Human Action Recognition using Salient Opponent-based Motion Features

Amir-Hossein Shabani
John S. Zelek, David A. Clausi

Vision and Image Processing Group
SYDE, University of Waterloo, ON

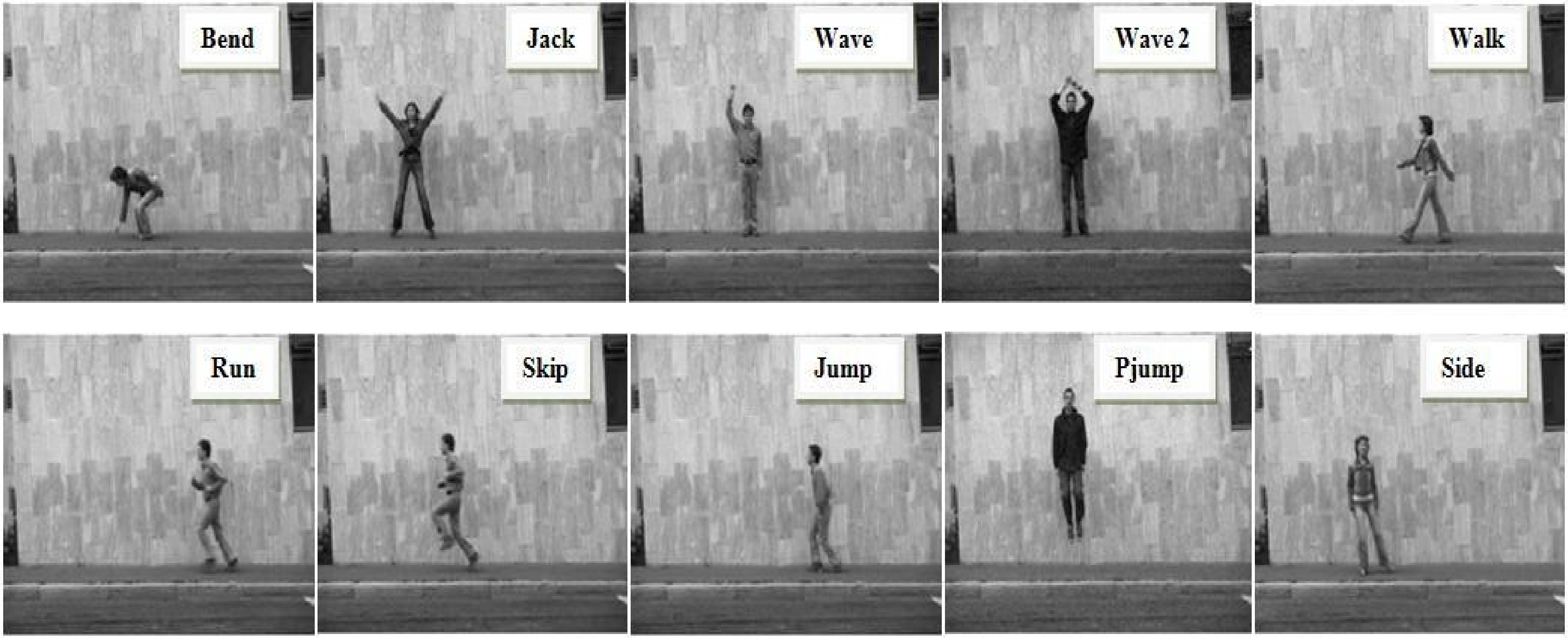
CRV- Ottawa, May 31, 2010

Applications

- Automated surveillance for scene analysis,
- Elderly home monitoring for assisted living,
- Content-based video retrieval,
- Human-computer interaction (HCI), …

Activity analysis

Human Action Recognition



Outline

- Applications
- Problem Statement of Human Action Recognition
- Bottom-up vs. Top-down Approaches
- Paper Contribution: Salient Opponent-based Motion Features
- Experimental Results

Problem Statement

- **Human action recognition refers to the labelling of the type of an unknown action.**
  - **type of actions: walking, running, jumping, hand waving, etc.**

Classification Problem

Existing Methods

1. Top - down methods

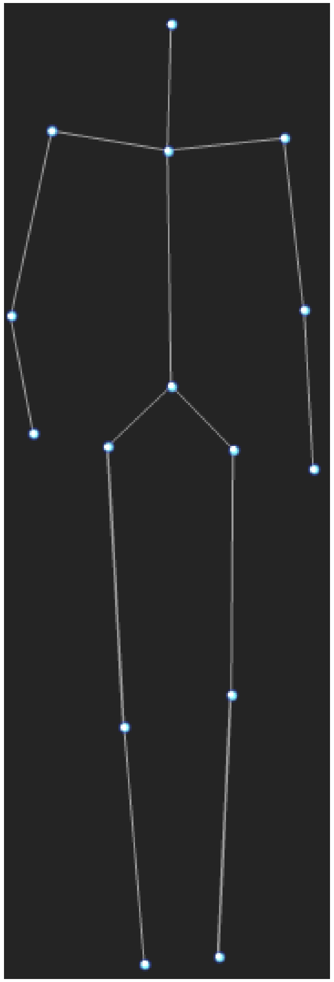
- Model-based approaches
  - shape models: stick-figure, 2D ribbon , 3D volume
- Video segmentation
  - tracking using motion model,
- Trajectory/eigen shapes for encoding

2. Bottom - up methods

- Model –free approaches
  - no shape/motion model
- No explicit segmentation
- Salient features for encoding

Why Bottom-Up Approach?

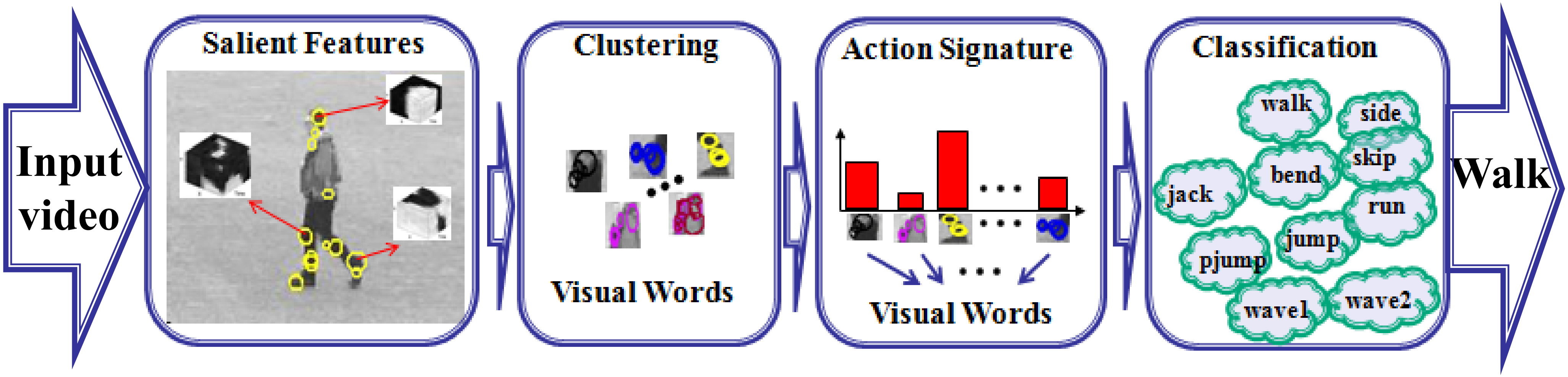
- Bottom-up approaches are model free and universal.
- Biological motion
- http://www.biomotionlab.ca/Demos/BMLwalker.html



- Object recognition

Bottom-Up Approach

Bag-of-words framework is a standard realization of bottom-up approach for human action recognition.



Salient Feature Extraction

- Salient feature extraction consists of three steps:
  - Video filtering at different spatio-temporal scales
  - Key point detection
  - Key point description using the characteristic of the point's surrounding volume.

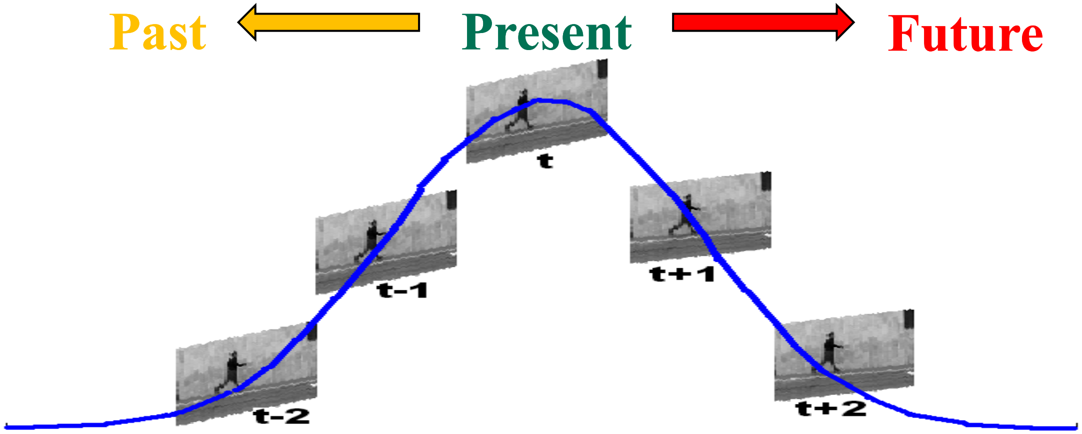
- Key point detection:

  (1) Saliency map construction

  (2) non-max suppression (and thresholding)

Problem of Existing Methods in Feature Extraction

- Temporal Gaussian/Gabor filter requires both prior and posterior frames.



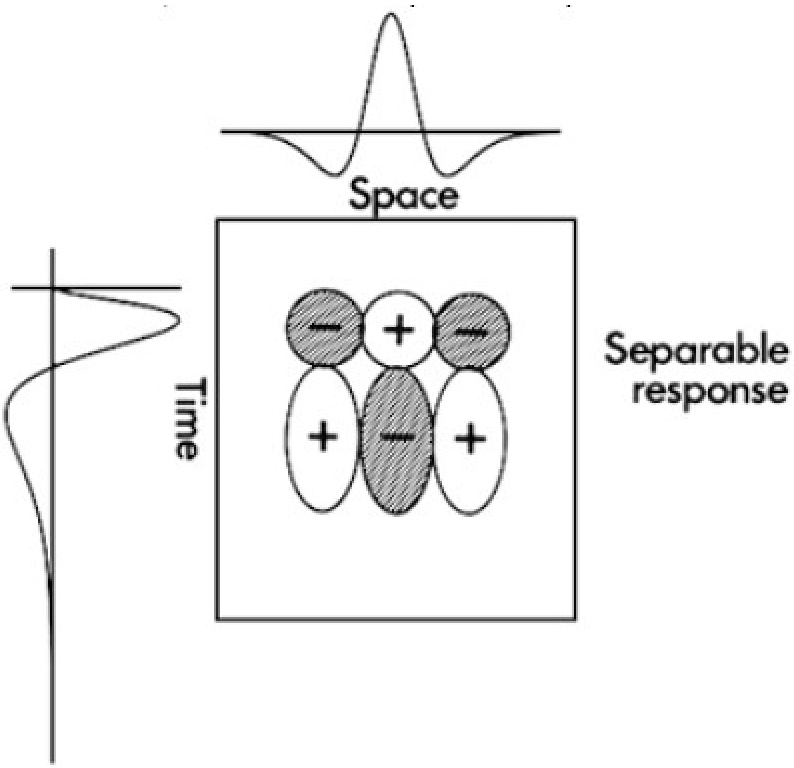
- Biological vision promotes causal filtering in motion perception.
  - Q: how we can address the time causality?

Biological Vision

Motion perception filtering should be:

- time causal
- contrast-polarity insensitive
- phase insensitive
- opponent-based

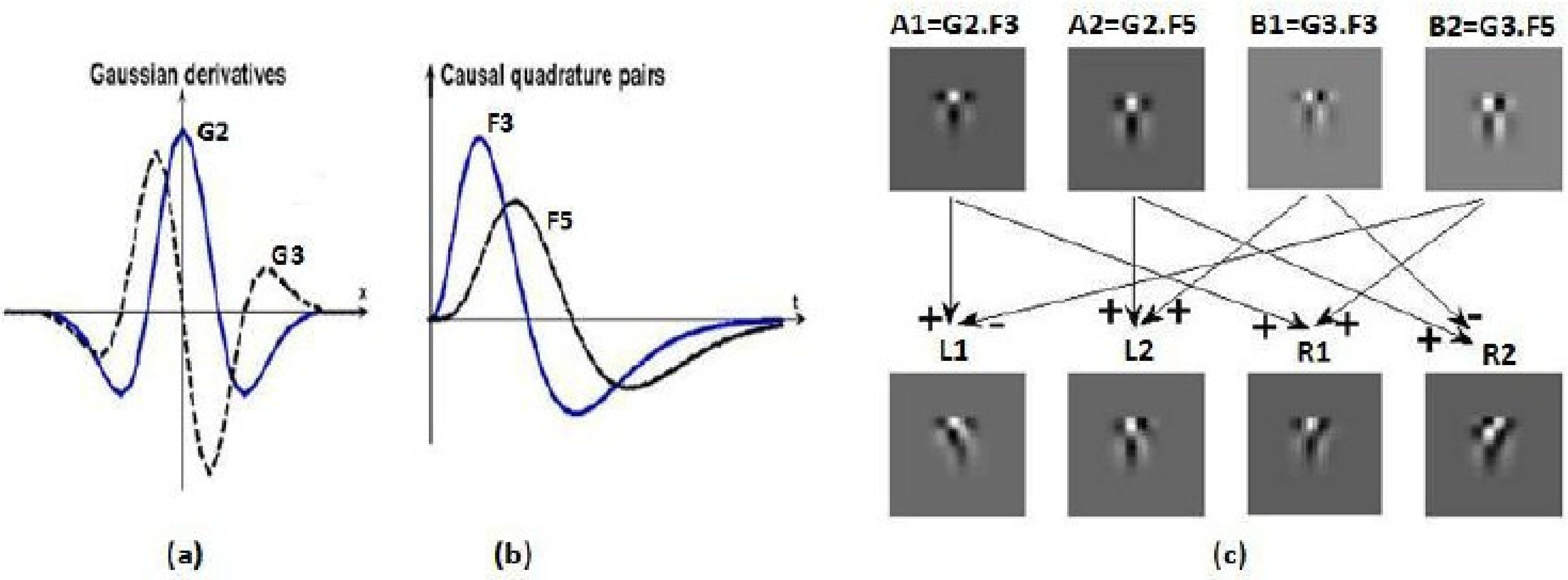
Linear Separable Spatial and Temporal Filters



$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

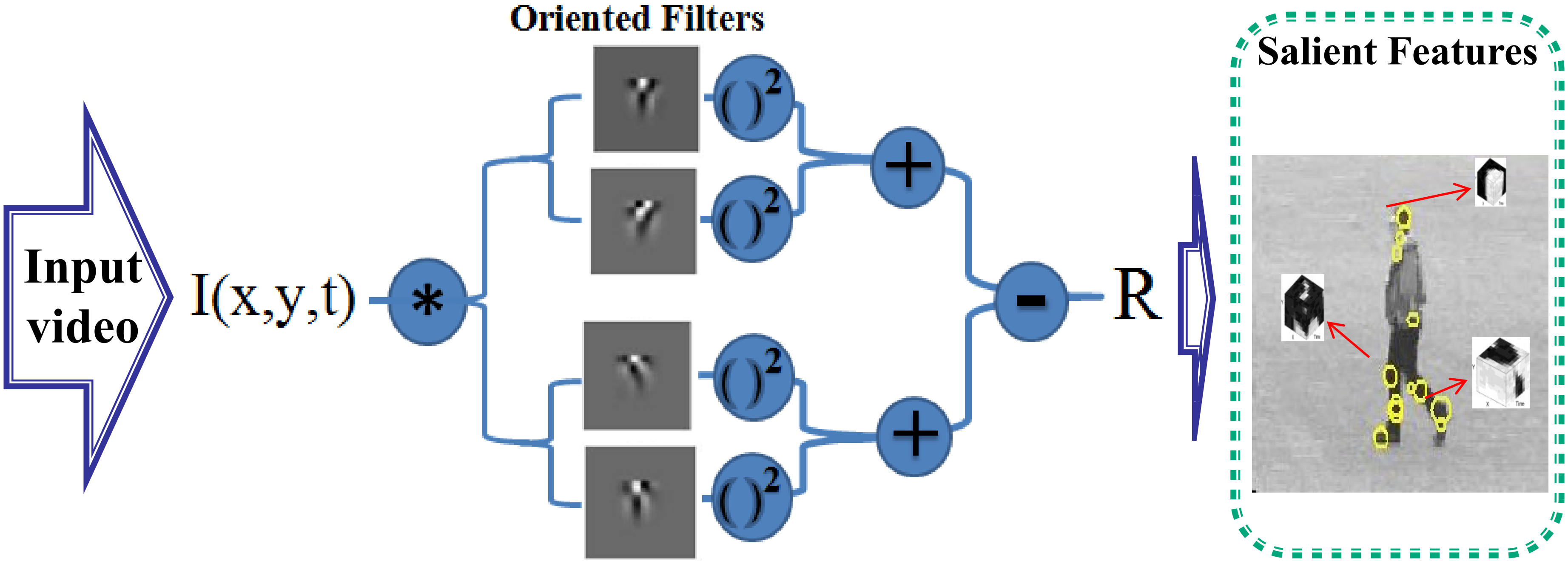
$$F_n(t) = [\frac{1}{n!} - \frac{(kt)^2}{(n+2)!}](kt)^n e^{-kt}$$

Oriented Motion Filters



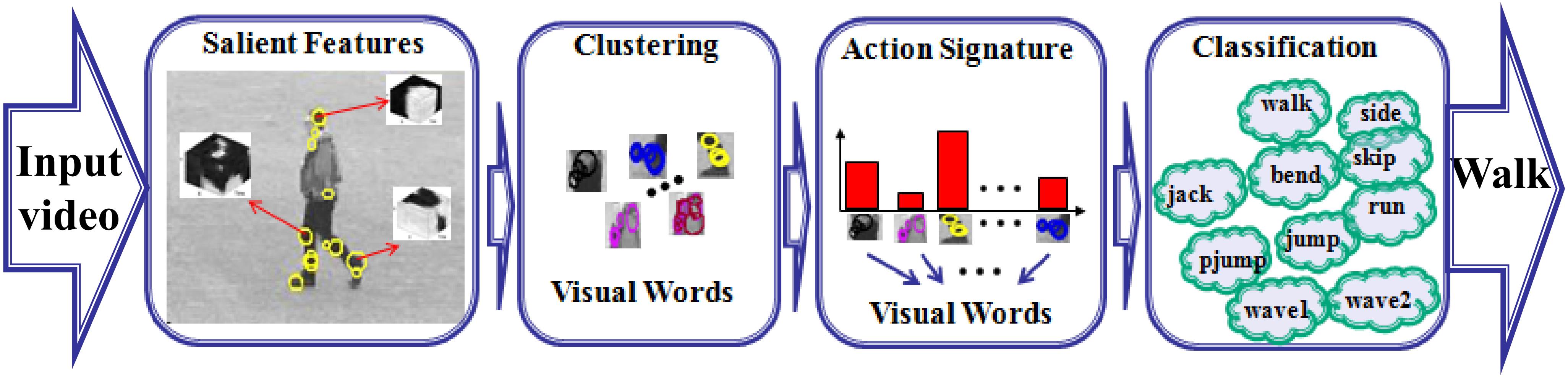
Salient Opponent-based Motion Features

- Oriented motion filtering
- Compute the opponent-based motion maps as the saliency map
- Non-maxima suppression => salient opponent-base motion features
- Use 3D SIFT descriptor



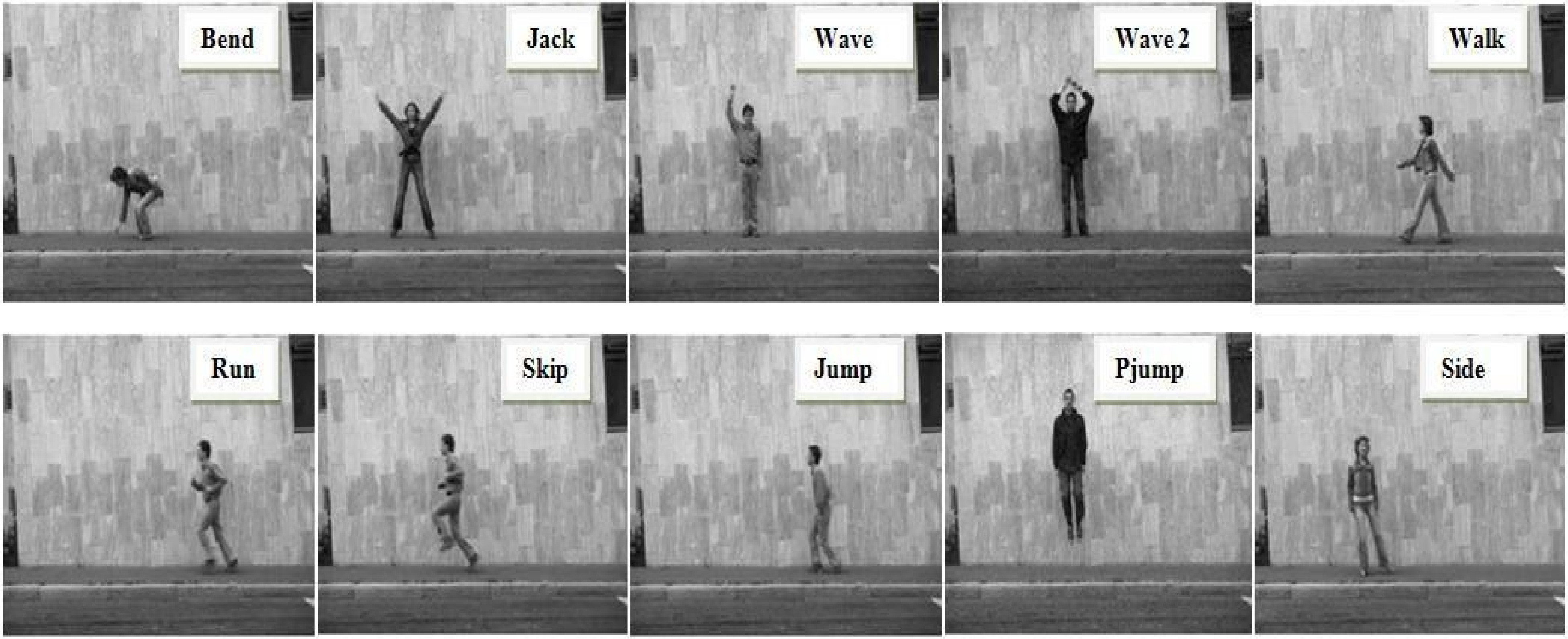
Bottom-Up Approach

- Bag-of-words framework is a standard realization of bottom-up approach for human action recognition.



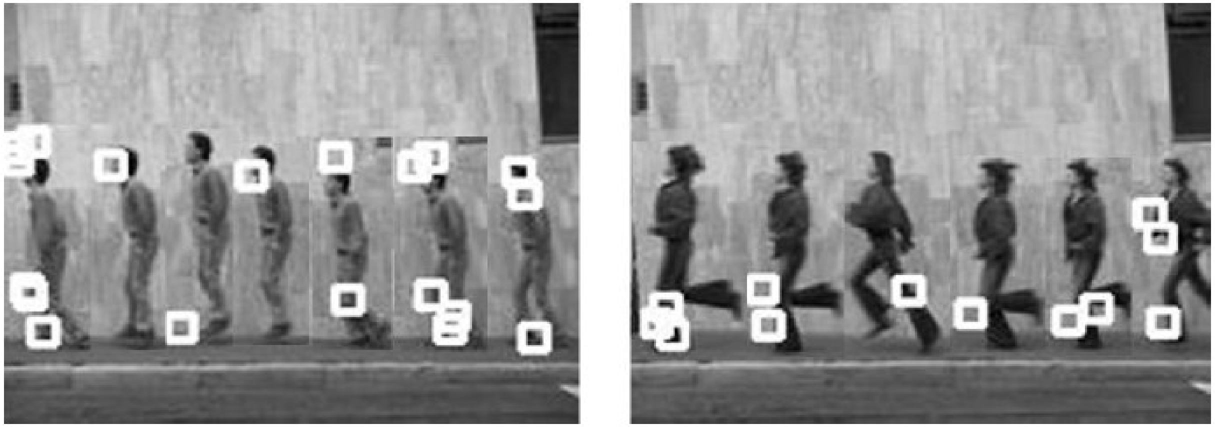
Weizmann Dataset

- Consists of ten different human actions performed by nine different people in front of a fixed camera.
- Each clip lasts about two seconds at 25Hz with an image frame size of 180 x 144.



Weizmann Dataset

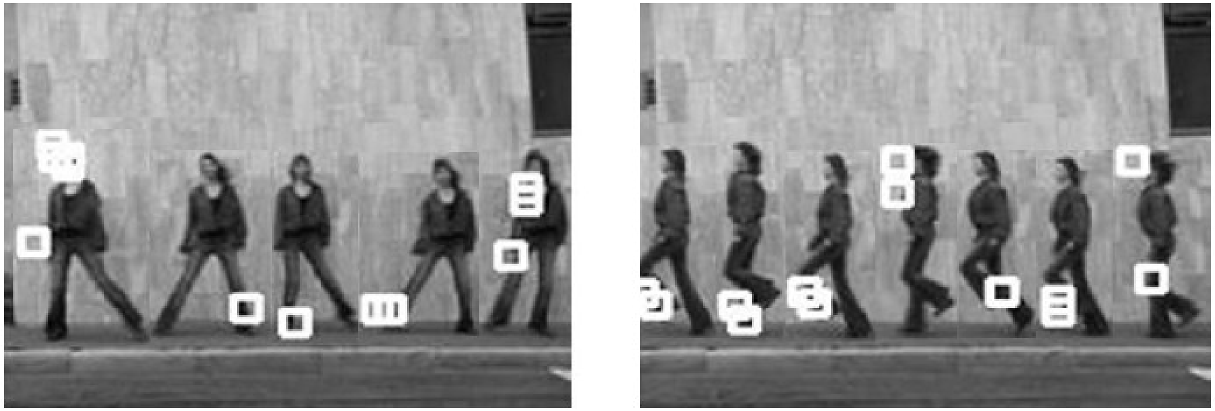
Jump



Run

(a) 2D projection of the salient motion events for jump action(left column) and run action (right column).

Gallop sideway

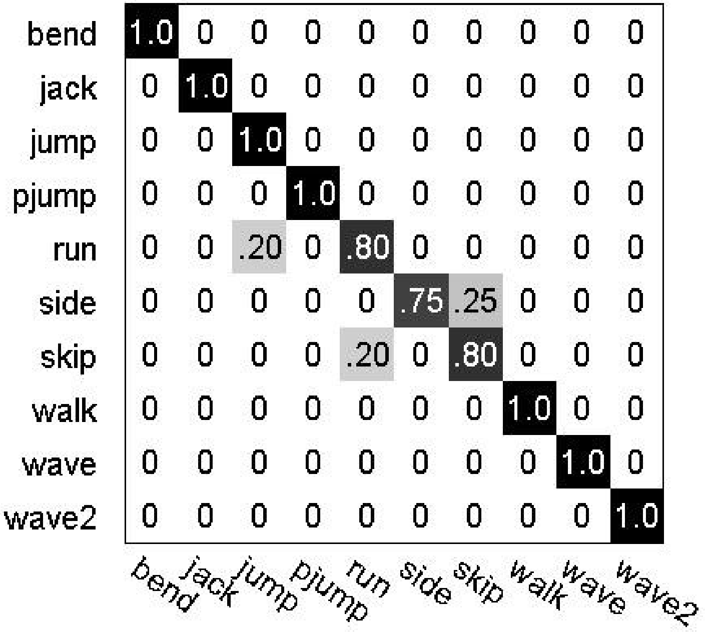


Skip

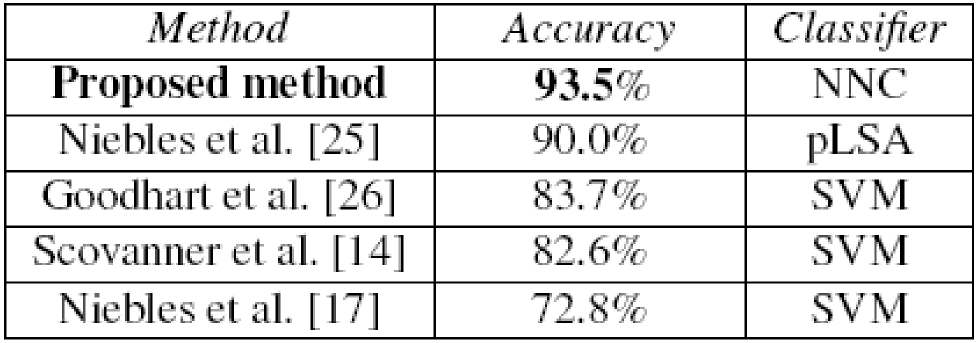
(b) 2D projection of the salient motion events for the side (gallop sideways) action (left column) and the skip action (right column).

Weizmann Dataset

- Human action recognition using salient opponent-based motion features in the bag-of-words framework



	bend	jack	jump	pjump	run	side	skip	walk	wave	wave2
bend	1.0	0	0	0	0	0	0	0	0	0
jack	0	1.0	0	0	0	0	0	0	0	0
jump	0	0	1.0	0	0	0	0	0	0	0
pjump	0	0	0	1.0	0	0	0	0	0	0
run	0	0	.20	0	.80	0	0	0	0	0
side	0	0	0	0	0	.75	.25	0	0	0
skip	0	0	0	0	.20	0	.80	0	0	0
walk	0	0	0	0	0	0	0	1.0	0	0
wave	0	0	0	0	0	0	0	0	1.0	0
wave2	0	0	0	0	0	0	0	0	0	1.0



Method	*Accuracy*	*Classifier*
Proposed method	**93.5%**	NNC
Niebles et al. [25]	90.0%	pLSA
Goodhart et al. [26]	83.7%	SVM
Scovanner et al. [14]	82.6%	SVM
Niebles et al. [17]	72.8%	SVM

Summary

- Bottom-up approaches are attractive and universal
  - as they do not require video segmentation OR any shape or motion model.
  - can be easily adapted for recognition of different entities.

- Robust and informative salient features are the key for the recognition task.

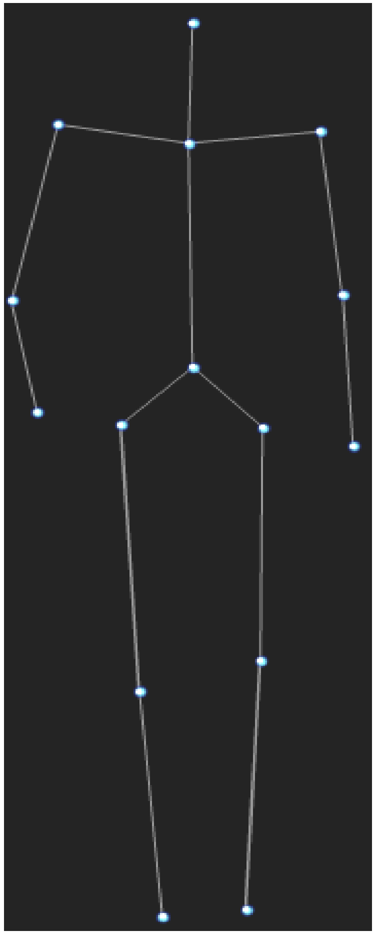
- Salient opponent-based motion features can provide a proper action encoding.

Thank you!

1. Ongoing Works

Hypothesis:

- **In a probabilistic recognition approach**
  - **use the structural constraints in clustering**
  - **helps a more intuitive action primitives.**

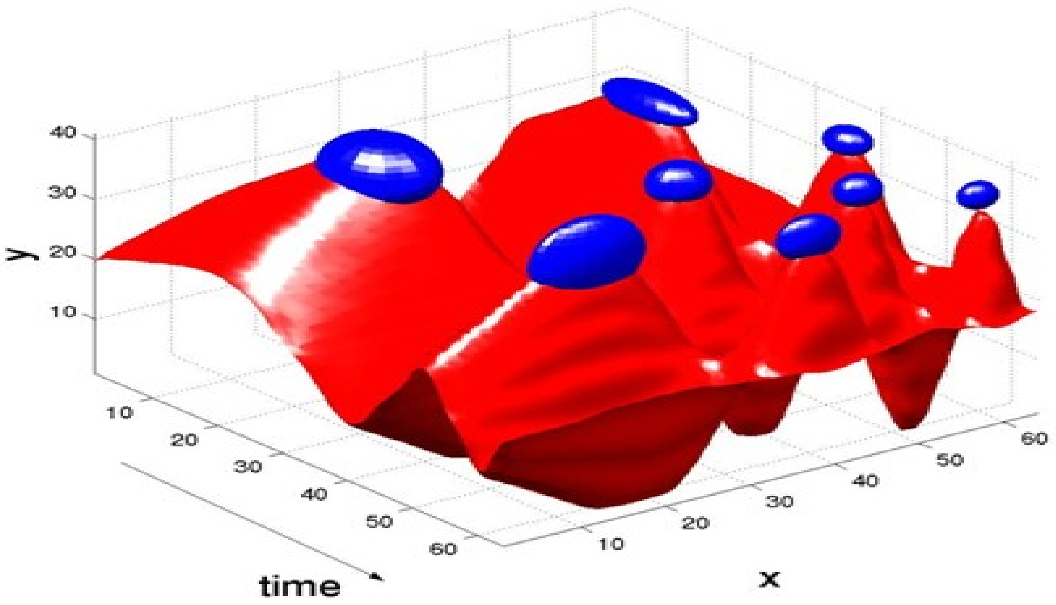


2. Ongoing Works

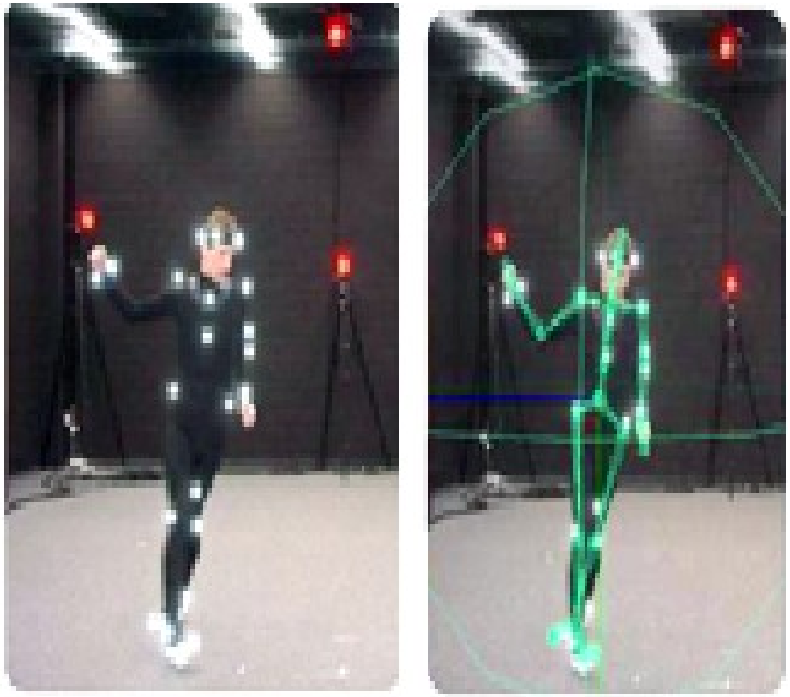
- Experiments on more unconstrained environments and challenging data sets
  - Hollywood movies, Sports data, YouTube videos ,…

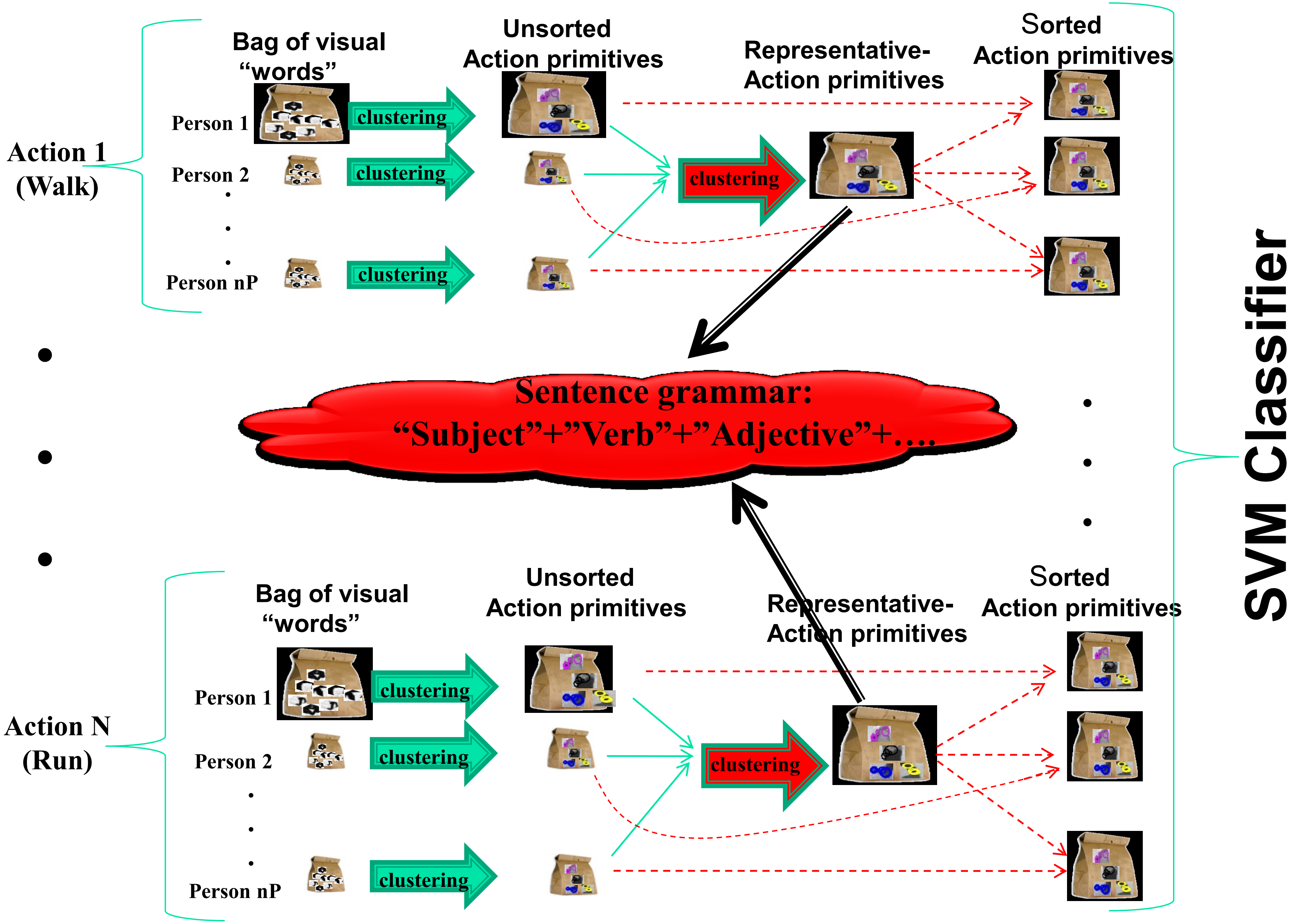
The Requirements & Experiments

1. Need robust multi-scale salient features.



2. Experiments for action recognition using two standard data sets.





Action 1
(Walk)

Bag of visual
"words"

Unsorted
Action primitives

Representative-
Action primitives

Sorted
Action primitives

Person 1

clustering

Person 2

clustering

Person nP

clustering

clustering

Sentence grammar:
"Subject"+"Verb"+"Adjective"+....

Action N
(Run)

Bag of visual
"words"

Unsorted
Action primitives

Representative-
Action primitives

Sorted
Action primitives

Person 1

clustering

Person 2

clustering

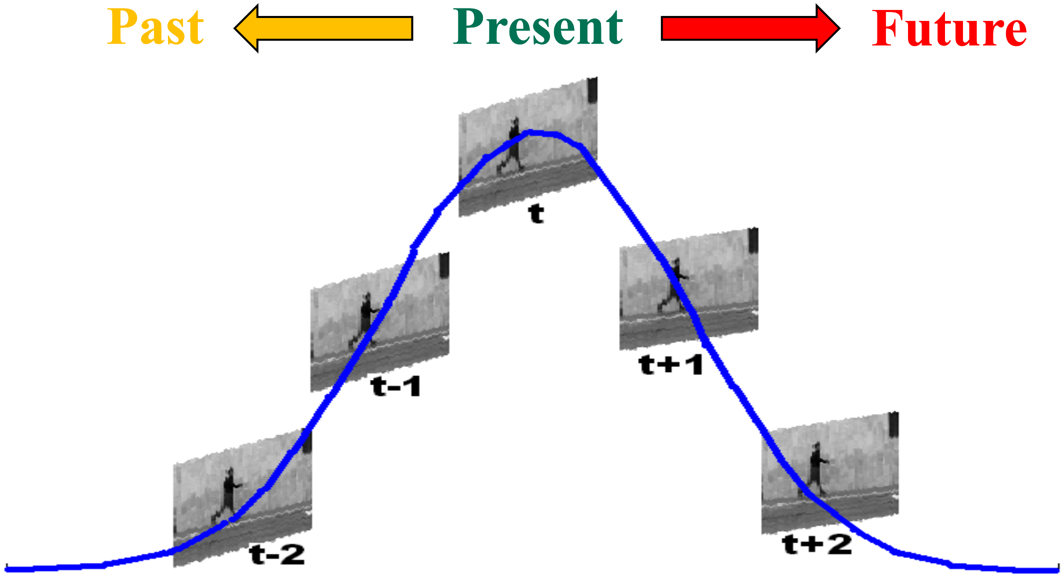
Person nP

clustering

clustering

SVM Classifier

Spatio-Temporal Salient Features



Past ← Present → Future

Challenges:

1) Temporal Gaussian/Gabor filter requires both prior and posterior frames.

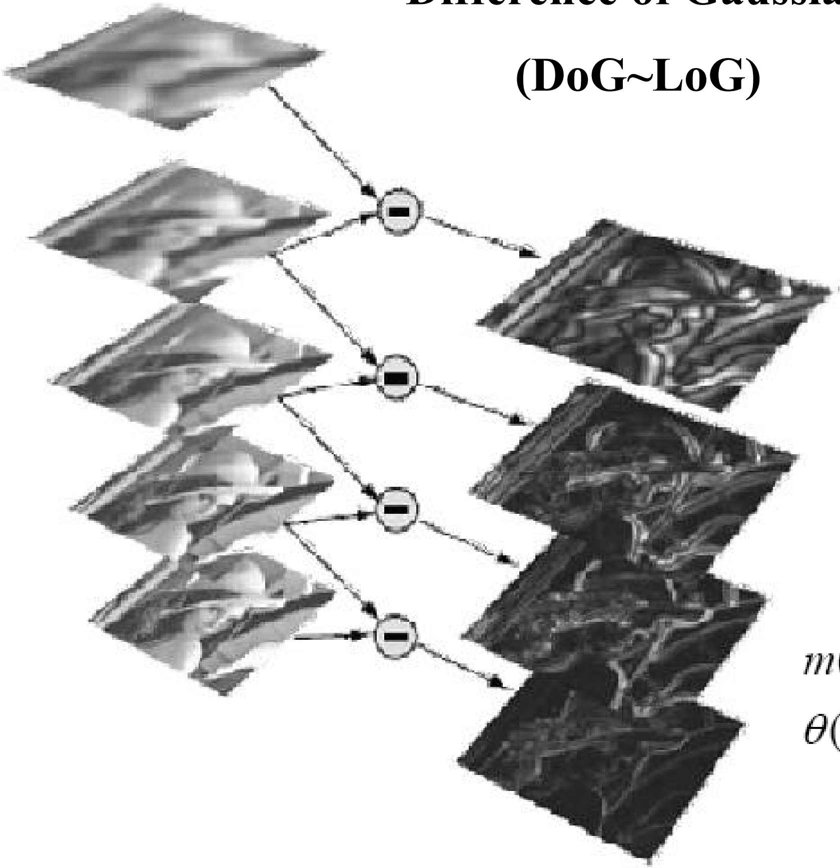
2) Gaussian filters dislocate the structures such as edges and salient motions.

3) How to model the uncertainty of the temporal correlation (due to unknown camera motion or jitter)?

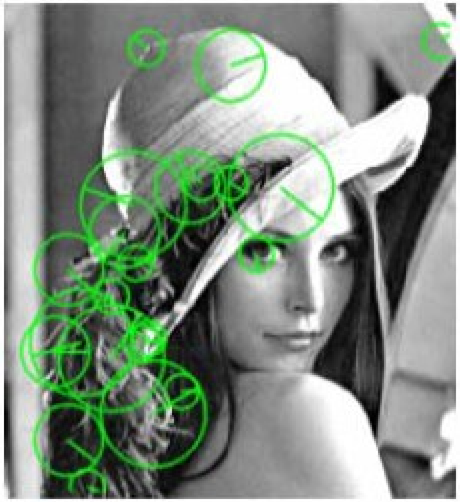
2 Dimensional Scale-Invariant Feature Transform

Gaussians (F)

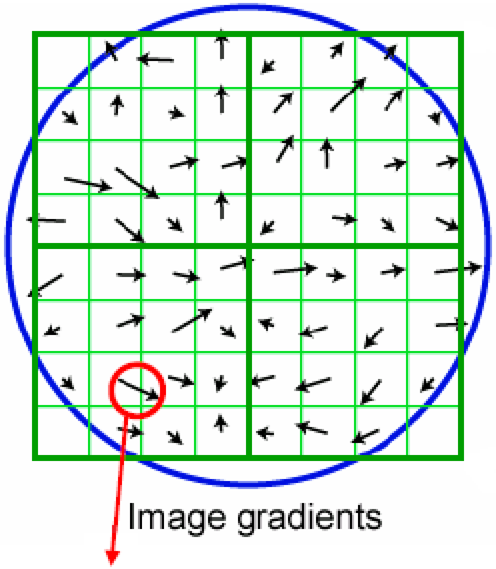
Difference of Gaussian (DoG~LoG)



2D SIFT features

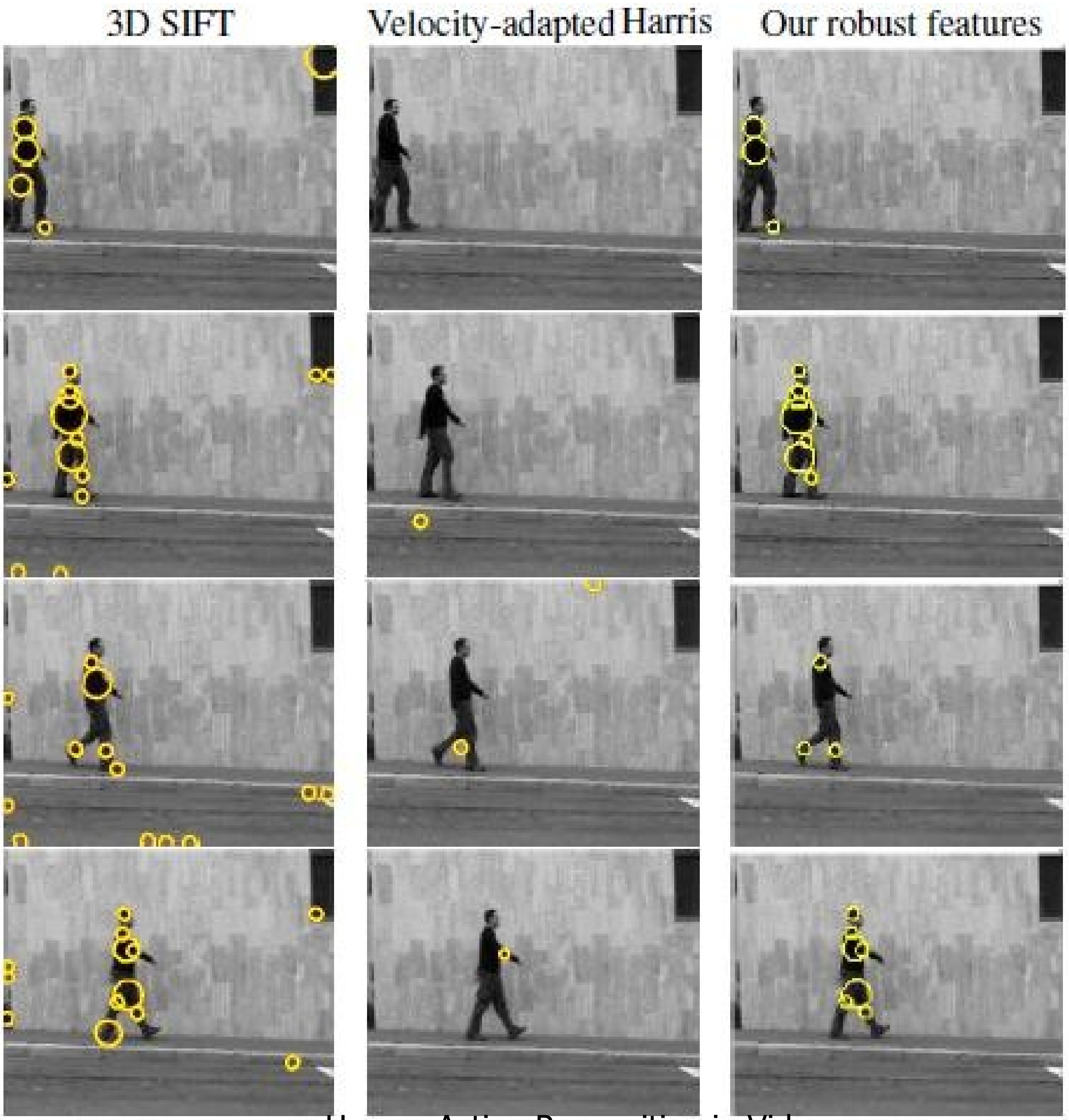


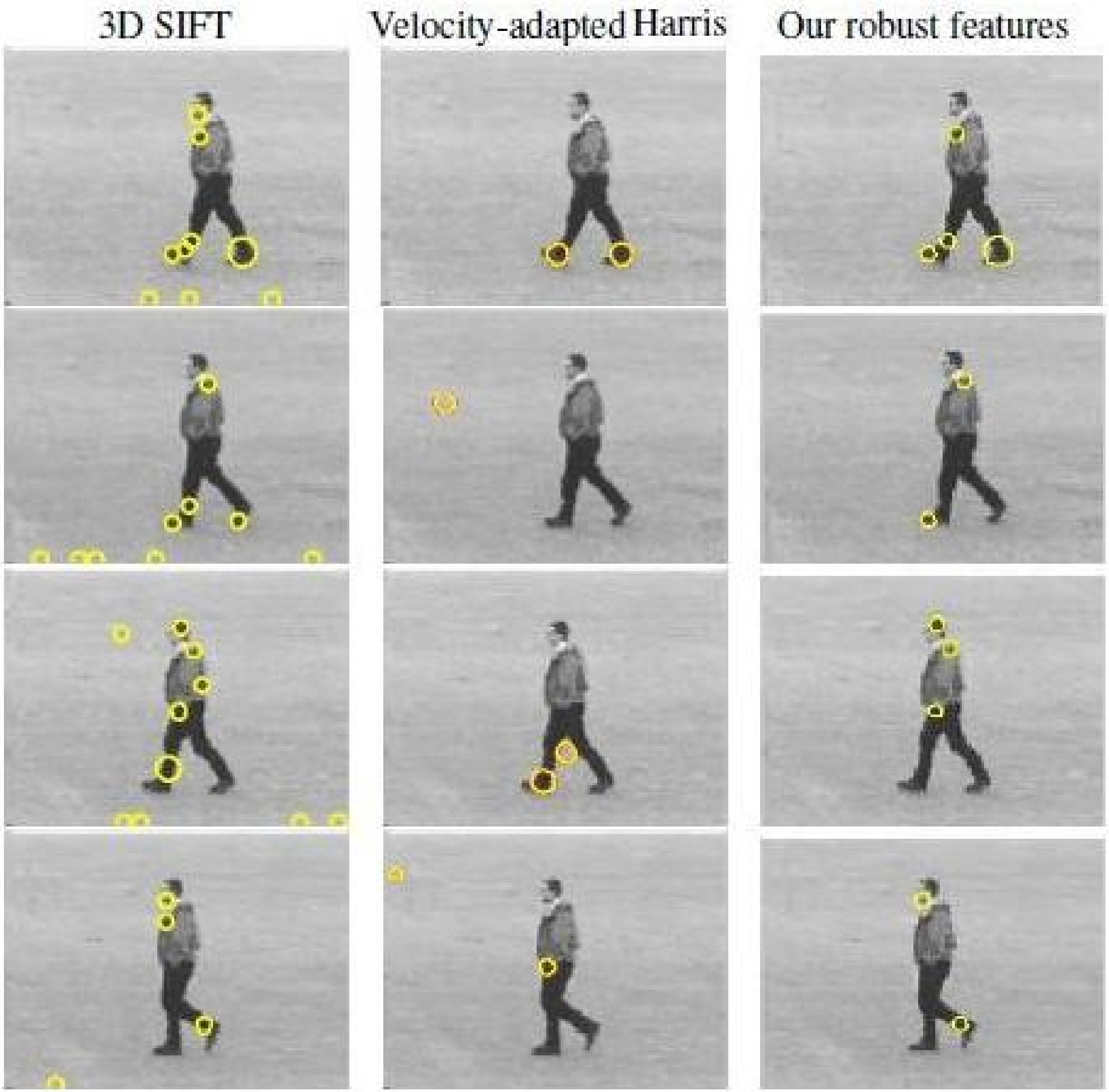
Oriented gradients



$$m(x,y) = \sqrt{(F(x+1,y)-F(x-1,y))^2 + (F(x,y+1)-F(x,y-1))^2}$$

$$\theta(x,y) = \text{atan}((F(x,y+1)-F(x,y-1))/(F(x+1,y)-F(x-1,y)))$$


3D SIFT
Velocity-adapted Harris
Our robust features


3D SIFT
Velocity-adapted Harris
Our robust features

Challenges

Main challenges in:

1. Detection

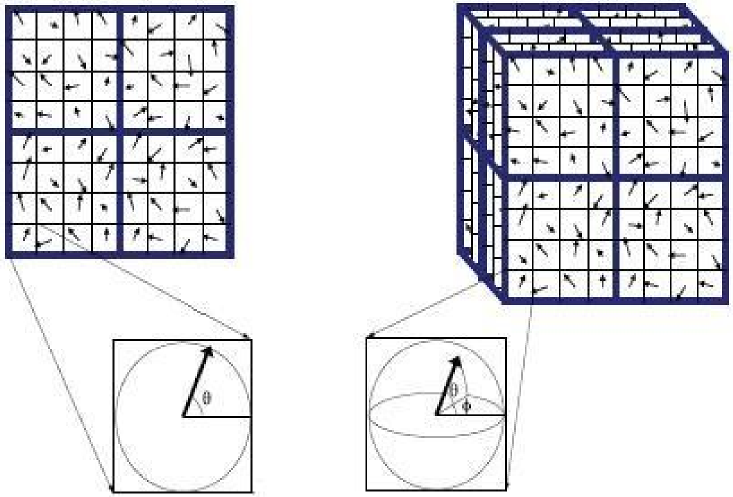
- Environment/illumination changes,
- Camera shaking/movement,
- Low video quality,
- Data association (occlusion/ clutter),
- Etc.

2. Recognition

- High intra-class variation:
  - *in the pattern of a given action*
- Low inter-class variation:
  - *in the pattern of different actions (e.g., running and jogging)*

Second objective: feature descriptor (ctd.)

- The closest idea is the 3D SIFT descriptor which uses
  - Histogram of (spatial) Oriented Gradients
  - implicit motion information



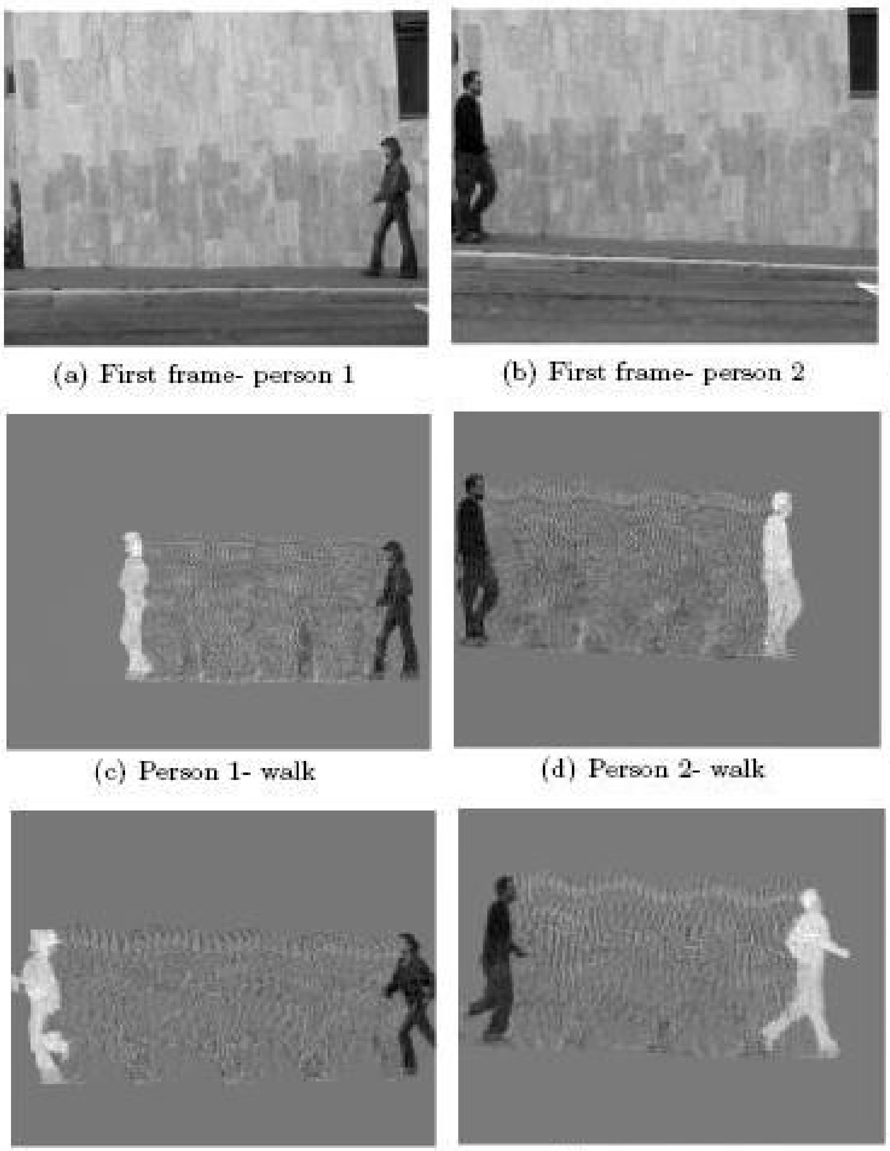
$$m_{3D}(x,y,t) = \sqrt{I_x^2 + I_y^2 + I_t^2}$$

$$\theta(x,y,t) = \arctan\left(I_y/I_x\right)$$

$$\phi(x,y,t) = \arctan\left(I_t/\sqrt{I_x^2 + I_y^2}\right)$$

$$hist(i_\theta, i_\phi) = m_{3D}(x',y',t')e^{\frac{-((x-x')^2+(y-y')^2+(t-t')^2)}{2\sigma^2}}$$

Space-time volume of different actions



(a) First frame- person 1

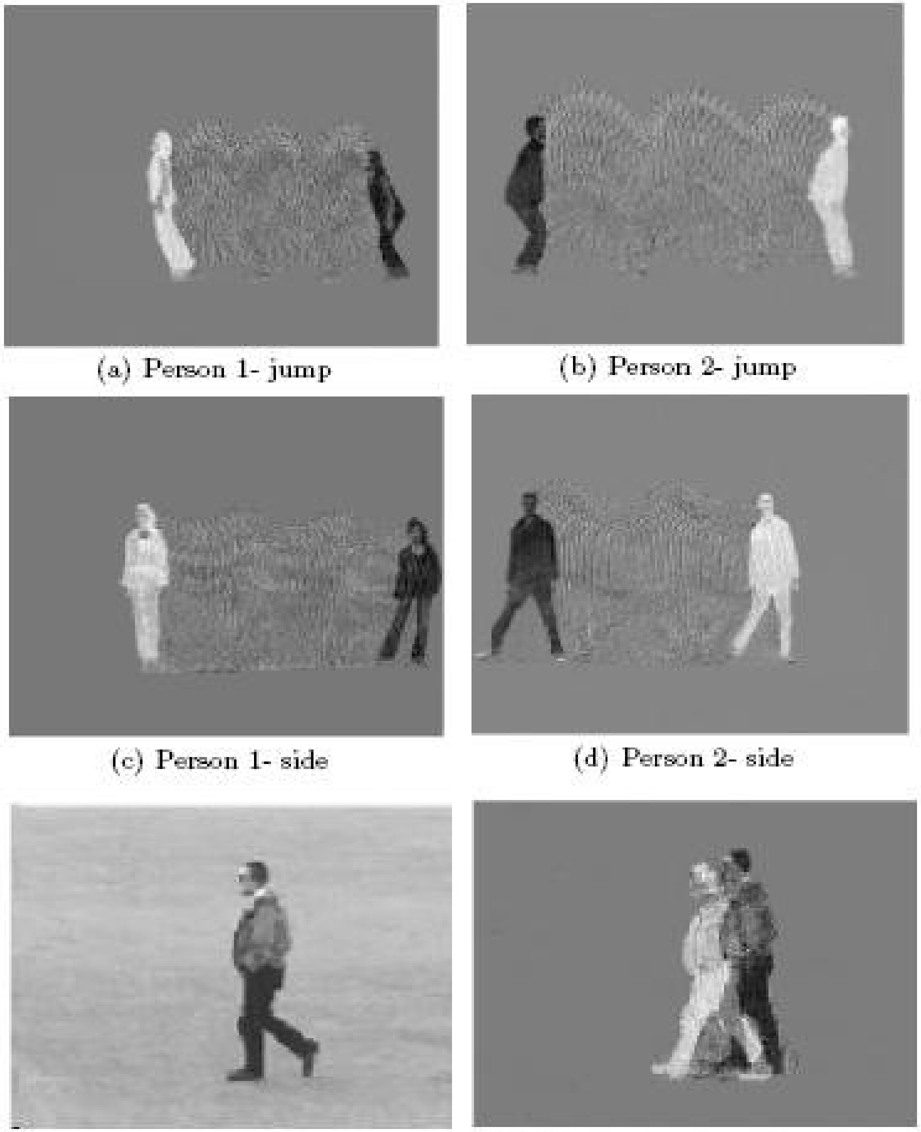
(b) First frame- person 2

(c) Person 1- walk

(d) Person 2- walk

(e) Person 1- run

(f) Person 2- run



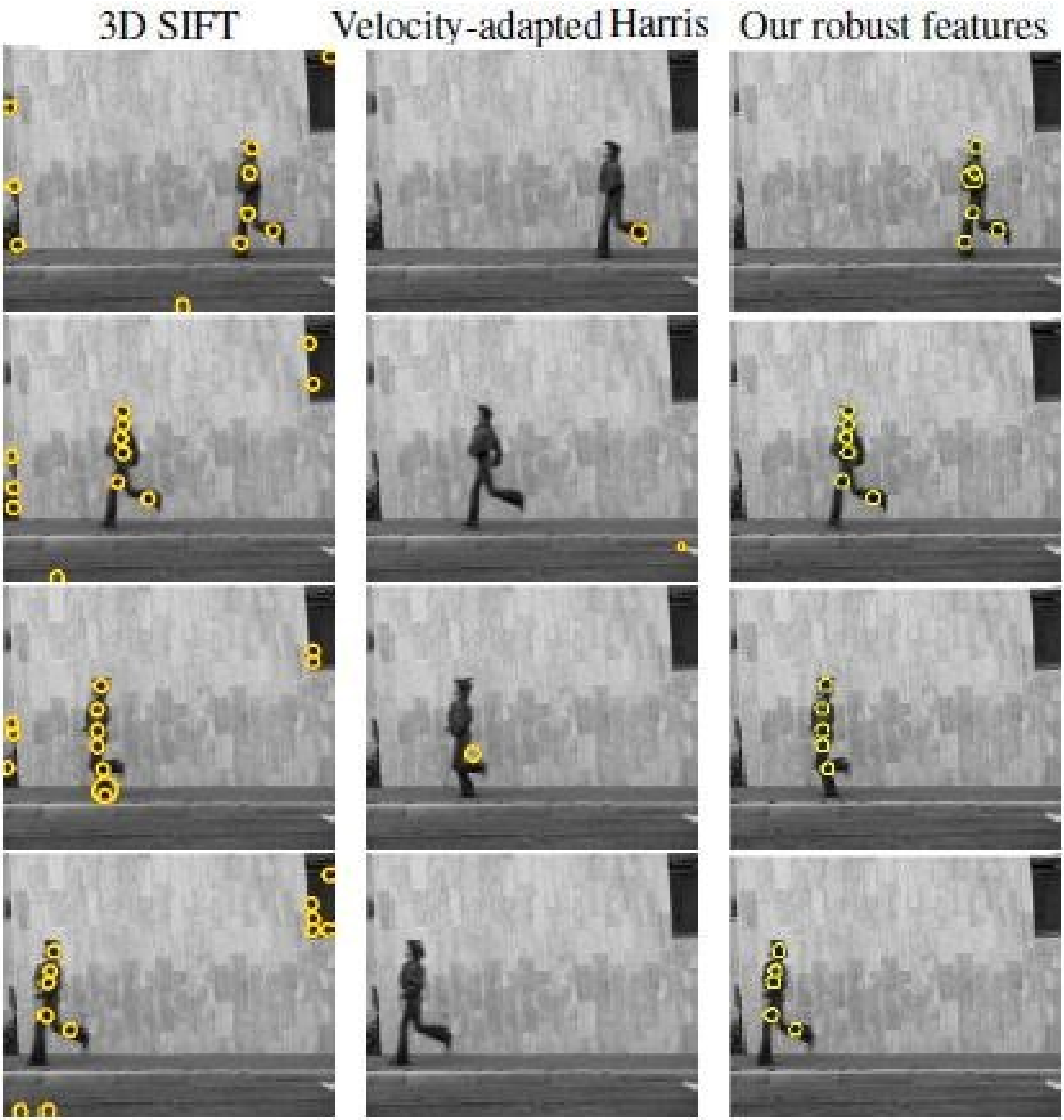
(a) Person 1- jump

(b) Person 2- jump

(c) Person 1- side

(d) Person 2- side

(f) Person 3 walk-moving camera



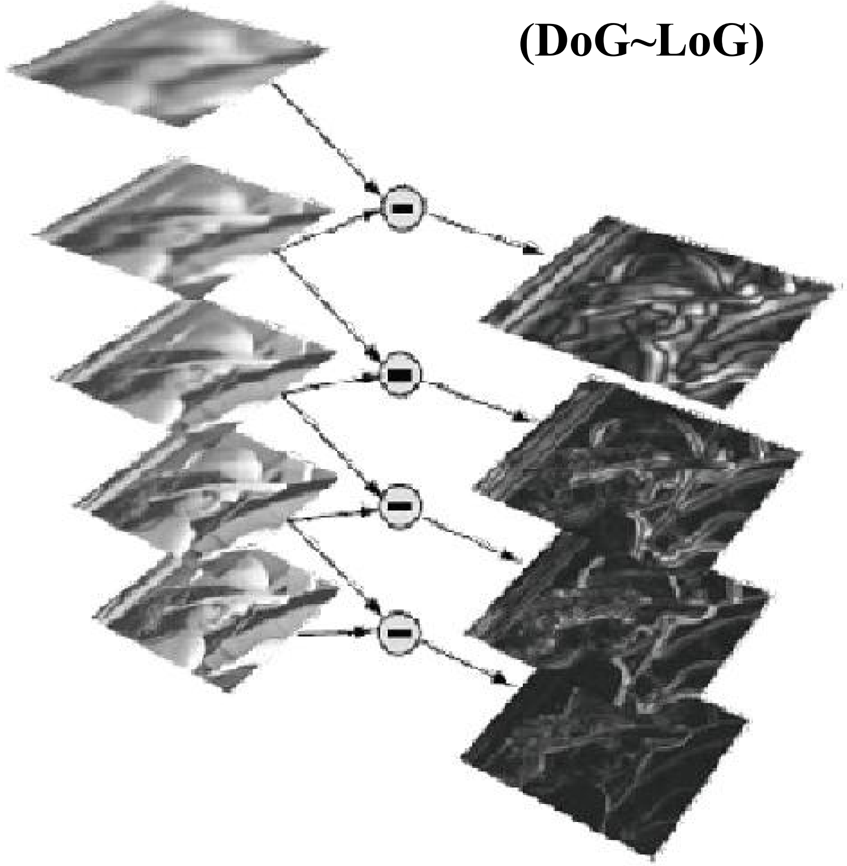
3D SIFT Velocity-adapted Harris Our robust features

2 Dimensional
Scale-Invariant Feature Transform

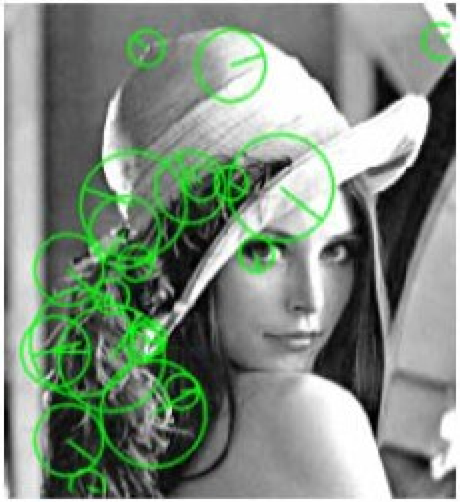
Gaussians (F)

Difference of Gaussian

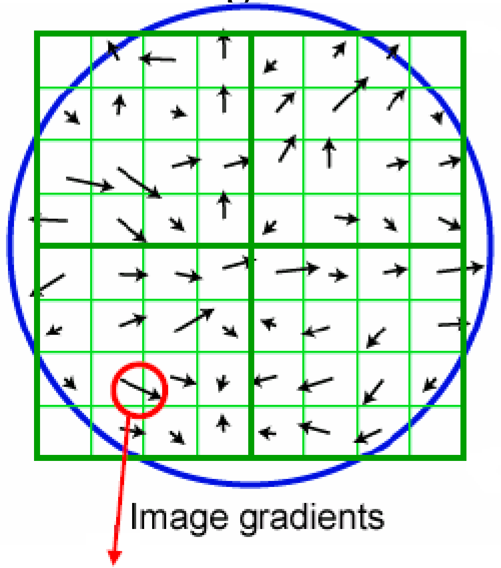
(DoG~LoG)



2D SIFT features



Oriented gradients



$$m(x,y)=\sqrt{(F(x+1,y)-F(x-1,y))^2+(F(x,y+1)-F(x,y-1))^2}$$

$$\theta(x,y)=\operatorname{atan}((F(x,y+1)-F(x,y-1))/(F(x+1,y)-F(x-1,y)))$$

$$hist(\theta)=m(x',y')e^{\frac{-((x-x')^2+(y-y')^2)}{2\sigma^2}}$$

Model-free approaches

  - Provide compact representation of an action by multi-scale salient features.
- Q1: Why multi-scale features?
  - no knowledge about the shape and motion pattern
- Q2: What is a salient feature?
  - should be: (a)distinct in a local neighbour, (b)distinct from other features
    - e.g., maximum entropy, corner, local maxima of LOG, etc.
  - verified by some psychophysical experiments
  - which saliency criteria human uses? No answer!

Model-free approaches

  - Compact representation of an action by multi-scale salient features.

- Q1: Why multi-scale features?
  - no knowledge about the shape and size of the moving person
  - no knowledge about how fast/slow a given action is performed

- Need scale-space filtering
  - e.g., Gaussian filtering $\partial I/\partial s = div(g\nabla I)$

$$I(x,y,s) = (G_{\sigma_s} * I_0)(x,y)$$

$$G_{\sigma_s}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{x^2+y^2}{2\sigma_s^2}}$$