LETTER

2D Log-Gabor Wavelet Based Action Recognition

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

SUMMARY The frequency response of log-Gabor function matches well the frequency response of primate visual neurons. In this letter, motion-salient regions are extracted based on the 2D log-Gabor wavelet transform of the spatio-temporal form of actions. A supervised classification technique is then used to classify the actions. The proposed method is robust to the irregular segmentation of actors. Moreover, the 2D log-Gabor wavelet permits more compact representation of actions than the recent neurobiological models using Gabor wavelet.

key words: *action recognition, log-Gabor function, motion-salient regions, visual neurons*

1. Introduction

Action recognition 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 research has focused on studying the computer vision-based features of moving objects such as the contour, interesting points, and 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. Primates outperform the best computer vision systems for action recognition, so building a system that emulates the primate visual cortex has always been an attractive idea.

Motivated by the object recognition mechanism [1], Jhuang et al. [2] and Ning et al. [3] use Gabor wavelet to emulate the representation of motion information in the primate visual cortex. The performance of these studies is better than typical computer vision-based algorithms for action recognition. However, Hawken et al. [4] suggest that the Gabor function 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.

Field [5] suggests that the log axis is the standard method for representing the spatial-frequency response of primate visual neurons. Theoretically, the frequency response of a log-Gabor function is symmetric on a log axis, while the Gabor function fails to capture the relative symmetry of the tuning curves on the axis. The frequency response has an extended tail at the high frequency, which spreads the information equally across the channels. Moreover, the bandwidth of the log-Gabor function increases with frequency. Therefore, the log-Gabor function permits a more compact representation than the Gabor function.

In this letter, an action sequence is firstly represented as the average motion energy (AME); the log-Gabor wavelet transform is then applied to take a multi-channel geometric analysis on the AME image; finally, the spatial motion-salient regions (MSR) is determined from the local energy images that show unique athletic property along different orientations. In the classification stage, a linear multi-class SVM classifier is applied.

In the experiment, by altering the configuration of filter parameter of the proposed approach, we prove the 2D log-Gabor wavelet allows more compact representation of actions than the 2D Gabor wavelet; by using the same filter configuration for our approach and neurobiological models that use Gabor wavelet to simulate object representation in the primate visual cortex, we prove the MSR of actions is more invariant to the unstable segmentation and deformation of actors.

The rest of the paper is organized as follows: Section 2 introduces the proposed approach. Experimental results are analyzed in Sect. 3. Conclusive remarks are addressed at the end of this letter.

2. The Proposed Approach

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 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), where x and y are the coordinates of pixels in the frames, and τ is the duration of a complete action. Figure 1 (a) shows the AME image for the example action "jack".

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

Manuscript received May 13, 2009.
Manuscript revised June 28, 2009.

[†]The authors are with the 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.2275

The AME image contains the body parts taking local movement (e.g. the limbs) and the parts taking global movement (e.g. the torso). The local movement provides the key clue to distinguish different action types. In the following section, the multi-channel geometric analysis on the AME image is proposed to extract the salient movement parts of actors.

2.2 Log-Gabor Wavelet Analysis on AME Image

Gabor function is usually used to simulate object perception in primate simple visual neurons. The study of Field [5] 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, therefore it can locate more precise local variation of natural image in the spatial domain. Another point in support of the log-Gabor function is that the frequency response of it has the Gaussian form when viewed on the logarithmic frequency scale, which is consistent with 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 (2) shows the 2D log-Gabor wavelet function in the frequency domain.

$$H_n^m(f,\theta) = \exp\left\{\frac{-\left[\ln(f/f_n)\right]^2}{2\left[\ln(\sigma_f/f_n)\right]^2}\right\} \times \exp\left\{\frac{-(\theta-\theta_m)^2}{2\sigma_\theta^2}\right\},$$
$$(n = 1\ldots N, m = 1\ldots M). \quad (2)$$

Where f_n represents the central frequency of the filter with the scaling index n; σ_f determines the bandwidth of the log-Gabor filter in the radial direction; σ_θ determines the bandwidth in the orientation direction; θ_m represents the orientation angle of the filter with the orientation angle index m. The central frequency is defined in Eq. (3):

$$f_n = \left(\lambda_1 \times s^{n-1}\right)^{-1}. \quad (3)$$

where the 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 orientation which is predefined empirically.

During the process of image analysis, the dot product between the log-Gabor filter set and the Fourier transform of an AME image is firstly calculated, and then a numerical inverse Fourier transform is performed to get the complex valued action images of different scales and orientations in the spatial domain. The complex valued action images are to be used to extract the motion-salient regions of actions.

2.3 Motion-Salient Regions

Primate visual system does not perceive input image as a whole entity, it processes the image as many elements in multiple orientations and scales [7]. Therefore, when we view the AME image of an action, in our visual cortex the spatio-temporal body shows different forms in accordance with orientations. These forms are named as local energy in this work.

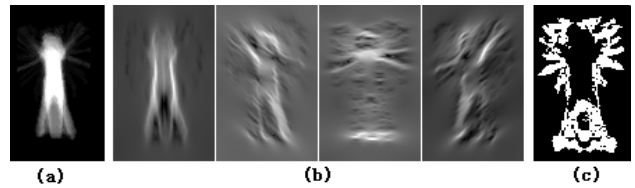

Fig. 1 The extraction of the MSR from the sample action "jack". (a) the AME image; (b) the local energy image in the orientation 0, $\pi/4$, $\pi/2$ and $3\pi/4$; (c) the MSR image obtained from the local energy images.

In each orientation, the local energy at the pixel (x, y) is calculated by adding the complex valued action images of multiple scales from head to tail, see Eq. (3):

$$E_m(x,y) = \sum_{n=1}^{N} A_n(x,y)$$
$$\times\left[\cos(\varphi_n(x,y) - \bar{\varphi}(x,y)) - |\sin(\varphi_n(x,y) - \bar{\varphi}(x,y))|\right]. \quad (4)$$

where $A_n(x,y)$ and $\varphi_n(x,y)$ respectively denote the amplitude and the phase angle of the complex valued image at scale n, and $\bar{\varphi}(x,y)$ denotes the phase angle of the local energy. Figure 1 (b) shows the local energy images for the action "jack" in 4 orientations, which simulate the representation of the AME image in the primate visual cortex. Each local energy image shows unique athletic property along different orientation.

The motion-salient regions (MSR) of an action is determined from these local energy images. The MSR consists of the body parts taking the most remarkable movement along different orientations, which means each component of the MSR has high intensity in certain orientation but does not show high intensity in other orientations in the local energy image. Therefore, the MSR is determined by taking groups of pixels having the highest intensity value from every local energy image. Figure 1 (c) shows the MSR image obtained from the local energy images.

2.4 Action Classification Using MSR

Features allow the association of an action to a point in a multidimensional feature space. Good features should have a large discriminatory capability and require a low computational time. Marco et al. [8] propose that a tradeoff between these requisites is the histograms of horizontal and vertical projections. In this letter, the horizontal and vertical histograms of a MSR image are firstly normalized by dividing every element by the total area, the normalized histograms are then aligned as a row vector.

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. For the multi-class classification problem we use the one-against-one strategy, where all pairs of classes are compared to each other. In this way, a multi-class problem is decomposed into multiple binary classification problems. To classify a testing action, we combine all the pairwise decision. That is, given an action dataset having N action categories, $N \times (N-1)/2$ tests need to be performed for one testing action; the action is then assign to the class that wins the most pairwise comparisons ("max-wins" rule).

3. Experiment Evaluation

3.1 Action Datasets

The Weizmann and Weiz.Robust action datasets are used in this letter. To increase the size of both datasets, each video is divided into 4 pieces. Figure 2 shows the example MSR images obtained from the datasets. The Weizmann provides 90 video sequences shown by nine subjects, each performing 10 types of natural actions repeatedly. The actions are "bend", "jack", "jump-forward", "jump-up-down", "run", "gallop-sideways", "skip", "walk", "wave-one-hand" and "wave-two-hands". We split the dataset as: 6 subjects are used as training set and the remaining 3 subjects are used for testing. The experiment is repeated by 25 random splits. The Weiz.Robust is designed for the robustness evaluation on recognition systems. It provides 10 walking actions present in various irregularities, which involve "walk with bag", "walk with case", "walk with dog", "knees-up", "limp", "moonwalk", "feet occluded by boxes", "normal walk", "body occluded by pole", "walk in skirt". Each type of walks is performed by only one subject. These "walk" actions are tested against the SVM classifier trained by the Weizmann dataset. Similar to the [2], we use a primitive attention mechanism: the input is a sequence of fixed-size image windows, centred at the person of interest.

3.2 The Benchmark Approaches

For benchmarking, we use the approaches proposed by Jhuang et al. [2] and Ning et al. [3]. The [2] speculates that neurons in intermediate visual areas of primate dorsal stream such as MT, MST respond to spatio-temporal features of target objects. Experimental results in [2] show that the sparse C3 feature using the space-time oriented based S1 units can best represent actions. This feature is used as the benchmark and is denoted as $StC3$ [2]. The code was graciously provided by Hueihan Jhuang. The [3] also proposes a neurobiological approach for action recognition, which is closely related to the feedforward template matching architecture for static object representation in the primate visual cortex. Both of the benchmarks use the Gabor wavelet as the key technique to analysis input frames.

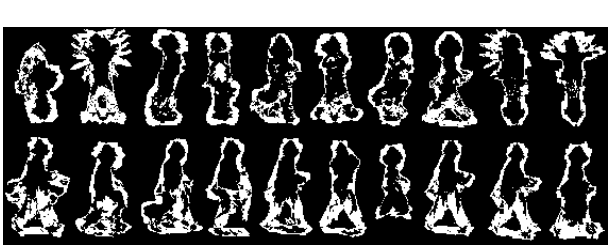

Fig. 2 First row: MSR images for 10 action categories in Weizmann. Second row: MSR images for 10 types of "walk" actions in Weiz.Robust.

3.3 Experiment Results

Theoretically, the frequency response of the log-Gabor filters should have minimal overlap necessary to achieve fairly even spectral coverage. In Eq. (3) the λ_1 is assigned 3 pixels; and in Eq. (2) $\sigma_\theta = \pi/(M \times 1.5)$, which results in approximately the minimum overlap needed to get even spectral coverage in the angular direction. Table 1 shows a series of parameter groups obtained experimentally, which result minimal overlap necessary to achieve fairly even spectral coverage in the radial direction. Figure 3 compares the action recognition results on the Weizmann dataset using the 4 parameter groups. The fourth group generates the highest recognition rate on each type of actions.

Field [5] concludes that when the parameter generates the bandwidth equals to about 1 octave, the log-Gabor wavelet has the same performance as the Gabor wavelet. Thus, the log-Gabor filter with the second parameter group can be used to simulate the Gabor filter. The [2] and [3] use 64 Gabor filters (4 orientations × 16 scales), thus in Table 2 the proposed approach uses 64 log-Gabor filters with the second parameter group for impartial comparison. In

Table 1 Parameter groups for 2D log-Gabor wavelet. s represents the scaling factor, BW denotes bandwidth.

Para. group	σ_f/f_i	s	BW (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	$BW \approx 2$

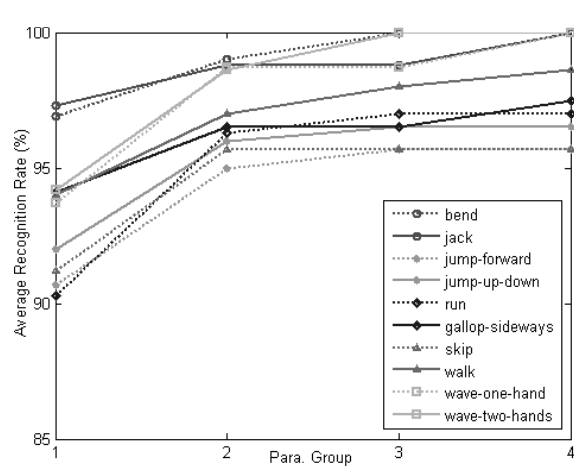

Fig. 3 Comparison of the recognition results obtained by the 2D log-Gabor wavelet using different parameter groups on Weizmann dataset.

Table 2 Comparison between the benchmarks using 64 Gabor filters and the proposed approach using 64 log-Gabor filters with the second parameter group.

Datasets	$StC3$ [2]	[3]	Proposed approach
Weizmann	**97.7**	97.3	97.5
Weiz.Robust	89.3	92.4	**94.2**

Table 3 Comparison between the benchmarks using 24 Gabor filters and the proposed approach using 24 log-Gabor filters with the fourth parameter group.

Datasets	$StC3$ [2]	[3]	Proposed approach
Weizmann	97.7	96.4	**98.1**
Weiz.Robust	87.5	90.0	**95.0**

Table 3, the proposed approach uses 24 log-Gabor filters (4 orientations $\times$ 6 scales) with the fourth parameter group (in the case of the fourth group, with the datasets providing the frame size about 120×80, a larger scale will not generate distinguishable output for different actions), therefore the benchmarks utilize 24 Gabor filters for fair comparison.

The advantage of the log-Gabor wavelet over the Gabor wavelet is then discussed. The recognition rates obtained from the proposed approach using 64 log-Gabor filters with the second parameter group in Table 2 are less than those obtained from the proposed approach using 24 log-Gabor filters with the fourth parameter group in Table 3. This directly proves the log-Gabor function covers much broader spectral bandwidth than the Gabor function.

The validity of the MSR feature is proved in Table 2, where the same number of filters and the same filter species are used for benchmarks and the proposed approach. In this way, the motion features are extracted based on the same criterion. Table 2 shows our approach has similar recognition rate with benchmarks on the Weizmann dataset, but performs the best on the Weiz.Robust dataset. Therefore, the MSR feature is more robust to the unstable segmentation and deformation of actors than the benchmark features.

Finally, the proposed approach is compared with the benchmarks on the Weizmann and Weiz.Robust dataset. In Table 3 the Gabor wavelet and the log-Gabor wavelet with the fourth parameter group are initialized by 4 filter orientation and 6 filter scales. Table 3 shows that the proposed approach outperforms the benchmarks especially on the Weiz.Robust dataset.

4. Conclusion

In this letter, the unique athletic property of a different type of action is represented by the motion-salient regions (MSR). The MSR is determined in terms of the local maximum of the 2D log-Gabor wavelet transform of the spatio-temporal form of actions. Experiments on publicly available action datasets demonstrate our approach outperforms some successful neurobiological approaches using Gabor wavelet. In the future work, more compact motion information will be extracted so that a snippet of action sequence is sufficient for action recognition. This would be more applicable for practical scenarios, where decisions have to be taken online.

References

[1] 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.

[2] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," IEEE 11th ICCV, pp.1–8, Oct. 2007.

[3] L. Ning and D. Xu, "Action recognition using visual neuron feature," IEICE Trans. Inf. & Syst., vol.E92-D, no.2, pp.361–364, Feb. 2009.

[4] M. Hawken and A. Parker, "Spatial properties of neurons in the monkey striate cortex," Proc. R. Soc., London Ser. B 231, pp.251–288, 1987.

[5] D.J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," J. Optical Society of America A., vol.4, no.12, pp.2379–2394, Dec. 1987.

[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. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require," 2008. IEEE Conference on CVPR, pp.1–8, June 2008.

[8] M. Leo, T. D'Orazio, and P. Spagnolo, "Human activity recognition for automatic visual surveillance of wide areas," Proc. 2nd ACM Inter. Workshop on Video Surveillance and Sensor Networks, pp.124–130, Oct. 2004.