Spatio-Temporal Salient Features

Amir H. Shabani

Vision and Image Processing Lab.,
University of Waterloo, ON

CRV Tutorial day- May 30, 2010
Ottawa, Canada

Applications

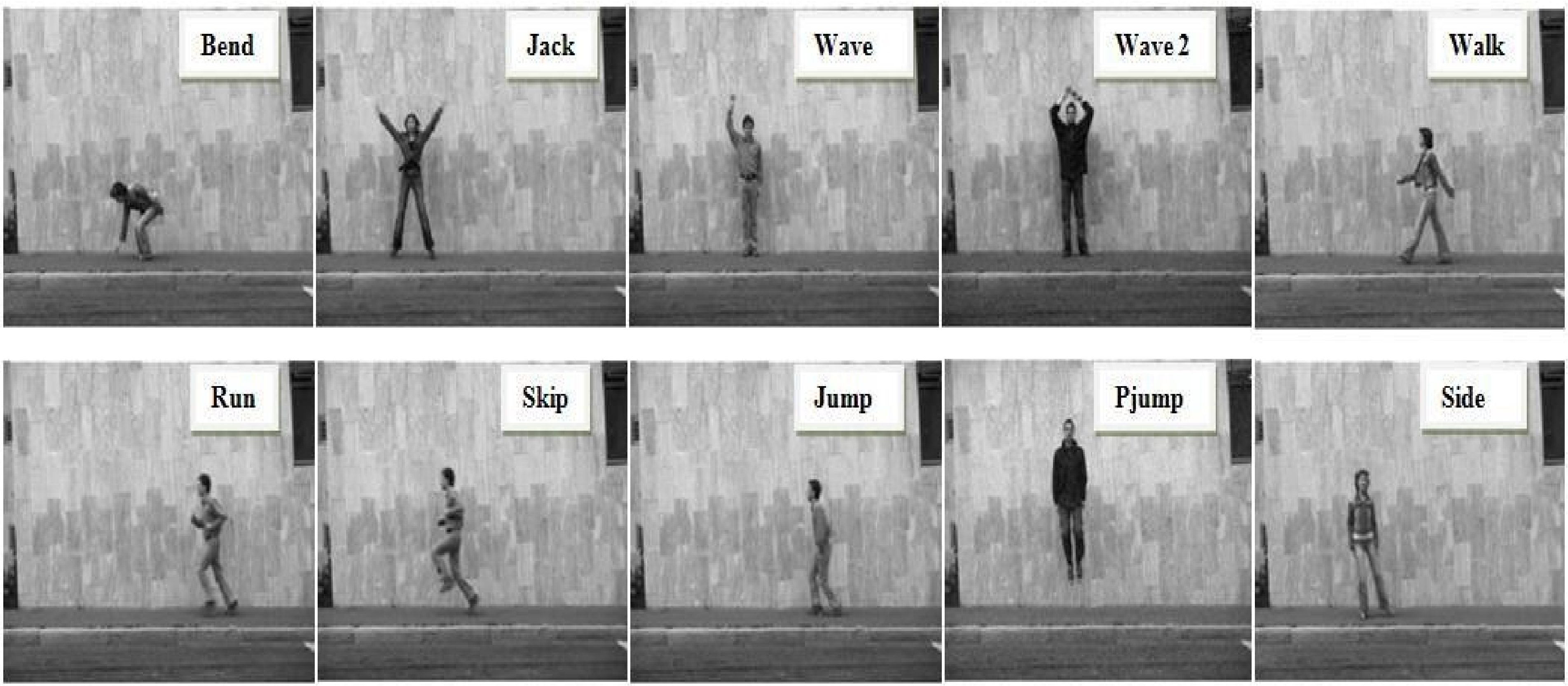
- Automated surveillance for scene analysis,
- Elderly home monitoring for assisted living,
- Content-based video retrieval,
- Human-computer interaction (HCI), …

Video analysis



Application

- Human Action Recognition (Weizmann data set)



Motivations

- Conventional video analysis methods require tracking the objects or features and motion estimation.
  - segmentation is hard in presence of non-stationary background or sensory.
  - appearance/ motion model varies from one object to another.

Motivations (2)

- Conventional video analysis methods require tracking the objects or features and motion estimation.
  - segmentation is hard in presence of non-stationary background or sensory.
  - appearance/ motion model varies from one object to another.

Q: is there any way to bypass these tasks or perform them implicitly?

  - there are key frames and key places in the video that carry most information of what is happening in the video.
- The key space-time points correspond to the video events.
- Salient features characterize the video events.

Motivation (3): from image to video

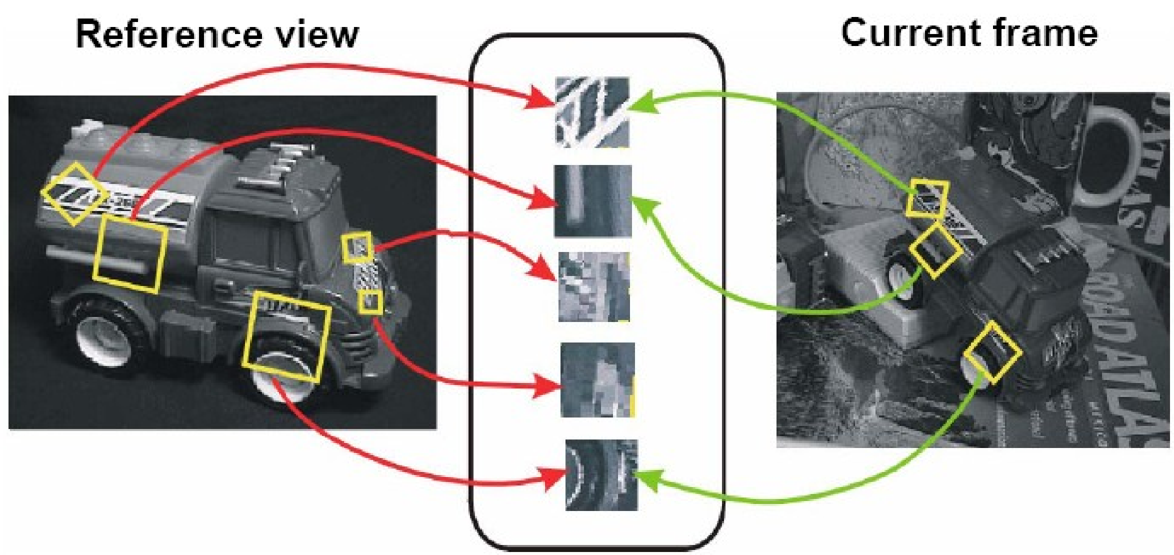
- Spatial Salient Features
- Applications: Tracking, Object Recognition,....

Two steps in detection:

Keypoint Position		Keypoint Descriptor					
x1	y1	g11	g12	g13	g14	...	g1n
x2	y2	g21	g22	g23	g24	...	g2n
x3	y3	g31	g32	g33	g34	...	g3n
x4	y4	g41	g42	g43	g44	...	g4n

1. (Off-line) Select

2. (On-line) Select and Match



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)

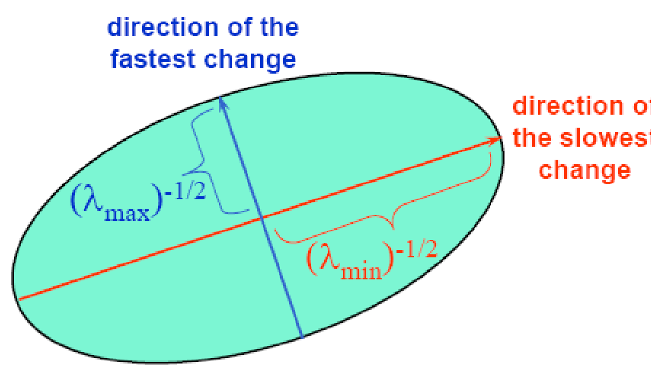
2D Harris-Affine Corner Detector

(1) compute gradient distribution matrix (M) in a local neighbourhood of a point.

$$M = \sigma_D^2 \; g(\sigma_I) \begin{bmatrix} I_{\mathrm{x}}^2(\mathrm{x}, \sigma_D) & I_x I_y(\mathrm{x}, \sigma_D) \\ I_x I_y(\mathrm{x}, \sigma_D) & I_{\mathrm{y}}^2(\mathrm{x}, \sigma_D) \end{bmatrix}$$

(2) find points for which both curvature are significant.

$$\begin{aligned} C &= \det(M) - k\, trace^2(M) \\ &= \lambda_1 \lambda_2 - k\,(\lambda_1 + \lambda_2)^2 \end{aligned}$$



- Corners are stable in arbitrary lighting condition.

Spatio-temporal Harris Corners

Q: Interest? high variation in both space and time.

=> Extend the Harris corner function into 3D spatio-temporal domain.

- Spatial and temporal Gaussian filtering in the computation of the autocorrelation matrix.

$$\mu = g(\cdot;\, \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix},$$

$$L_\xi(\cdot;\, \sigma_l^2, \tau_l^2) = \partial_\xi (g(\cdot;\, \sigma_l^2, \tau_l^2) * f)$$

Spatio-temporal Harris Corners

Q: Interest? high variation in both space and time.

=> Extend the Harris corner function into 3D spatio-temporal domain.

- Spatial and temporal Gaussian filtering in the computation of the autocorrelation matrix.

$$\mu = g(\cdot;\, \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix},$$

$$L_\xi(\cdot;\, \sigma_l^2, \tau_l^2) = \partial_\xi(g(\cdot;\, \sigma_l^2, \tau_l^2) * f)$$

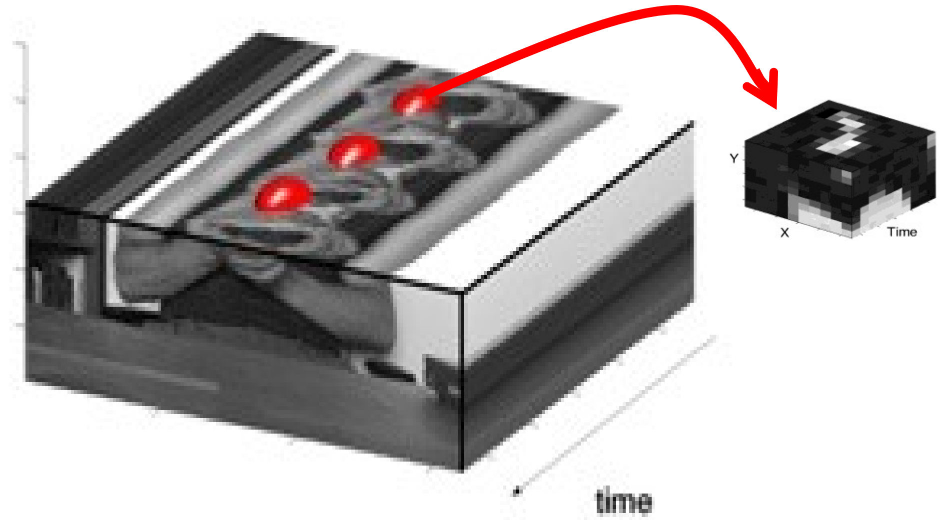
(1) Compute the Harris saliency map.

$$H \;=\; \det(\mu) - k\,\mathrm{trace}^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3,$$

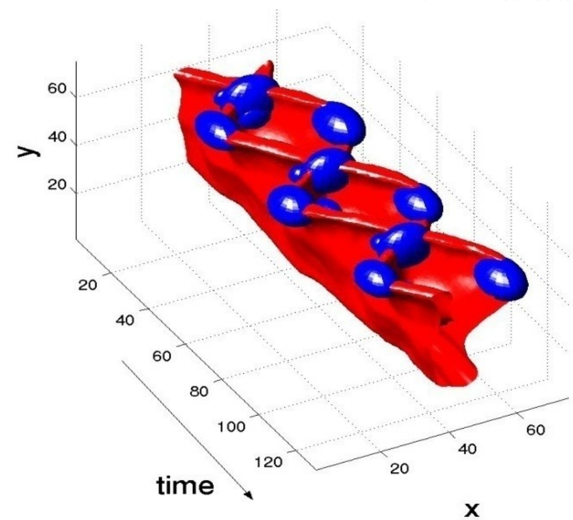
(2) Points with maximum saliency in a local volume => 3D corners.

Spatio-temporal Harris Corners

Hand clapping

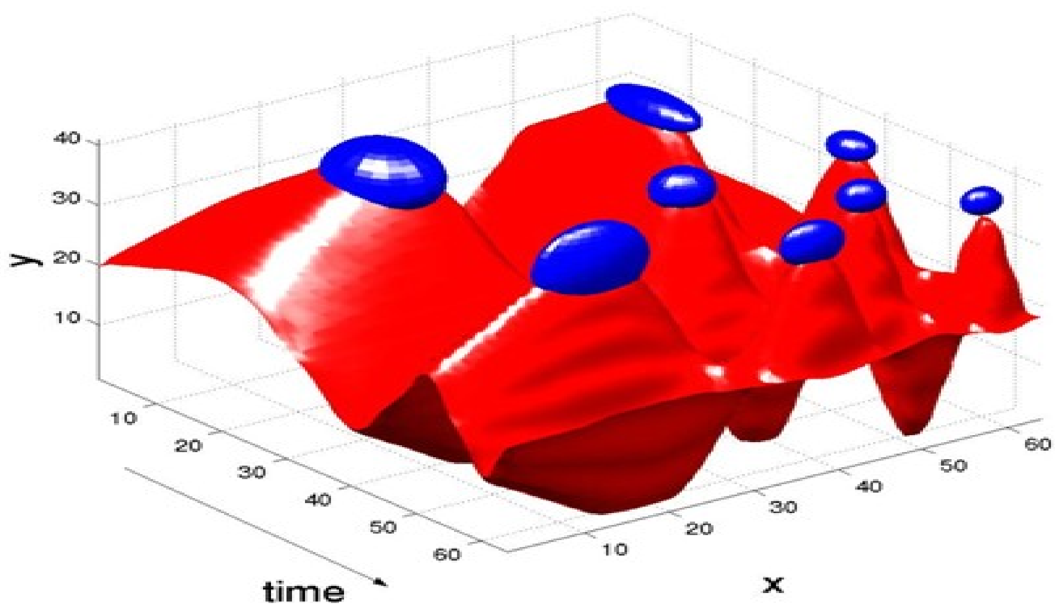


Walking



Why Multi-scale Salient Features?

- Spatial and temporal scale events require salient features at different spatial and temporal scales.



Spatio-temporal Hessian Blobs

Extension of the Hessian blob detector into 3D spatio-temporal domain:

- Spatial and temporal Gaussian filtering in the computation of the Hessian matrix .

$$H(\cdot;\sigma^2,\tau^2) = \begin{pmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{pmatrix}$$

$$L_{x^k y^l t^m}(\cdot;\sigma^2,\tau^2) = \partial_{x^k y^l t^m} g(\cdot;\sigma^2,\tau^2) * f(\cdot)$$

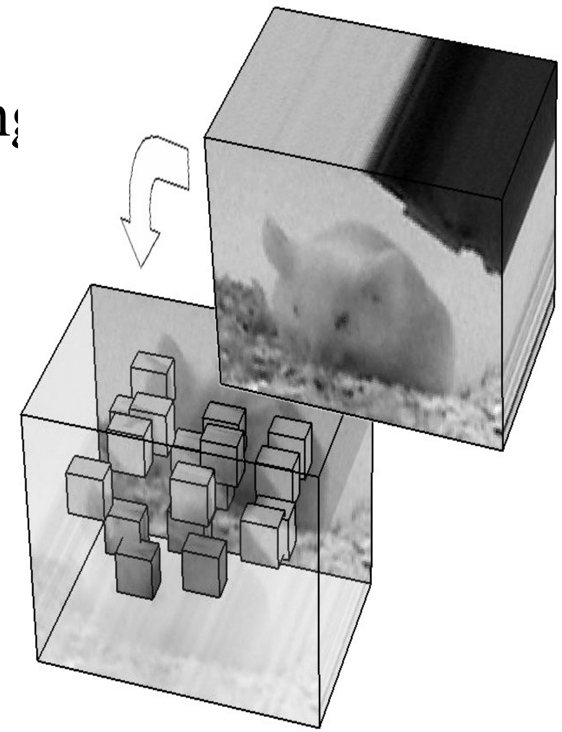
(1) Compute the determinant of Hessian matrix as the saliency map.

$$S = |det(H)|$$

(2) Points with maximum saliency in a local volume => center of 3D blobs.

Cuboids

- Spatial Gaussian along with temporal Gabor filtering



$$h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$$

$$h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$$

(1) Compute the energy of the filter response

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

(2) Points with maximum energy in a local volume => center of cuboids.

Other Spatio-temporal key Points

- Entropy-based interest points
- Dense sampling interest points
- 3D Scale-Invariant Feature Transform
- Salient opponent-based motion features (our paper in CRV 2010)
- ...

Spatio-temporal Feature Description

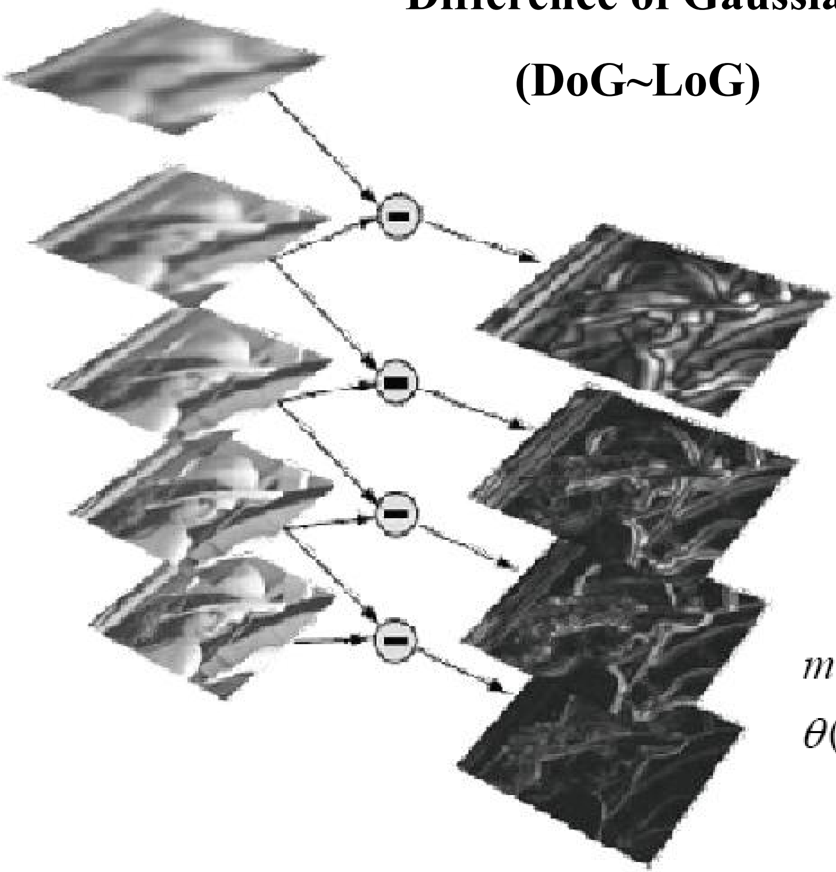
So far:

- For event detection in video, we utilize the salient features.
  - Video filtering at different spatio-temporal scales
  - Key point detection
  - Key point description using the characteristic of the point's surrounding volume.

2 Dimensional Scale-Invariant Feature Transform

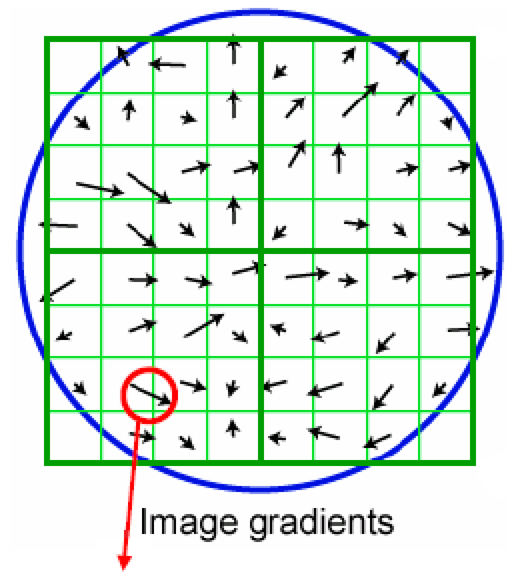
Gaussians (F)

Difference of Gaussian (DoG~LoG)



2D SIFT features

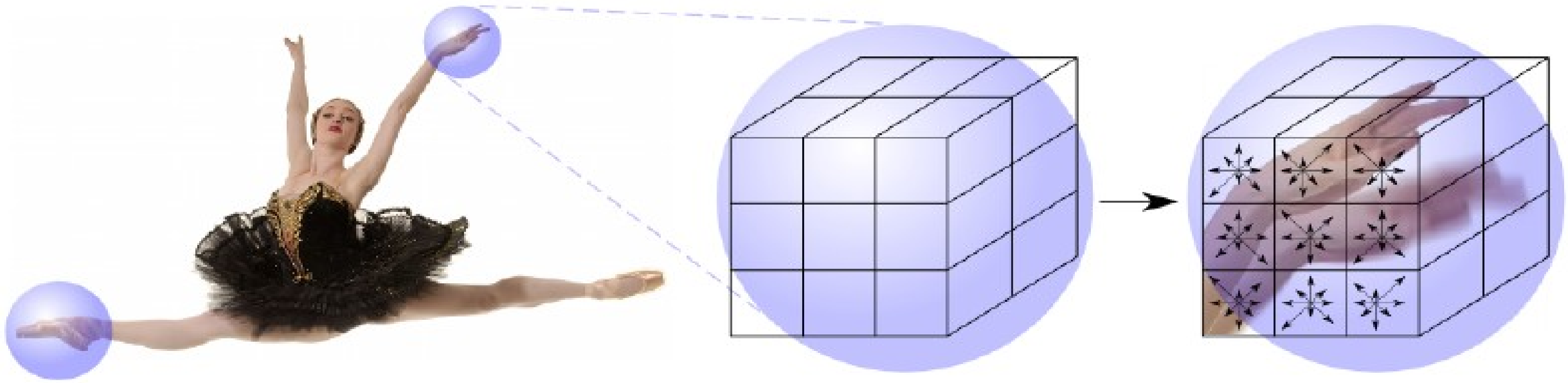
Oriented gradients

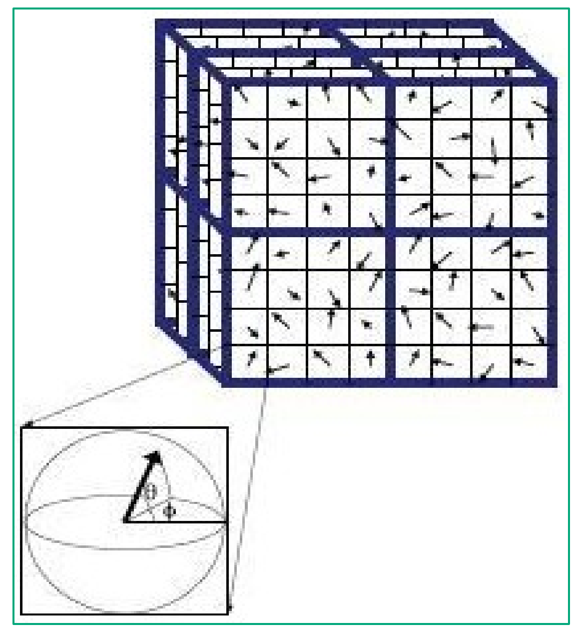


$$m(x,y)=\sqrt{(F(x+1,y)-F(x-1,y))^2+(F(x,y+1)-F(x,y-1))^2}$$

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

3D SIFT Descriptor





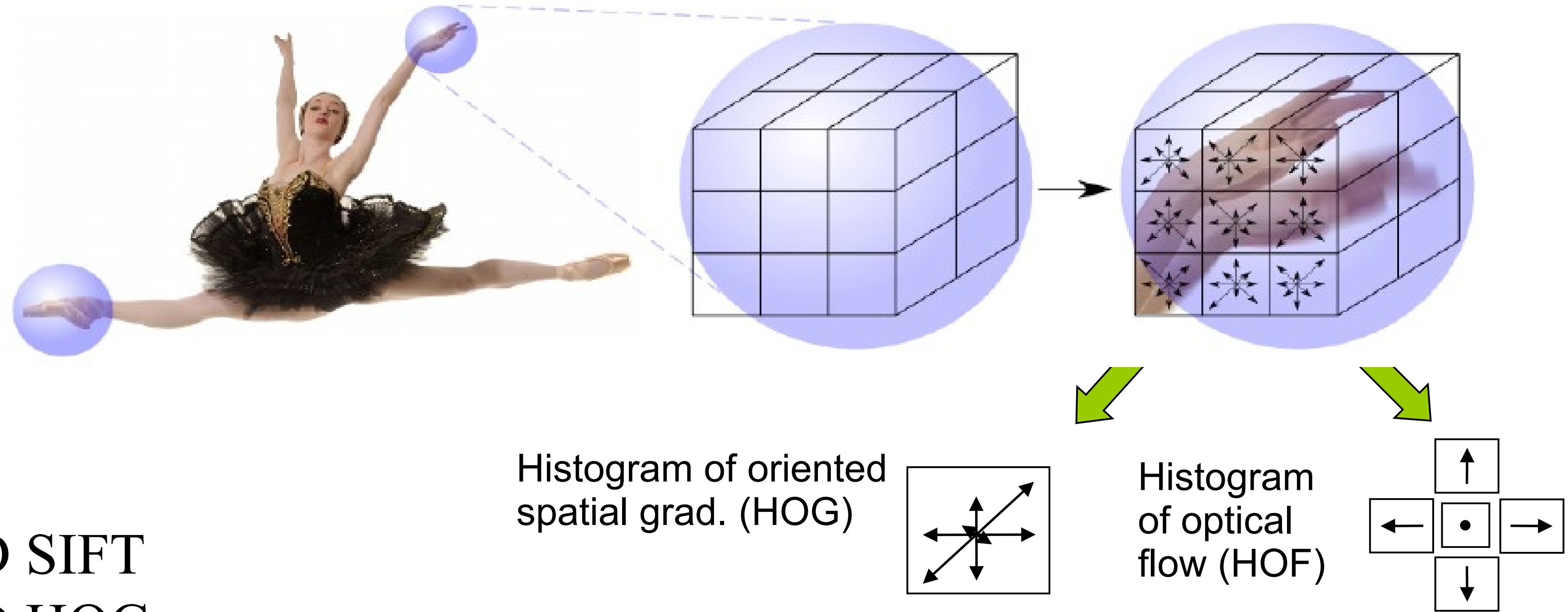
$$m_{3D}(x,y,t) = \sqrt{I_x^2 + I_y^2 + I_t^2},$$

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

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

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

Other Feature Descriptors



- 3D SIFT
- 3D HOG
- HOG + HOF
- Local jets , ...

Summary

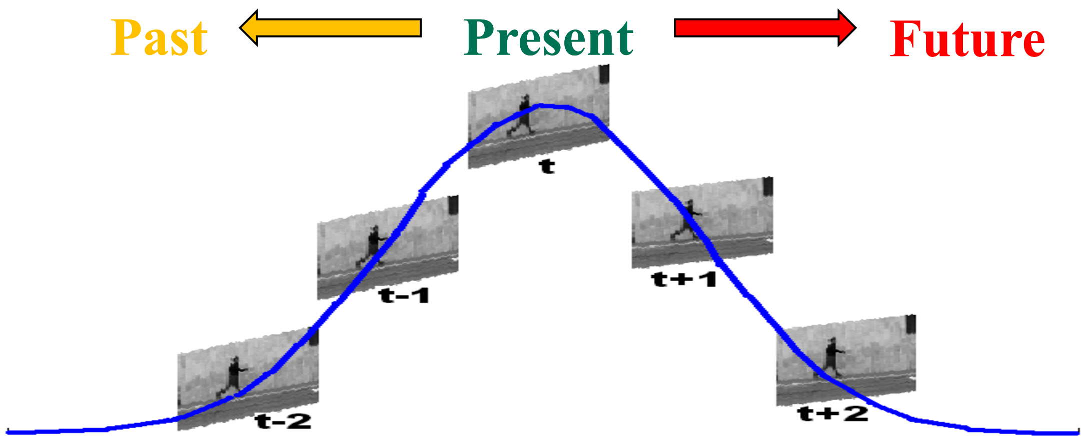
- For video analysis we can study the events characterised by spatio-temporal salient features.

- Spatial and temporal scale events require salient features at different spatial and temporal scales.

- Salient feature extraction requires scale-space filtering, interest point detection, and feature description.

A Challenge: Time Causality

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



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

  A: come to our paper presentation on Wednesday ☺

References

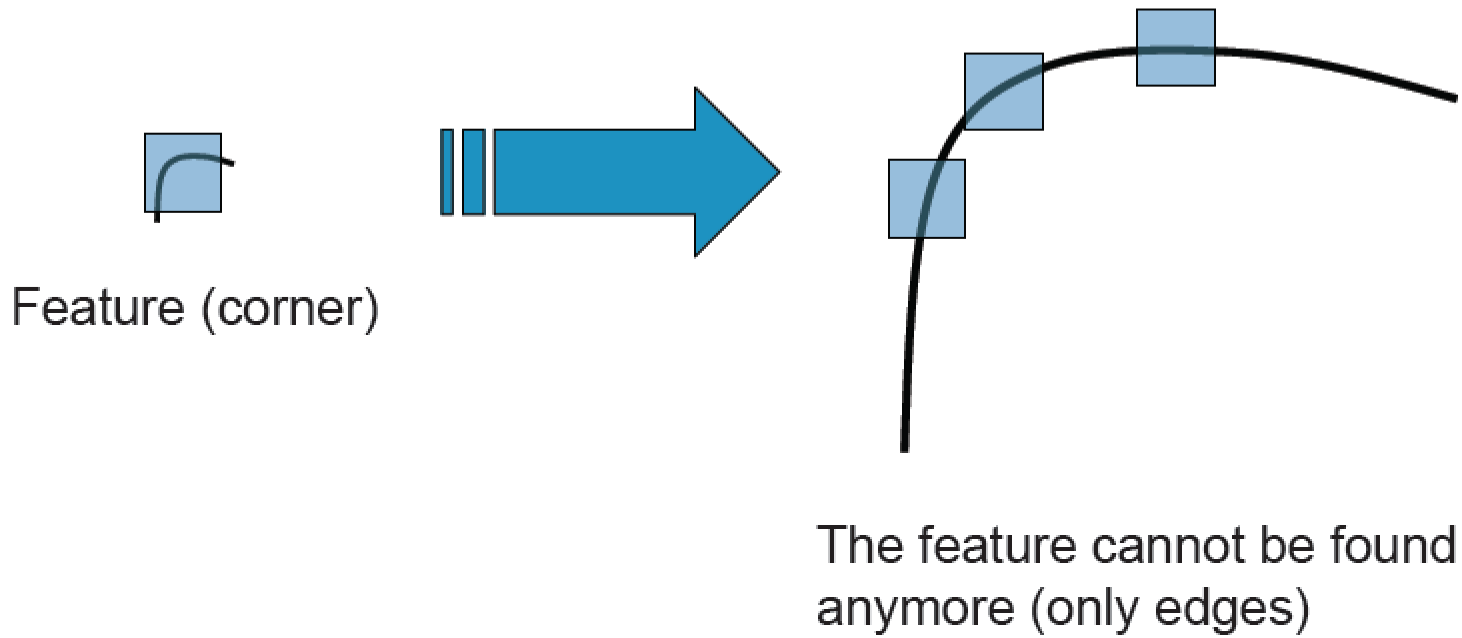
- **Evaluation of local spatio-temporal features for action recognition,** Wang et al., BMCV, 2009
- **Space–time interest points,** Laptev and Lindeberg, ICCV 2003.
- **Behavior recognition via sparse spatio-temporal filters,** Dollar et al., IEEE Workshop VS-PETS, 2005.
- **An efficient dense and scale-invariant spatio-temporal interest point detector,** Willems et al., ECCV 2008.
- **A comparison of affine covariant region detectors,** IJCV 2006, Mikolajczyk et al.

Thank you

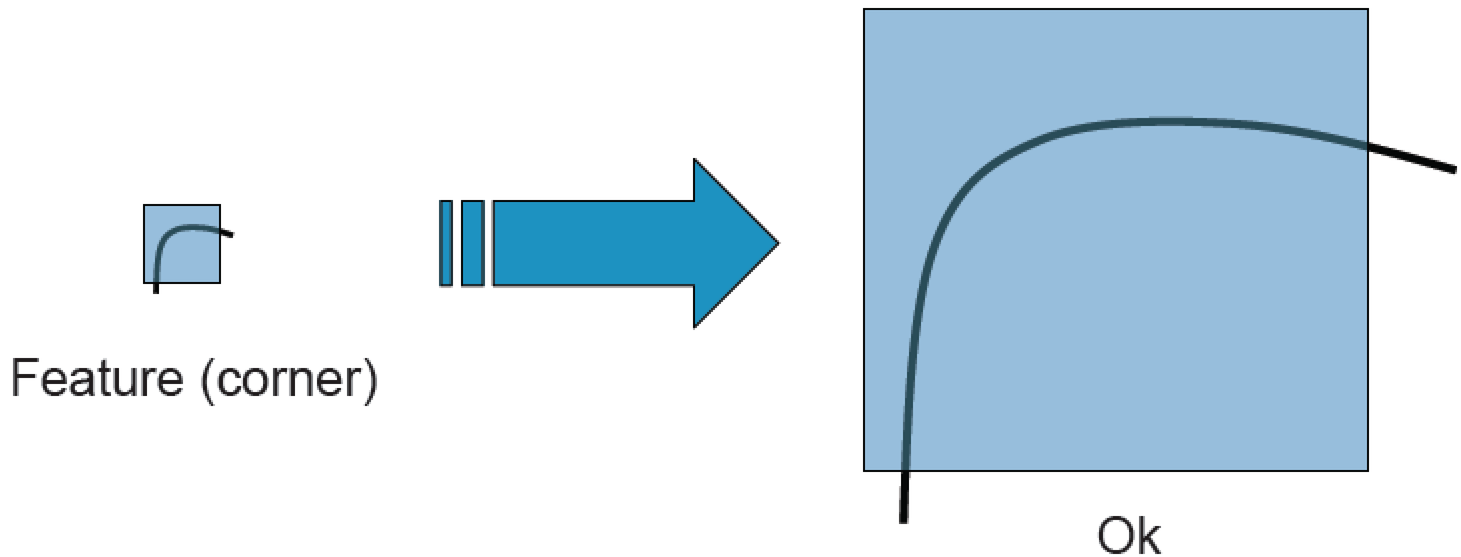
Definition

- Interest point:
  - Distinctive in a local region
    - rich in local image structure, with its surrounding
  - Repeatable : reproducible under different views
  - Stable to noise, geometric/photometric deformation
  - Has well-defined position in the image space

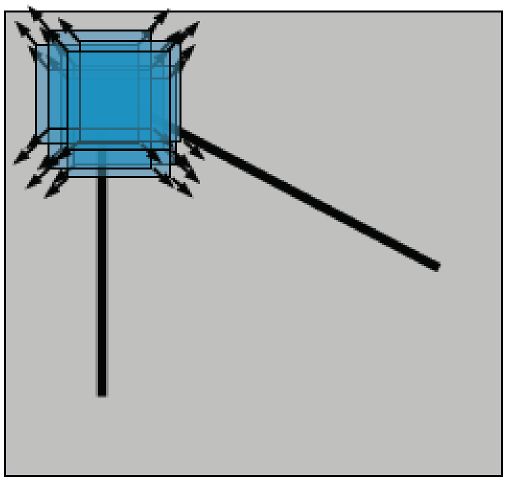
Covariant to Scale Change



Covariant to Scale Change



Some Requirements



Corner: significant changes in all directions

Scale change

Geometric transformation

Photometric transformation

Salient Features

- Scale-Invariant feature Transform (SIFT)
- Harris-affine corner detection
- Hessian-affine corner detection
- Edge-based Regions (EBR)
- Intensity-based Regions (IBR)
- Maximally Stable Extremal Regions (MSER)
- Entropy-based salient regions

Scale- Invariant Feature Transform (SIFT)

Scale-space Theorem:

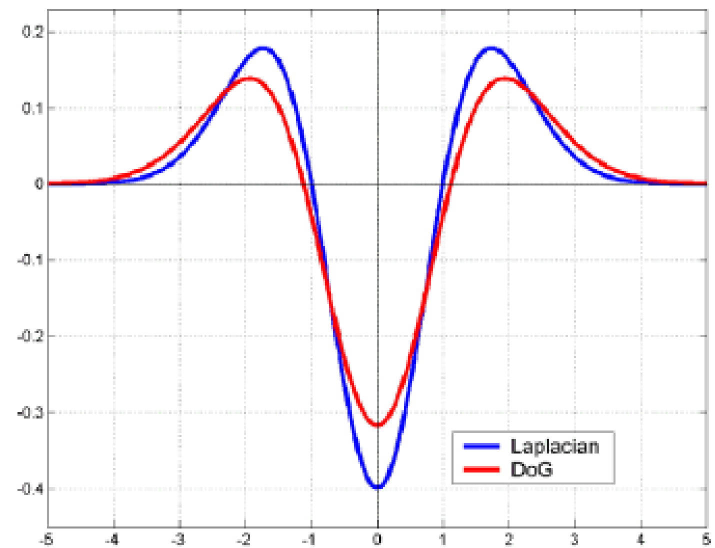
a local (3D) maximum of |NLoG| in (x,y,σ) is a point that can be identified with different size in images ---> it is a scale-invariant keypoint.

Laplacian Kernel

$$NLoG(x, y, \sigma) = \sigma^2 \nabla^2 G$$

Difference of Gaussians Kernel

$$DoG(x, y, \sigma) = G(x, y, k\sigma) - G(x, y, \sigma)$$



NOTE: both kernels find features invariant to *scale* and *rotation*

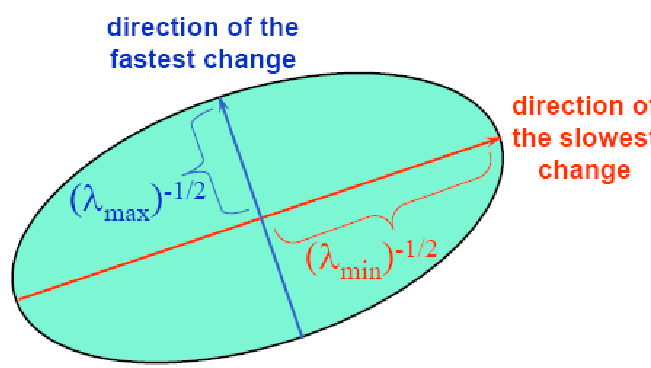
Harris-Affine Corner Detector

(1) compute gradient distribution matrix (M) in a local neighbourhood of a point.

$$M = \sigma_D^2 \; g(\sigma_I) \begin{bmatrix} I_{\mathrm{x}}^2(\mathrm{x}, \sigma_D) & I_x I_y(\mathrm{x}, \sigma_D) \\ I_x I_y(\mathrm{x}, \sigma_D) & I_{\mathrm{y}}^2(\mathrm{x}, \sigma_D) \end{bmatrix}$$

(2) find points for which both curvature are significant.

$$C = \det(M) - k\, trace^2(M)$$
$$= \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2$$



- Corners are stable in arbitrary lighting condition.