Bio-Inspired Model with Dual Visual Pathways for Human Action Recognition

Bolun Cai*, Xiangmin Xu* and Chunmei Qing*
*School of Electronic and Information Engineering, South China University of Technology
Guangzhou, China
Email: caibolun@gmail.com, xmxu@scut.edu.cn, qchm@scut.edu.cn

***Abstract*—Spatio-temporal interesting points feature is a key technology for a wide class of computer vision approaches to recognize human actions. In this paper, a novel bio-inspired model based spatio-temporal interesting points (BIM-STIP) framework is proposed. Different from traditional STIP framework, the introduction of bio-inspired model provides a biological theory for interest points detection and spatio-temporal descriptor construction. Moreover, unlike existing bio-inspired model for action recognition with a single pathway, a dual pathway joint model with dorsal stream and ventral stream is constructed. Finally, this paper demonstrate how these feature to be used in a standard STIP framework for action recognition. Promising experimental results show that the proposed framework outperforms most of existing algorithms in action recognition, which encourages us to develop the BIM-STIP framework to other applications in future.**

Keywords*—Bio-Inspired Model, dual visual pathways, human action recognition.

I. Introduction

Human action recognition (HAR) in video continues attracting significant attention from the computer vision community. However, it still remains a challenge in computer vision due to different appearances of people, unsteady background, moving camera, illumination changes. On the other hand, HAR is an interdisciplinary research, which not only contains computer vision, pattern recognition and machine learning, but also biology, cognitive science, psychology.

Nowadays many motion information extraction and classification methods are adopted to analyze human action in the video. The bulk of the research focus primarily on spatio-temporal interesting points framework, where approaches derived from bag-of-words dominate. Paul *et al.* [1] used a bag of 3D-SIFT descriptors approach to represent videos. In [2], researchers proposed a direct 3D interest point detectors and devised a recognition algorithm based on spatio-temporally windowed data. Willems *et al.* [3] proposed the extended SURF (ESURF) descriptor which extends the image SURF descriptor to videos. These feature descriptors, process the temporal dimension information and the spatial dimension information independently. Therefore, the small distance between the different behaviors could not be described well. In addition, probability network methods [4], human limb movements feature [5], neural network [6] and so on are being applied in HAR field gradually. However, design of these models relies on subjective experience excessively.

As is well known, the primate biological visual system can achieve the unification of invariance and distinguish between different objects easily. As far as the visual system is concerned, one can make a similar observation asking whether motion is also essential to recognize actions. If biological visual perception mechanism was applied to the traditional signal processing model, a novel feature could be extracted. In this paper, a bio-inspired model based spatio-temporal interesting points (BIM-STIP) framework is proposed for human action recognition. Our contributions are as follows: (1) A novel spatio-temporal interesting points detector based on BIM, which simulates the dorsal stream of primate visual system, that performs well in terms of multi-scale and robustness. (2) Formulation of BIM-STIP descriptor, which includes a dorsal stream feature (DSF) and a ventral stream feature (VSF), that captures local characteristics of human actions accurately. (3) Comparative analysis of our BIM-SIFT framework with previous STIP descriptors used for the same purpose.

This paper is organized as follows: Section II summarizes the related work. Section III presents the proposed BIM-STIP framework for HAR in detail, which mainly consists of BIM-STIP detector and BIM-STIP descriptor. Section IV provides the evaluation of the proposed framework and experimental results. Section V gives the conclusion of this paper.

II. Related Work

Primate visual system [7] is a complex system, which involves hypothalamus, brain and the other neural activities. The optical signal is transformed into a nerve pulse signal on retina cells, and then the signal turns to LGN in thalamus. After that, the visual information is transmitted to the visual cortex, which is organized in two different pathways: a ventral stream and a dorsal stream. These two pathways originate in the primary visual cortex (V1), which extracts location, multi-scale and multi-orientation feature. Complex visual cells (V2) are the relay store, which preserves visual information and determines the flow of stimulation signal. A part of visual information is tuned to spatial orientations and project to cells of the ventral stream, and the order is sensitive to directions of motions and project to MT area in the dorsal stream.

Jhuang *et al.* [8] presented a biologically-motivated system, which is based on hierarchical feed-forward architectures and extends a neurobiological model in the visual cortex. The bio-inspired system is constructed modeling the form-processing pathway at the level of the V1 area with simple and complex cells essentially. As far as the visual system is concerned, one can make a similar observation asking whether motion is also essential to recognize actions. Ungerleider and Mishkin [9] found that there are two different pathways in the visual cortex of a monkey, a motion-processing pathway (dorsal stream) and a form-processing pathway (ventral stream).

978-1-4799-2581-0/14/$31.00 ©2014 IEEE

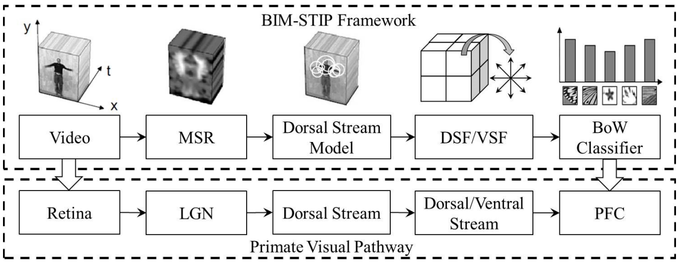


Figure 1. Bio-inspired Model corresponding to primate visual pathway

Escobar *et al.* [10] propose a new motion-based feature only modeling the dorsal stream, focusing on cortical areas V1 and MT. As we know, the primate visual cortex is a multi-pathway complex system: low-level nerve stimulation will be integrated into high-level nerve cells for comprehensive discrimination. Riesenhuber *et al.* [11] found that the ventral stream and the dorsal stream are combined to complete object and action recognition. In [12], authors show that combining motion feature (HoF) with shape feature (HoG) outperforms a single feature.

In this article our goal is to construct a novel spatio-temporal interesting points framework for action recognition, inspired by visual system processing. To achieve this goal, this framework models a motion-processing pathway and a form-processing pathway. At the phase of STIP detection, a STIP detector is designed based on the dorsal stream model. In dorsal stream model, MT cells pool the shape information from V1 to the speed and direction of moving visual stimuli, so that MT plays a significant role in the processing of visual motion. The bio-inspired feature proposed in this article will be defined from the dorsal stream and the ventral stream. The purpose of ventral stream can be thought of as being similar to many spatially local and complex Fourier transforms to extract shape feature, and the purpose of dorsal stream can be thought of as being similar to a spatio-temporal transform to extract motion feature.

III. BIM-STIP FRAMEWORK

In this section, we will present the proposed BIM-STIP framework. According to primate visual pathway, a proposed bio-inspired framework proposed is shown Fig. 1. In our framework, video corresponds to retinal stimulation, motion salient region (MSR) simulates LGN, a multi-layer dorsal and ventral model simulates the visual dorsal and ventral stream, and a bag of words (BoW) classifier is used to simulate prefrontal cortex (PFC).

There are two major parts for human action recognition in image sequence as illustrated in Fig. 2. In the first part, a BIM-STIP detection method is proposed to find interesting points in the image sequence. Then multi-scale spatio-temporal domains are constructed by interesting points for BIM-STIP descriptors (VSF and DSF) extraction in Part 2. Finally, a visual vocabulary is similar to the memory effect of the cerebral cortex, so BoW classifier [13] is regarded as the memory pool of PFC.

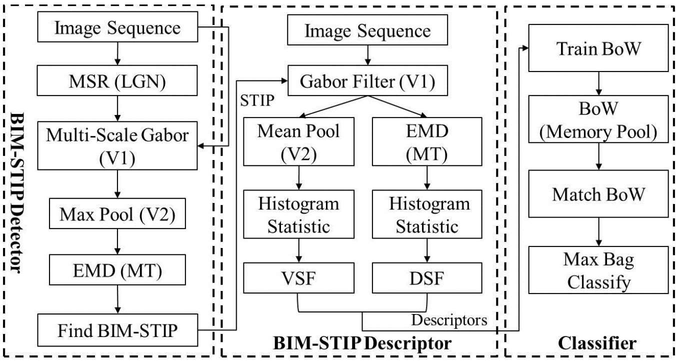


Figure 2. BIM-STIP Framework

A. *BIM-STIP detector*

In order to detect BIM-STIP, a method simulating LGN and the dorsal stream is proposed in Fig. 1. Motion salient region (MSR) is used for simulating the LGN in finding the coarse interesting region and reduces the computational amount of the dorsal stream model. The multi-layer dorsal stream model, including V1 layer, V2 layer and MT layer, extracts interesting points accurately.

1) LGN - spatial attention regulation: LGN, shaped like a saddle, is located in the rear end of the optic tract and lateral hypothalamus pillow. Through the macaque visual cortex research, McAlonan and Cavanaugh *et al.* [14] show that LGN plays an important role about spatial attention regulation. In our framework, MSR [15] is extracted to analyze the image sequence in LGN. Because interesting point extraction will be refined in the dorsal stream model, a simple method is used to calculate locality motion energy (LME) for MSR extraction. In this framework, pixel change probability map (PCPM) [16] is used as follows:

$$\begin{aligned} P\left(x,y,t\right) = \eta \times P\left(x,y,t-1\right) + \left(1-\eta\right) \times \\ \left|I\left(x,y,t\right) - I\left(x,y,t-1\right)\right| \end{aligned} \tag{1}$$

, where the forgetting factor is denoted as η and the original grayscale image as $I\left(x,y,t\right)$. On the other hand, it relies on integral images [17] $PI\left(x,y,t\right)$ showed as (2) to reduce the computation time of LME.

$$PI\left(x,y,t\right) = \sum_{(x,y)=(0,0)}^{(x,y)} P\left(x,y,t\right) \tag{2}$$

LME is calculated as (3)

$$\begin{aligned} LME\left(x,y,t,w,h\right) = \frac{1}{wh}[PI\left(x+w,y+h,t\right) + \\ PI\left(x,y,t\right) - PI\left(x+w,y,t\right) PI\left(x,y+h,t\right)] \end{aligned} \tag{3}$$

LME statistics method extracts MSR, and then the dorsal stream model detects STIP at the interesting region for further refinement. MSR extraction can reduce the computation amount of the dorsal stream model significantly as well as the computational complexity of the overall framework.

2) V1 - primary visual feature extraction: In the primary visual cortex (V1), simple cell receptive field has the basic characteristics of multi-directional, multi-scale and multi-frequency selection [18]. Therefore, they can be approximated by a series of two-dimensional filters to meet the time-frequency uncertainty principle. In [19], the authors suggested that several properties of simple or complex cells in V1 can be described by energy filters, in particular by Gabor patches.

The V1 responses are obtained by applying a series of Gabor filters to the input image , which can be described by the following equation:

$$G_{even}\left(\cdot,\theta,s\right)=\exp\left(-\frac{X^2+\gamma^2Y^2}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}X\right) \tag{4}$$

, where $X = xcos\theta + ysin\theta$,$Y = -xcos\theta + ysin\theta$, Gabor patches' coordinate(x, y), orientation θ, scales s with 3 parameters(effective width δ, wavelength λ and aspect-ratio γ).

According to [20], this model selects a series of Gabor filters, which sizes from 7 to 37 , to model the receipt field (RF) of V1 cells ξ as Table I. The filters come in 4 orientations θ and 16 scales s ($16 \times 4 = 64$ patches) that are arranged in 8 bands ε. The responding result of V1 can be described by (5).

$$V_1\left(\cdot,t,\theta,s\right)=I\left(\cdot,t\right)*G_{even}\left(\cdot,\theta,s\right) \tag{5}$$

, where $I\left(\cdot,t\right)$ is the original gray-scale image of the video.

3) V2 - scale, shift and orientation invariance: The complex visual cells (V2) receive the response of simple cells from V1, and play the role of primary linear feature integration. The competition mechanism of V2 layer's cells can be divided into scale, orientation and shift. Ilan *et al.* [21] researched the spatial integration properties of complex cells and proposed which responses can be described by a maximum operation. Corresponds to complex cells which show some invariance to scale, shift and orientation: complex cells have more broadly spatial frequency (scale invariance), larger receptive fields(shift invariance) and respond to multiple texture structure (orientation invariance).

Because the shift and orientation information will be used in MT, the scale competition is only done in V2 under our framework as follow:

$$V_2\left(x,y,t,\theta,\varepsilon\right)=\underset{s\in\{2\varepsilon,2\varepsilon-1\}}{\mathrm{Max}}V_1\left(\cdot,t,\theta,s\right) \tag{6}$$

The shift and orientation competition will be done after MT respond extraction.

4) MT - higher order motion analysis: The middle temporal (MT) of visual cortex plays an important role in local motion perception and in guiding eye movements. The electrophysiological characteristics research of the MT neurons shows that MT cells have a stronger stimulus response about motion speed and direction. An elementary motion detector (EMD) [22] proposed by Reichardt *et al.* turned out to be similar to the motion detection unit of the amitcontextprimate visual pathway. The research by Ron *et al.* [23] prove that realistic Reichardt correlators can provide accuracy velocity estimation in a natural visual environment. In our framework, EMDs are used to simulate MT area, which combines spatial and temporal information well.

EMD is a 2D motion detector, which is described by (7).

$$R\left(x,t\right)=F_A\left(t-\tau\right)F_B\left(t\right)-F_A\left(t\right)F_B\left(t-\tau\right) \tag{7}$$

, where$F_A\left(t\right)=F\left(x,t\right)$, $F_B\left(t\right)=F\left(x+\Delta\Phi,t\right)$ with spatial shifts $\Delta\Phi$, $F\left(x,t\right)$ is the input signal and τ is the temporal delay. The process result from V2 include 8 bands and 4 orientations, which will be tuned to MT layer. 4 EMDs are choosed corresponding to the different orientations of V2. When the offset direction of EMD is orthogonal to the Gabor filter direction, EMD is the most sensitive for motion. Therefore, the MT response and input signal can be represented as follow:

$$\begin{cases} MT(\cdot,t,\theta,\varepsilon)=F'_A(t-\tau)F'_B(t)-F'_A(t)F'_B(t-\tau) \\ F'_A\left(t\right)=F'(\cdot,t,\theta,\varepsilon,0) \\ F'_B\left(t\right)=F'(\cdot,t,\theta,\varepsilon,\Delta\Phi) \end{cases} \tag{8}$$

, where $F'=V_2\left(x+\Delta\Phi\sin\theta,y+\Delta\Phi\cos\theta,t,\theta,\varepsilon\right)$.

5) BIM-STIP detection strategy: Through the process of the dorsal stream model, a MT biologically inspired response map $MT\left(x,y,t,\theta,\varepsilon\right)$ will be obtained. A series of detection strategy is proposed to find spatio-temporal interesting points, which include shift and orientation competition, location refine and global optima. The BIM-STIP detection strategy will be show in Alg. 1.

Firstly, the orientation competition will be done for the MT response, and the orientation competition (OC) equation is showed by (9).After that, interesting points are detected by local maximum in the different scale-space to simulate the nervous shift competition. In order to refine interesting points, local refinement and global optima operation will be done in the end.

$$OC\left(x,y,t,\varepsilon\right)=\underset{\theta}{\mathrm{Max}}\left\{MT\left(x,y,t,\theta,\varepsilon\right)\right\} \tag{9}$$

Algorithm 1 BIM-STIP detection strategy

```
Input: $OC(x,y,t,\varepsilon)$, $LME(x,y,w,h)$
Output: vector<Keypoint> $points$
  \\Local maximum
  while $(x,y)$ do
    if $LME(x,y,\Sigma(\varepsilon),\Sigma(\varepsilon)) < lmeThres$ then
      continue
    end if
    $keypoint \leftarrow \underset{(x,y)\in\Sigma(\varepsilon)}{\mathrm{argmax}}\ OC(x,y,t,\varepsilon)$
    $points$.push[$keypoint$]
  end while
  \\Locality refine
  while $(i,j)$ do
    if $\|points[i]-points[j]\|_2{}^{1/2} < dThres$ then
      $points.\mathrm{pop}\left[\underset{k\in\{i,j\}}{\mathrm{argmin}}\,points[k].respond\right]$
    end if
  end while
  \\Global optima
  $P_{(1)} < P_{(2)} < ... < P_{(n)} \leftarrow points.respond$
  $points \leftarrow P_{(n)}, P_{(n-1)}, ..., P_{(n-N)}$
```

B. BIM-STIP descriptor

To describe spatio-temporal interesting points, a BIM-STIP descriptor is proposed showed as Fig. 2. The construction of BIM-STIP descriptor is based on the dorsal stream model and the ventral stream model. A motion feature, called dorsal

TABLE I. SUMMARY OF PARAMETERS USED IN BIM-STIP FRAMEWORK

Band ε	1		2		3		4		5		6		7		8	
Scale s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
RF ξ	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	37
δ	2.8	3.6	4.5	5.4	6.3	7.3	8.2	9.2	10.2	11.3	12.3	13.4	14.6	15.8	17.0	18.2
λ	3.5	4.6	5.6	6.8	7.9	9.1	10.3	11.5	12.7	14.1	15.4	16.8	18.2	19.7	21.2	22.8
γ	0.23	0.28	0.32	0.37	0.41	0.46	0.51	0.55	0.60	0.64	0.69	0.74	0.78	0.83	0.87	0.92
Σ	8		12		16		20		24		28		32		36	
θ	$\pi/4$ $\pi/2$ $3\pi/4$ π															

stream feature(DSF), simulates V1 and MT area to describe the movement characteristic of HAR. Meanwhile, a shape feature, called ventral stream feature (VSF), simulates V1 and V2 area to describe the appearance characteristic of HAR. Therefor, BIM-STIP descriptor integrate the motion and shape feature to improve the representation ability compared to other bio-inspired frameworks.amitcontext

The previous operations via BIM-STIP detector have assigned a location and scale to each interesting point. These parameters provide invariance to describe the local video region. The next step is to compute DSF and VSF for the local video region that is highly distinctive yet is as invariant as possible to remaining variations, such as change in illumination or 3D viewpoint. Here space-time volumes are defined in the neighborhood of detected points. The size of each volume $(\Delta x, \Delta y, \Delta t)$ is related to the detection scales by $\Delta x = \Delta y = \Sigma$ (Σ is selected as Table I) and Δt equates 8 frames in this paper. Each volume is subdivided into $a(n_x, n_y, n_t)$ grid of cuboids; for each cuboid computing the histograms of DSF and VSF. Histograms are concatenated into DSF and VSF descriptor vectors, which are similar in spirit to the well known SIFT descriptor [24]. Parameter values are selected here as $n_x = n_y = n_t = 2$.

1) DSF - BIM motion feature: DSF is extracted through the simulation of V1 and MT area. Each cuboid is filled by a weighted sum of uniformly sampled responses of dorsal stream model, which is calculated similarly to spatio-temporal interesting points detection.

Firstly, V1 response is obtained in the space-time cube. In contrast to previous interesting points scale detection methods, for extracting detailed feature, the smallest scale space ($s = 1$ and $\xi = 7$) is selected for simulating V1 respond as (10).

$$V_1R_{even}(\cdot, t, \theta) = I(\cdot, t) * G_{even}(\cdot, \theta, s = 1) \qquad (10)$$

Second, the primary visual feature is turned to MT function to obtain the dorsal stream response showed by (11). To reserve detail, the time offset is set to a minimum value ($\tau = 1$) and the space offset is set to the main lobe width of smallest scale Gabor filter ($\Delta\Phi = 4$).

$$MTR(\cdot, t, \theta) = F'_A(t - \tau)F'_B(t) - F'_A(t)F'_B(t - \tau) \qquad (11)$$

,where $F'_A(t) = V_1R_{even}(x, y, t, \theta)$ and $F'_B(t) = V_1R_{even}(x + 4\sin\theta, y + 4\cos\theta, t, \theta)$. For EMDs, the symbol of MT response represents the motion direction. Finally, a 64-dimensional histogram is calculated according to 8 cuboids, 4 orientations and 2 response symbols.

2) VSF - BIM shape feature: VSF is a shape feature extracted by the simulation of V1 and V2 area. Same as DSF extraction, the ventral stream respond is also calculated in the neighborhood of each interesting point.

For the motion feature, the movement direction of the texture is more concerned; while for the shape feature, the gradient direction of the texture is more important. Therefore, different from the V1 response calculation of DSF, an other Gabor filter patch is selected which can extract not only the texture intensity, but also the texture gradient direction. As a result, V1 area is simulated by a Gabor even function DSF extraction, and by a Gabor odd function as (12) in extracting VSF. The response symbol represents the gradient direction of the texture.

$$G_{odd}(\cdot, \theta, s) = \exp\left(-\frac{X^2 + \gamma^2 Y^2}{2\sigma^2}\right) \sin\left(\frac{2\pi}{\lambda} X\right) \qquad (12)$$

V1 respond result can be described by (13).

$$V_1R_{odd}(x, y, t, \theta) = I(\cdot, t) * G_{odd}(\cdot, \theta, s = 1) \qquad (13)$$

The maximum pool operation can increase the feature robustness, but also cause the loss of useful information. In order to keep more effective information, a mean pool is choosed here to simulate V2 instead of maximum pool as showed as (14). Finally, as the histogram statistic of DSF, a 64-dimensional histogram is got at each interesting point according to 8 cuboids, 4 orientation and 2 response symbols.

$$V_2R(x, y, t, \theta) = \underset{\Sigma}{\mathrm{Mean}} \left\{V_1R_{odd}(x, y, t, \theta)\right\} \qquad (14)$$

IV. EXPERIMENTAL RESULTS

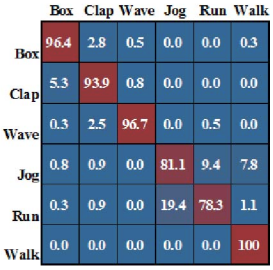
We conduct an extensive set of experiments to evaluate the performance of the proposed action recognition system on two publicly available human action databases: KTH [25] and Weizmann [26].

A. Results of human action databases

KTH database contains 25 subjects performing 6 actions: boxing, handclapping, handwaving, jogging, running and walking. The sequences are separated in four different scenarios: outdoors, outdoors with scale variations, outdoors with different vestment, and indoor with lighting variations. A five actions KTH database [10] without running is utilized to ignore the confusion of jogging and running.

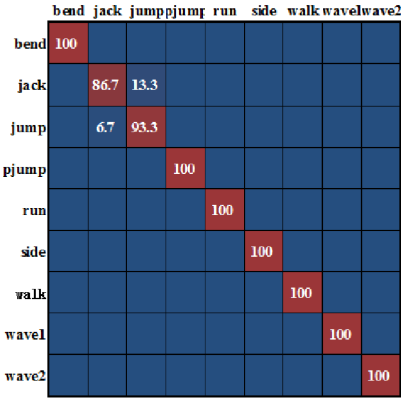
Weizmann database contains 9 subjects performing 9 actions: bending (bend), jumping jack (jack), jumping forward on two legs (jump), jumping in place on two legs (pjump), running (run), galloping sideways (side), walking (walk), waving one hand (wave1) and waving two hands (wave2).

For benchmark we get KTH examination result using [10], [25], and Weizmann examination result using [8], [10]. According to the subjects, the KTH database is divided into a training set (16 persons) and a test set (9 persons). The presented recognition results are obtained on the test set by



	Box	Clap	Wave	Jog	Run	Walk
Box	96.4	2.8	0.5	0.0	0.0	0.3
Clap	5.3	93.9	0.8	0.0	0.0	0.0
Wave	0.3	2.5	96.7	0.0	0.5	0.0
Jog	0.8	0.9	0.0	81.1	9.4	7.8
Run	0.3	0.9	0.0	19.4	78.3	1.1
Walk	0.0	0.0	0.0	0.0	0.0	100

Figure 3. Confusion matrices obtained for KTH dataset



	bend	jack	jump	pjump	run	side	walk	wave1	wave2
bend	100								
jack		86.7	13.3						
jump		6.7	93.3						
pjump				100					
run					100				
side						100			
walk							100		
wave1								100	
wave2									100

Figure 4. Confusion matrices obtained for Weizmann dataset

10 random cross validations. In Weizmann database, we select six random subjects as a training set ($6 \times 9 = 54$ videos) and use the remaining three subjects as a testing set ($3 \times 9 = 27$ videos). The confusion matrices of recognition performance are shown in Fig. 3 and Fig. 4 for KTH and Weizmann database, respectively.

B. Comparisons with other bio-inspired methods

Table II demonstrates the experimental results in comparison with other bio-inspired methods. With the double visual pathways the proposed approach shown as DSF/VSF can obtain the recognition accuracy up to 91.1%, 96.9% and 97.8% for KTH(6 actions), KTH(5 actions) and Weizmann database, respectively. Our method outperforms bio-inspired feature (BIF) [10] in any database. Different from the benchmark, bio-inspired system (BIS) [8] obtains 91.7% in KTH database for each scenario separately, and it is only slightly higher than our approach for all scenarios. Moreover BIS needs to preprocess the database to identify the foreground pixels of each frame. In summary, BIM-STIP framework **outperforms existing bio-inspired methods**.

TABLE II. COMPARISON TO BIO-INSPIRED METHODS

Method	KTH (6 actions)	KTH (5 actions)	Weizmann
BIS [8]	**91.7%**	-	96.3%
BIF [10]	83.8%	92.4%	95.3%
DSF	88.8%	95.7%	90.4%
VSF	88.7%	94.5%	91.9%
DSF/VSF	91.1%	**96.9%**	**97.8%**

TABLE III. COMPARISON TO THE STATE-OF-ART ON KTH

Method		KTH (6 actions)	KTH (5 actions)
STIP	LF [25]	71.7%	81.8%
	VF [30]	63.0%	81.76%
	Dollár [2]	81.2%	88.6%
	E-SURF [3]	84.3%	90.0%
	HOG/HOF [12]	91.8%	96.2%
	STW [31]	83.3%	91.6%
State-of-art	Unified [32]	87.3%	93.6%
	Speech [33]	90.3%	96.2%
	SMT [27]	91.7%	93.4%
	HOG-OF [28]	**94.3%**	93.6%
	MTP [29]	92.5%	96.5%
Ours	DSF	88.8%	95.7%
	VSF	88.7%	94.5%
	VSF/DSF	91.1%	**96.9%**

C. Comparisons with representative algorithms

The experimental results of different representative algorithms are illustrated in Table III. On KTH (6 actions) we just obtain 91.1% recognition accuracy, which is not best but comparable to 91.7% in [27], 94.5% in [28] and 92.5% in [29]. It should be noticed that on this database those methods use a complex classifier or multiple descriptors combined to obtain better performance.

In this paper, our main contribution is a new bio-inspired interesting points detector and a novel spatio-temple feature extractor. Therefore, we mainly concern STIP framework using the standard setting here. Our method obtains 91.1% accuracy, which is slightly lower than HOG/HOF [12] (91.8%). Compared with HOG/HOF, our framework has the following advantages. Firstly, our framework does not have parameter selection problem. However, HOG/HOF method must to select different parameters and set grid according to different databases to achieve best performance. And experiments demonstrate that parameter selection can cause a significant impact on the result. Secondly, BIM-STIP detector introduce a coarse-to-fine idea and the DSF extraction is more simple than optical flow feature than HOG/HOF extraction, which means the computation cost of BIM-STIP will be highly reduced.

From the above analysis, it can be seen that the proposed Framework can get comparable performance in most actions (including boxing, handclapping, hand waving and walking). For bio-inspired model, jogging and running have a high similarity and are hard to be distinguished. Following the experiments in [10], a KTH sub-database (5 actions) is used to assist the evaluation. From Table III, it can be seen that the proposed approach obtains the best performance. Optimizing the speed characterization of DSF to improve the recognition rate between jogging and running will be our future work.

V. CONCLUSION

In this paper, a novel spatio-temporal interest points framework is proposed based on bio-inspired model with double visual pathways. Firstly,it is analyzed that by using the determinant of the biologically inspired model, the point localization and scale-selection can be combined in a direct way. Secondly, an integrate shape and motion feature is proposed from the two visual pathways. Finally, BIM-STIP framework have been developed for action recognition. Extensive experiments support the conclusion that the proposed algorithm outperforms all representative existing techniques with lower computing cost and higher processing speed.

In biological visual system, location interest points are some of the key feature; besides, dense trajectories [34], structured information [35] also play an important role. Further, we will develop a joint framework, which allows for an efficient computation of scale-invariant feature.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No.61171142), the Science and Technology Planning Project of Guangdong Province of China (No.2011A010801005, 2010A080402015), and the Fundamental Research Funds for the Central Universities (No.2013ZM0081).

REFERENCES

[1] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 357–360.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[3] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 650–663.

[4] X. Li, "Hmm based action recognition using oriented histograms of optical flow field," *Electronics Letters*, vol. 43, no. 10, pp. 560–561, 2007.

[5] J. W. Davis and S. R. Taylor, "Analysis and recognition of walking movements," in *16th International Conference on Pattern Recognition*, vol. 1. IEEE, 2002, pp. 315–318.

[6] W. Hu, D. Xie, T. Tan, and S. Maybank, "Learning activity patterns using fuzzy self-organizing neural network," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 3, pp. 1618–1626, 2004.

[7] D. C. Van Essen, C. H. Anderson, D. J. Felleman *et al.*, "Information processing in the primate visual system: an integrated systems perspective," *Science*, vol. 255, no. 5043, pp. 419–423, 1992.

[8] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[9] M. Mishkin, L. G. Ungerleider, and K. A. Macko, "Object vision and spatial vision: two cortical pathways," *Trends in neurosciences*, vol. 6, pp. 414–417, 1983.

[10] M.-J. Escobar and P. Kornprobst, "Action recognition via bio-inspired features: The richness of center–surround interaction," *Computer Vision and Image Understanding*, vol. 116, no. 5, pp. 593–605, 2012.

[11] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*. IEEE, 2008, pp. 1–8.

[13] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Machine learning: ECML-98*. Springer, 1998, pp. 4–15.

[14] K. McAlonan, J. Cavanaugh, and R. H. Wurtz, "Guarding the gateway to cortex with attention in visual thalamus," *Nature*, vol. 456, no. 7220, pp. 391–394, 2008.

[15] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1997, pp. 568–574.

[16] Y.-H. Qin, H.-L. Li, G.-H. Liu, and Z.-N. Wang, "Human action recognition using pem histogram," in *2010 International Conference on Computational Problem-Solving (ICCP)*. IEEE, 2010, pp. 323–325.

[17] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 404–417.

[18] J. G. Daugman *et al.*, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Optical Society of America, Journal, A: Optics and Image Science*, vol. 2, no. 7, pp. 1160–1169, 1985.

[19] H. Knutsson, R. Wilson, and G. Granlund, "Anisotropic nonstationary image estimation and its applications: Part i–restoration of noisy images," *IEEE Transactions on Communications*, vol. 31, no. 3, pp. 388–397, 1983.

[20] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.

[21] I. Lampl, D. Ferster, T. Poggio, and M. Riesenhuber, "Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex," *Journal of neurophysiology*, vol. 92, no. 5, pp. 2704–2713, 2004.

[22] W. Reichardt, "Autocorrelation, a principle for the evaluation of sensory information by the central nervous system," *Sensory communication*, pp. 303–317, 1961.

[23] R. O. Dror, D. C. O'Carroll, and S. B. Laughlin, "Accuracy of velocity estimation by reichardt correlators," *JOSA A*, vol. 18, no. 2, pp. 241–252, 2001.

[24] D. G. Lowe, "Object recognition from local scale-invariant features," in *The proceedings of the seventh IEEE international conference on Computer vision, 1999.*, vol. 2. Ieee, 1999, pp. 1150–1157.

[25] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36.

[26] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005.*, vol. 2. IEEE, 2005, pp. 1395–1402.

[27] I. Jargalsaikhan, S. Little, C. Direkoglu, and N. E. O'Connor, "Action recognition based on sparse motion trajectories," 2013.

[28] F. Baumann, "Action recognition with hog-of features," in *Pattern Recognition*. Springer, 2013, pp. 243–248.

[29] T. P. Nguyen, A. Manzanera, and M. Garrigues, "Motion trend patterns for action modelling and recognition," in *Computer Analysis of Images and Patterns*. Springer, 2013, pp. 360–367.

[30] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005.*, vol. 1. IEEE, 2005, pp. 166–173.

[31] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[32] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen, and D. Zhao, "A unified framework for locating and recognizing human actions," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 25–32.

[33] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, "Human action recognition using star skeleton," in *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*. ACM, 2006, pp. 171–178.

[34] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3169–3176.

[35] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.