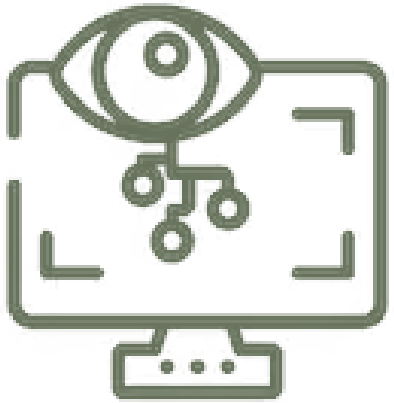


CS 554
Computer Vision

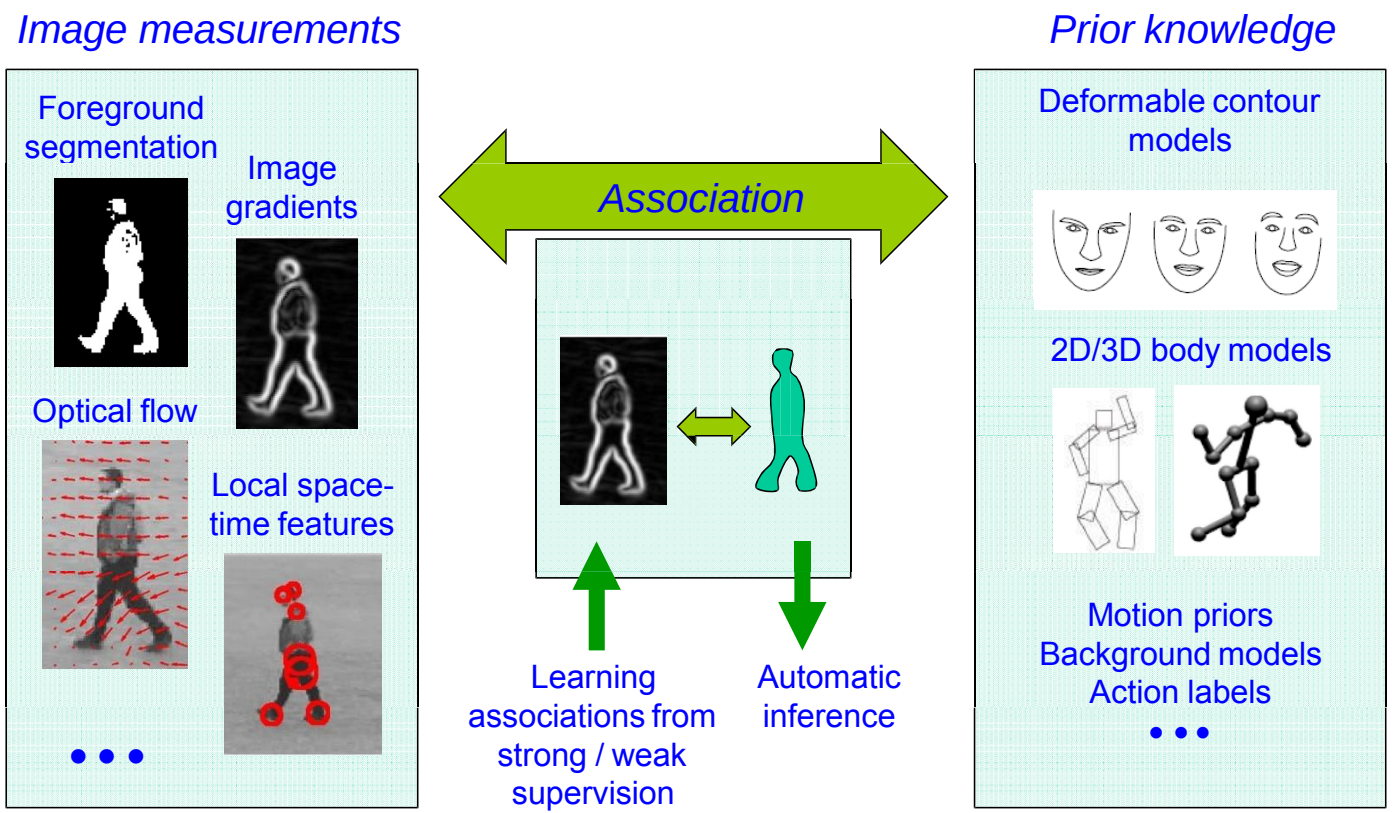
Action Recognition
& Motion Magnification

Hamdi Dibeklioğlu

Slide Credits: I. Laptev, C. Schmid, H-Y. Wu

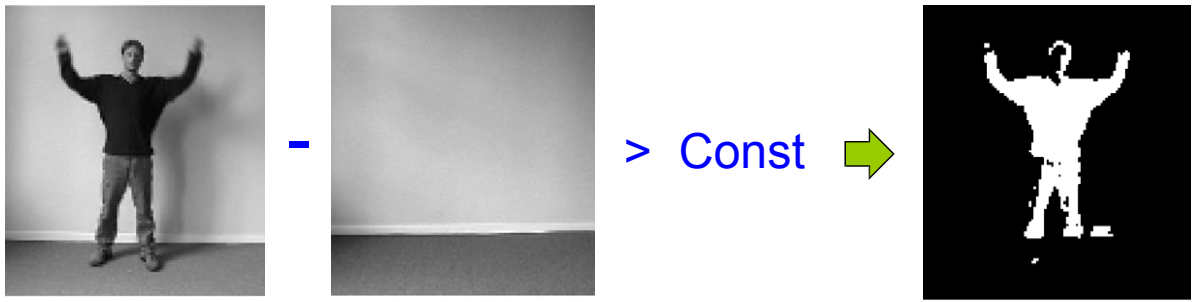
How to recognize actions?

Action understanding: Key components



Foreground segmentation

Image differencing: a simple way to measure motion/change



Better Background / Foreground separation methods exist:

- Modeling of color variation at each pixel with Gaussian Mixture
- Dominant motion compensation for sequences with moving camera
- Motion layer separation for scenes with non-static backgrounds

Temporal Templates

$D(x, y, t)$ $t = 1, ..., T$

Idea: summarize motion in video in a *Motion History Image (MHI)*:

$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } \Psi(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t-1) - \delta) & \text{otherwise} \end{cases}$$

Descriptor: Hu moments of different orders

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q \rho(x, y) dx dy$$

[Bobick and Davis, PAMI 2001]

Temporal Templates

Descriptor: Hu moments of different orders

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q \rho(x, y) dx dy$$

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})^p (y - \bar{y})^q \rho(x, y) d(x - \bar{x}) d(y - \bar{y}) ,$$

$$\text{where} \quad \begin{aligned} \bar{x} &= m_{10}/m_{00} , \\ \bar{y} &= m_{01}/m_{00} . \end{aligned}$$

[Bobick and Davis, PAMI 2001]

Moments invariant to translation, scale, and orientation

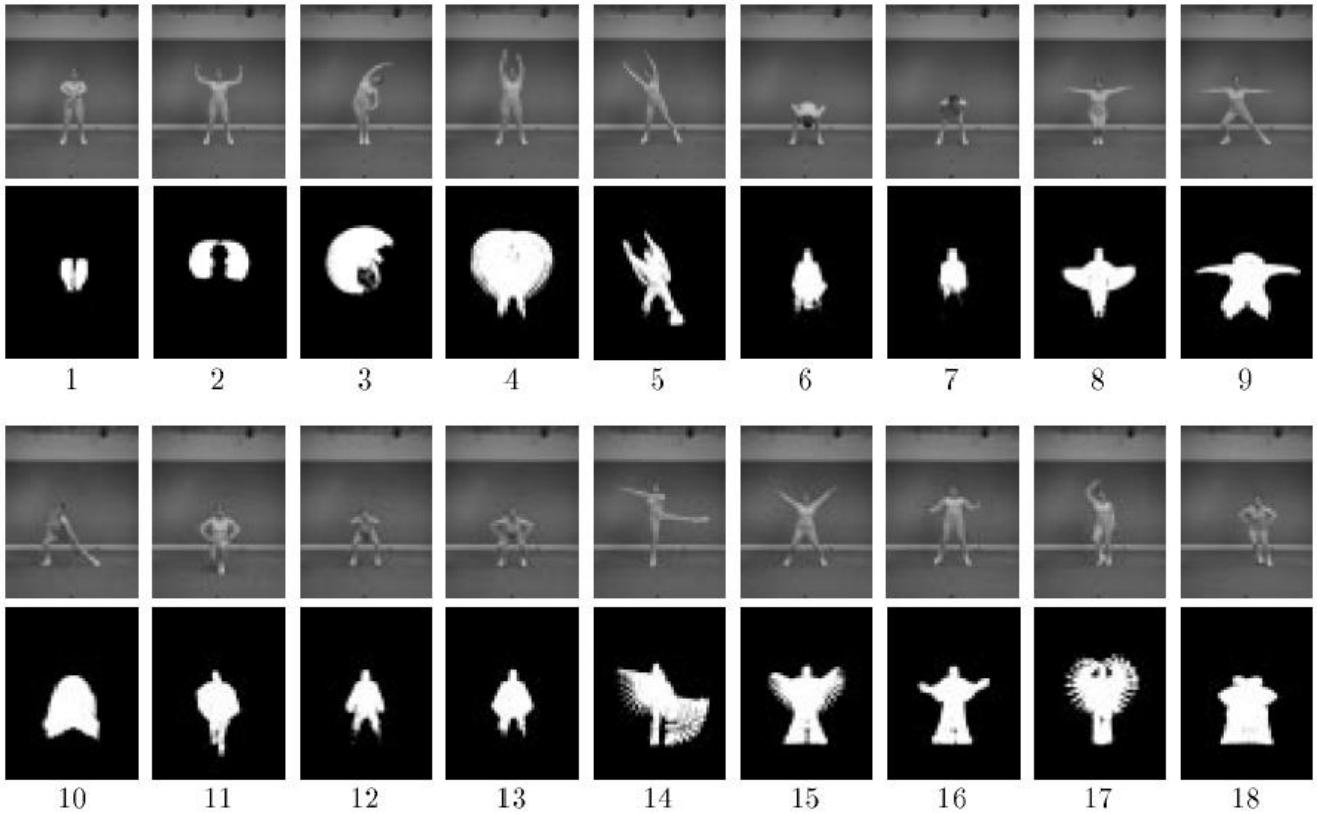
- For the second and third order moments, we have the following seven translation, scale, and orientation moment invariants:

$$
\begin{aligned}
\nu_1 &= \mu_{20} + \mu_{02} \\
\nu_2 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2 \\
\nu_3 &= (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \\
\nu_4 &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \\
\nu_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
&\quad +(3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \\
&\quad \cdot[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\
\nu_6 &= (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\
&\quad +4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \\
\nu_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
&\quad -(\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]
\end{aligned}
$$

- These moments can used for pattern identification

[Bobick and Davis, PAMI 2001]

Aerobics dataset



Nearest Neighbor classifier: 66% accuracy

[Bobick and Davis, PAMI 2001]

Temporal Templates: Summary

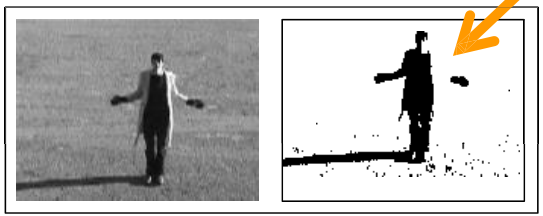
Pros:

+ Simple and fast
+ Works in controlled settings

Cons:

- Prone to errors of background subtraction

Not all shapes are valid
➡ Restrict the space
of admissible silhouettes



Variations in light, shadows, clothing…

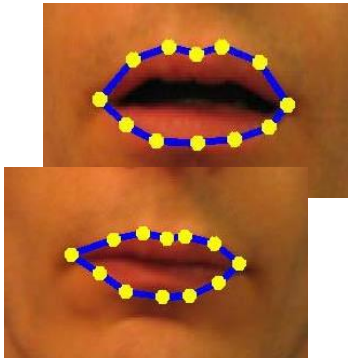
What is the background here?

- Does not capture *interior* motion and shape

Silhouette tells little about actions

Shape-based Representation

- Shape parameters can be used for activity/action representation:
  - Active appearance models
  - Active shape models
  - …

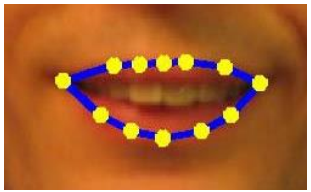


Active Shape Models

Point Distribution Model

- Represent the shape of samples by a set of corresponding points or *landmarks*

$$\mathbf{x} = (x_1, \ldots, x_n, y_1, \ldots, y_n)^T$$



- Assume each shape can be represented by the linear combination of basis shapes

$$\mathbf{\Phi} = (\phi_1 | \phi_2 | \ldots | \phi_t)$$

such that $\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{\Phi b}$

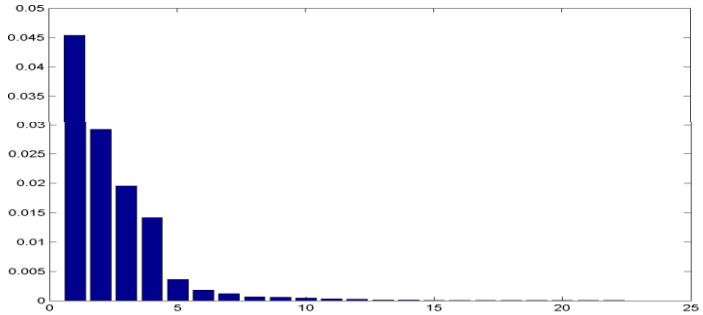
for the mean shape $\bar{\mathbf{x}} = \frac{1}{s} \sum_{i=1}^{s} \mathbf{x}_i$

and some parameter vector $\mathbf{b}$

[Cootes et al. 1995]

Active Shape Models

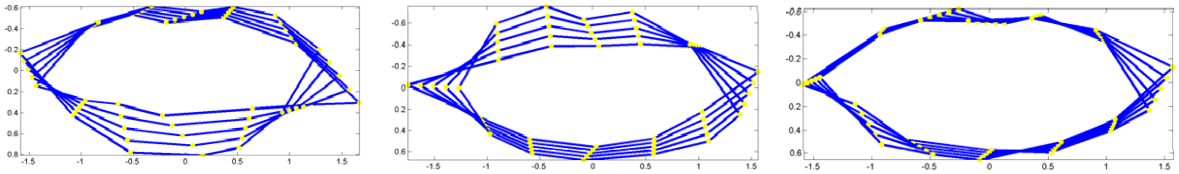
- Distribution of eigenvalues of $\mathbf{S}: \lambda_1, \lambda_2, \lambda_3, \ldots$



A small fraction of basis shapes accounts for the most of shape variation (=> landmarks are redundant) (eigenvecors)

- Three main modes of lips-shape variation:

$\mathbf{b} = (\mu\lambda_1, 0, 0, \ldots)^\top$ $\mathbf{b} = (0, \mu\lambda_2, 0, 0, \ldots)^\top$ $\mathbf{b} = (0, 0, \mu\lambda_3, 0, 0, \ldots)^\top$



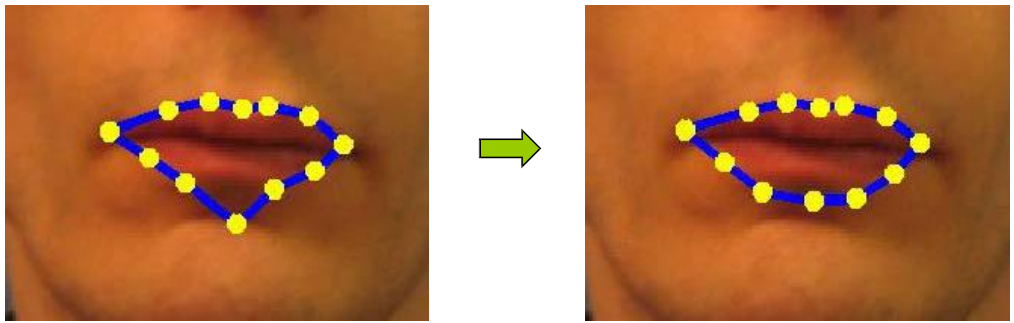
$\mu = -3, 1.5, 0, 1.5, 3$

Active Shape Models

Effect of regularization

- Projection onto the shape-space serves as a regularization

$$\mathbf{x} \Rightarrow \mathbf{b} = \Phi^\top(\mathbf{x} - \bar{\mathbf{x}}) \Rightarrow \mathbf{x}' = \bar{\mathbf{x}} + \Phi\mathbf{b}$$



Active Shape Models: Summary

Pros:

+ Shape prior helps overcoming segmentation errors
+ Fast optimization
+ Can handle interior/exterior dynamics

Cons:

- Optimization gets trapped in local minima
- Re-initialization is problematic

Possible improvements:

- Learn and use motion priors, possibly specific to different actions

Motion priors

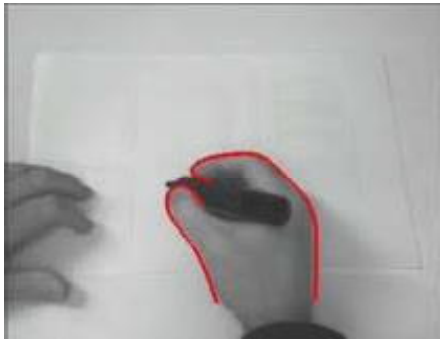
- Accurate motion models can be used both to:
  - ❖ Help accurate tracking
  - ❖ Recognize actions

- Goal: formulate motion models for different types of actions and use such models for action recognition

Example:

Drawing with 3 action modes

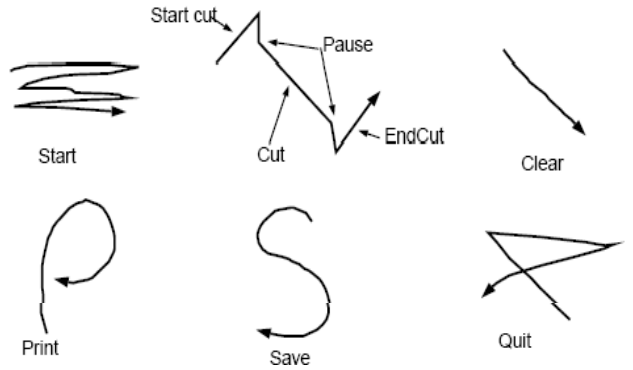
- line drawing
- scribbling
- idle

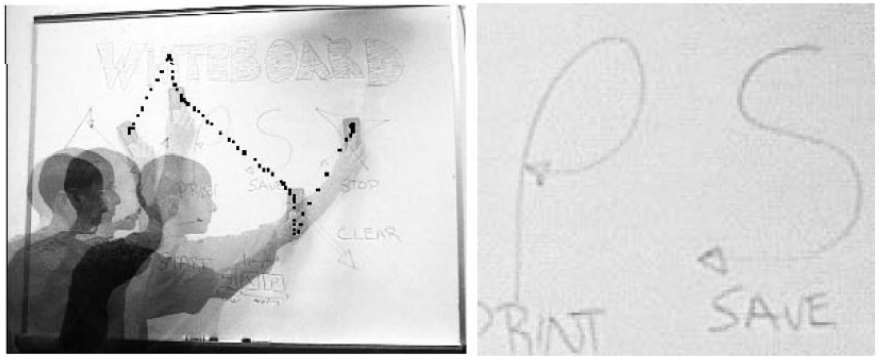


[Isard and Blake, ICCV 1998]

Dynamics with discrete states

Example: Gesture recognition in the context of a visual black-board interface





[Black and Jepson, ECCV 1998]

Motion priors & Tracking: Summary

Pros:

+ more accurate tracking using specific motion models
+ Simultaneous tracking and motion recognition with discrete state dynamical models

Cons:

- Local minima is still an issue
- Re-initialization is still an issue

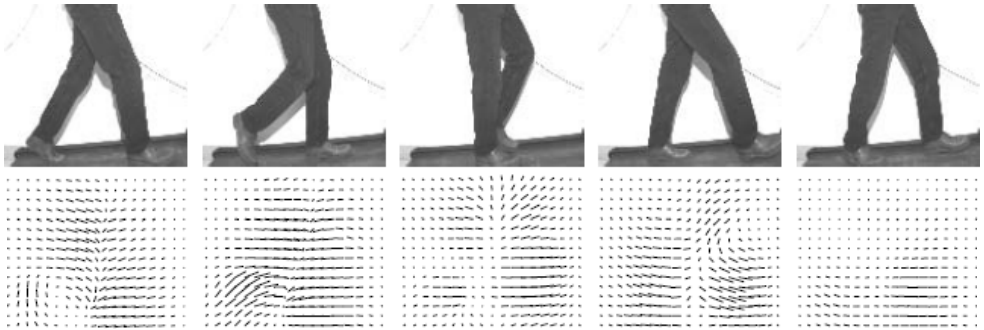
Shape and Appearance versus Motion

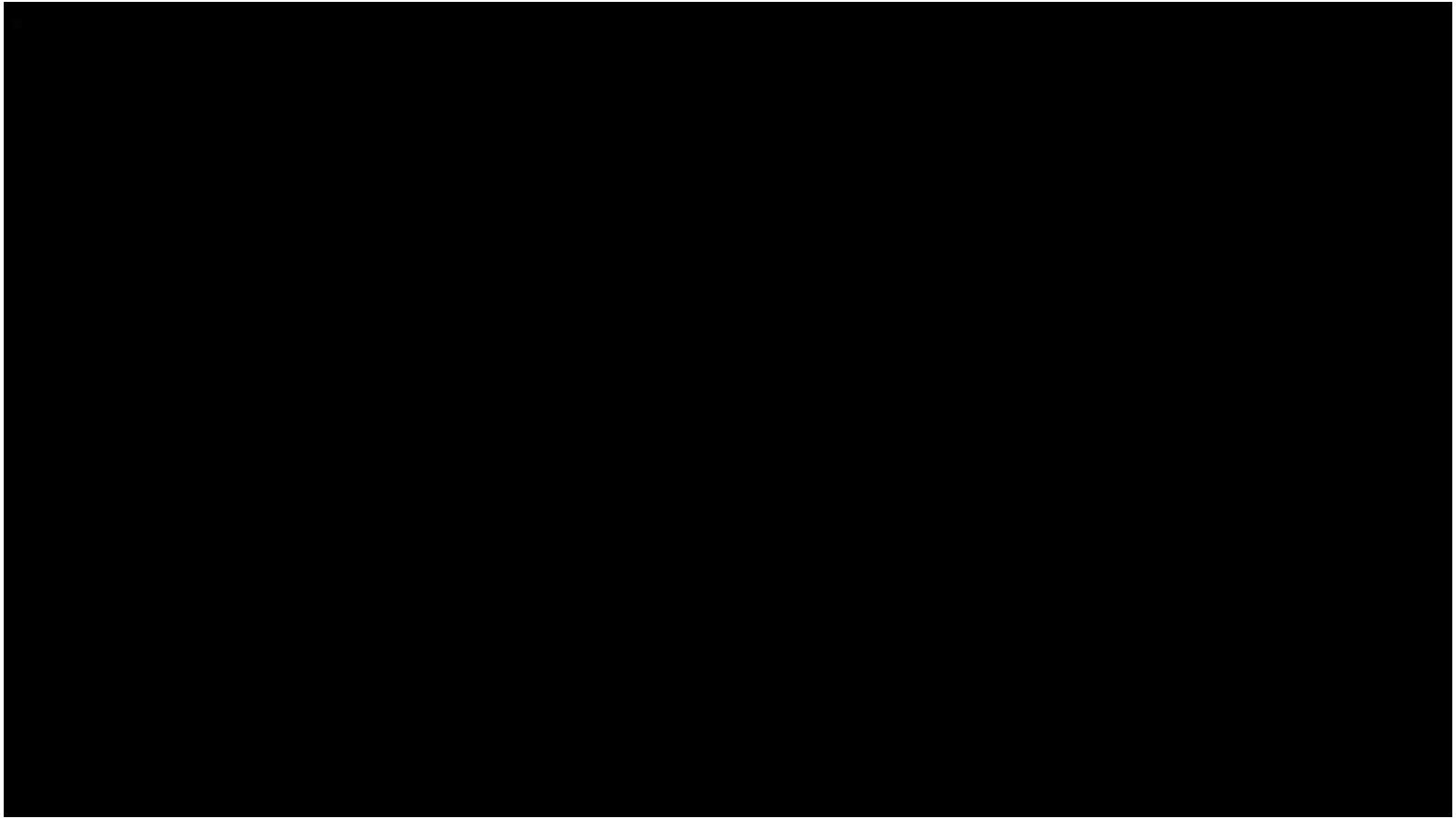
- Shape and appearance in images depends on many factors: clothing, illumination contrast, image resolution, etc…

Shape and Appearance versus Motion

- Shape and appearance in images depends on many factors: clothing, illumination contrast, image resolution, etc…

- Motion field (in theory) is invariant to shape and can be used directly to describe human actions





Gunnar Johansson, **Moving Light Displays,** 1973

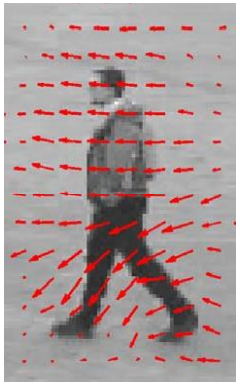
Motion estimation: Optical Flow (recap)

- Classic problem of computer vision [Gibson 1955]

- Goal: estimate motion field

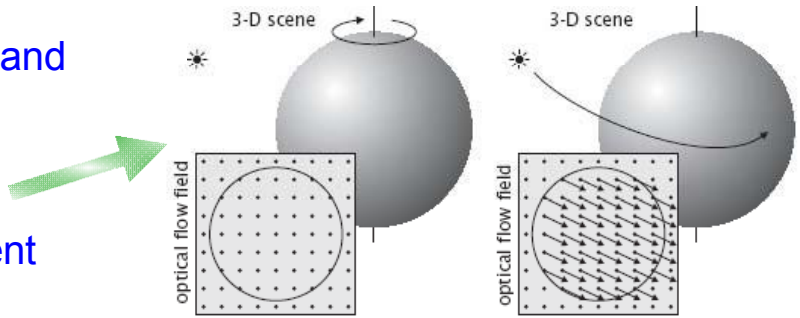
  How? We only have access to image pixels

  ➡ Estimate pixel-wise correspondence between frames = Optical Flow



- Brightness Change assumption: corresponding pixels preserve their intensity (color)

  - ❖ Useful assumption in many cases
  - ❖ Breaks at occlusions and illumination changes
  - ❖ Physical and visual motion may be different



Parameterized Optical Flow

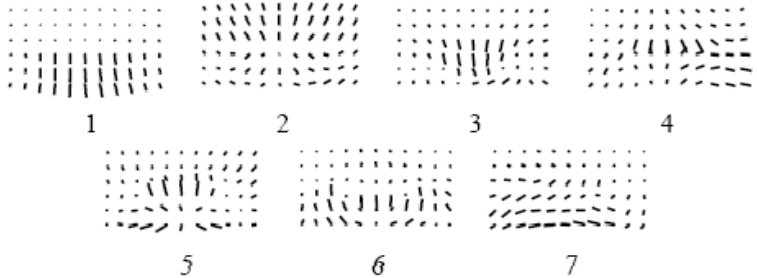
1. Compute standard Optical Flow for many examples
2. Put velocity components into one vector

$$\mathbf{w} = (v_x^1, v_y^1, v_x^2, v_y^2, ..., v_x^n, v_y^n)^\top$$

3. Do PCA on $\mathbf{w}$ and obtain most informative PCA flow basis vectors

Training samples

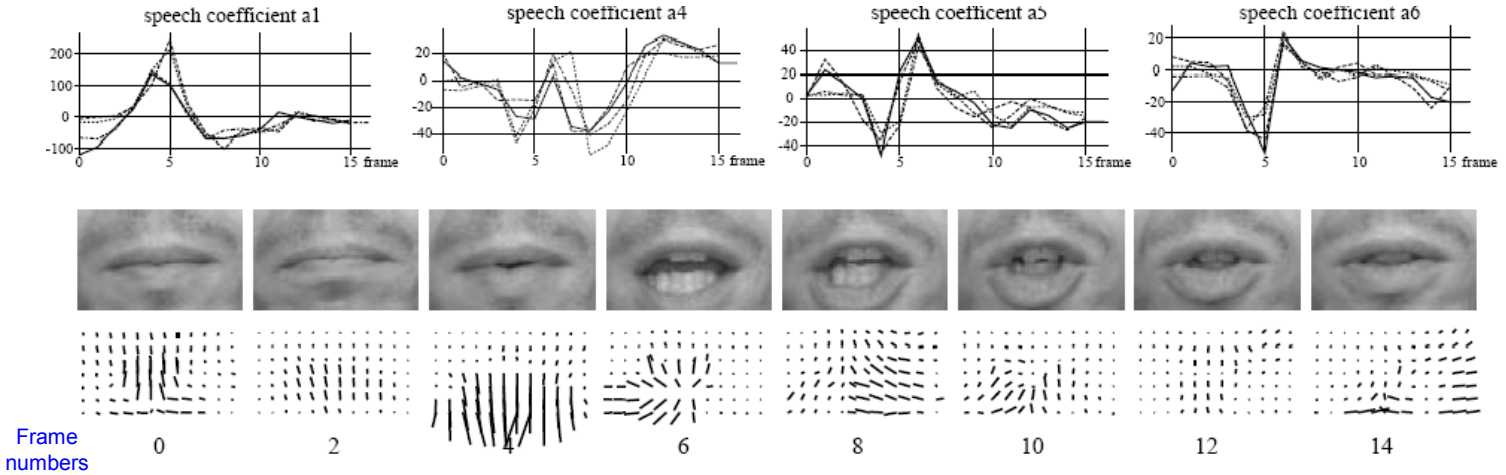
PCA flow bases



[Black, Yacoob, Jepson, Fleet, CVPR 1997]

Parameterized Optical Flow

- Estimated coefficients of PCA flow bases can be used as action descriptors



➡ Optical flow seems to be an interesting descriptor for motion/action recognition

[Black, Yacoob, Jepson, Fleet, CVPR 1997]

Spatial Motion Descriptor

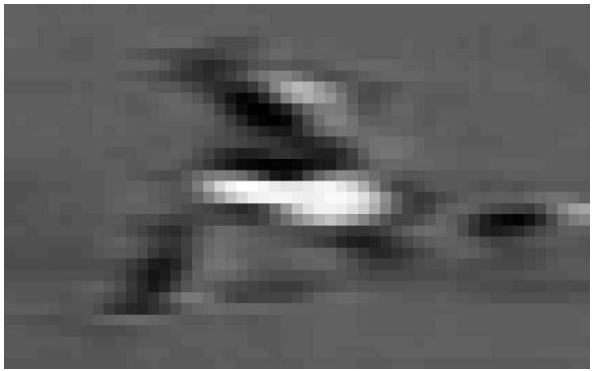
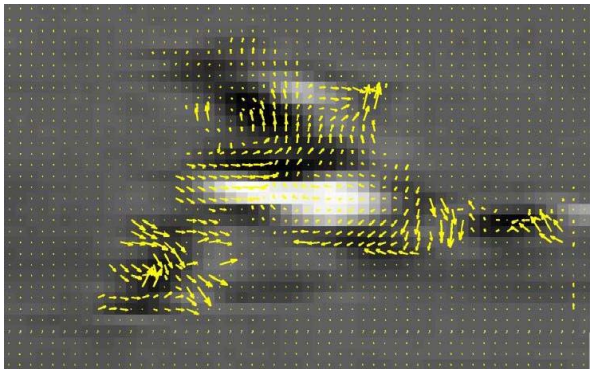
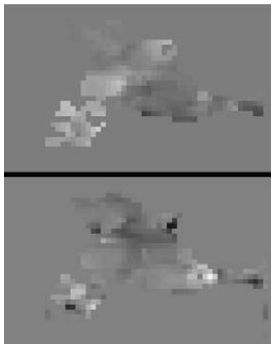


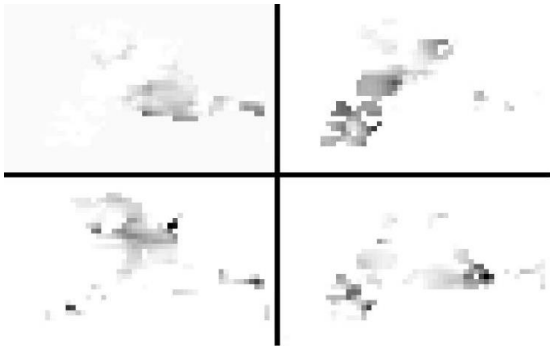
Image frame



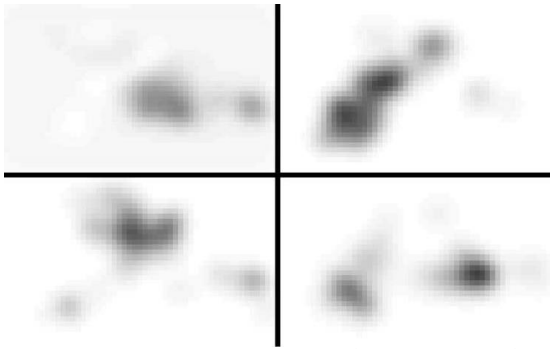
Optical flow $F_{x,y}$



F_x, F_y



$F_x^-, F_x^+, F_y^-, F_y^+$



blurred $F_x^-, F_x^+, F_y^-, F_y^+$

[Efros, Berg, Mori and Malik, ICCV 2003]

Football Actions: matching

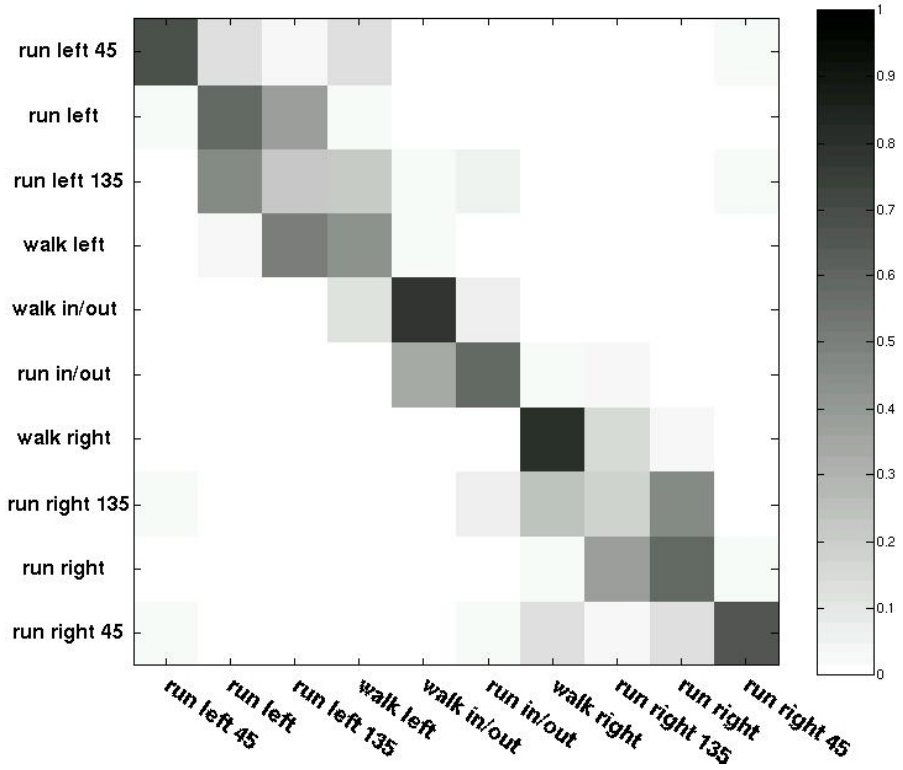
Input Sequence

Matched Frames

input matched

[Efros, Berg, Mori and Malik, ICCV 2003]

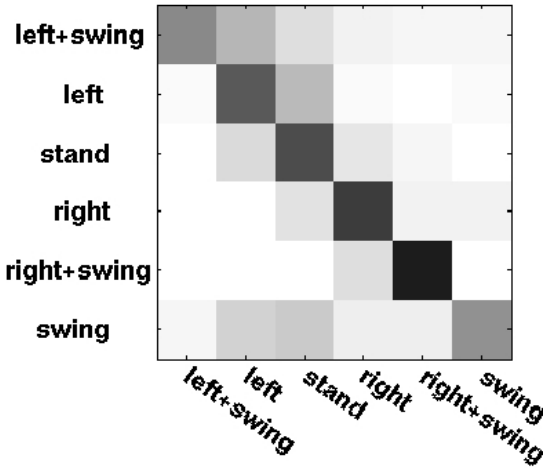
Football Actions: classification



10 actions; 4500 total frames; 13-frame motion descriptor

[Efros, Berg, Mori and Malik, ICCV 2003]

Tennis Actions: classification



6 actions; 4600 frames; 7-frame motion descriptor
Woman player used as training, man as testing.

[Efros, Berg, Mori and Malik, ICCV 2003]

Classifying Tennis Actions

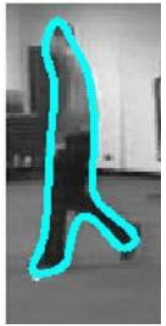
Red bars illustrate classification confidence for each action

[A. A. Efros, A. C. Berg, G. Mori, J. Malik, ICCV 2003]

Where are we so far ?

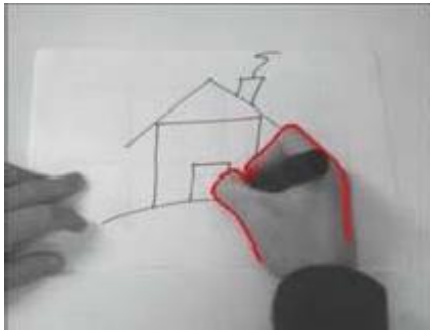
Temporal templates:

+ simple, fast
- sensitive to segmentation errors



Active shape models:

+ shape regularization
- sensitive to initialization and tracking failures

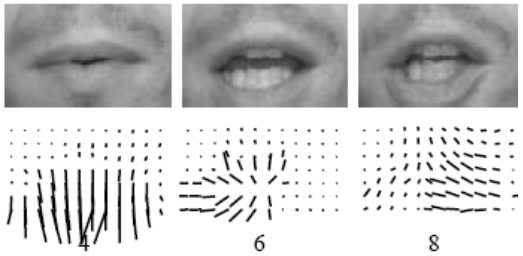


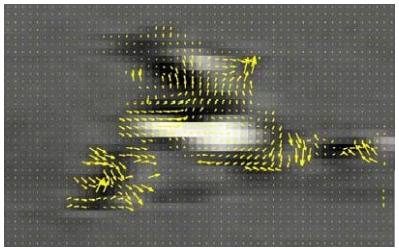
Tracking with motion priors:

+ improved tracking and simultaneous action recognition
- sensitive to initialization and tracking failures

Motion-based recognition:

+ generic descriptors; less depends on appearance
- sensitive to localization/tracking errors





How to handle real complexity?

Common methods:

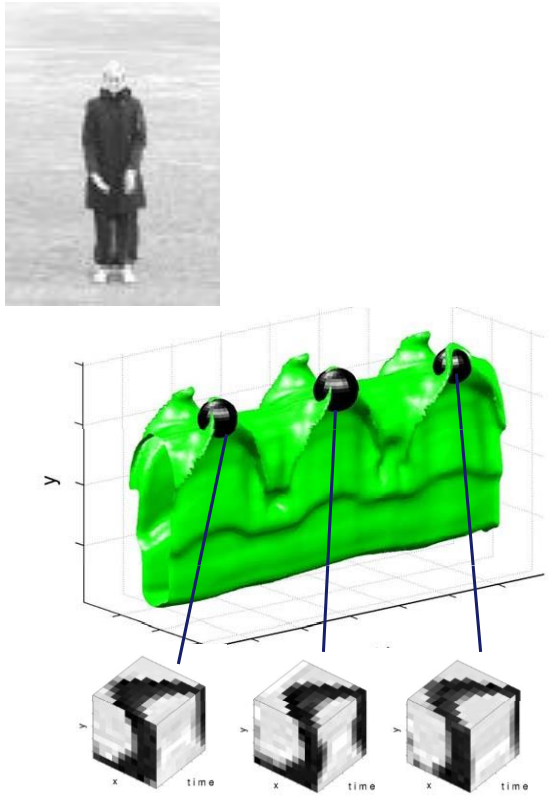
- Camera stabilization
- Segmentation ?
- Tracking ?

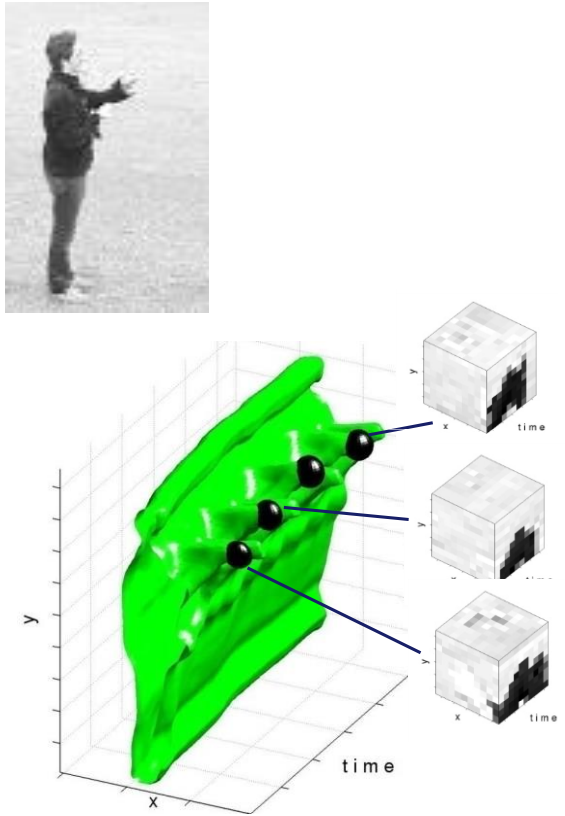
Common problems:

- Complex & changing background
- Changes in appearance

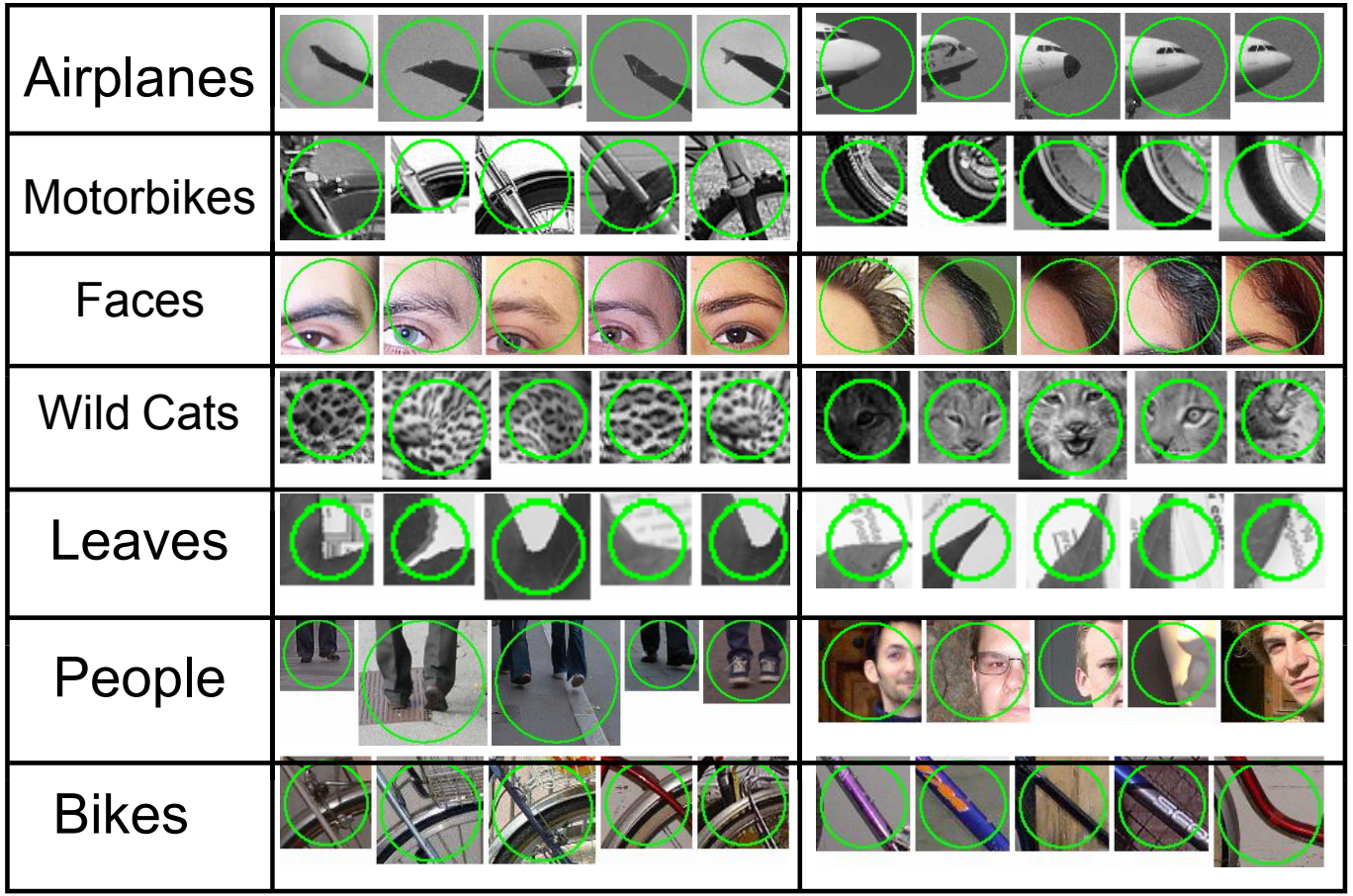
➡ Avoid global assumptions!

No global assumptions → Local evidence





Relation to local image features



Space-Time Interest Points

What neighborhoods to consider?

Distinctive neighborhoods $\Rightarrow$ High image variation in space and time $\Rightarrow$ Look at the distribution of the gradient

Definitions:

$f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ Original image sequence

$g(x, y, t; \Sigma)$ Space-time Gaussian with covariance

$L_\xi(\cdot; \Sigma) = f(\cdot) * g_\xi(\cdot; \Sigma)$ Gaussian derivative of f

$\nabla L = (L_x, L_y, L_t)^T$ Space-time gradient

$$\mu(\cdot; \Sigma) = \nabla L(\cdot; \Sigma)(\nabla L(\cdot; \Sigma))^T * g(\cdot; s\Sigma) = \begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix}$$

Second-moment matrix

[Laptev, IJCV 2005]

Space-Time Interest Points

Properties of $\mu(\cdot;\ \Sigma)$

$\mu(\cdot;\ \Sigma)$ defines second order approximation for the local distribution of ∇L within neighborhood Σ

$\text{rank}(\mu) = 1 \quad \Rightarrow$ 1D space-time variation of f e.g. moving bar

$\text{rank}(\mu) = 2 \quad \Rightarrow$ 2D space-time variation of f e.g. moving ball

$\text{rank}(\mu) = 3 \quad \Rightarrow$ 3D space-time variation of f e.g. jumping ball

Large eigenvalues of μ can be detected by the local maxima of H over (x,y,t):

$$\begin{aligned} H(p;\ \Sigma) &= \det(\mu(p;\ \Sigma)) + k\,\text{trace}^3(\mu(p;\ \Sigma)) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3 \end{aligned}$$

(similar to Harris operator [Harris and Stephens, 1988])

[Laptev, IJCV 2005]

Space-Time Interest Points

Large eigenvalues of μ can be detected by the local maxima of H over (x,y,t):

$$\begin{aligned} H(p;\Sigma) &= \det(\mu(p;\Sigma)) + k\,\mathrm{trace}^3(\mu(p;\Sigma)) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3 \end{aligned}$$

To show how positive local maxima of H correspond to points with high values of $\lambda_1, \lambda_2, \lambda_3$ ($\lambda_1 \le \lambda_2 \le \lambda_3$), we define the ratios $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$ and re-write H as

$$H = \lambda_1^3(\alpha\beta - k(1+\alpha+\beta)^3).$$

Spatio-temporal interest points can be found by detecting local positive spatio-temporal maxima in H.

[Laptev, IJCV 2005]

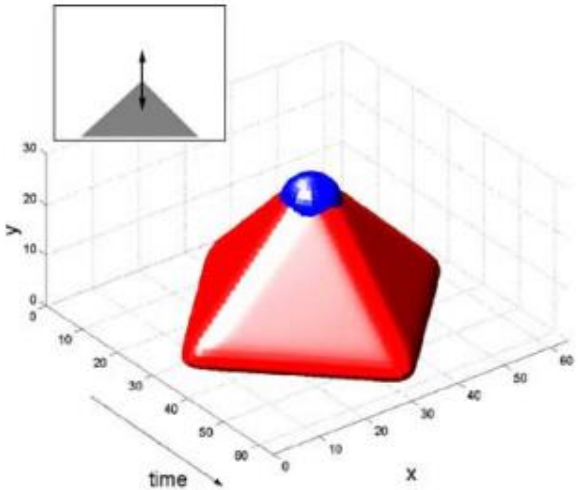
Space-Time Interest Points: Examples

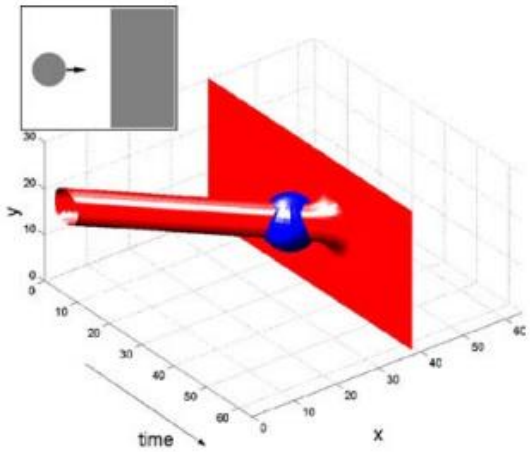
Motion event detection: synthetic sequences

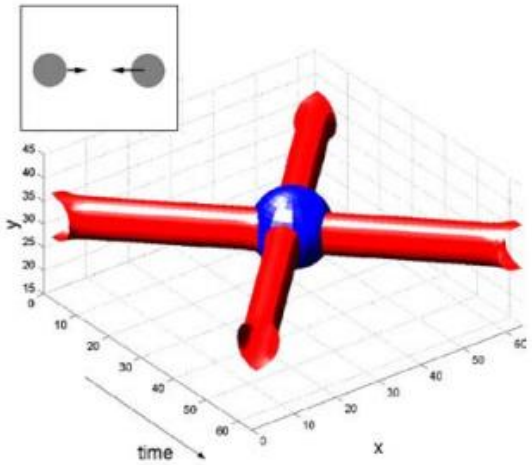
accelerations

appearance/
disappearance

split/merge







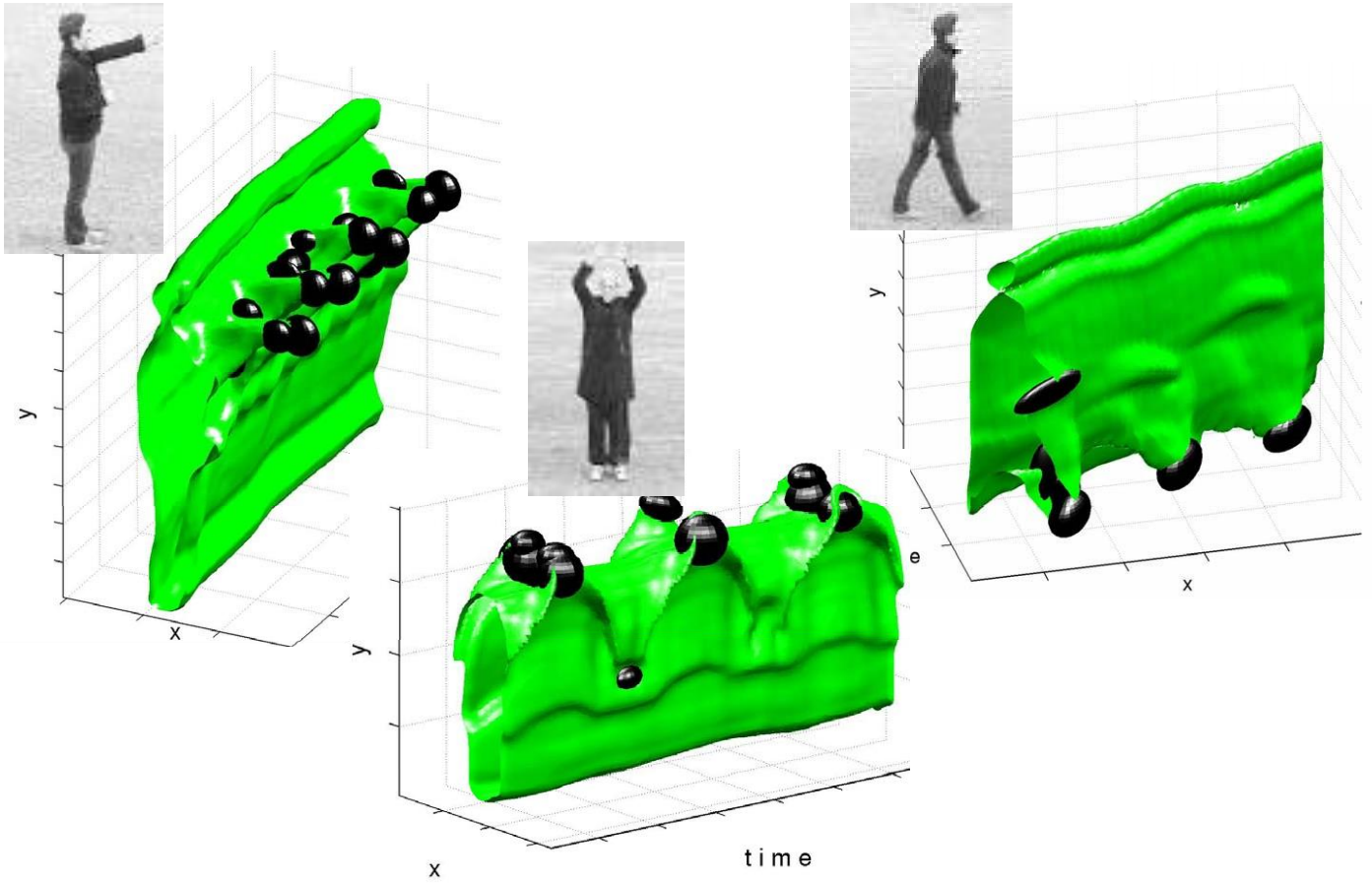
[Laptev, IJCV 2005]

Space-Time Interest Points: Examples

Motion event detection: complex background

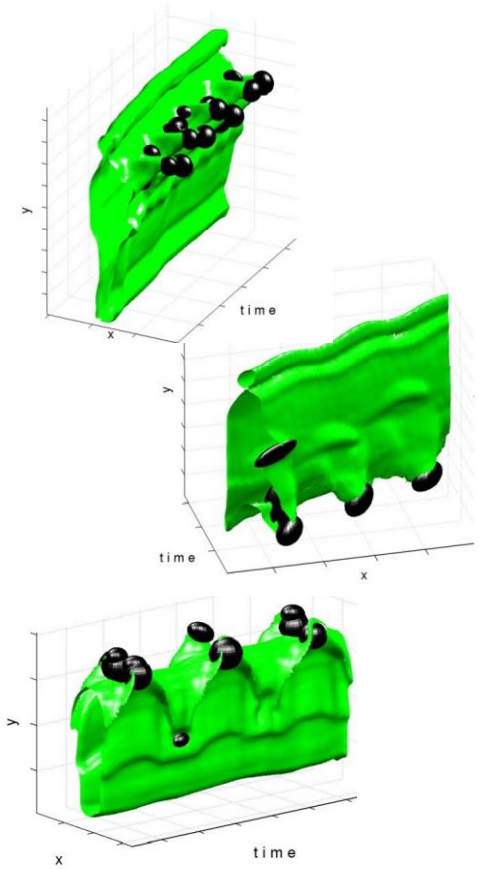
[Laptev, IJCV 2005]

Features from human actions

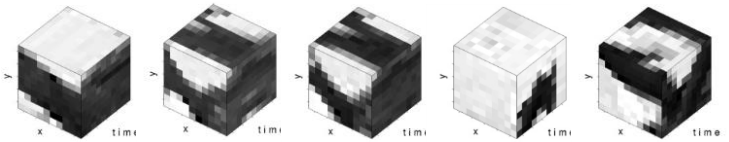


[Laptev, IJCV 2005]

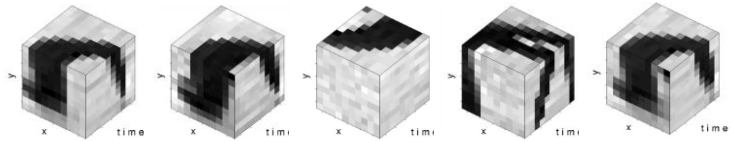
Features from human actions



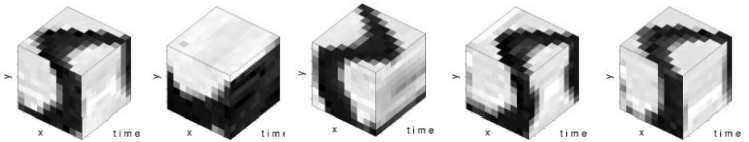
boxing



walking



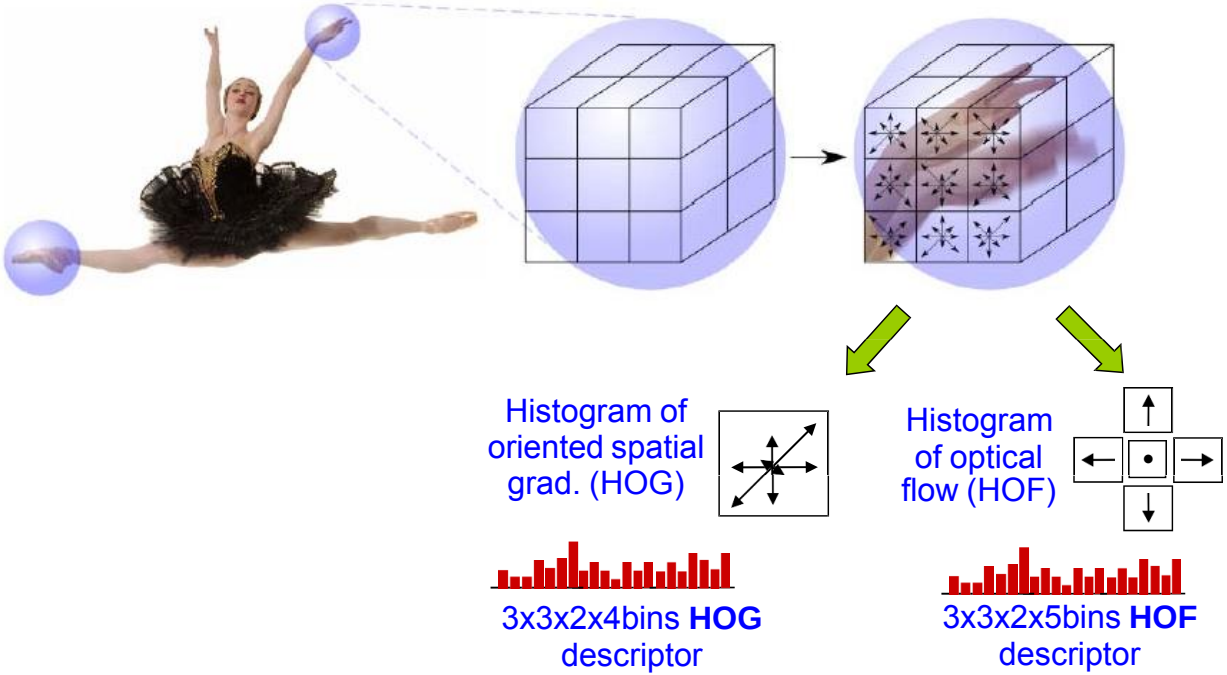
hand waving



[Laptev, IJCV 2005]

Space-Time Features: Descriptor

Multi-scale space-time patches
from corner detector

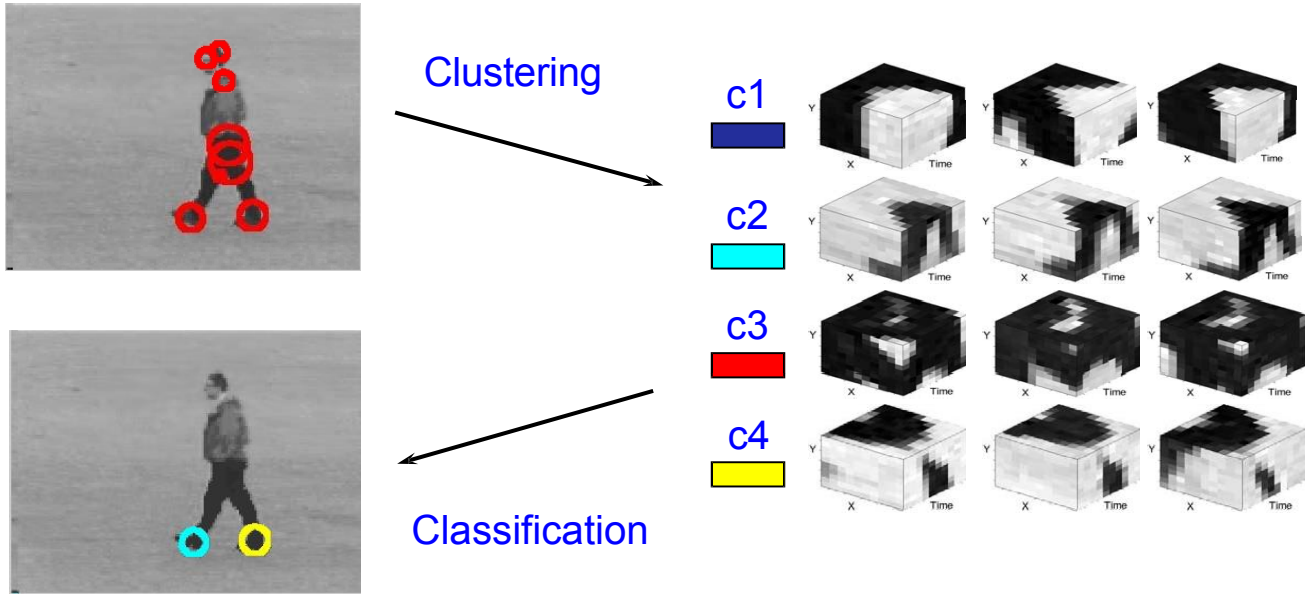


Public code available at
https://www.di.ens.fr/~laptev/actions/

[Laptev, Marszałek, Schmid, Rozenfeld, CVPR 2008]

Visual Vocabulary: K-means clustering

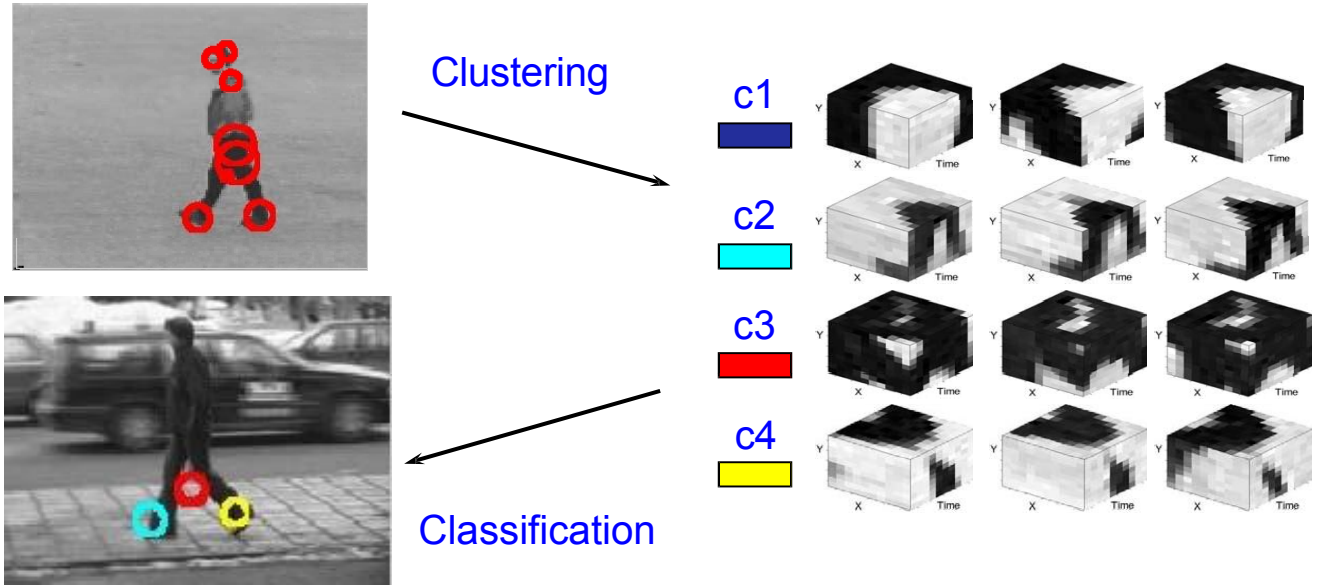
- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters



[Laptev, IJCV 2005]

Visual Vocabulary: K-means clustering

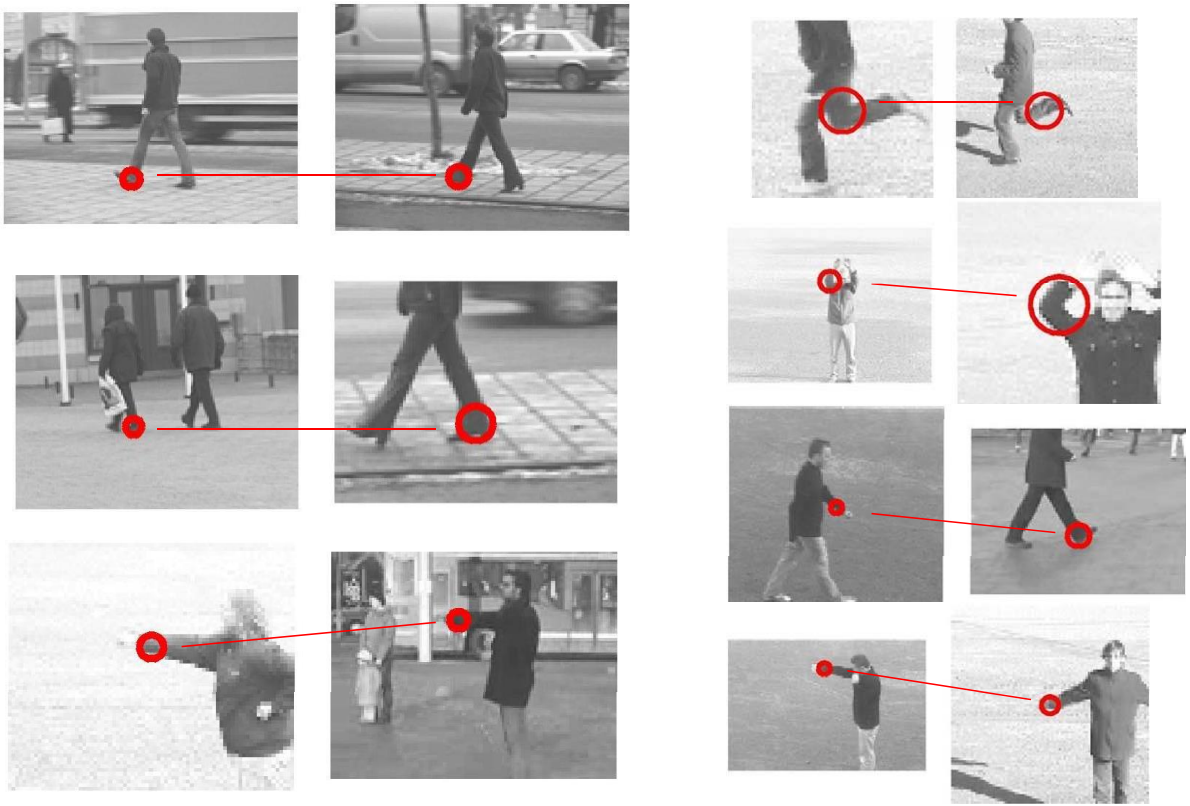
- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters



[Laptev, IJCV 2005]

Local Space-time features: Matching

- Find similar events in pairs of video sequences



What are Human Actions?

Actions in recent datasets:

Is it just about kinematics?

Should actions be associated with the *function* and the *task?*

Kinematics + Objects

What are Human Actions?

Actions in recent datasets:

Is it just about kinematics?

Should actions be associated with the *function* and the *task?*

Kinematics + Objects + Scenes

Action recognition in realistic settings

Standard
action
datasets

Actions “In the Wild”:






Learning Actions from Movies

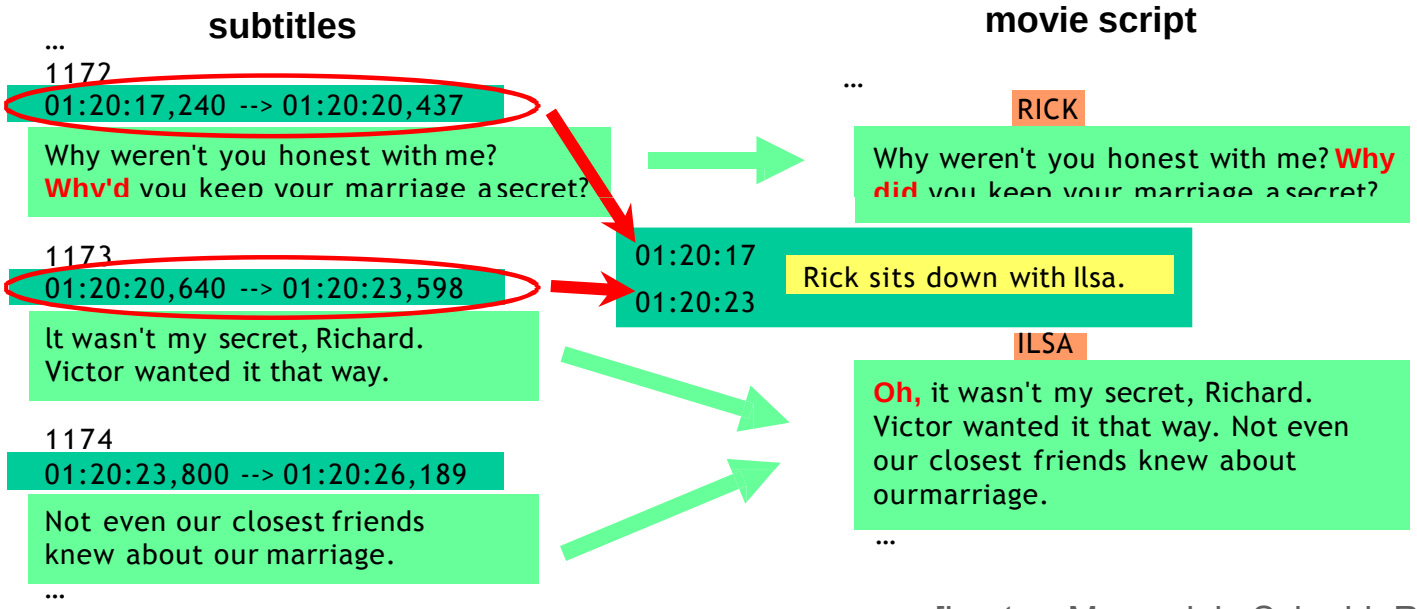
- Realistic variation of human actions
- Many classes and many examples per class

Problems:

- Typically only a few class-samples per movie
- Manual annotation is very time consuming

Automatic video annotation with scripts

- Scripts available for thousands of movies (no time synchronization)
  www.dailyscript.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment



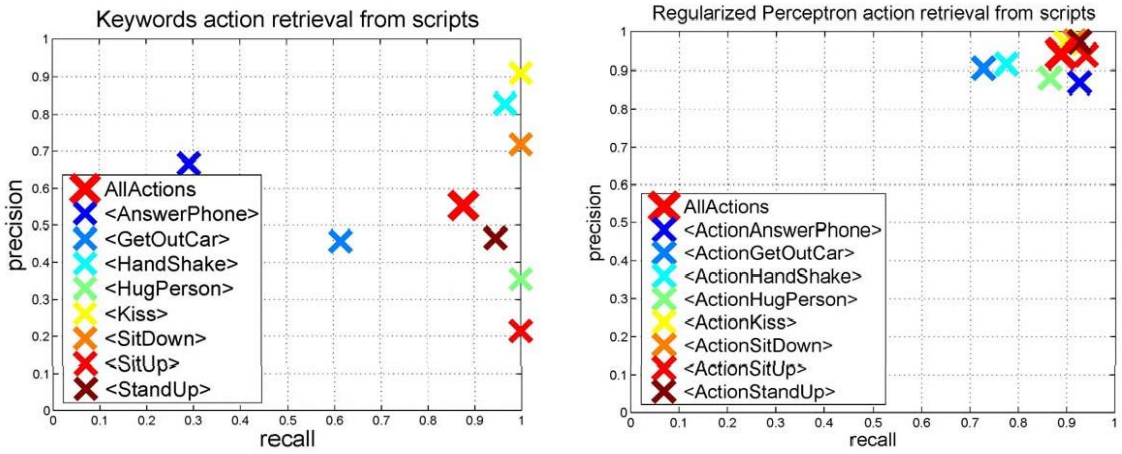
[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Text-based action retrieval

- Large variation of action expressions in text:

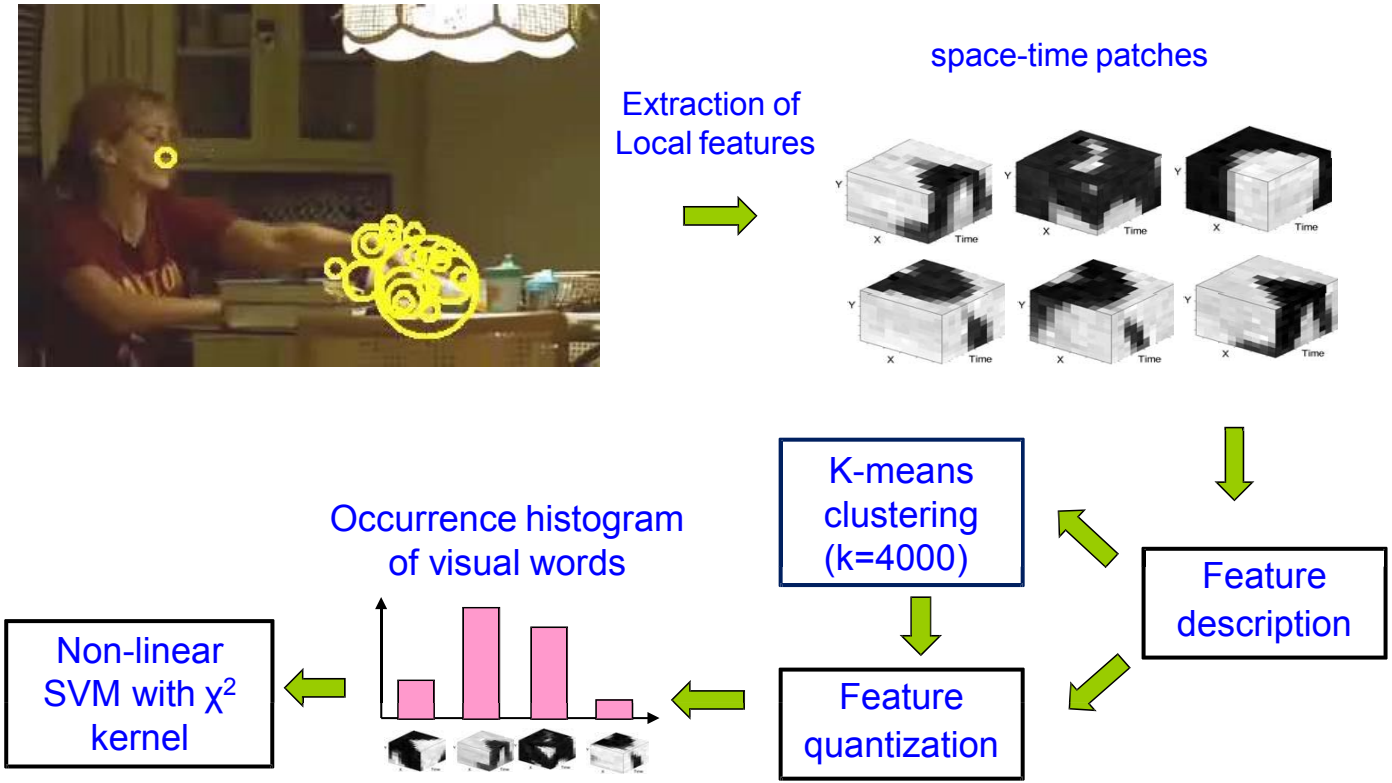
GetOutCar action:	*"… Will gets out of the Chevrolet. …"* *"… Erin exits her new truck…"*
Potential false positives:	*"…About to sit down, he freezes…"*

- => Supervised text classification approach



[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Bag-of-Features Recognition (BoF)

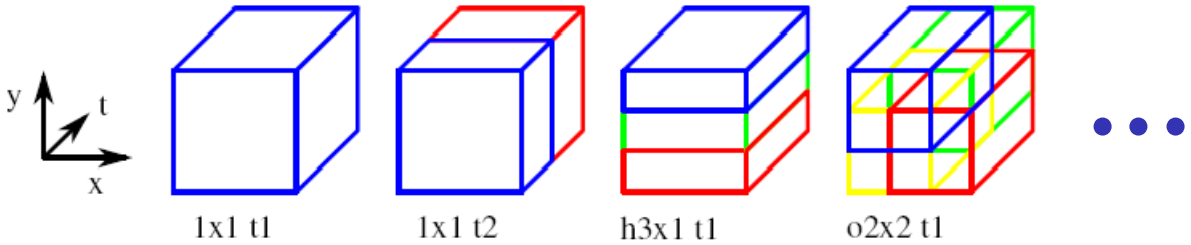


[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Spatio-temporal bag-of-features

Use global spatio-temporal grids

- In the spatial domain:
  - 1x1 (standard BoF)
  - 2x2, o2x2 (50% overlap)
  - h3x1 (horizontal), v1x3 (vertical)
  - 3x3
- In the temporal domain:
  - t1 (standard BoF), t2, t3



[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Multi-channel chi-square kernel

Use SVMs with a multi-channel chi-square kernel for classification

$$K(H_i, H_j) = \exp\Big(-\sum_{c \in \mathcal{C}} \frac{1}{A_c} D_c(H_i, H_j)\Big)$$

where $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$

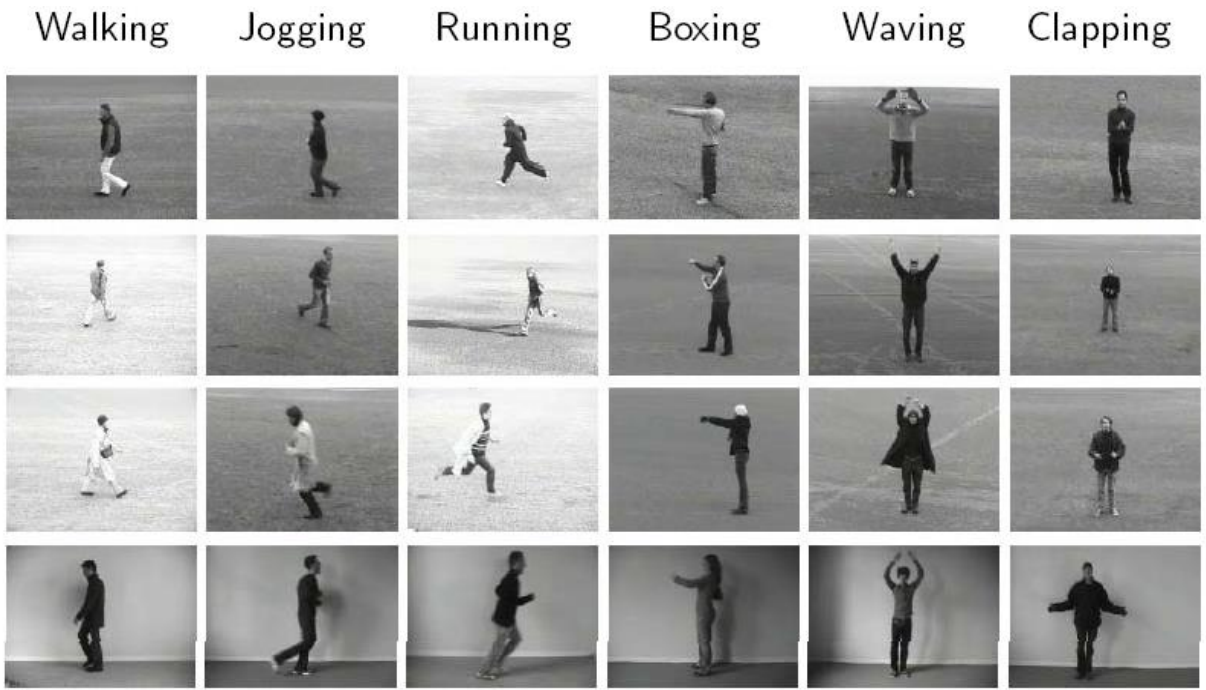
$$D_c(H_i, H_j) = \frac{1}{2} \sum_{n=1}^{V} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}$$

χ^2 distance

- Channel c is a combination of a detector, descriptor and a grid
- $D_c(H_i, H_j)$ is the chi-square distance between histograms
- A_c is the mean value of the distances between all training samples
- The best set of channels C for a given training set is found in a greedy manner

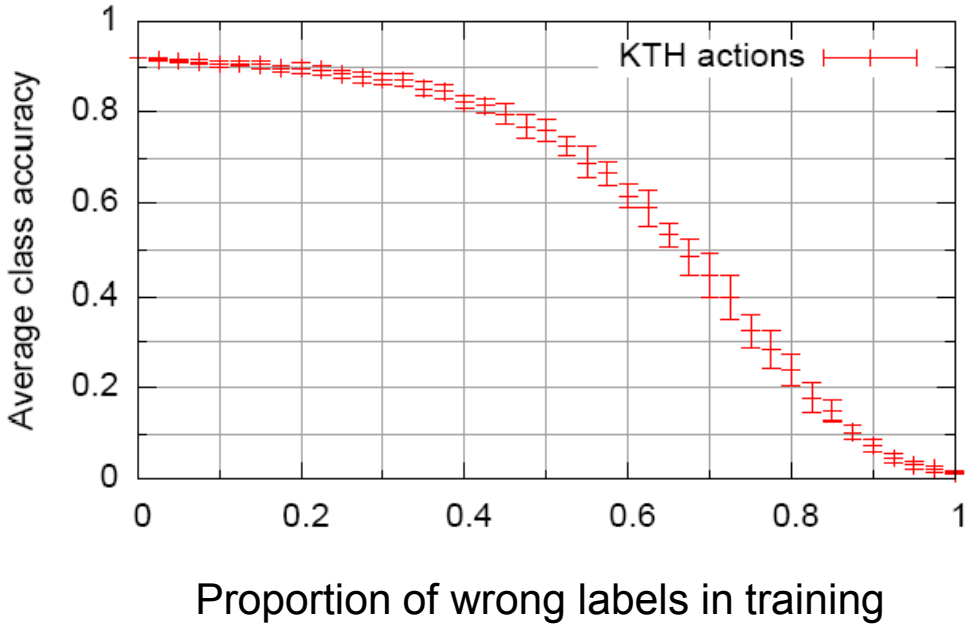
[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Results: KTH actions dataset



Sample frames from KTH action dataset for six classes (columns) and four scenarios (rows)

Results: Robustness to noise in training on KTH Dataset



- Up to p=0.2 the performance decreases insignificantly
- At p=0.4 the performance decreases by around 10%

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Results: Action recognition in movies

- Real data is hard!
- False Positives (FP) and True Positives (TP) often visually similar
- False Negatives (FN) are often particularly difficult

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Actions in Context

- Human actions are frequently correlated with particular scene classes

  Reasons: *physical properties* and *particular purposes* of scenes

Eating -- *kitchen*

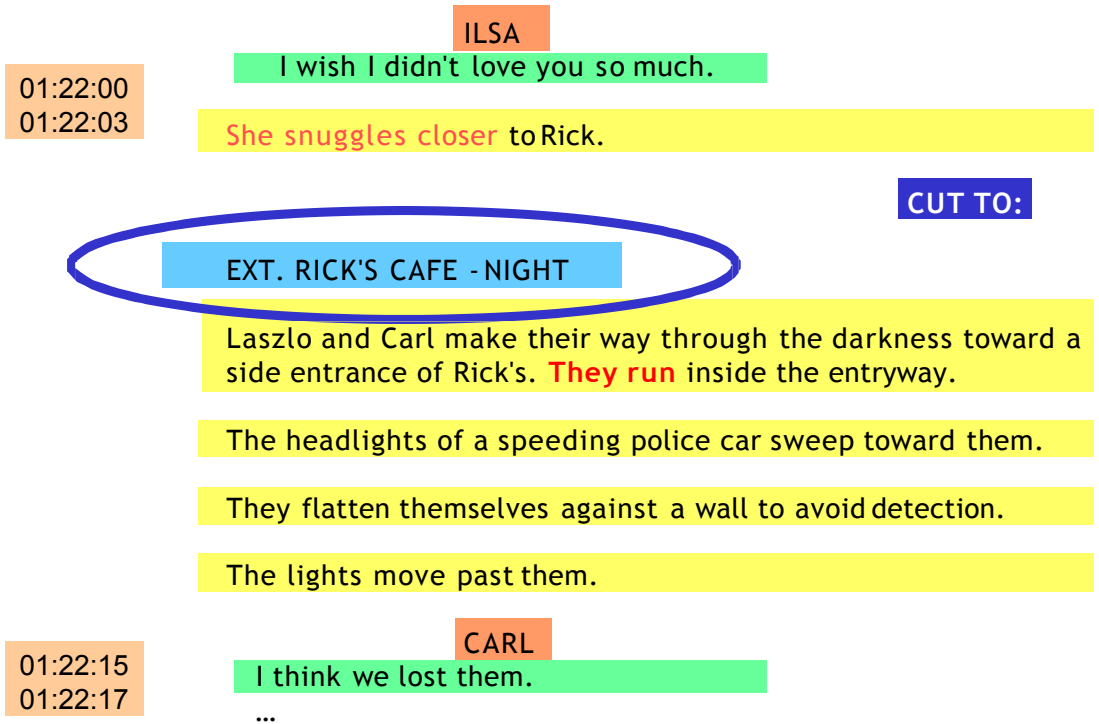
Eating -- *cafe*

Running -- *road*

Running -- *street*

[Marszałek, Laptev, Schmid, 2009]

Mining scene captions



[Marszałek, Laptev, Schmid, 2009]

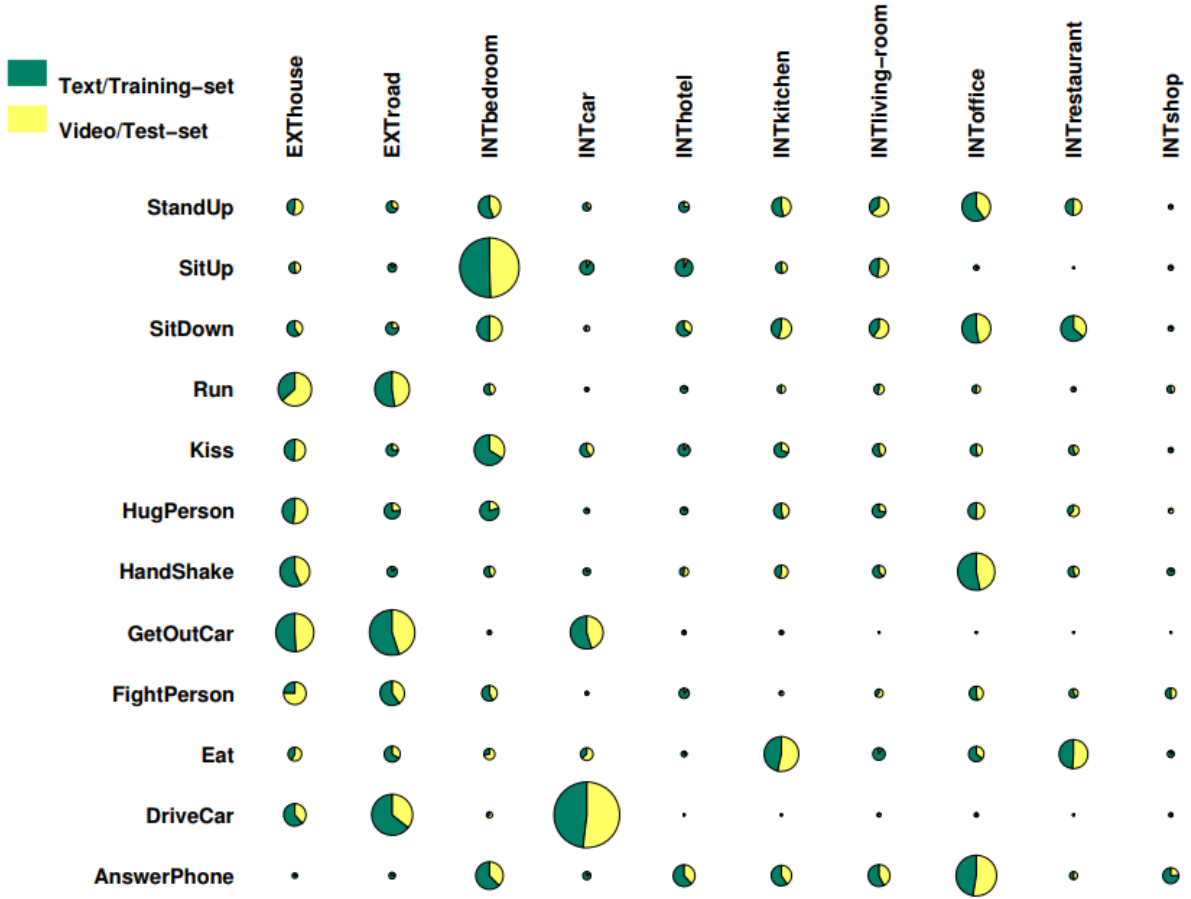
Mining scene captions

INT. TRENDY RESTAURANT - NIGHT
INT. MARSELLUS WALLACE S DINING ROOM MORNING
EXT. STREETS BY DORA'S HOUSE - DAY.
INT. MELVIN'S APARTMENT, BATHROOM – NIGHT
EXT. NEW YORK CITY STREET NEAR CAROL'S RESTAURANT – DAY
INT. CRAIG AND LOTTE'S BATHROOM - DAY

- Maximize word frequency ➡ street, living room, bedroom, car ….
- Merge words with similar senses using WordNet: taxi → car, cafe → restaurant
- Measure correlation of words with actions (in scripts) and
- Re-sort words by the entropy for $P = p(\text{action} \mid \text{word})$ $S = -\sum P_i \ln P_i$

[Marszałek, Laptev, Schmid, 2009]

Co-occurrence of actions and scenes in text vs. video



[Marszałek, Laptev, Schmid, 2009]

Classification with the help of context

$$a_i'(\boldsymbol{x}) = a_i(\boldsymbol{x}) + \tau \sum_{j \in \mathcal{S}} w_{ij} s_j(\boldsymbol{x})$$

$a_i(\boldsymbol{x})$ Action classification score

$s_j(\boldsymbol{x})$ Scene classification score

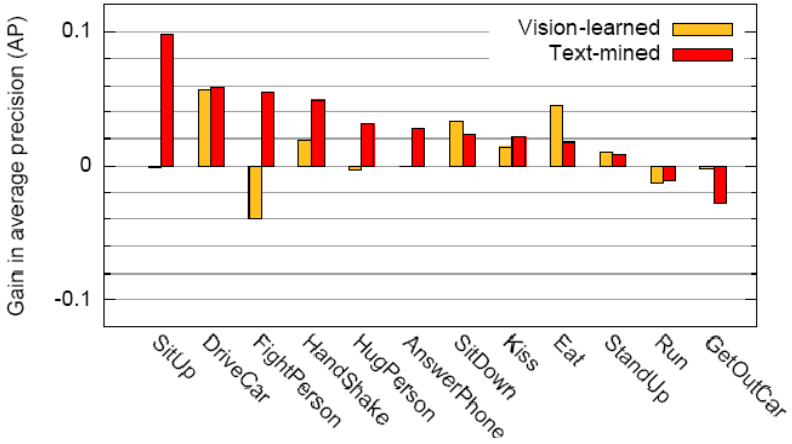
w_{ij} Weight, estimated from text: $p(Scene|Action)$

$a_i'(\boldsymbol{x})$ New action score

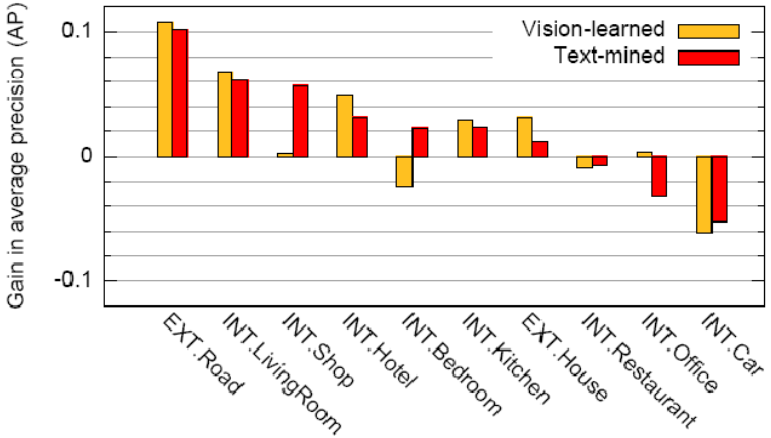
[Marszałek, Laptev, Schmid, 2009]

Results: actions and scenes (jointly)

Actions in the context of Scenes



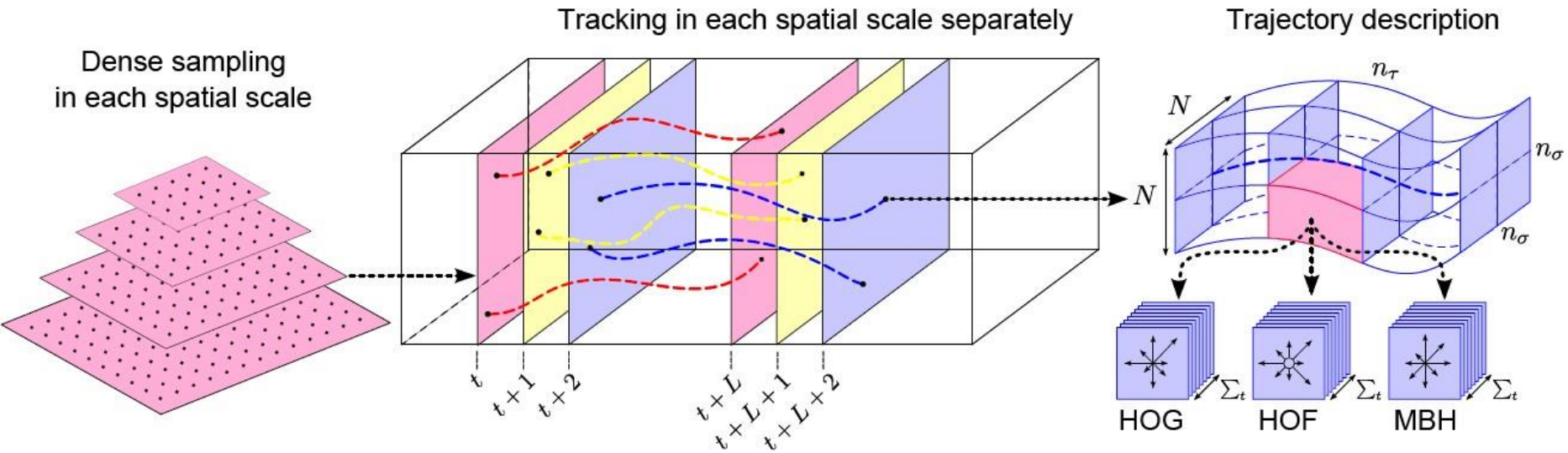
Scenes in the context of Actions



[Marszałek, Laptev, Schmid, 2009]

Dense trajectories

Action recognition by dense trajectories
Wang et al., 2011



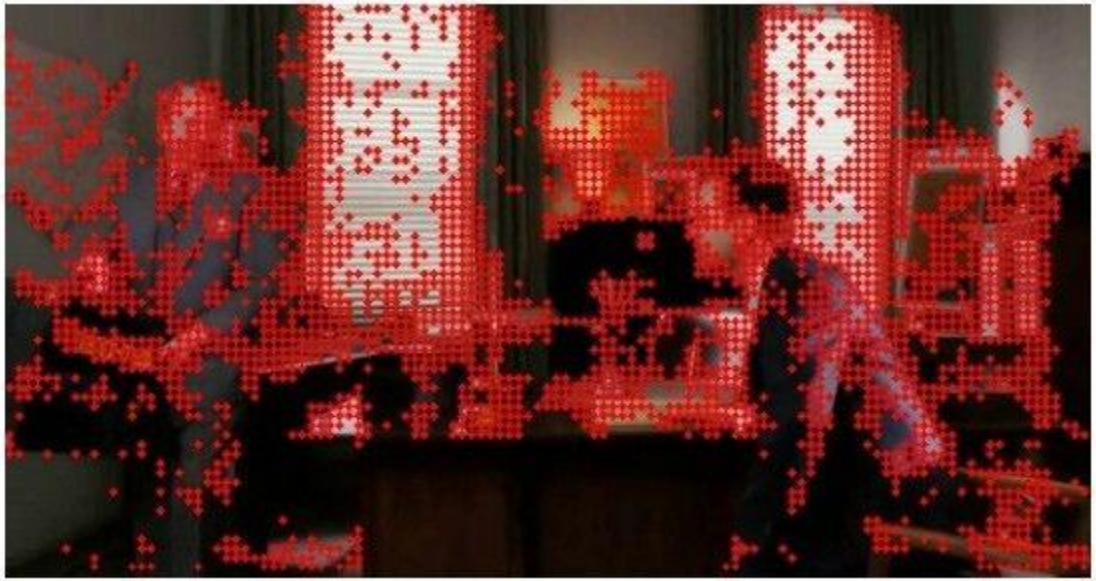
detect feature points

track features with optical flow

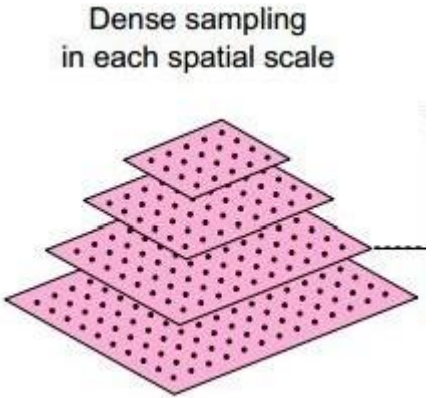
extract HOG/HOF/MBH features in the (stabilized) coordinate system of each tracklet

Dense trajectories

Action recognition by dense trajectories
Wang et al., 2011



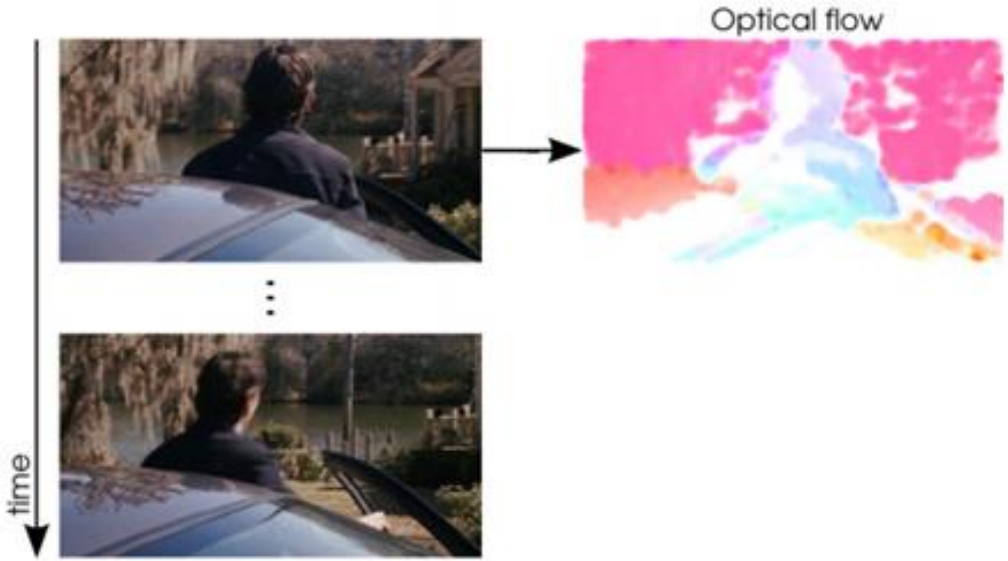
detected feature points



[J. Shi and C. Tomasi, “Good features to track,” CVPR 1994]
[Ivan Laptev 2005]

Dense trajectories

Action recognition by dense trajectories
Wang et al., 2011

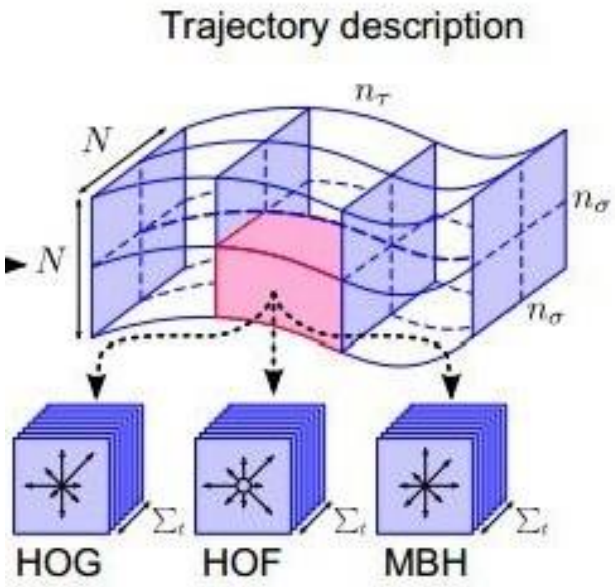


Track each keypoint using **optical flow**

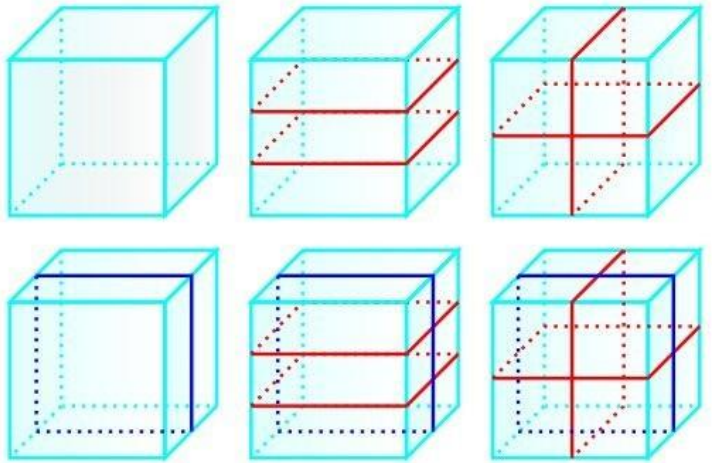
[G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” 2003]

Trajectory descriptors

Action recognition by dense trajectories
Wang et al., 2011



Extract features in the local coordinate system of each tracklet.



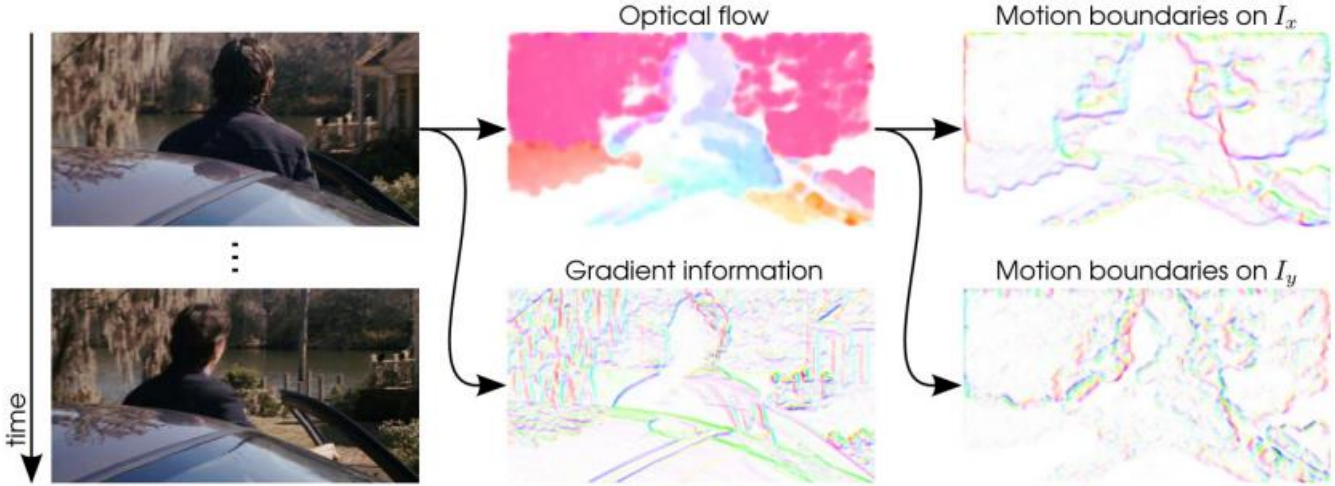
Accumulate into histograms, separately according to multiple spatio-temporal layouts.

Trajectory descriptors

Action recognition by dense trajectories
Wang et al., 2011

Motion Boundary Descriptor

- spatial derivatives are calculated separately for optical flow in x and y, quantized into a histogram
- relative dynamics of different regions
- suppresses constant motions (camera motions)



Dense trajectories

Action recognition by dense trajectories
Wang et al., 2011

Advantages:

- Captures the intrinsic dynamic structures in videos
- MBH is robust to certain camera motion

Disadvantages:

- Generates irrelevant trajectories in background due to camera motion
- Motion descriptors are modified by camera motion, e.g., HOF, HOG

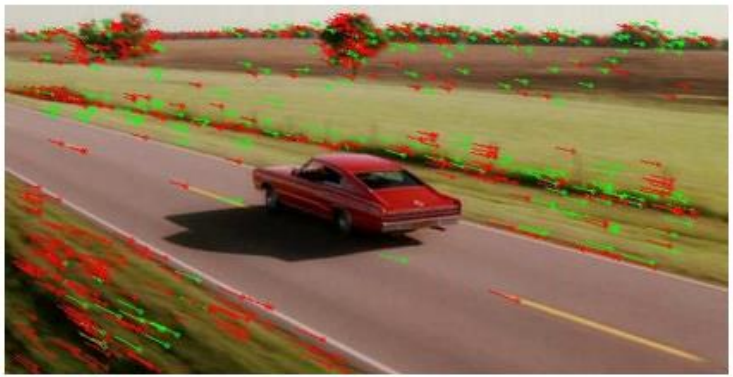
→ Improved dense trajectories

Improved dense trajectories

Action recognition with improved trajectories
Wang et al., 2013

Camera motion estimation

- Match feature points between frames using SURF descriptors (green) and dense optical flow (red)
- The combination of SURF and optical flow results in a more balanced distribution
- Use RANSAC to estimate a Homography from all feature matches

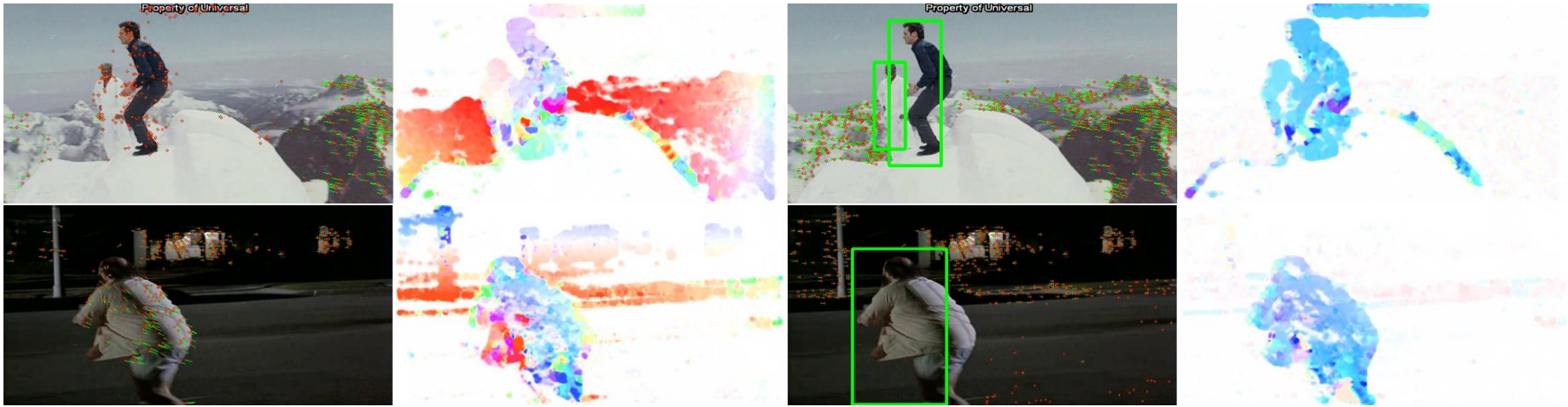


Improved dense trajectories

Action recognition with improved trajectories
Wang et al., 2013

Remove inconsistent matches due to humans

- Human motion is not constrained by camera motion
- Apply a human detector in each frame, and remove feature matches inside the human bounding box

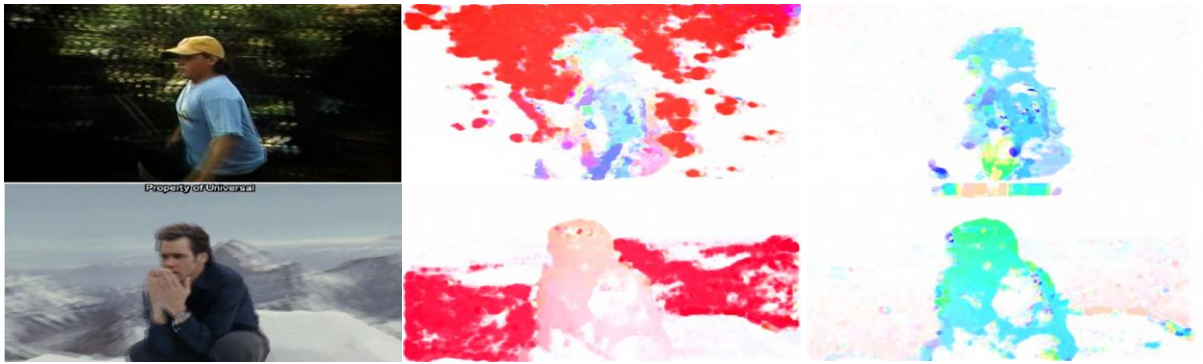


Improved dense trajectories

Action recognition with improved trajectories
Wang et al., 2013

Warp optical flow

- Warp the second frame of two consecutive frames with the homography and recompute the optical flow
- For the HOF descriptor, the warped flow removes irrelevant camera motion, thus only encodes foreground motion
- For the MBH descriptor, it also improves, as the motion boundaries are enhanced



Improved dense trajectories

Action recognition with improved trajectories
Wang et al., 2013

Remove background trajectories

- Remove trajectories due to camera motion, which has similar effects as sampling features with visual saliency
- It works well under various camera motions, such as pan, zoom, tilt

Removed trajectories: white | Foreground ones: green

Action recognition - Tasks

- Action classification: assigning an action label to a video clip

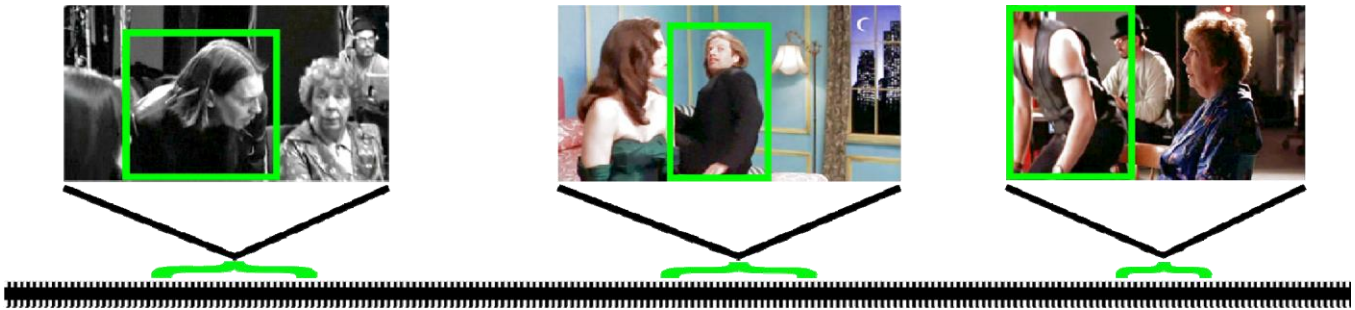
Making sandwich: present
Feeding animal: not present
…

Action recognition - Tasks

- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present
…

- Action localization (temporal): search temporal locations of an action in a video



Action recognition - Tasks

- Action localization (spatio-temporal) + interaction with an object, human, etc.

[Prest et al., PAMI 13]

Why automatic action localization?

- Query for specific videos in professional Archives and YouTube
- Analyze and describe content of videos
- Produce audio descriptions for visual impaired

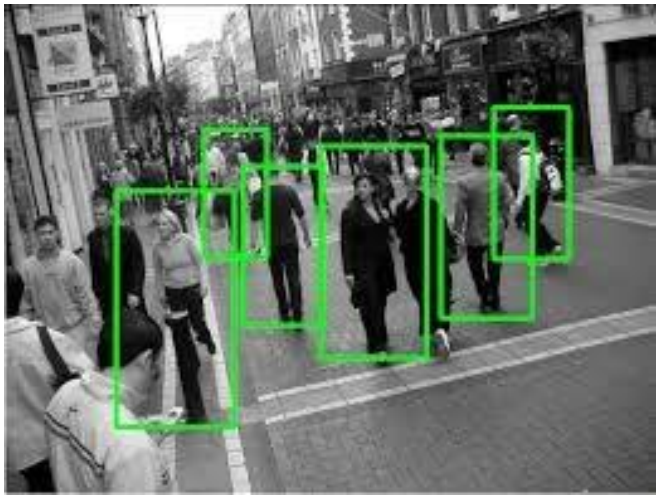
Education: How do I make a pizza?

Sociology research: Influence of character smoking in movies

Why automatic action localization?

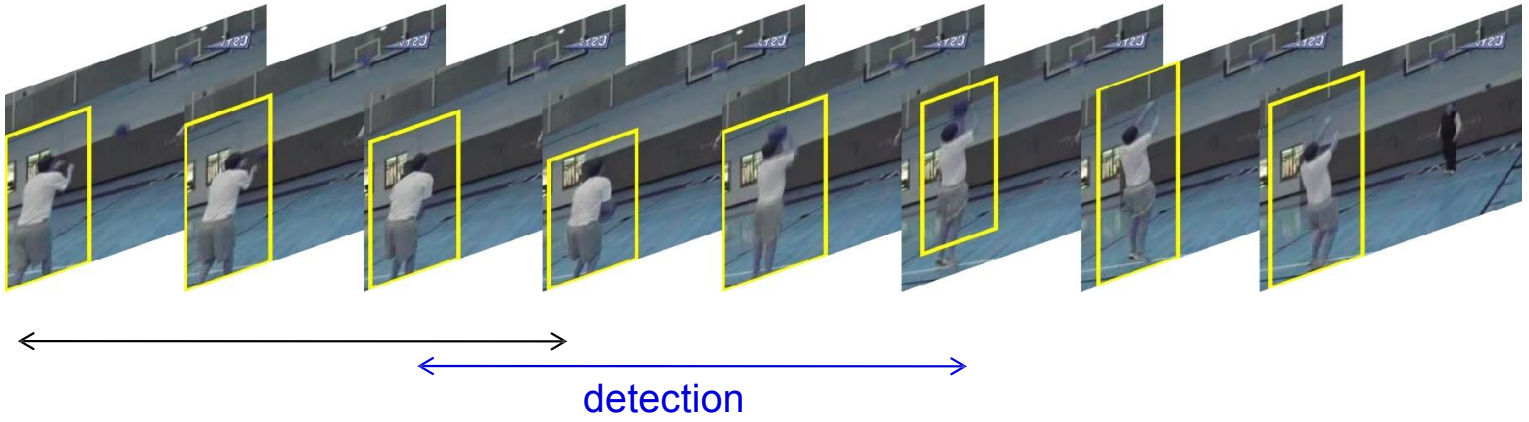
- Car safety & self-driving and video surveillance
- Detection of humans (pedestrians) and their motion, detection of unusual behavior

Courtesy Volvo

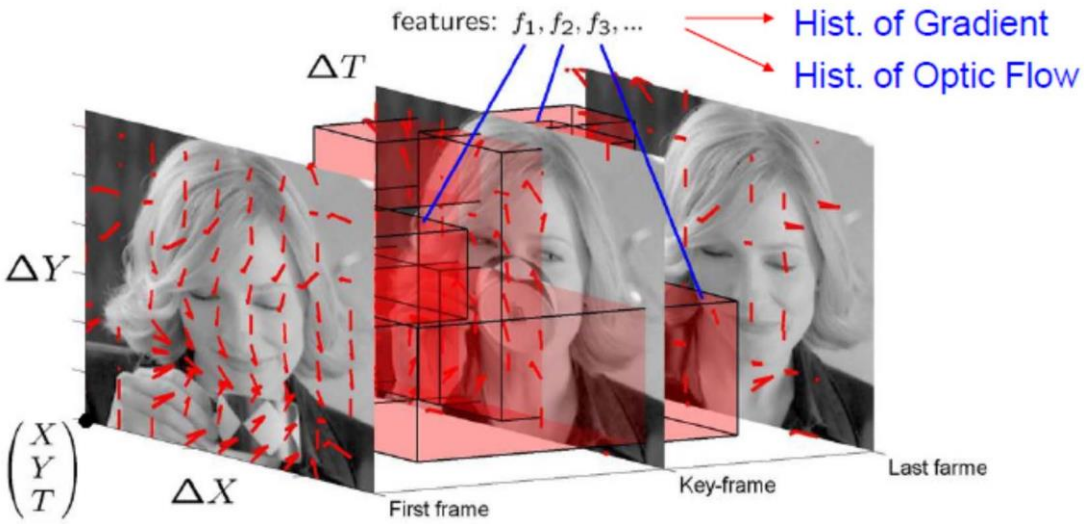

Courtesy Embedded Vision Alliance

Temporal action localization

- Temporal sliding window
- Shot detection
  - Spatio-temporal volume proposals, etc.

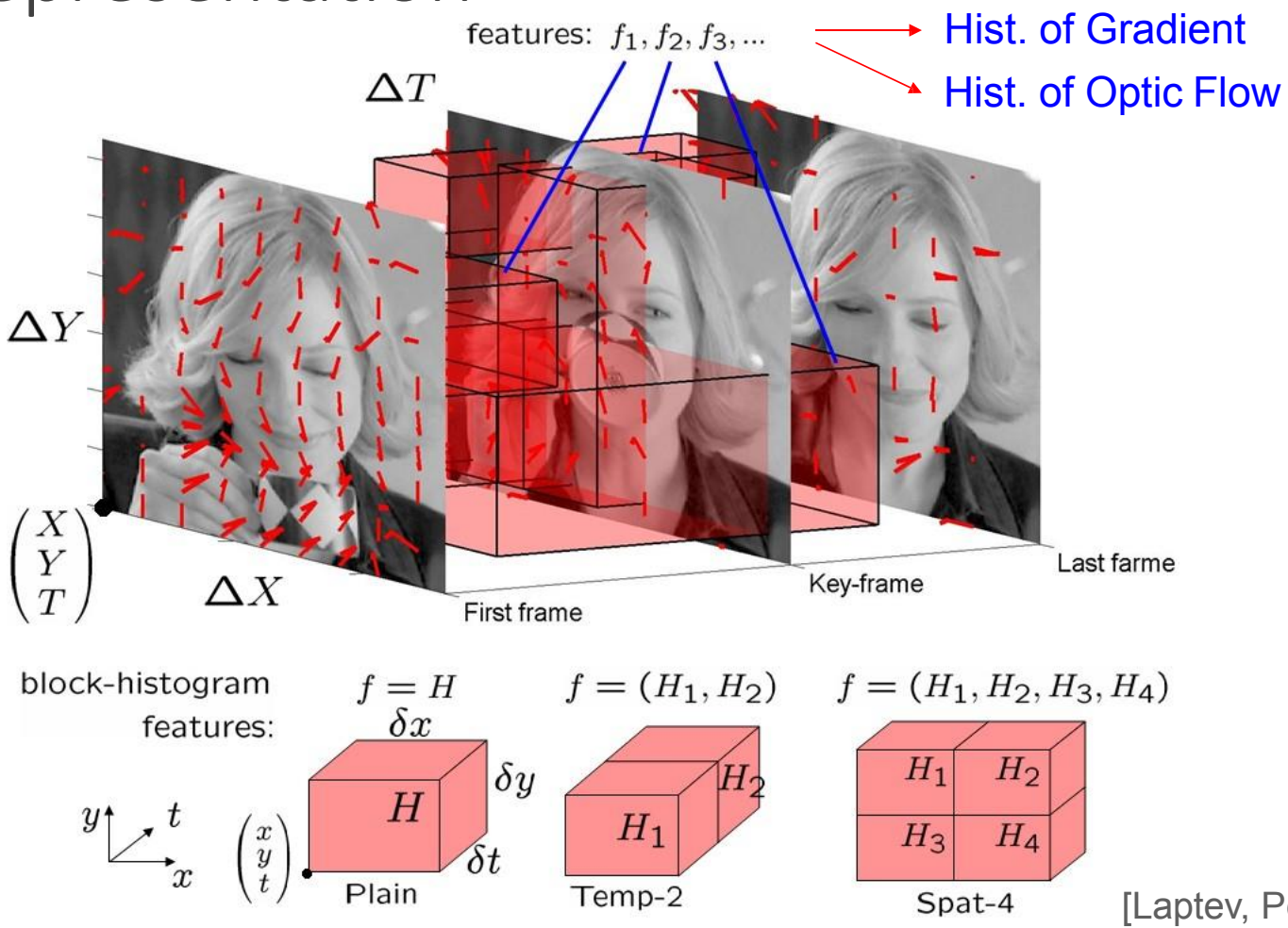


Temporal action localization



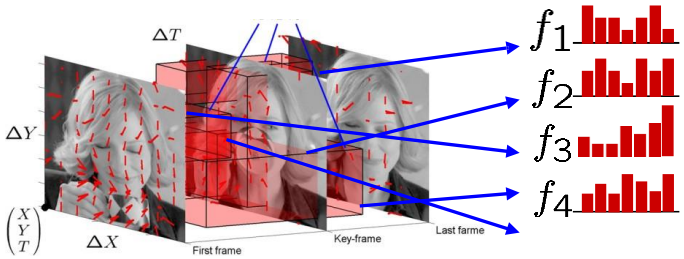
[Retrieving actions in movies, I. Laptev and P. Pérez, ICCV'07]

Action representation



[Laptev, Perez 2007]

Action learning



boosting

selected features

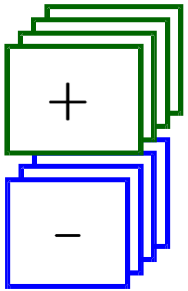
$$H(z) = \mathrm{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(f_t)\right)$$

weak classifier

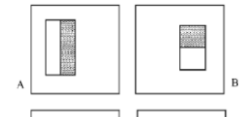
AdaBoost:

- Efficient discriminative classifier [Freund&Schapire'97]
- Good performance for face detection [Viola&Jones'01]

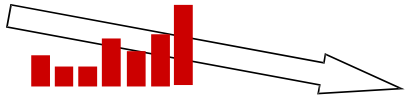
pre-aligned samples

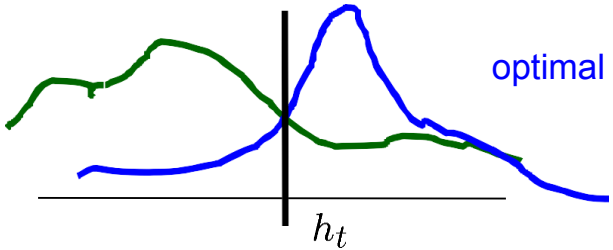


Haar features

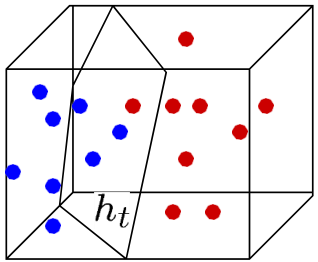


Histogram features





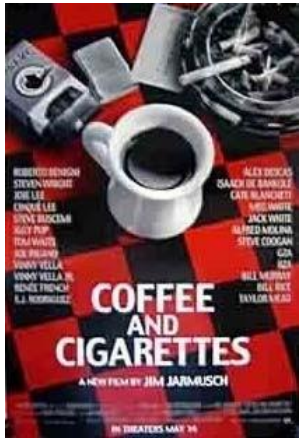
optimal threshold



Fisher discriminant

[Laptev, Perez 2007]

Dataset for action localization

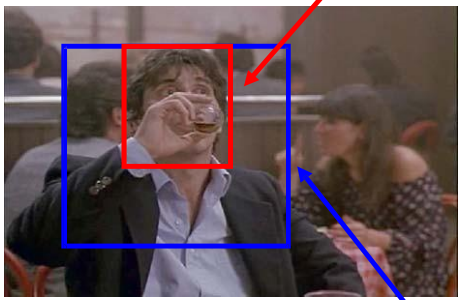


Manual annotation of drinking actions in movies:
“Coffee and Cigarettes”; “Sea of Love”

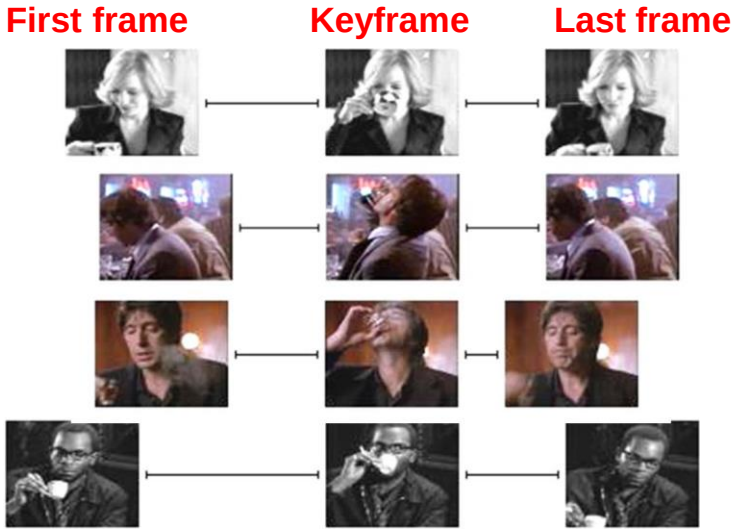
“*Drinking*”: 159 annotated samples

“*Smoking*”: 149 annotated samples

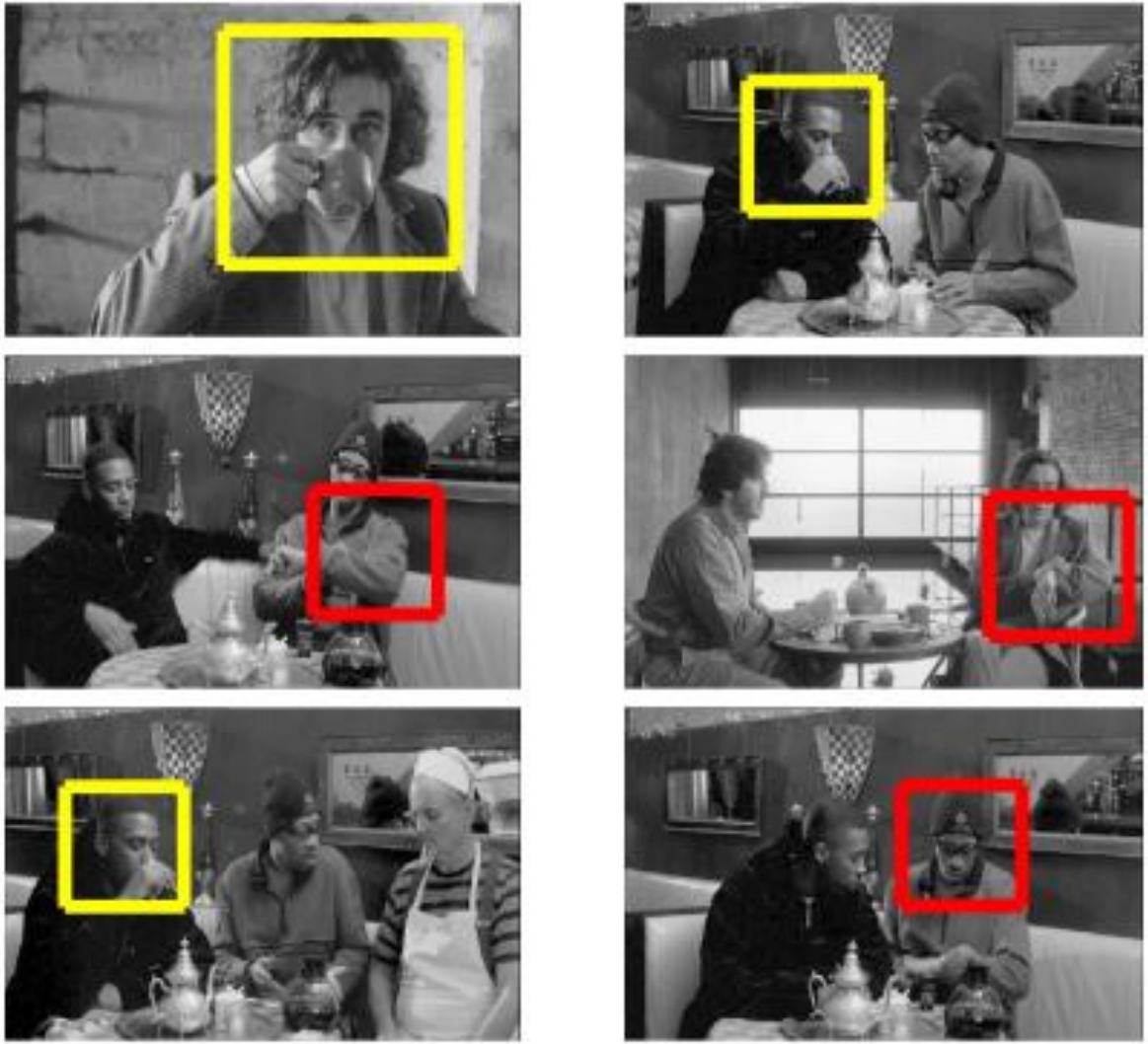
Spatial annotation



Temporal annotation



Results

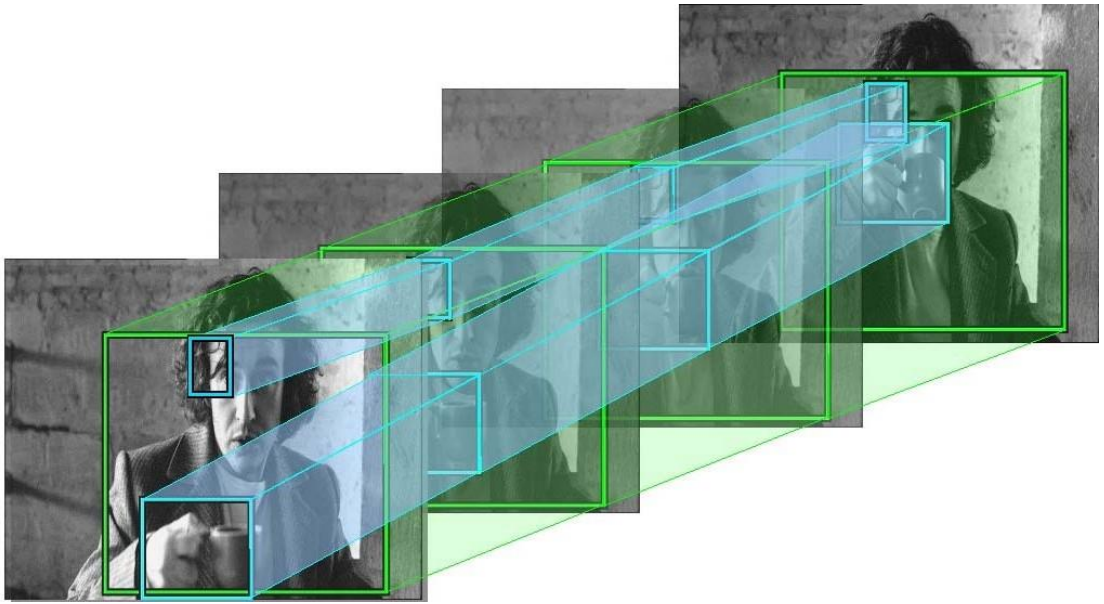


[Laptev, Perez 2007]

Action localization: Modeling temporal human-object interaction

[Explicit modeling of human-object interactions in realistic videos, Prest et al., PAMI 13]

Tracking humans and objects

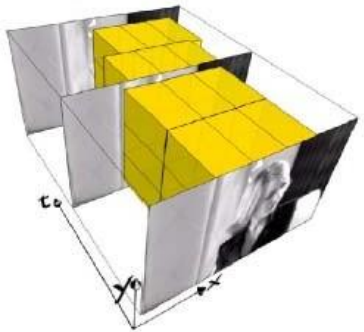


- Fully automatic human tracks: state of the art detector + optical flow tracker
- Object tracks: detector learnt from annotated training images + optical flow tracker
- Extraction of a large number of human-object track pairs

Action descriptors

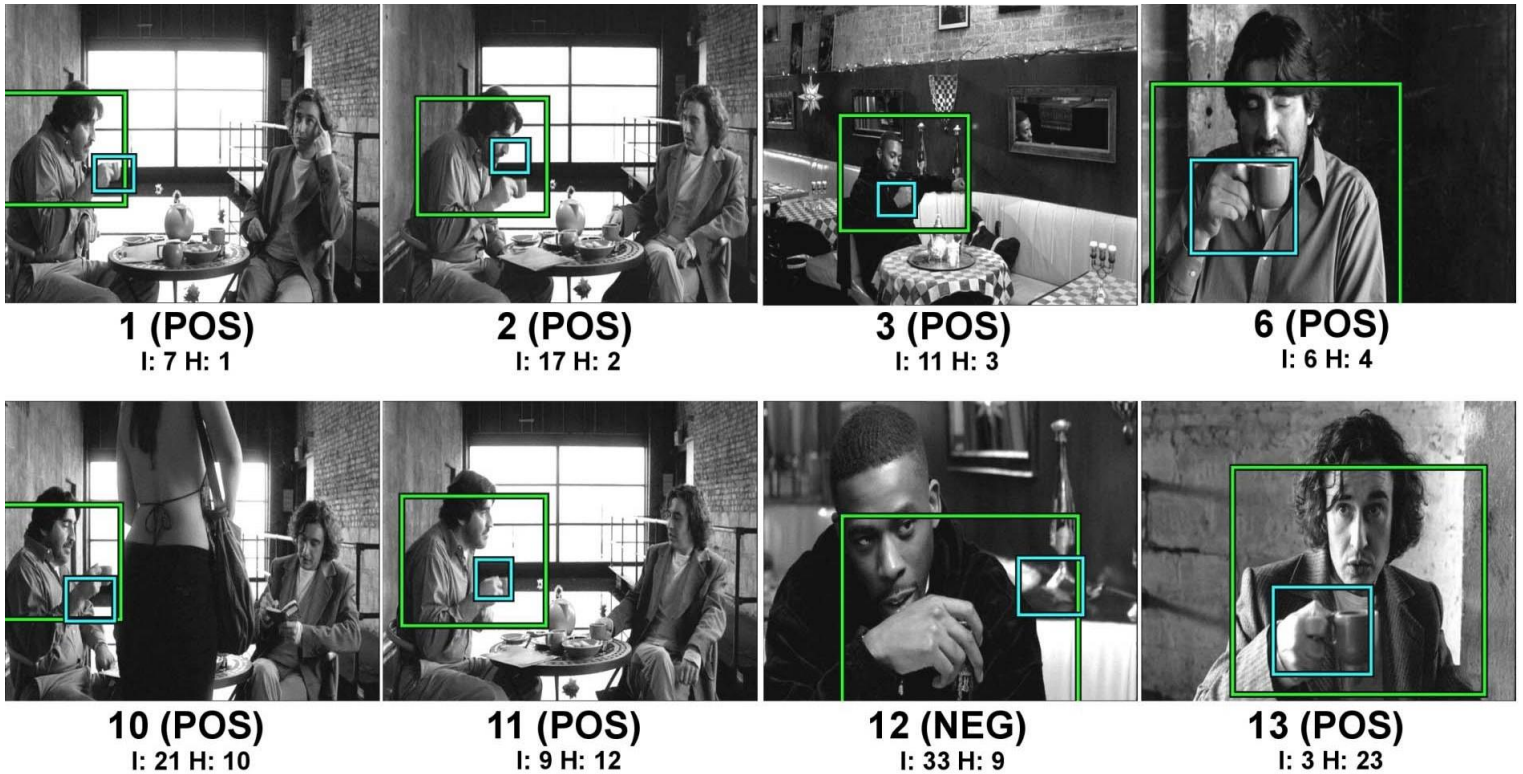
- Interaction descriptor: relative location, area and motion between human and object tracks

- Human track descriptor: 3DHOG-track [Klaeser et al.'10]



Results

Drinking



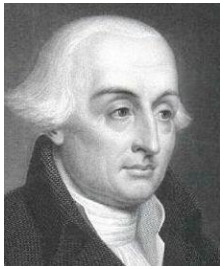
Analysis of subtle motion/actions

Imperceptible Changes in the World

Pulse and blood flow

Respiratory motion

Lagrangian and Eulerian Perspectives: Fluid Dynamics

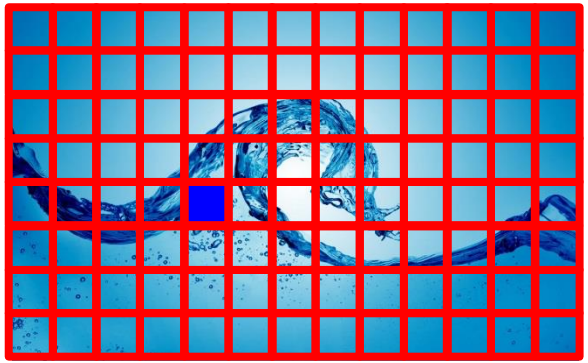


Lagrangian

Eulerian

The fluid properties are defined as functions of space and times. The flow is determined by the analyzing the behavior of the functions.

Pieces of the fluid are “tagged”. The fluid flow properties are determined by tracking the motion and properties of the particles as they move in time.



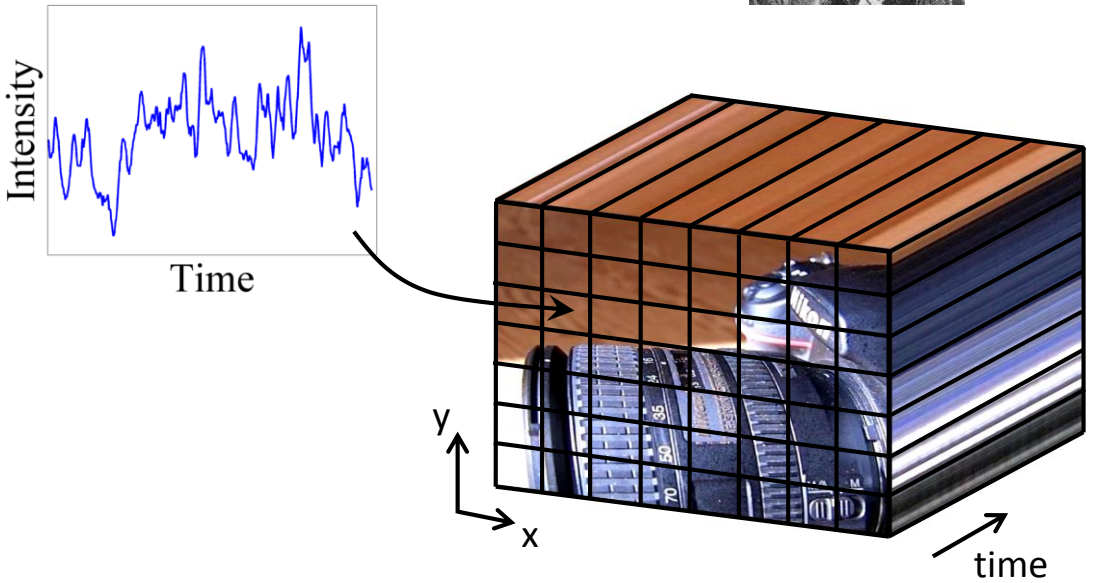
Eulerian Perspective: Videos

[H-Y. Wu et al., 2012]

Each pixel is processed independently

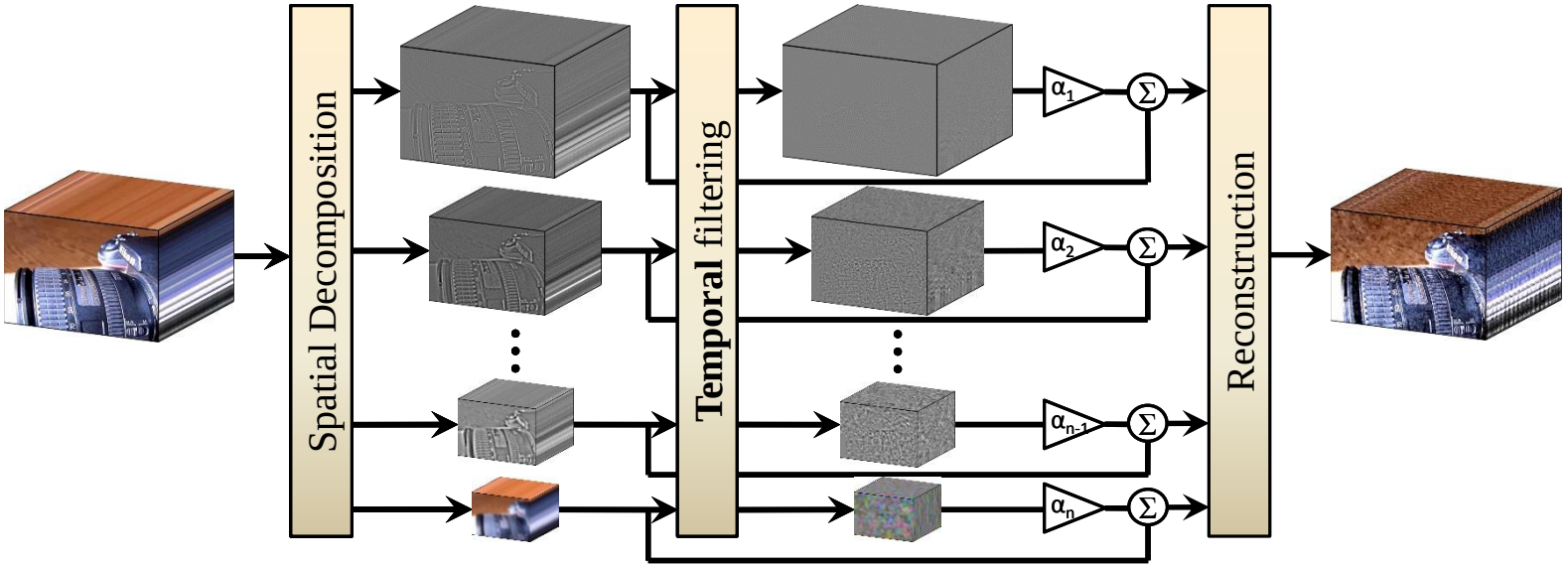
We treat each pixel as a time series and apply signal processing to it

Eulerian



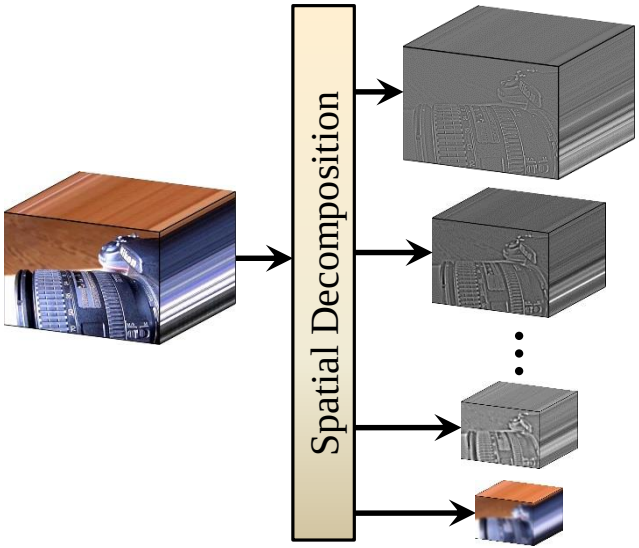
Eulerian Motion Magnification

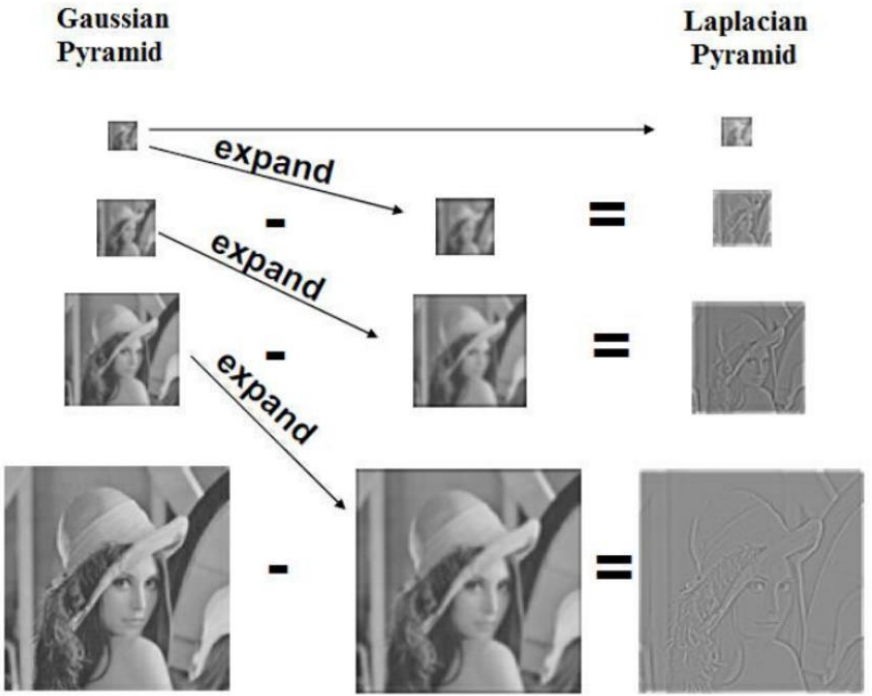
[H-Y. Wu et al., 2012]



Spatial Decomposition

[H-Y. Wu et al., 2012]



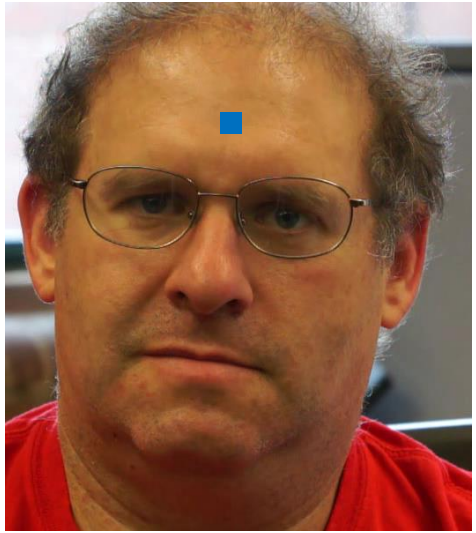


Subtle Color Variations

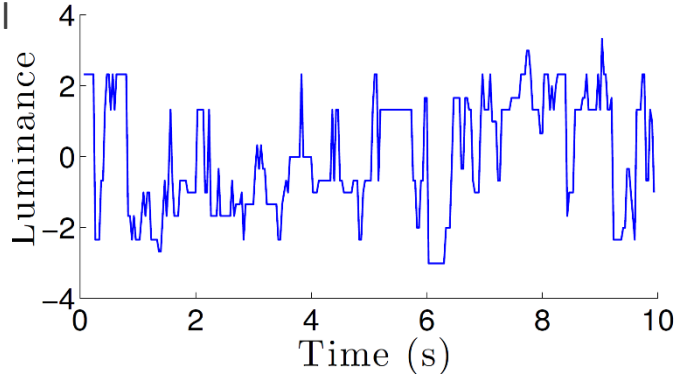
[H-Y. Wu et al., 2012]

The face gets slightly redder when blood flows

Unfortunately, usually below the per pixel noise level



Input frame

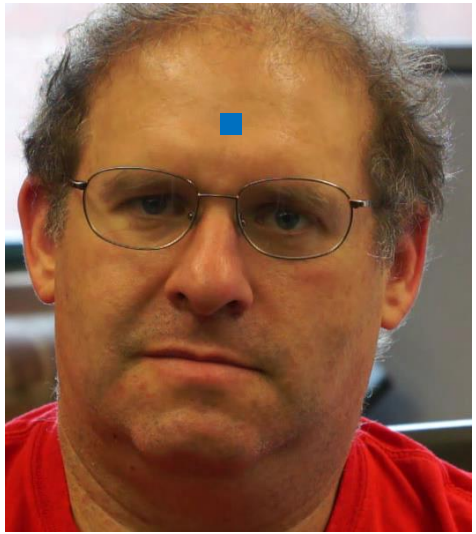


Luminance trace (zero mean)

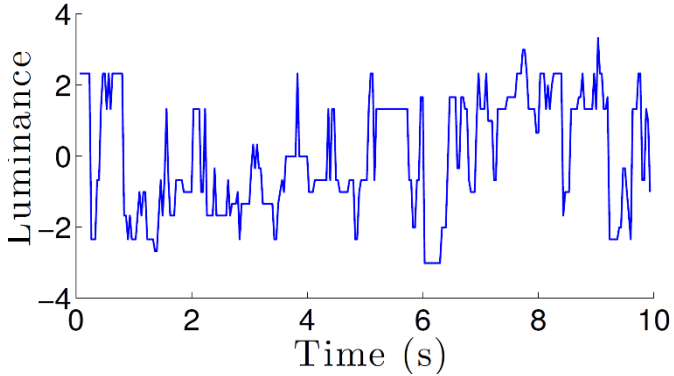
Subtle Color Variations

[H-Y. Wu et al., 2012]

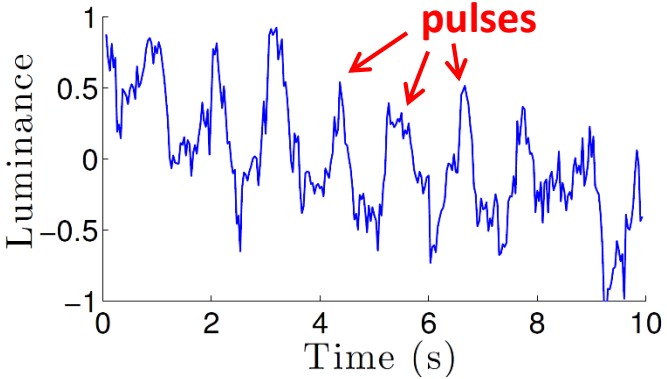
1. Average spatially to overcome sensor and quantization noise



Input frame



Luminance trace (zero mean)

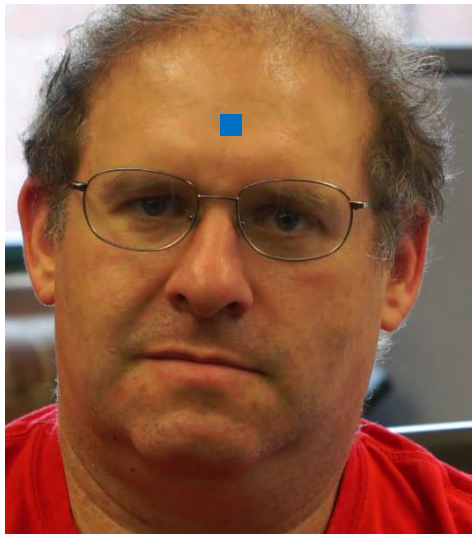


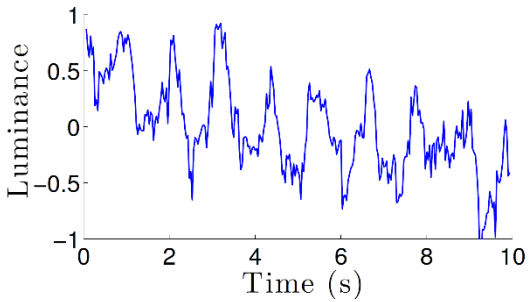
Spatially averaged luminance trace

Amplifying Subtle Color Variations

[H-Y. Wu et al., 2012]

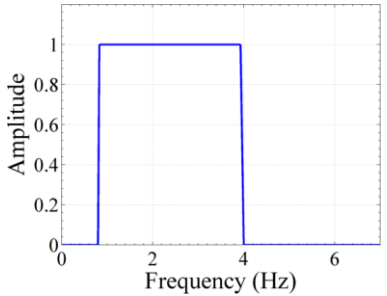
2. Filter temporally to extract the signal of interest





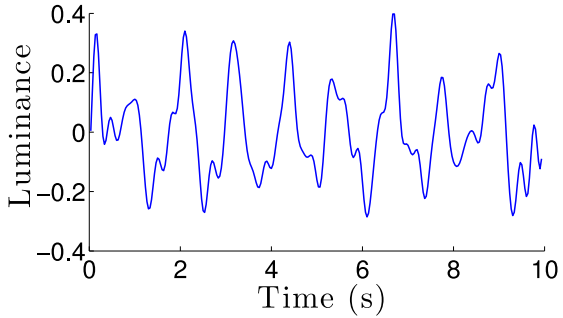
Spatially averaged luminance trace

⊗



Temporal filter

=



Temporally bandpassed trace

Color Amplification Results

[H-Y. Wu et al., 2012]

Source

Color-amplified (x100)
0.83-1 Hz (50-60 bpm)

Color Amplification Results

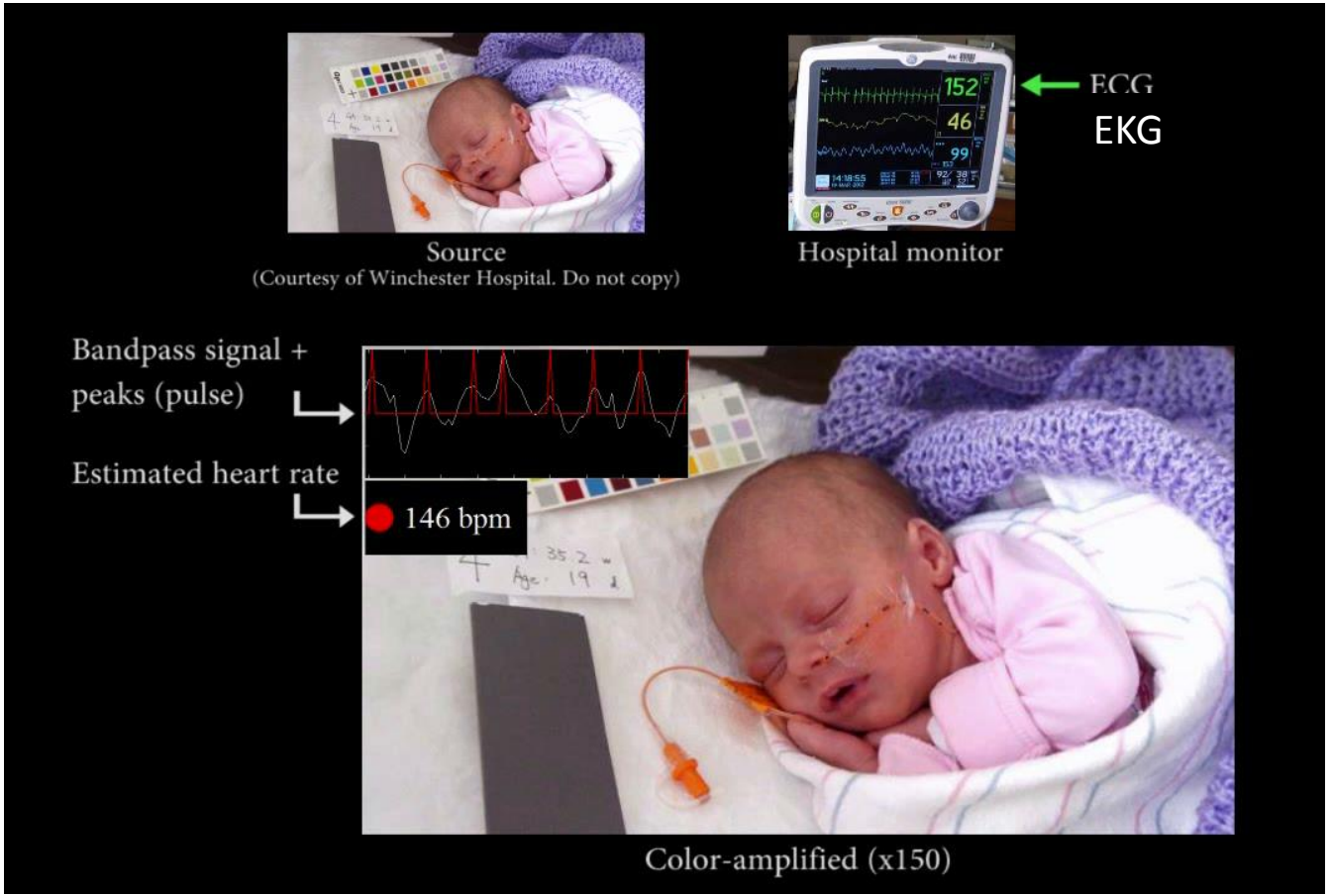
[H-Y. Wu et al., 2012]

Source

Color-amplified (x120)
0.83-1 Hz (50-60 bpm)

Heart Rate Extraction

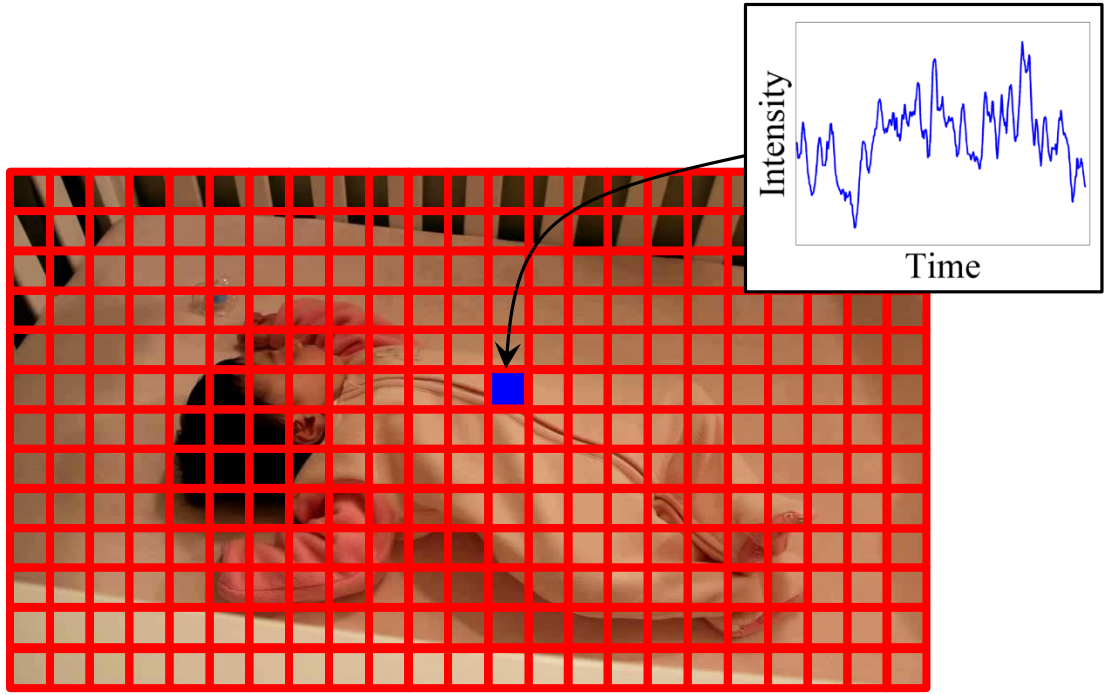
[H-Y. Wu et al., 2012]



2.33-2.67 Hz (140-160 bpm)

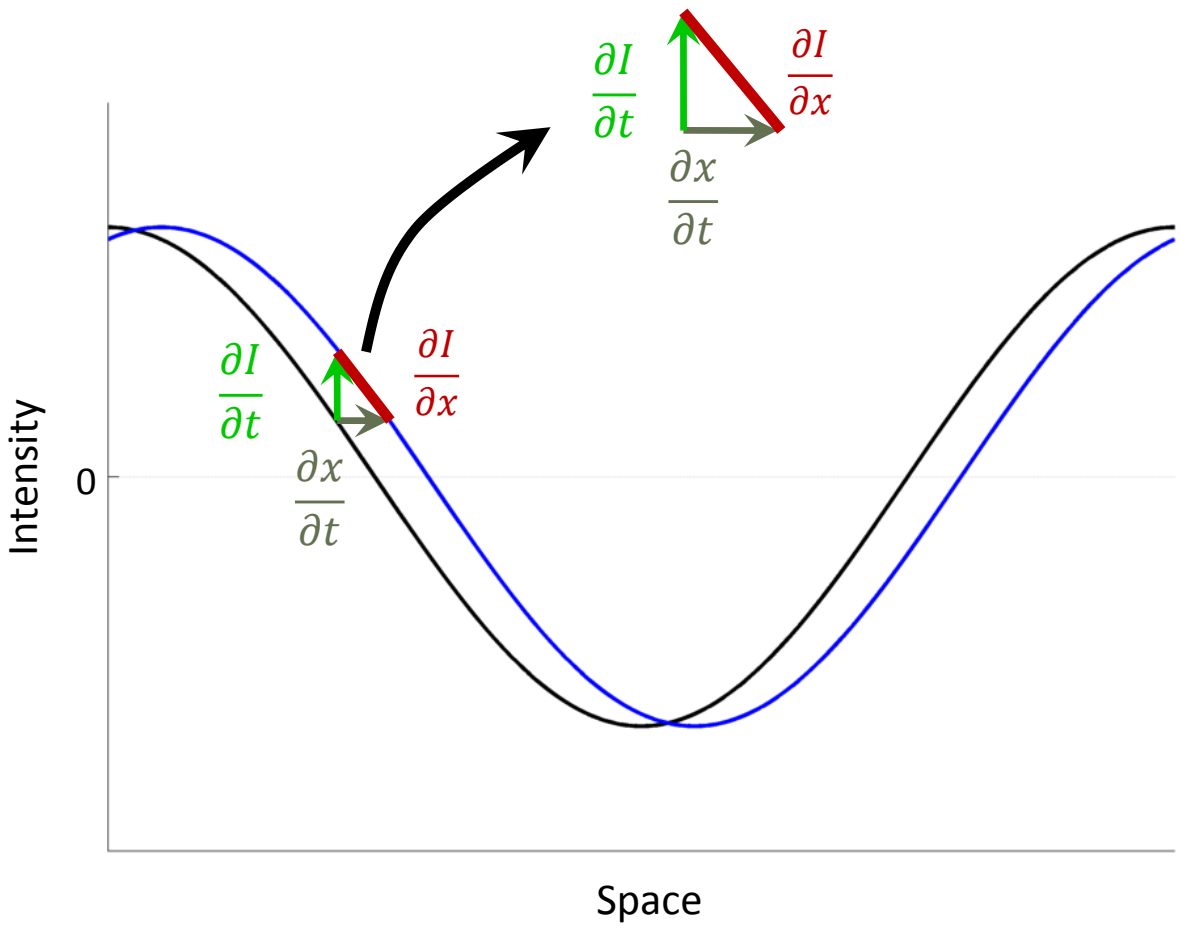
Courtesy of Dr. Donna Brezinski and the Winchester Hospital staff

Why/How It Amplifies Motion



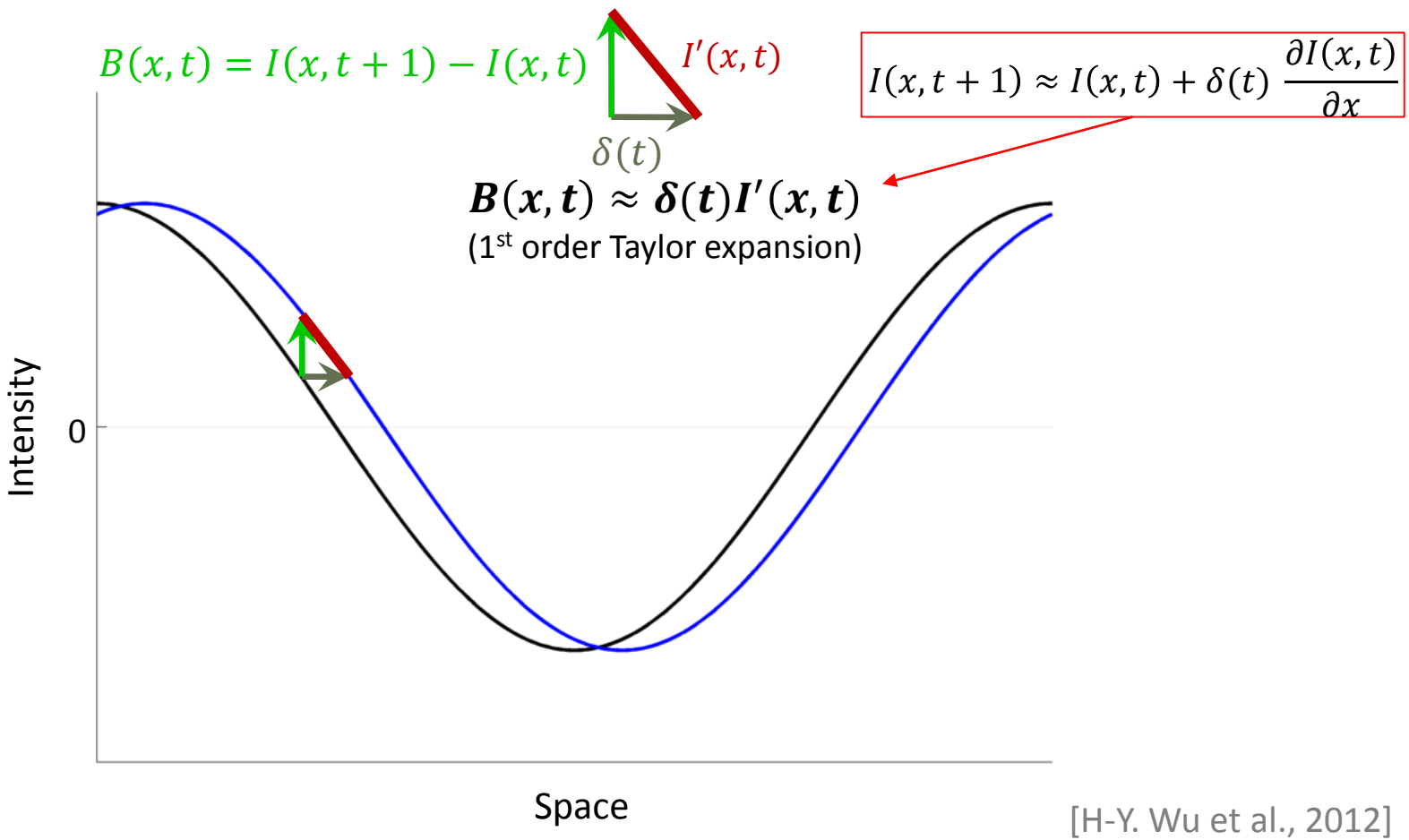
[H-Y. Wu et al., 2012]

Differential Brightness Constancy



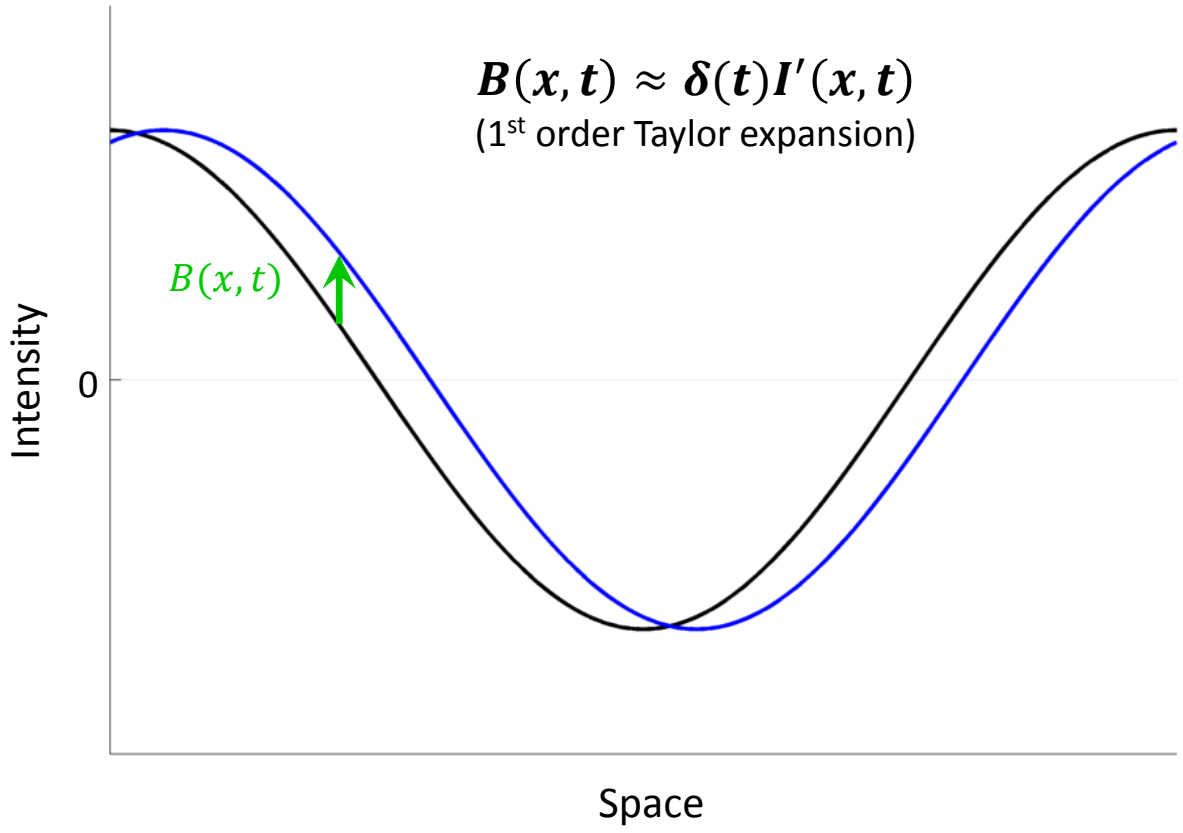
[H-Y. Wu et al., 2012]

Relating Temporal and Spatial Changes



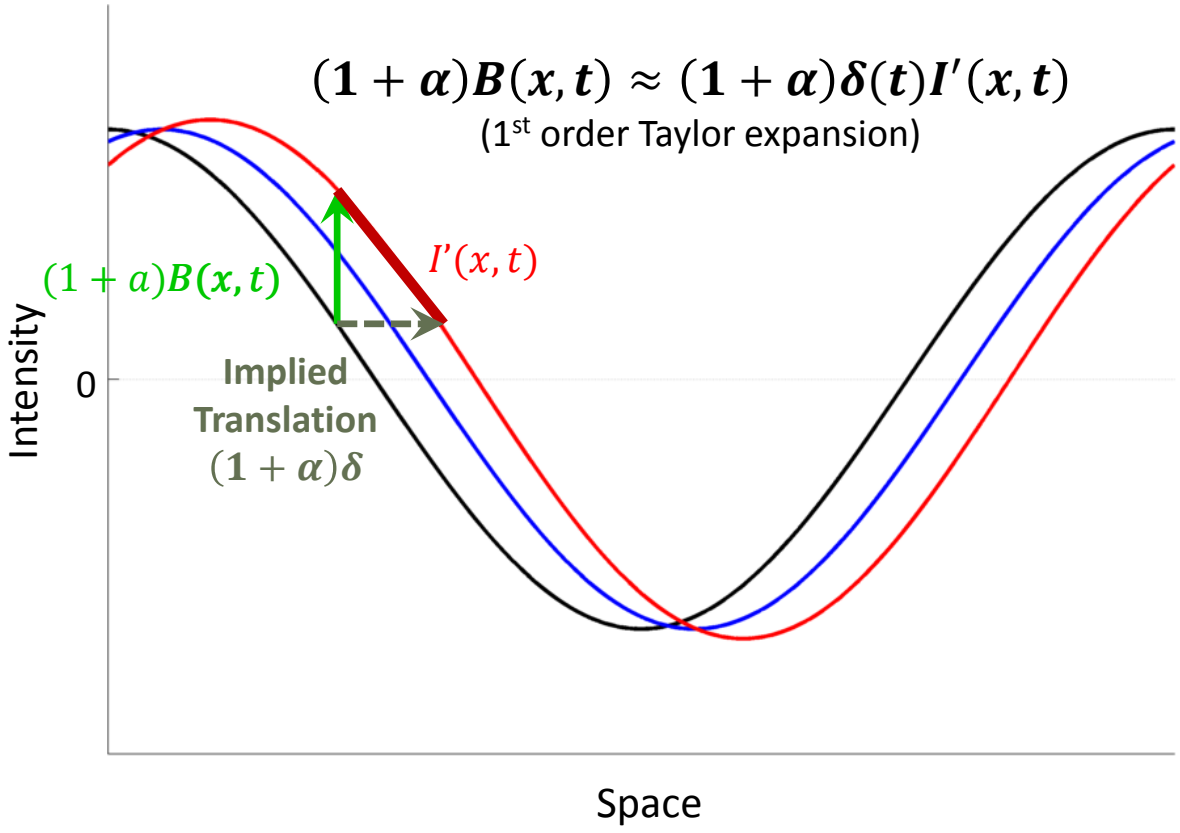
[H-Y. Wu et al., 2012]

Relating Temporal and Spatial Changes



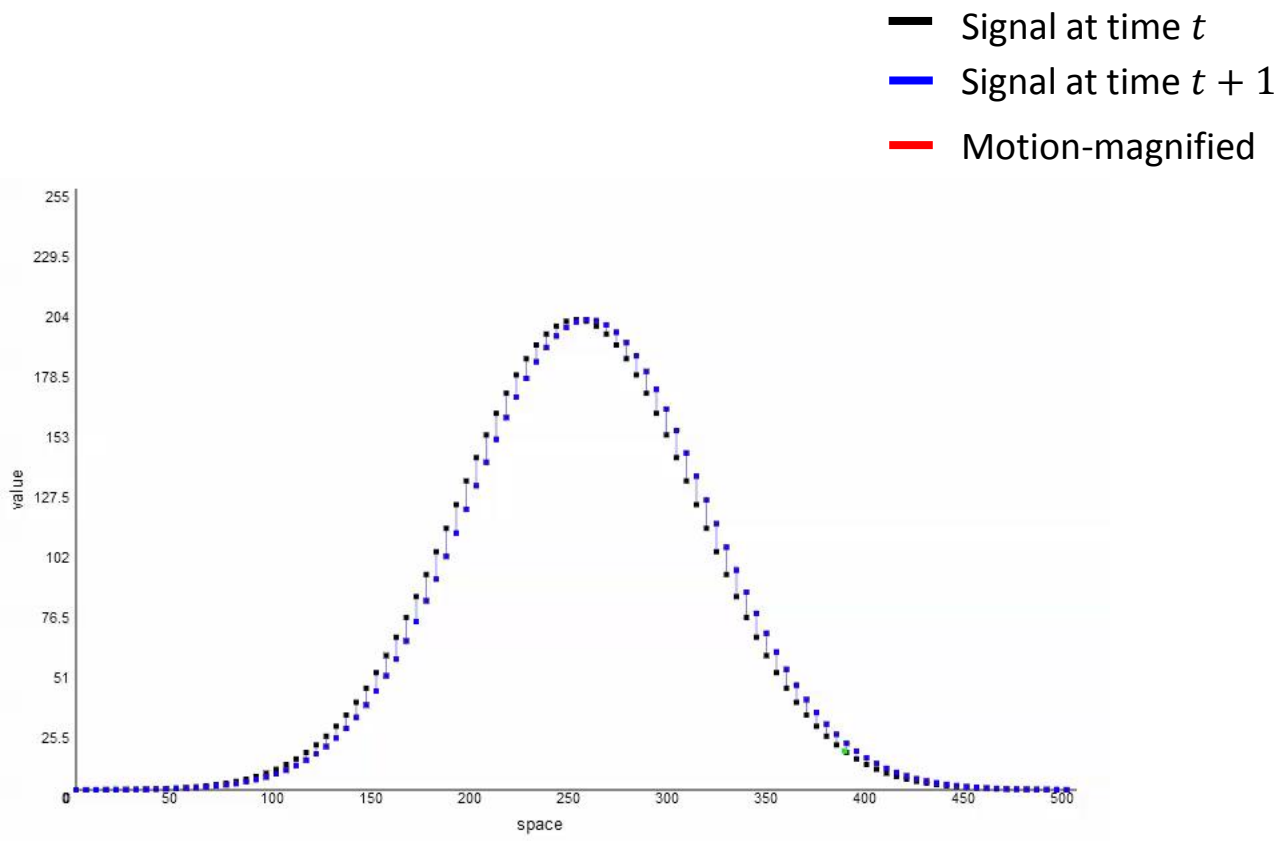
[H-Y. Wu et al., 2012]

Relating Temporal and Spatial Changes



[H-Y. Wu et al., 2012]

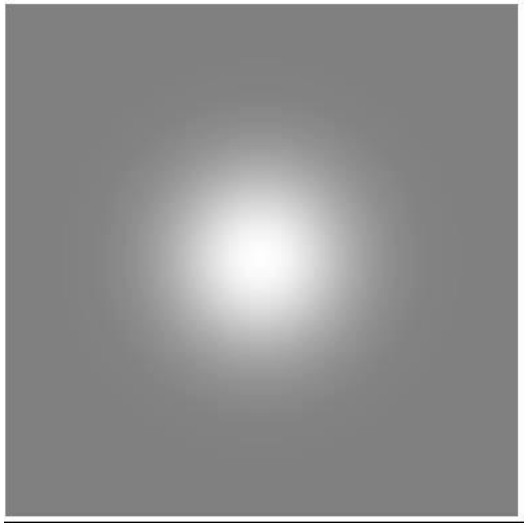
Relating Temporal and Spatial Changes



Courtesy of Lili Sun

[H-Y. Wu et al., 2012]

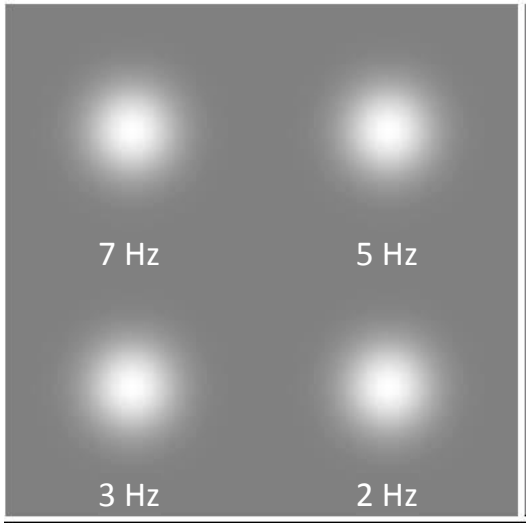
Synthetic 2D Example



Source

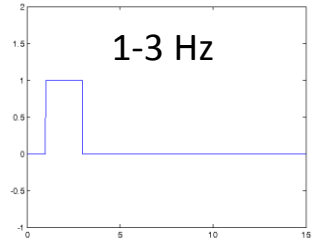
[H-Y. Wu et al., 2012]

Selective Motion Magnification



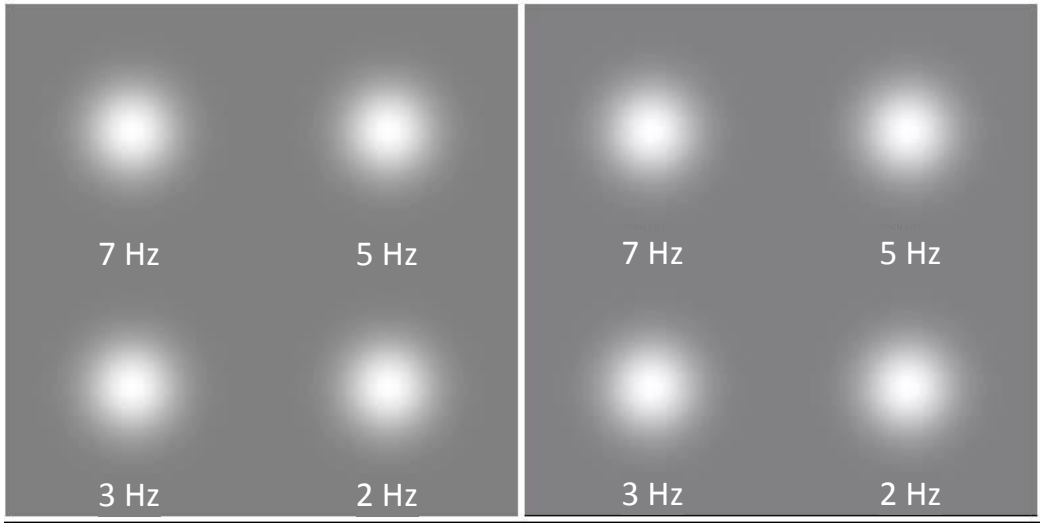
Source
(Single video with 4 blobs)

Temporal filter:



[H-Y. Wu et al., 2012]

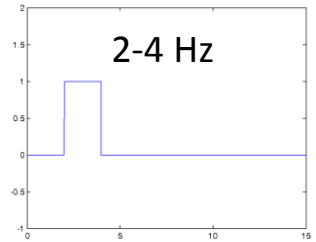
Selective Motion Magnification



Source
(Single video with 4 blobs)

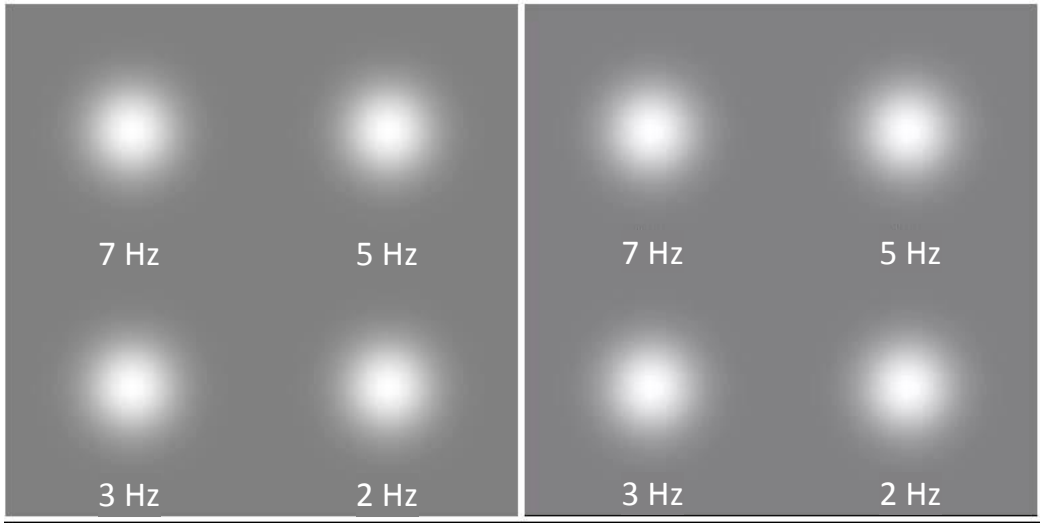
Motion-magnified (3 Hz)

Temporal filter:



[H-Y. Wu et al., 2012]

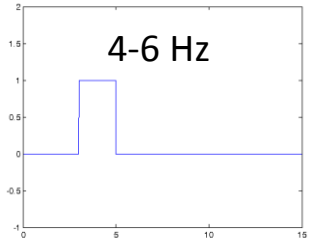
Selective Motion Magnification



Source
(Single video with 4 blobs)

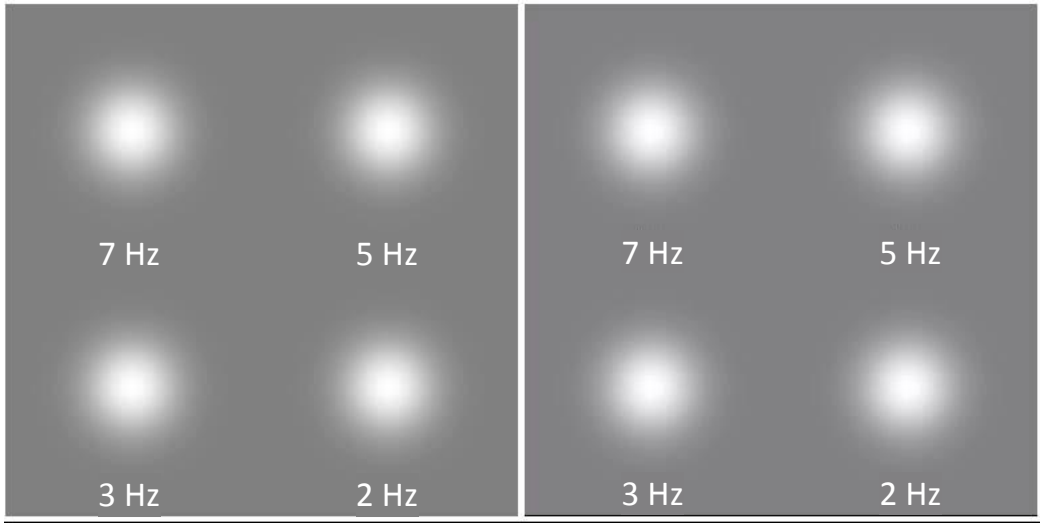
Motion-magnified (5 Hz)

Temporal filter:



[H-Y. Wu et al., 2012]

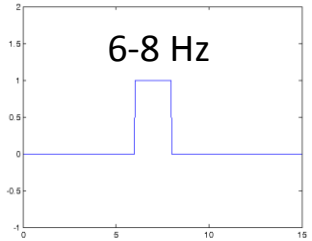
Selective Motion Magnification



Source
(Single video with 4 blobs)

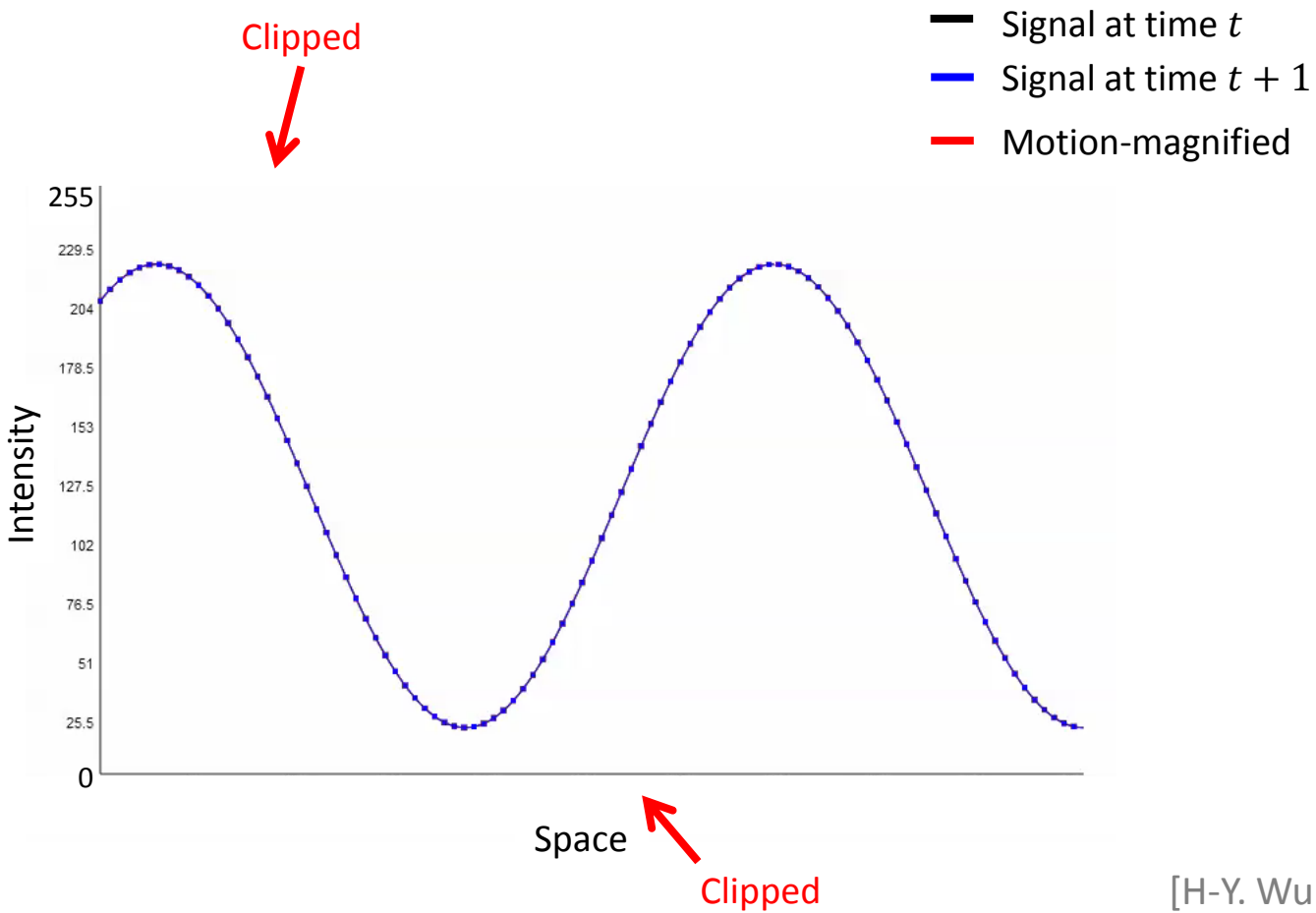
Motion-magnified (7 Hz)

Temporal filter:



[H-Y. Wu et al., 2012]

When Does It Break?

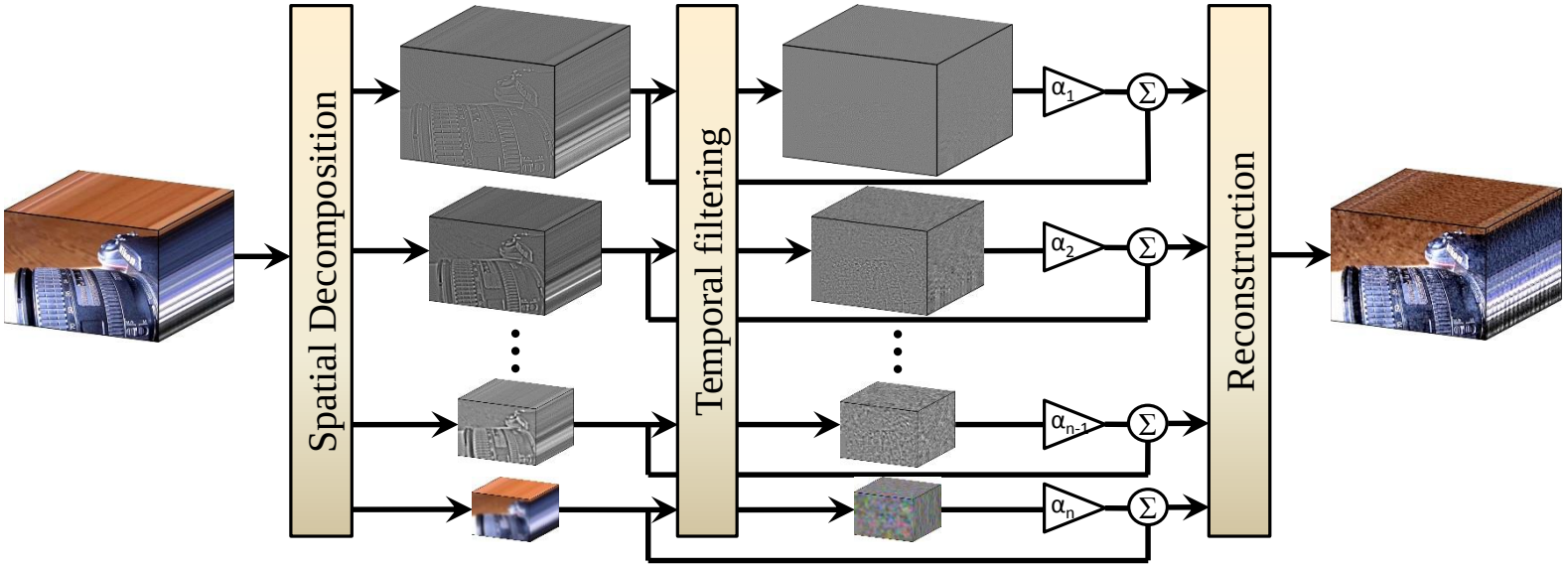


[H-Y. Wu et al., 2012]

Motion Magnification Artifacts

[H-Y. Wu et al., 2012]

Multi-scale Processing



[H-Y. Wu et al., 2012]

Motion Magnification Results

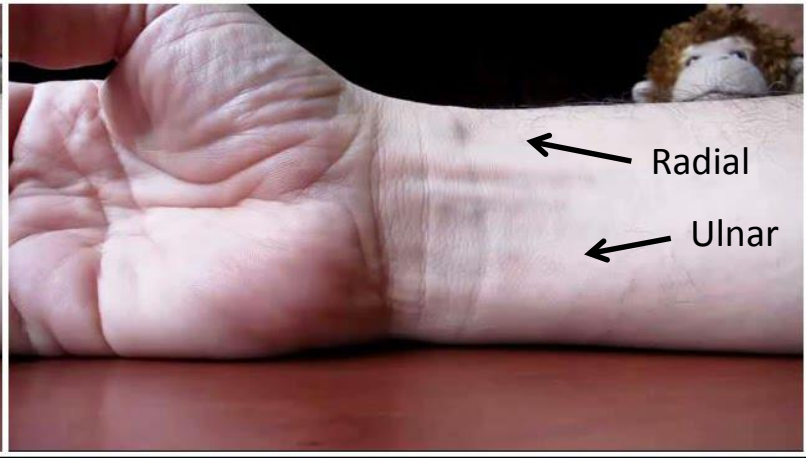
Source

Motion-magnified (0.4-3 Hz, x10)

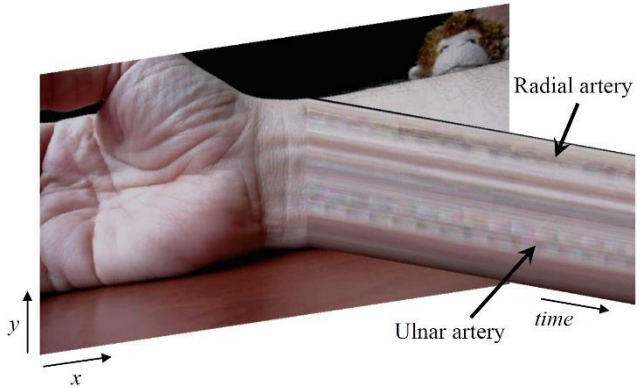
[H-Y. Wu et al., 2012]

Motion Magnification Results

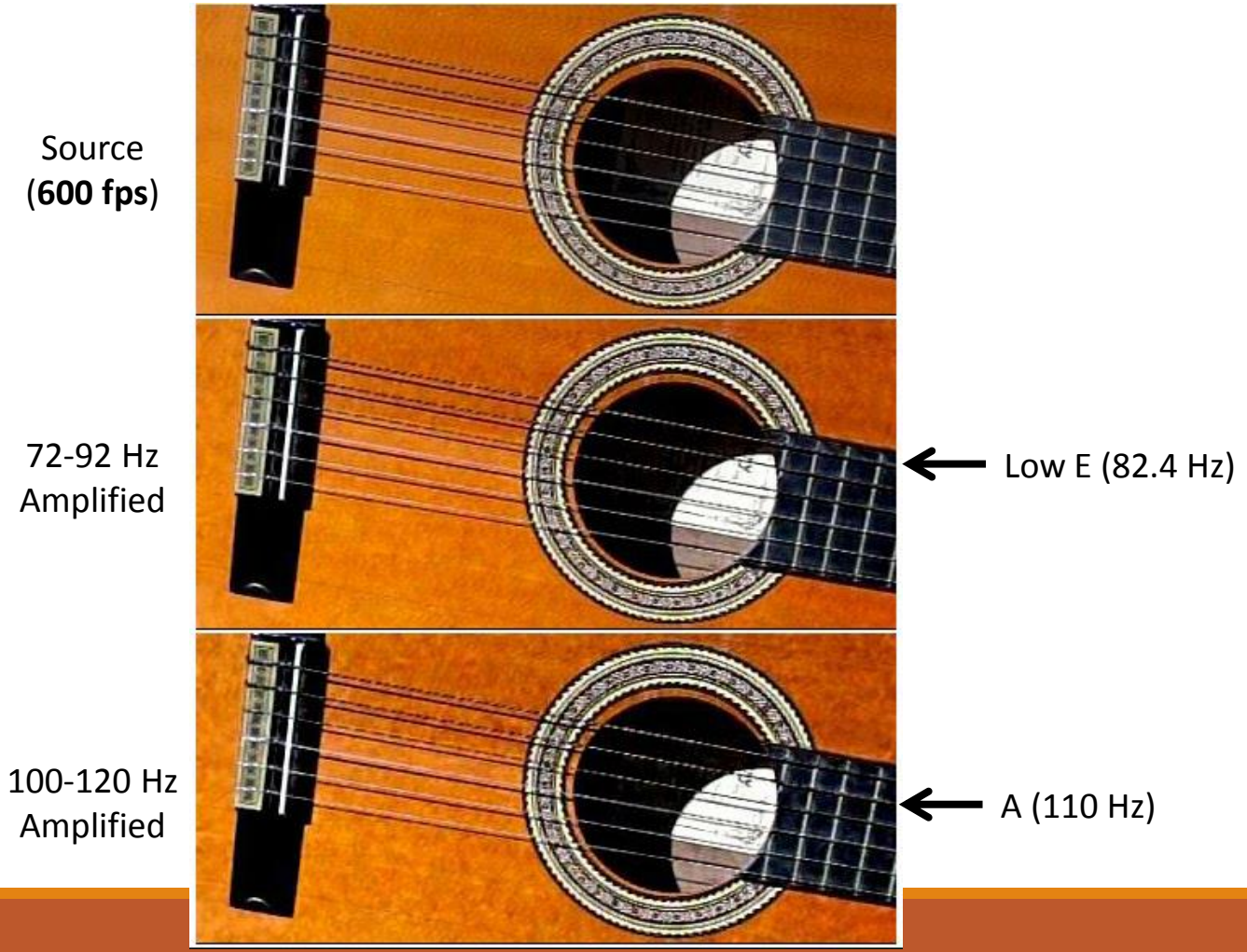
Source



Motion-magnified (0.4-3 Hz, x10)



Selective Motion Magnification in Natural Videos



Magnification of Non-periodic Motions

Source
(blend between two still images
taken 15 seconds apart)

Motion-magnified (0.5-10 Hz, x5)

[H-Y. Wu et al., 2012]

Motion Magnification Results

Source (**300 fps**) Motion-magnified (45-100 Hz, x100)

[H-Y. Wu et al., 2012]

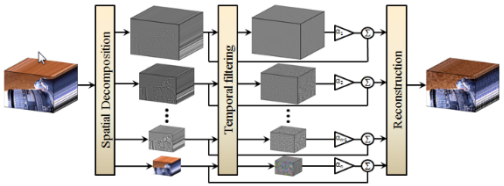
Eulerian Motion Magnification

Eulerian processing of standard videos can enhance changes that are hard to see with the naked eye

- Simple, Robust, Fast!

Unified framework that can amplify both color changes and motion

- No motion estimation or tracking required



[H-Y. Wu et al., 2012]

Reading material:

- A. A. Efros, A. C. Berg, G. Mori, and J. Malik. "Recognizing action at a distance." In *International Conference on Computer Vision*, 2003.
- A. F. Bobick and J. W. Davis. "The recognition of human movement using temporal templates." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, no. 3 (2001): 257-267.
- I. Laptev. "On space-time interest points." *International Journal of Computer Vision* 64, no. 2-3 (2005): 107-123.
- I. Laptev and P. Pérez. "Retrieving actions in movies." In *International Conference on Computer Vision*, 2007.
- A. Prest, V. Ferrari, and C. Schmid. "Explicit modeling of human-object interactions in realistic videos." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, no. 4 (2012): 835-848.
- H. Wang and C. Schmid. "Action recognition with improved trajectories." In *International Conference on Computer Vision*, 2013.
- H-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman. "Eulerian video magnification for revealing subtle changes in the world." *ACM Transactions on Graphics* 31, no. 4 (2012): 1-8.