined empirically.

In the $S1$ layer, the dot product between the log-Gabor function set and the Fourier transform of an input MSR image is firstly calculated, and then a numerical inverse Fourier transform is performed to get the $S1$ images of different scales and orientations in the spatial domain. The parameter configuration and the advantage of the proposed 2D log-Gabor function over the 2D Gabor function are discussed in subsection 5.3.2.

4.1.2 Corner Information for VNT Extraction

In the C1 layer, suppose there are N action categories in a given action dataset, then we define N template action groups. In each group, we randomly choose n instances, where the action has been represented by the MSR image. Each template group is used to randomly extract d prototype patches. This leads to d patches extracted per action category and a total number of $M = d \times N$ stored patches for the global VNT. In this step, we modify the extraction of VNT proposed in [11] by replacing the extraction of patches from *random positions* of template images with the selective extraction of patches from *corner points* of template images.

Corner points exploit the useful correlation of their neighboring structures, therefore corner points can effectively eliminate the redundant information induced by homogeneous region. Theoretically, the patches extraction strategy in [11] can result in much overlapping information, while the corner-point-based method can avoid this phenomenon. In this paper, the Harris Corner Detector proposed in [26] is used to detect the corner points. The number of corner points is chosen by users. For instance, to represent a 95 × 120 image, about 50 corner points can be sufficient. Figure 5 shows the corner point images of the $C1$ images for the action "run". The advantage of the corner-points based VNT extraction in the proposed approach over the VNT extraction in [11] is illustrated in Figure 7.

In the second step, the global VNT obtained in the first step is used to determine the VNF for input actions. This step thoroughly passes the $S1$, $C1$, $S2$ and $C2$ layers, which is the same as the extraction of VNF for static objects introduced in Section 3. The dimension of the VNF equals to the total number (M) of prototype patches contained in the global VNT database. The VNF of low dimension may not be sufficient to characterize an action. On the other hand, redundant dimensions will increase the computational intensity. The impact of various dimensional VNF will be discussed in Subsection 5.3.3.

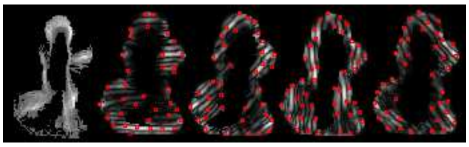


Fig. 5. Corner point images (denoted by the rectangles) for the $C1$ images of the action "run". From left to right: the MSR image for the action "run", the corner point images of the corresponding $C1$ images at band 1 in the orientation 0°, 45°, 90° and 135°.

4.2 Action Classification

The classification stage is performed using a linear multi-class SVM classifier. We apply the tool kit "LIBSVM" provided by Chih-Chung Chang and Chih-Jen Lin (`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`) to the stage. We use the one-against-one strategy for the multi-class classification problem, where all pairs of classes are compared to each other. In this way, a multi-class problem is decomposed into multiple binary classification problems [27]. To classify a test action, we combine all the pairwise decisions. That is, given an action dataset having N action categories, $N \times (N-1)/2$ tests need to be performed for one test action. Then, the test action is assigned to the class that wins the most pairwise comparisons ("max-wins" rule).

5 EXPERIMENTS

5.1 Action Datasets

As for the evaluation on different action recognition approaches, so far there are not benchmark datasets. For the shape-based neurobiological approach (SBNA), two popular action datasets are tested: Weizmann and KTH. Details about the dataset are given below.

1. **Weizmann:** The Weizmann provides 90 video sequences shown by nine subjects, each performing 10 types of natural actions. The actions are "bend", "jack", "jump-forward (jump1)", "jump-up-down (jump2)", "run", "gallop-sideways", "skip", "walk", "wave-one-hand" and "wave-two-hands". The subjects present in different scales in the scene. In the video sequences, the actions are repeated several times. In order to increase the size of data, we averagely segment every video sequence into 3 pieces. In the experiments, the cross-validation rule is adopted to compute an unbiased estimate of the true recognition rate. We split the dataset as follows. 6 subjects are used as templates (training set) and the remaining 3 subjects are used for testing. Thus, the size of each template action group (TAG) is $n = 18$ (6 subjects × 3 pieces). The system performance is evaluated by the average of five random splits.

2. KTH: The KTH human action dataset contains six types of human actions: "walk", "jog", "run", boxing, hand waving and hand clapping. These actions are performed several times by twenty-five subjects in four different conditions: outdoors ($s1$), outdoors with scale variation ($s2$), outdoors with different clothes ($s3$) and indoors with lighting variation ($s4$). The database contains 2 391 sequences which are taken over homogeneous backgrounds with a static camera. The dataset is split as: actions of 16 randomly drawn subjects for training and actions of the remaining 9 subjects for testing. The system performance is based on the average of five random splits.

The binary silhouette (foreground mask) sequences have been given by both Weizmann and KTH dataset. Similar to [10], we use a primitive attention mechanism: the input is a sequence of fixed-size image windows, centred at the person of interest. For the KTH, a bounding box of full height and half width of the frame is extracted for each foreground mask; for the Weizmann, bounding boxes have been extracted from the foreground masks by the dataset.

5.2 Benchmark Approaches

For benchmarking, we use the approaches proposed by Jhuang et al. [10] and Mokhber et al. [28].

The paper [10] extends the neuroscience model for static object recognition in [11] by experimenting with three types of $S1$ units that are sensitive to motion direction: space-time gradient-based units, optical flow based units and space-time oriented units. They essentially simulate the representation of motion in primate dorsal stream. [10] then experiments with the addition of two new layers ($S3$ and $C3$) that encodes actions with a higher degree of temporal invariance. The experimental results in [10] show that there is small improvement using the $C3$ features over the $C2$ features. Moreover, the sparse feature space appears to perform better than the dense one on average. Therefore, the sparse $C3$ features based on these three types of $S1$ units are used as benchmarks. They are denoted as $GrC3$, $OfC3$ and $StC3$, respectively. The code for [10] was graciously provided by Hueihan Jhuang.

The paper [28] is a typical action recognition model using the classical feature. It is designed to be invariant to the translation and deformation of actors. In the action representation stage, the silhouette sequence of an action is represented by a spatio-temporal volume that is characterized by a vector of its 3D geometric moments. We duplicate the code in the light of the procedure introduced by [28].

For the action classification stage of the benchmarks, we use the "LIBSVM" classifier to perform the same multi-label classification as our approach.

5.3 Experimental Results

In this part, we first find an uniform threshold to extract the motion-salience region (MSR) images for all actions. The design of the 2D log-Gabor function in the $S1$ layer is then discussed. The impact of the dimension of the corner-points based visual-neuron feature (VNF) on the proposed action recognition approach is analyzed. We also demonstrate the robustness of our approach to the challenging situation in real-world scenarios, where the unstable contour of subjects occurs. Finally, our approach is evaluated by comparing with the benchmarks.

5.3.1 Choice of the Threshold for MSR Images

In this experiment, the Weizmann dataset is used to determine the threshold. The thresholds are chosen from 10 % to 90 % of the highest pixel intensity in the average motion energy image. The MSR images are empirically represented by the $10\,000-D$ corner-points based VNF. The first parameter group in Table 1 is selected for the 2D log-Gabor function. Figure 6 shows the average recognition rates obtained using different thresholds.

Figure 6 illustrates the threshold that is equal to 50 % of the highest intensity can effectively grasp the spatio-temporal property of actions. Higher thresholds ignore much useful motion information and thus can hardly distinguish different action types while lower ones can not eliminate the body region which contribute little to the representation of actions, and thus increase the computational intensity.

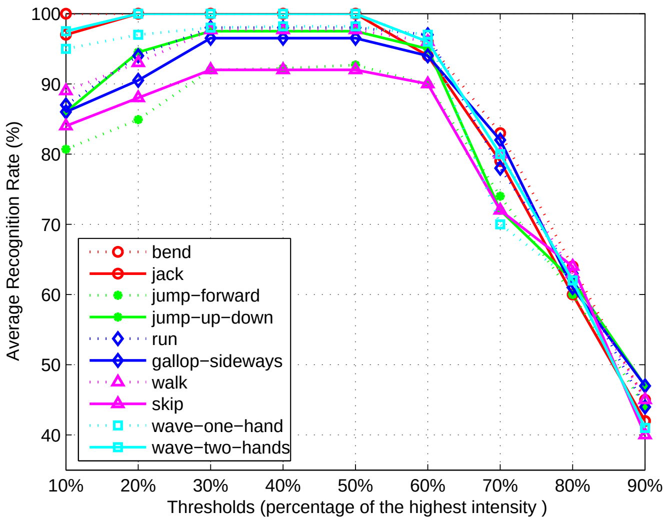


Fig. 6. Average recognition rate on the Weizmann dataset based on different thresholds

5.3.2 Design of 2D Log-Gabor Function

The log-Gabor filter bank does not form an orthogonal basis set and hence there is no unique or ideal arrangement of the filters. Theoretically, the frequency response of the log-Gabor filters should have minimal overlap to achieve fairly even spectral coverage. In Equation (8) we empirically set $J = 4$ filter orientations which equally spread in the spectrum space, and $I = 8$ filter scales (with the Weizmann dataset providing the bounding box with the size about 120×80, a larger scale number will not generate distinguishable output for different actions); in Equation (9) the λ_1 is assigned 3 pixels. In this work $\sigma_\theta = \pi/(J \times 1.5)$, which results in approximately the minimum overlap needed to get even spectral coverage in the angular direction. Table 1 gives a series of parameter groups experimentally, that result fairly even spectral coverage in the radial direction.

Para. group	σ_f/f_i	Scaling factor s	Bandwidth (octave)
1	0.85	1.3	$BW < 1$
2	0.75	1.6	$BW \approx 1$
3	0.65	2.1	$1 < BW < 2$
4	0.55	3	$BW \approx 2$

Table 1. Parameter groups for 2D log-Gabor function set

In this experiment, the $10\,000 - D$ corner-points based VNF is used to detect the best design of the log-Gabor function. Table 2 compares the action recognition results on the Weizmann datasets using the 4 parameter groups and the Gabor function set proposed by [11]. To make an fair comparison, we only use 8 filter scales and 4 orientations for the Gabor function. Table 2 shows the recognition rates increase with the bandwidth, and the 4th group for the 2D log-Gabor function generates the highest average recognition rate. The 1st and 2nd parameter groups have similar recognition result as the Gabor function, which is related to the conclusion made by Field [25] that when the parameter groups generate the bandwidth ≤ 1 octave, the 2D log-Gabor function has the same performance as the 2D Gabor function. The experiment result indirectly proves that the log-Gabor function permits a more compact image representation than the Gabor function when the bandwidth > 1 octave. In the following experiment, the 4th parameter group is chosen to evaluate the performance of the proposed method.

Para. group 1	Para. group 2	Para. group 3	Para. group 4	Gabor fun.
86.3	87.7	97.7	98.1	86.9

Table 2. Average recognition rate (%) on Weizmann using various filter parameter sets and the Gabor function

5.3.3 Feature Selection

The SBNA with visual-neuron template (VNT) extracted from *random position* of $C1$ images is denoted as $SBNA_{rand}$, and the one with VNT extracted from *corner points* is denoted as $SBNA_{cp}$. The low-dimensional visual-neuron feature (VNF) may not be sufficient to characterize the spatio-temporal shape of target objects, while redundant dimensions can increase the computational intensity of the proposed approach. The following experiment studies and compares the performance of the $SBNA_{rand}$ and $SBNA_{cp}$ for the action recognition based on various dimensions of VNF.

In this part, we calculate the average recognition rate on the Weizmann dataset. The dataset contains 10 types of action categories, therefore M patches of $C1$ units are extracted from 10 template action groups to compute the $C2$ features. Figure 7 shows the impact of various dimensions of VNF on the $SBNA_{rand}$ and $SBNA_{cp}$. In the case of the $SBNA_{cp}$ it is possible to reduce the dimension of $C2$ features quiet remarkably from $D = 10\,000$ down to $D = 300$ with minimal loss in accuracy; while in the case of the $SBNA_{rand}$, because the VNT extracted from random position of $C1$ images may include much overlapping information, the average recognition rate decreases rapidly when a few number of patches are extracted. Consequently, in the subsequent experiments, we apply the $300 - D$ VNF determined by the corner-point-based VNT to represent the MSR images of actions.

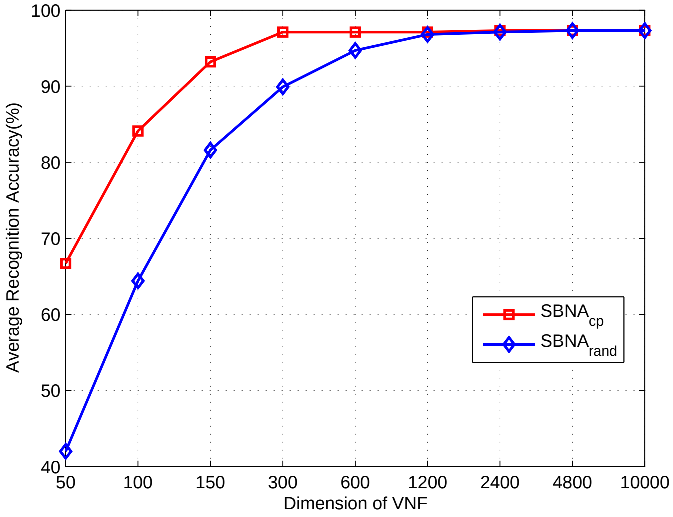


Fig. 7. The average recognition rate obtained by the $SBNA_{rand}$ and $SBNA_{cp}$ on the Weizmann using variant-dimensional VNF

5.3.4 Robustness

In this experiment, we demonstrate the robustness of our method to the unstable segmentation of foreground subjects. The Weizmann dataset also provides 10 walking actions present in various irregularities, which involve "normal walk", "moonwalk", "limp", "walk with bag", "walk with case", "walk with dog", "walk in skirt", "knees-up", "feet occluded by boxes", and "body occluded by pole". Figure 8 shows the MSR images for these "walk" actions. We still use the training set in previous experiments. To increase the size of the testing data, each "walk" video is divided into three segments. Table 3 shows the classification result obtained using the proposed $SBNA_{cp}$ based on the 300-D VNF. We list three classification results which denote the first, second and third mostly chosen action categories. The test videos are sorted to their first mostly chosen action category.

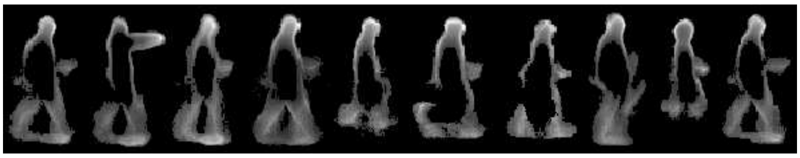


Fig. 8. MSR images for the "walk" actions in various irregularities. From left to right, the actions are: "normal walk", "moonwalk", "limp", "walk with bag", "walk with case", "walk with dog", "walk in skirt", "knees-up", "feet occluded by boxes" and "body occluded by pole"

Table 3 shows that most test actions in various irregularities are classified correctly as the "walk" action except for the "knees-up" action. In our opinion, it is even hard for human being to recognize the MSR image of the "knees-up" as the "walk" action. The experiments demonstrate strong robustness of the proposed method to partial occlusions and deformation of actors.

5.3.5 Comparison

Finally, we compare the performance of our approach with the benchmarks on the KTH dataset. We run experiment on these approaches using the same setup of training and testing data. To make comparison on an equal footing, we apply the $300 - D$ visual-neuron feature (VNF) to the proposed $SBNA_{cp}$ and [10].

Table 4 shows there is small improvement using the proposed $SBNA_{cp}$ over the $StC3$ (the $StC3$ in general can best represent moving objects among the three motion sensitive features proposed by [10], see 5.2). Both of them adopt the concept of spatio-temporal representation for actions at the $S1$ layer. The $StC3$ requires two template matching stages, while the $SBNA_{cp}$ only requires one. Thus our approach is more compact than the $StC3$. The experiment on the KTH shows the $SBNA_{cp}$ has similar performance with the $StC3$ under the $s1$ and $s4$ condition, while the $SBNA_{cp}$ is more robust to changes in scaling and appearance of actors than the

TestActions	**1stresult**	**2ndresult**	**3rdresult**
Normal walk	Walk 3	Non	Non
Moonwalk	Walk 3	Non	Non
Limp	Walk 3	Non	Non
Walk with bag	Walk 3	Non	Non
Walk with case	Walk 3	Non	Non
Walk with dog	Walk 3	Non	Non
Walk in skirt	Walk 3	Non	Non
Knees-up	Skip 2	Walk 1	Non
Feet occluded by boxes	Walk 3	Non	Non
Body occluded by pole	Walk 3	Non	Non

Table 3. Experiment results for the evaluation of robustness. The leftmost column shows the testing videos. For each testing video, the action category it is classified to and the corresponding times are recorded in the second, third and forth columns. The video is sorted to its first mostly chosen action category.

$StC3$ (under the $s2$ and $s3$ condition). Therefore, the action recognition approach which combines the motion-salience region of actors and the shape representation in ventral stream outperforms the approach simulating the mechanism of motion representation in dorsal stream.

The proposed approach also outperforms the benchmark [28] on the KTH, which shows that the neurobiology based feature can represent actions more successfully than the classical feature. Therefore the proposed SBNA can be used as a suggestive substitute for the action recognition models using the classical features of target objects such as the contour, interest points, local or global spatio-temporal volume, etc.

We run the programs in Matlab 7.0 under Dual Xeon X5570 2.93 GHz (L3 Cache 8 MB). For the KTH human database, suppose the action videos consist of 30 frames, then [10] takes about 60 seconds per video, most of the running time is taken up by the S2 + C2 computations; our approach, on the other hand, takes about 40 seconds per video, more than half of the running time is taken up by extracting the MSR region. The reason why our approach is much faster that of [10] is that [10] analyzes action videos frame by frame while our approach treats the video as one spatio-temporal shape and thus remarkably reduces the amount of data.

6 CONCLUSIONS AND FUTURE WORK

Humans and primates outperform the best computer vision systems for the recognition of action. Therefore building a system that can efficiently simulate the recognition of action in primate visual cortex is the research emphasis of this paper. With the extraction of features inspired by the primate visual cortex, the action recognition model can be designed under two types of frameworks: the framework relies on a model of the dorsal stream and the framework relies on a model of the ventral

KTH	$GrC3$	$OfC3$	$StC3$	[28]	$SBNA_{cp}$
$s1$	91.7	92.5	**96**	92.1	95.5
s.e.m.	±3.3	±3.1	±2.2	±2.3	±2.7
$s2$	87.5	81.4	86.5	77.8	**88.7**
s.e.m.	±3.7	±4.2	±4.7	±2.9	±3.1
$s3$	89.7	91.7	89.7	86.9	**92.1**
s.e.m.	±3.2	±3.6	±3.3	±1.7	±3.4
$s4$	93.4	92.3	**95**	90.4	94.7
s.e.m.	±2.1	±2.3	±2.1	±1.7	±2.4
Avg.	90.6	89.5	91.8	86.8	**92.8**
s.e.m.	±3.1	±3.3	±3.1	±2.2	±2.9

Table 4. Comparison between the proposed $SBNA_{cp}$ and the benchmark algorithms (denoted as $GrC3$, $OfC3$, $StC3$, and [28]). The numbers recorded are the average recognition rates (%) and the standard error of the mean (%).

stream. The ventral stream processes spatial information and the dorsal stream deals with temporal information.

Motivated by the recent work [16] which has shown the benefit of using shape features in addition to motion features for the recognition of actions, we propose a shape-based neurobiological approach for action recognition. In this paper, the motion-salience regions of actions are firstly extracted, and then they are represented by simulating the mechanism of primate visual system. During the implementation of the mechanism, we improve the computational model for static object representation in primate ventral stream proposed by [11]. The improvements are:

1. replace the 2D Gabor function by the 2D log-Gabor function in the $S1$ layer;
2. involve the corner point information for the VNT extraction in $C1$ layer.

Experiments show that the shape representation method in our approach is more representative and compact than in [11]. Although our approach does not remarkably improve the recognition rate of the approach simulating the mechanism of motion representation in dorsal stream [10], our approach results in substantial reduction in computational intensity. Moreover, our approach is more robust to the deformation of actors, imperfect extraction of silhouettes, and partial occlusion than the action recognition models using the classical features of target objects such as the contour, interest points, local or global spatio-temporal volume, etc. Therefore our approach can be used as a suggestive substitute for these models.

Actions can be well represented in terms of limited number of key postures in machine vision [29]. Schindler et al. [30] suggest that very short snippets (1–7 frames) are sufficient for basic action recognition; with rapidly diminishing returns, more frames are added. The snippet is essentially in line with the definition of key frames. In our action representation method, as long as there are key frames (postures) included in the action sequence, the action can be effectively represented. Specifically, to include key frames, in the case of rapid actions, e.g. the "run", 10–20 frames is enough to include key frames; in the case of slow actions, e.g. the "bend", 30–40 frames can be adequate. Therefore, although the proposed action recognition approach does not require complete sequences, it sill lags behind [30] when very slow actions are present. This would be inpracticable for practical scenarios, where decisions have to be taken online. The future work will try to extract more compact motion information from single frame so that a snippet of action sequence is sufficient to be recognized.

Acknowledgement

This work is supported by National Nature Science Foundation of China (60803072, 90820013), National High Technology Research and Development Program of China (2007AA01Z168) and Science Foundation of Beijing Jiaotong University (2007XM008).

REFERENCES

[1] Riesenhuber, M.—Poggio, T.: Hierarchical Models of Object Recognition in Cortex. Nat. Neurosci., Vol. 2, 1999, No. 11, pp. 1019–1025.

[2] Goodale, M. A.—Milner, A. D.: Separate Pathways for Perception and Action. Trends in Neuroscience, Vol. 15, 1992, No. 1, pp. 20–25.

[3] Simoncelli, E.—Heeger, D.: A Model of Neural Responses in Visual Area MT. Vis. Res., No. 38, 1998, pp. 743–761.

[4] Rosa, M. G. P.—Tweedale, R.: Visual Areas of the Lateral and Ventral Extrastriate Cortices of the Marmoset Monkey. The Journal of Comparative Neurology, Vol. 422, 2000, No. 4, pp. 621–651.

[5] Saleem, K. S.—Suzuki, W.—Tanaka, K.—Hashikawa, T.: Connections Between Anterior Inferotemporal Cortex and Superior Temporal Sulcus Regions in the Macaque Monkey. J. Neurosci, Vol. 20, 2000, pp. 5083–5101.

[6] Saito, H.: Brain Mechanisms of Perception and Memory. Oxford Univ. Press, 1993, pp. 121–140.

[7] Wersing, H.—Korner, E.: Learning Optimized Features for Hierarchical Models of Invariant Recognition. Neural Computation, Vol. 15, 2003, No. 7, pp. 1559–1588.

[8] Blake, R.—Shiffrar, M.: Perception of Human Motion. Ann. Rev. Psychol., No. 58, 2007, pp. 47–73.

[9] Giese, M.—Poggio, T.: Neural Mechanisms for the Recognition of Biological Movements. Nat. Rev. Neurosci, No. 4, 2003, pp. 179–192.

[10] Jhuang, H.—Serre, T.—Wolf, L.—Poggio, T.: A Biologically Inspired System for Action Recognition. Proceedings of the 11$^{\text{th}}$ IEEE Inter. Conf. Computer Vision (ICCV), 14–21 Oct. 2007, pp. 1–8.

[11] Serre, T.—Wolf, L.—Bileschi, S.—Riesenhuber, M.—Poggio, T.: Object Recognition With Cortex-Like Mechanisms. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), Vol. 29, 2007, No. 3, pp. 411–426.

[12] Davis, J. W.—Bobick, A. F.: The Representation and Recognition of Action Using Temporal Templates. IEEE Inter. Conf. Computer Vision and Pattern Recognition (CVPR), 1997.

[13] Laptev, I.—Lindeberg, T.: Space-Time Interest Points. Proceedings of the 9th IEEE Inter. Conf. Computer Vision (ICCV), 13–16 Oct. 2003, Vol. 1, pp. 432–439.

[14] Gorelick, L.—Blank, M.—Shechtman, E.—Irani, M.–Basri, R.: Actions as Space-Time Shapes. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), Vol. 29, 2007, No. 12, pp. 2247–2253.

[15] Weinland, D.—Boyer, E.: Action Recognition using Exemplar-based Embedding. IEEE Inter. Conf. Computer Vision and Pattern Recognition (CVPR), 23–28 June 2008, pp. 1–7.

[16] Niebles, J.—Fei-Fei, L.: A Hierarchical Model of Shape and Appearance for Human Action Classification. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 17–22 June 2007, pp. 1–8.

[17] Serre, T.—Kouh, M.—Cadieu, C.—Knoblich, U.—Kreiman, G.—Poggio, T.: A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex. AI Memo 2005-036/CBCL Memo 259, Massachusetts Inst. of Technology, Cambridge, 2005.

[18] Dollar, P.—Rabaud, V.—Cottrell, G.—Belongie, S.: Behavior Recognition Via Sparse Spatio-Temporal Feature. 2nd Joint IEEE Inter. Workshop, Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 15–16 Oct. 2005, pp. 65–72.

[19] Ye, Q. Z.: The Signed Euclidean Distance Transform and Its Applications. The 9th IEEE Inter. Conf. Pattern Recognition, 14–17 Nov. 1988, Vol. 1, pp. 495–499.

[20] Gorelick, L.—Galun, M.—Sharon, E.—Brandt, A.—Basri, R.: Shape Representation and Classification Using the Poisson Equation. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), Vol. 28, 2006, No. 12, pp. 1991–2005.

[21] Han, J.—Bhanu, B.: Statistical Feature Fusion for Gait-Based Human Recognition. Proceedings of the 2004 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR), 27 June–2 July 2004, Vol. 2, pp. 842–847.

[22] Serre, T.—Louie, J.—Riesenhuber, M.—Poggio, T.: On the Role of Object-Specific Features for Real World Recognition in Biological Vision. In Biologically Motivated Computer Vision, Second International Workshop (BMCV 2002), Tuebingen, Germany, 2002, pp. 387–397.

[23] Tanaka, K.: Inferotemporal Cortex and Object Vision. Ann. Rev. Neurosci, No. 19, 1996, pp. 109–139.

[24] Hawken, M.—Parker, A.: Spatial Properties of Neurons in the Monkey Striate Cortex. Proc. R. Soc., London Ser. B 231, 1987, pp. 251–288.

[25] Field, D. J.: Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells. J. of The Optical Society of America A., Dec. 1987, Vol. 4, No. 12, pp. 2379–2394.

[26] Chang, P. S.–Jian-Jiun, D.: Improved Harris' Algorithm for Corner and Edge Detections. IEEE Inter. Conf. Image Processing (ICIP), Vol. 3, Sept. 16–Oct. 19, 2007, pp. III-57–III-60.

[27] Allwein, E. L.–Schapire, R. E.—Singer, Y.: Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. Journal of Machine Learning Research, No. 1, 2001, pp. 113–141.

[28] Mokhber, A.—Achard, C.—Milgram, M.: Recognition of Human Behavior by Space-Time Silhouette Characterization. Pattern Recognition Papers, Vol. 29, 2008, No. 1, pp. 81–89.

[29] Carlsson, S.—Sullivan, J.: Action Recognition by Shape Matching to Key Frames. In Proc. Workshop on Models versus Exemplars in Computer Vision, 2001.

[30] Schindler, K.—van Gool, L.: Action Snippets: How Many Frames Does Human Action Recognition Require? IEEE Inter. Conf. Computer Vision and Pattern Recognition (CVPR), 23–28 June 2008, pp. 1–8.

Ning Li has been working towards the Ph. D. degree in Institute of Computer Science and Engineering, Beijing Jiaotong University, Beijing. His current research interests include action recognition and visual perception.

De Xu is now a Professor at the Institute of Computer Science and Engineering, Beijing Jiaotong University, Beijing. His research interest includes database systems and multimedia processing.

Lu Liu has been working towards the Ph. D. degree in School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing. Her current research interests include image annotation and retrieval.