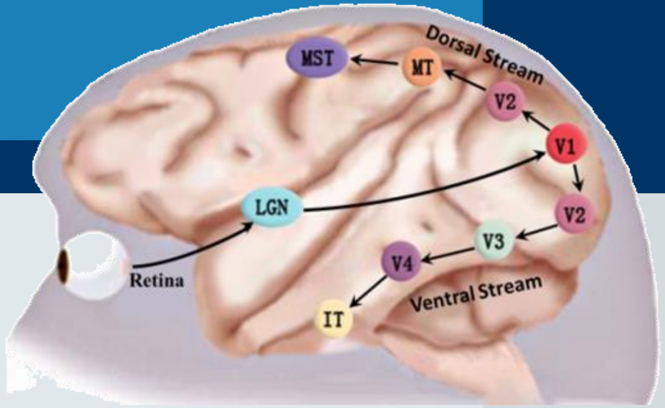
Bio-inspired Model with Dual Visual Pathways for Human Action Recognition

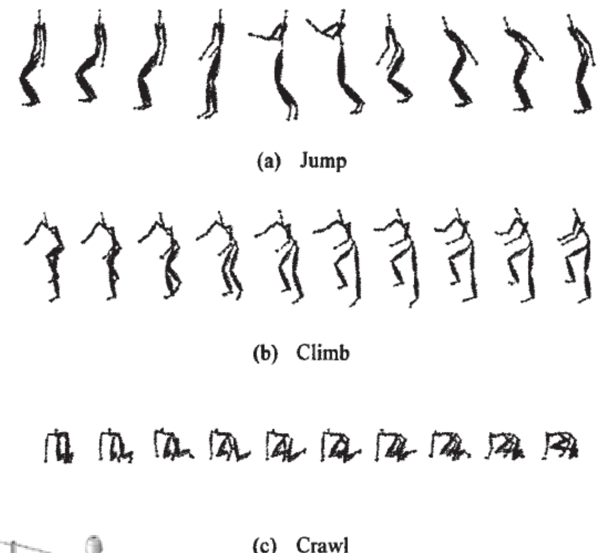
Bolun Cai, Xiangmin Xu and Chunmei Qing

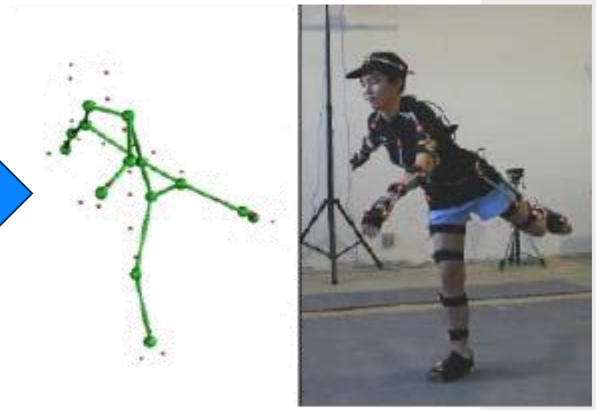
South China University of Technology

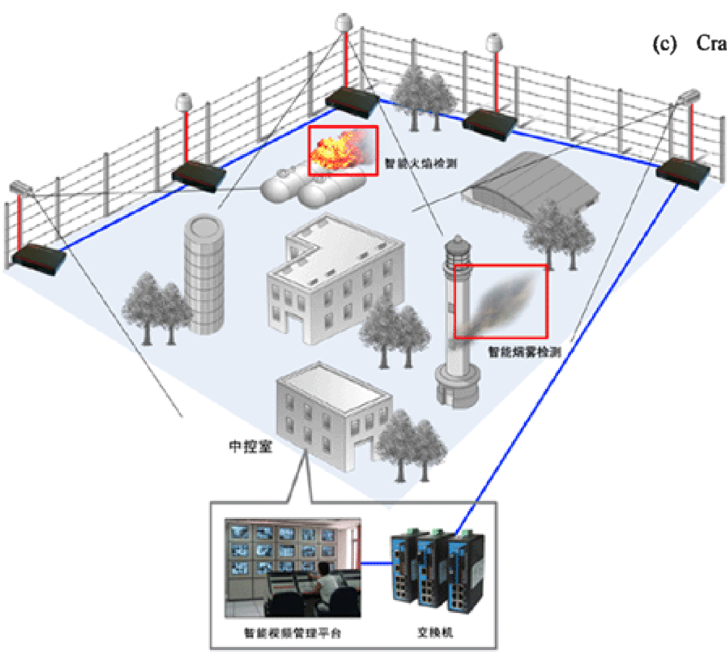


This work is supported by National Natural Science Foundation of China (No.61171142), the Science and Technology Planning Project of Guangdong Province of China (No.2011A010801005, 2010A080402015), and the Fundamental Research Funds for the Central Universities (No.2013ZM0081)

Human Action Recognition (HAR)

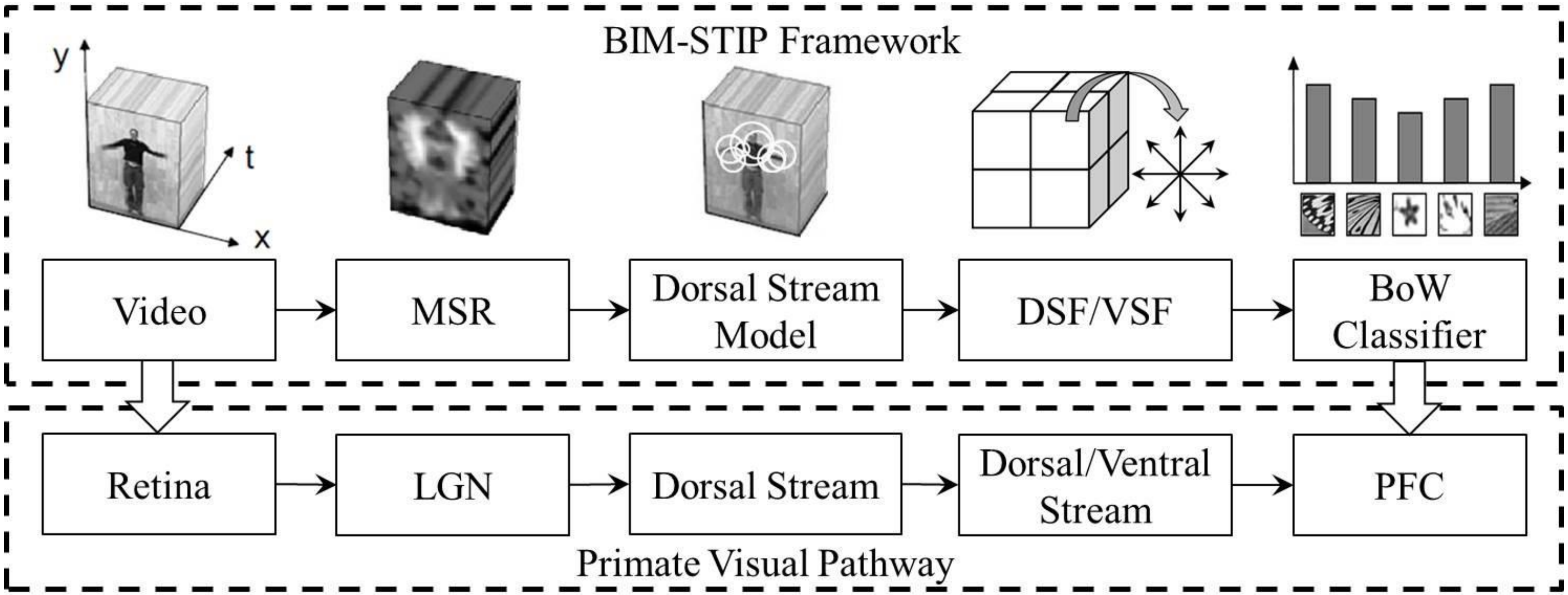


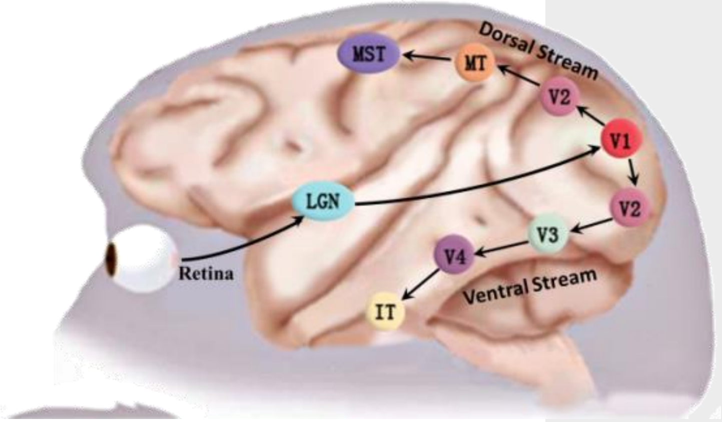




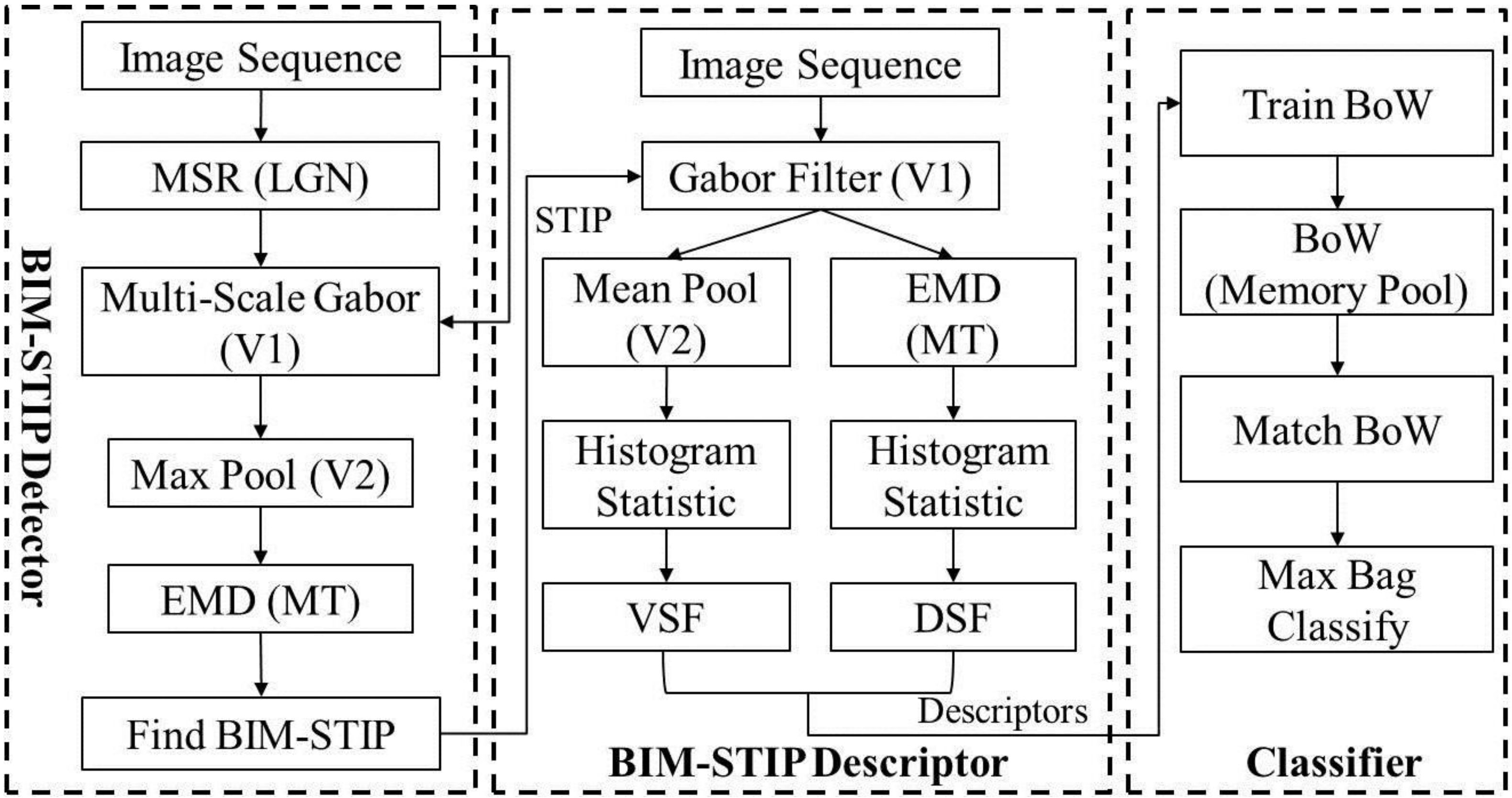
flickr
Google
YouTube

BIM correspond to visual pathway



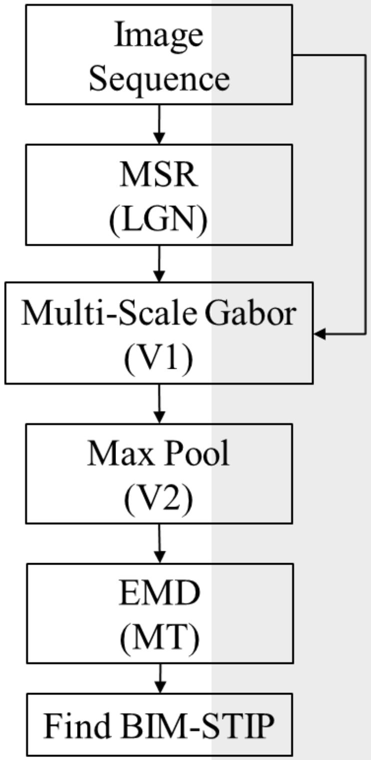


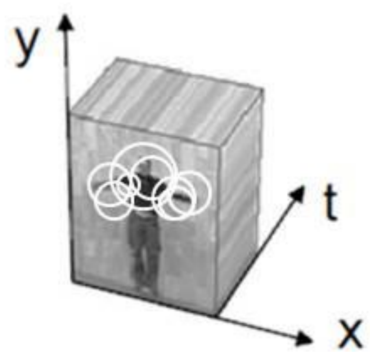
BIM-STIP Framework



BIM-STIP Framework

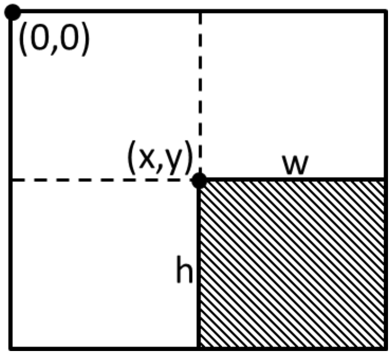
BIM-STIP DETECTOR





LGN - *spatial attention regulation*

- Pixel Change Probability Map (PCPM)
  - $P(x, y, t) = \eta P(x, y, t-1) + (1-\eta)|I(x, y, t) - I(x, y, t-1)|$
- Integral Images
  - $PI(x, y, t) = \sum_{(x,y)=(0,0)}^{(x,y)} P(x, y, t)$
- Locality Motion Energy (LME)
  - $LME(x, y, t, w, h) = (1/wh)\,[PI(x+w, y+h, t) + PI(x, y, t) - PI(x+w, y, t) - PI(x, y+h, t)]$



V1 and V2

- V1 - *primary visual feature extraction*
  - Even Gabor filters

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

$$X = x\cos\theta + y\sin\theta \ , Y = -x\cos\theta + y\sin\theta$$

  - V1 Respond

$$V1(\cdot,t,\theta,s) = I(\cdot,t) * G_{even}(\cdot,\theta,s)$$

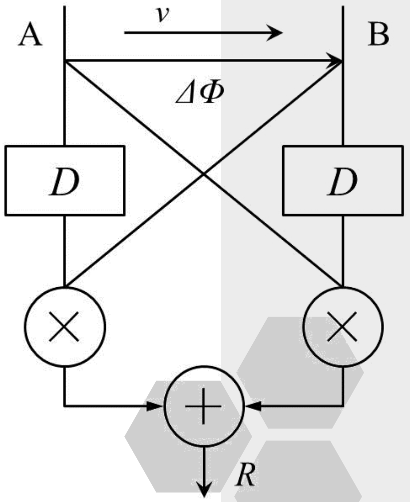
- V2 - *scale, shift and orientation invariance*

$$V2(x,y,t,\theta,\varepsilon) = \underset{s\in\{2\varepsilon,2\varepsilon-1\}}{\mathrm{Max}} V1(\cdot,t,\theta,s)$$

ε	1		2		3		4		5		6		7		8	
s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
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	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	0 $\frac{\pi}{4}$ $\frac{\pi}{2}$ $\frac{3\pi}{4}$															

MT - *higher order motion analysis*

- 2D EMD
  - $R(x,t) = F_A(t-\tau)F_B(t) - F_A(t)F_B(t-\tau)$
    $F_A(t) = F(x,t), F_B(t) = F(x+\Delta\Phi, t)$
- 3D EMD
  - $$\begin{cases} MT(\cdot,t,\theta,\epsilon) = F'_A(t-\tau)F'_B(t) - F'_A(t)F'_B(t-\tau) \\ F'_A(t) = F'(\cdot,t,\theta,\varepsilon,0) \\ F'_B(t) = F'(\cdot,t,\theta,\varepsilon,\Delta\Phi) \end{cases}$$
    $F' = V2(x+\Delta\Phi\sin\theta, y+\Delta\Phi\cos\theta, t, \theta, \varepsilon)$



BIM-STIP detection strategy

- Orientation Competition

$$OC(x,y,t,\varepsilon) = \max_{\theta}\{MT(x,y,t,\theta,\varepsilon)\}$$

- Shift Competition (Local Maximum)
- Locality Refine
- Global Optima

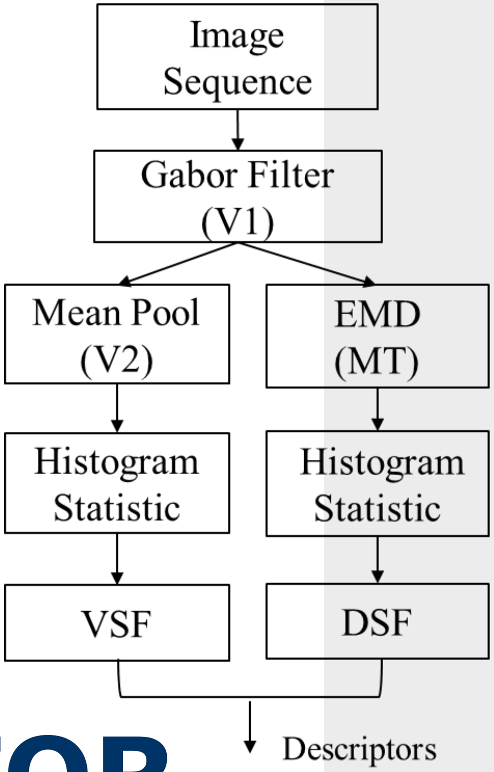
Algorithm 1 BIM-STIP detection strategy

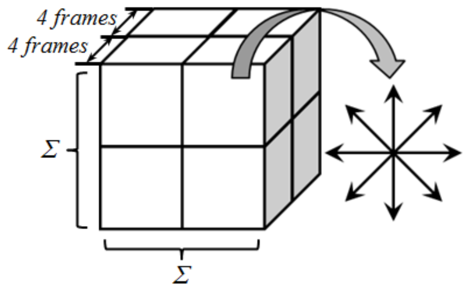
Input: $OC(x,y,t,\varepsilon)$, $LME(x,y,w,h)$
Output: vector<Keypoint> $points$

```
\\Local maximum
while (x, y) do
    if LME(x, y, Σ(ε), Σ(ε)) < lmeThres then
        continue
    end if
    keypoint ← argmax_{(x,y)∈Σ(ε)} OC(x, y, t, ε)
    points.push[keypoint]
end while
\\Locality refine
while (i, j) do
    if ||points[i] − points[j]||_2^{1/2} < dThres then
        points.pop[ argmin_{k∈{i,j}} points[k].respond ]
    end if
end while
\\Global optima
P_(1) < P_(2) < ... < P_(n) ← points.respond
points ← P_(n), P_(n−1), ..., P_(n−N)
```

BIM-STIP Framework

BIM-STIP DESCRIPTOR





DSF - *BIM motion feature*

V1 Respond

- For extracting detailed feature, the smallest scale space ($s = 1$ and $\xi = 7$) is selected

$$V1R_{even}(\cdot, t, \theta) = I(\cdot, t) * G_{even}(\cdot, \theta, s = 1)$$

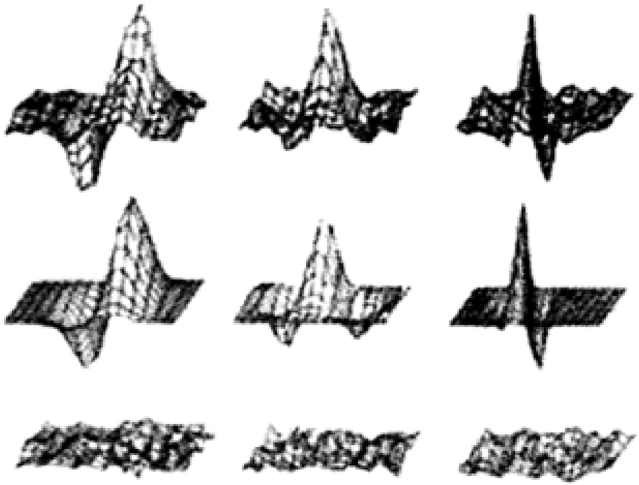
MT Respond

- The time offset is set to a minimum value ($\tau = 1$)
- The space offset is set to the main lobe width of smallest scale Gabor filter ($\Delta\Phi = 4$)

$$MTR(\cdot, t, \theta) = F'_A(t - \tau)F'_B(t) - F'_A(t)F'_B(t - \tau)$$
$$F'_A = V1R_{even}(x, y, t, \theta),$$
$$F'_B = V1R_{even}(x + 4\sin\theta\,, y + 4\cos\theta\,, t, \theta, \varepsilon)$$

VSF - *BIM shape feature*

- Odd Gabor Filter
  - $G_{odd}(\cdot,\theta,s) = \exp\left(-\frac{X^2+\gamma^2Y^2}{2\sigma^2}\right)\sin\left(\frac{2\pi}{\lambda}X\right)$
- V1 Respond
  - $V1R_{odd}(\cdot,t,\theta) = I(\cdot,t) * G_{odd}(\cdot,\theta,s=1)$
- V2 Respond
  - $V2R(x,y,t,\theta) = \underset{\Sigma}{\mathrm{Mean}}\{V1R_{odd}(x,y,t,\theta)\}$



Results of human action databases

	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

Confusion matrices of Weizmann

	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

Confusion matrices of KTH

Result Comparisons

- Comparisons with other 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%**

- Comparisons with representative algorithms

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

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%**

Thank You !

Email：caibolun@gmail.com