Action recognition in videos

Cordelia Schmid

Action recognition - goal

- Short actions, i.e. answer phone, shake hands

answer phone

hand shake

Action recognition - goal

- Activities/events, i.e. making a sandwich, doing homework

Making sandwich

Doing homework

TrecVid Multi-media event detection dataset

Action recognition - goal

- Activities/events, i.e. birthday party, parade

Birthday party

Parade

TrecVid Multi-media event detection dataset

Action recognition - tasks

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

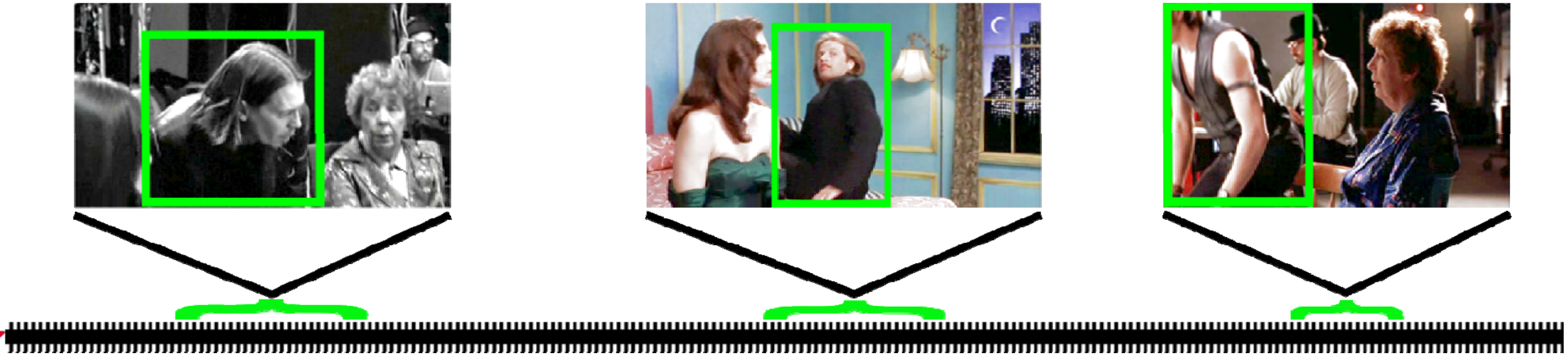
Making sandwich: present
Feeding animal: not present
...

Action recognition - tasks

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

Making sandwich: present
Feeding animal: not present
...

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



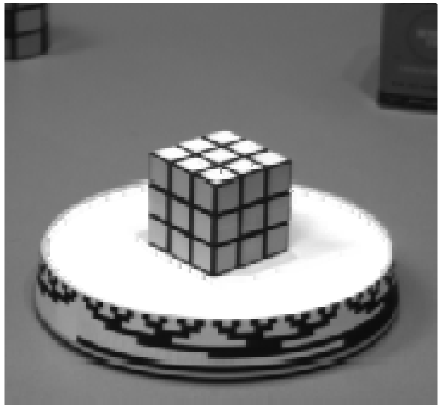
Overview

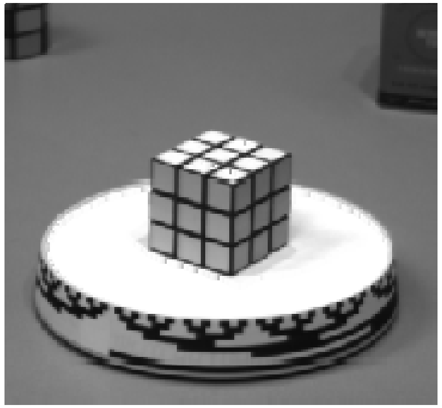
- *Optical flow*

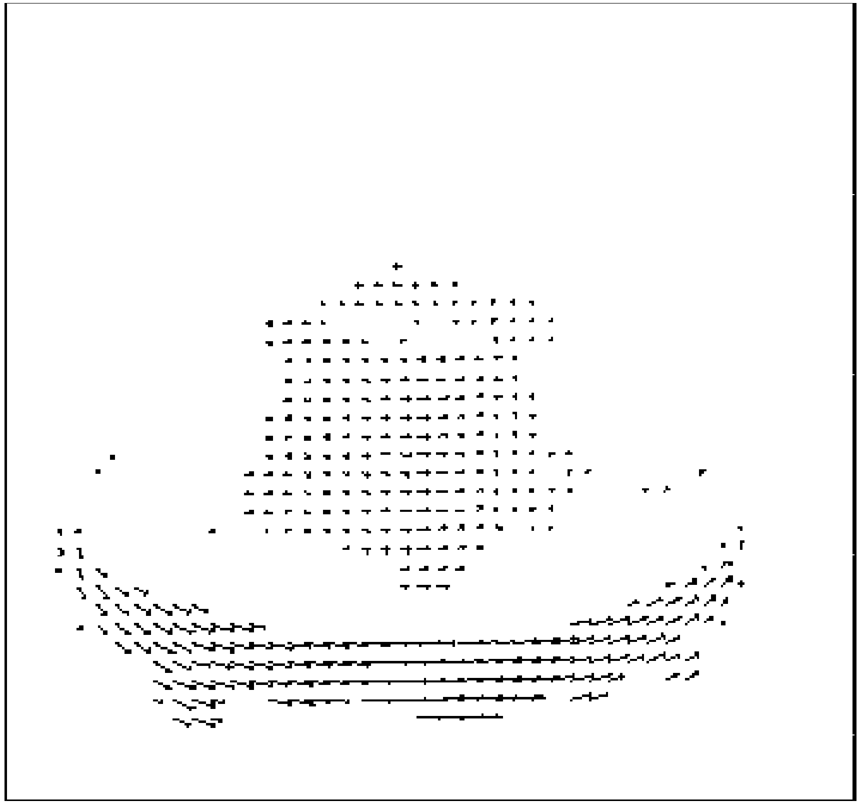
- Trajectory-based low level features for action recognition

Motion field

- The motion field is the projection of the 3D scene motion into the image



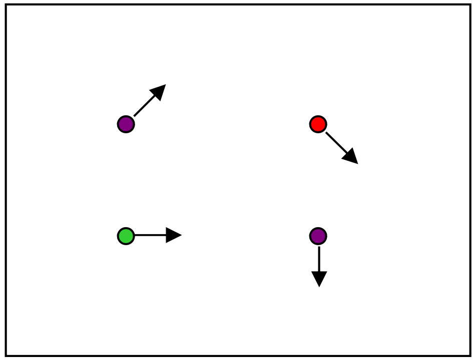




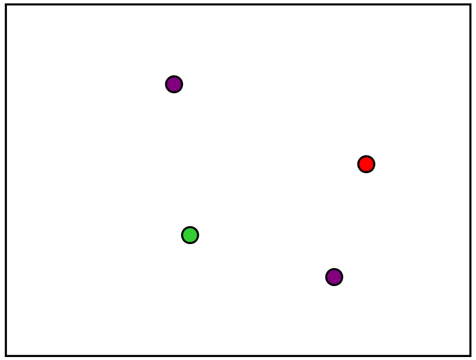
Optical flow

- Definition: optical flow is the *apparent* motion of brightness patterns in the image
- Ideally, optical flow would be the same as the motion field
- Have to be careful: apparent motion can be caused by lighting changes without any actual motion
  - Think of a uniform rotating sphere under fixed lighting vs. a stationary sphere under moving illumination

Estimating optical flow



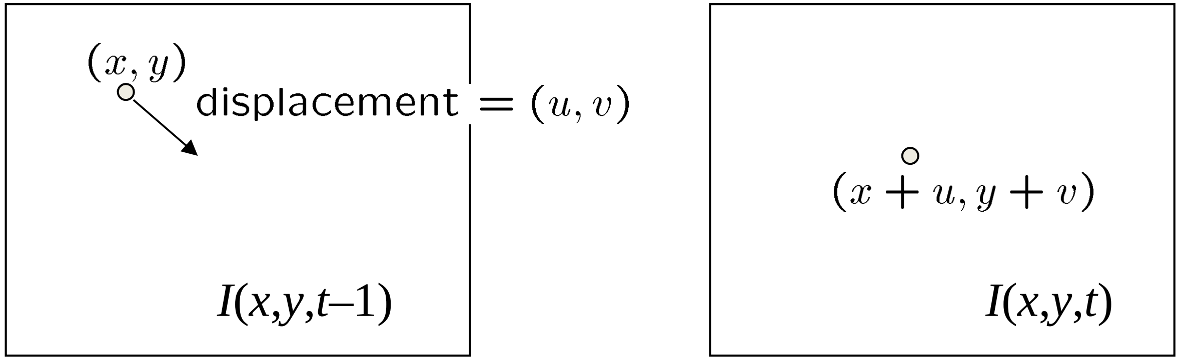
$I(x,y,t-1)$



$I(x,y,t)$

- Given two subsequent frames, estimate the apparent motion field $u(x,y)$ and $v(x,y)$ between them
- Key assumptions
  - Brightness constancy: projection of the same point looks the same in every frame
  - Small motion: points do not move very far
  - Spatial coherence: points move like their neighbors

The brightness constancy constraint



Brightness Constancy Equation:

$$I(x, y, t-1) = I(x + u(x, y), y + v(x, y), t)$$

Linearizing the right side using Taylor expansion:

$$I(x, y, t-1) \approx I(x, y, t) + I_x u(x, y) + I_y v(x, y)$$

Hence, $I_x u + I_y v + I_t \approx 0$

The brightness constancy constraint

$$I_x u + I_y v + I_t = 0$$

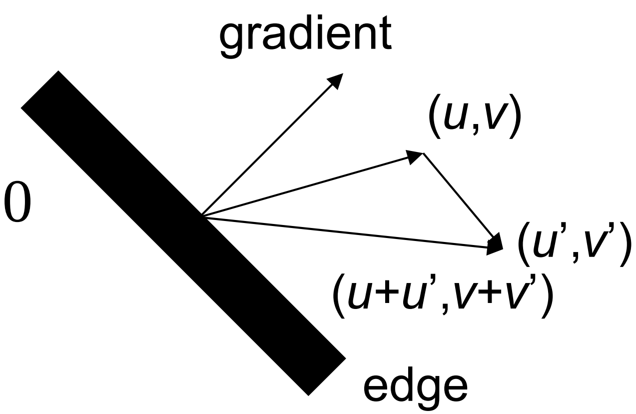
- How many equations and unknowns per pixel?
  - One equation, two unknowns

- What does this constraint mean?

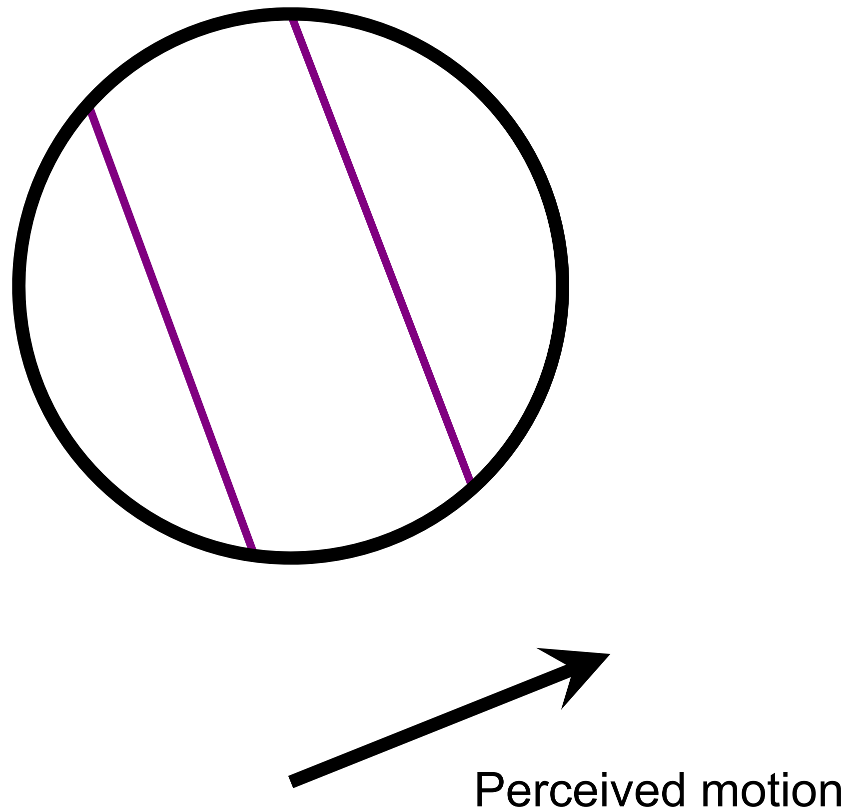
$$\nabla I \cdot (u,v) + I_t = 0$$

- The component of the flow perpendicular to the gradient (i.e., parallel to the edge) is unknown

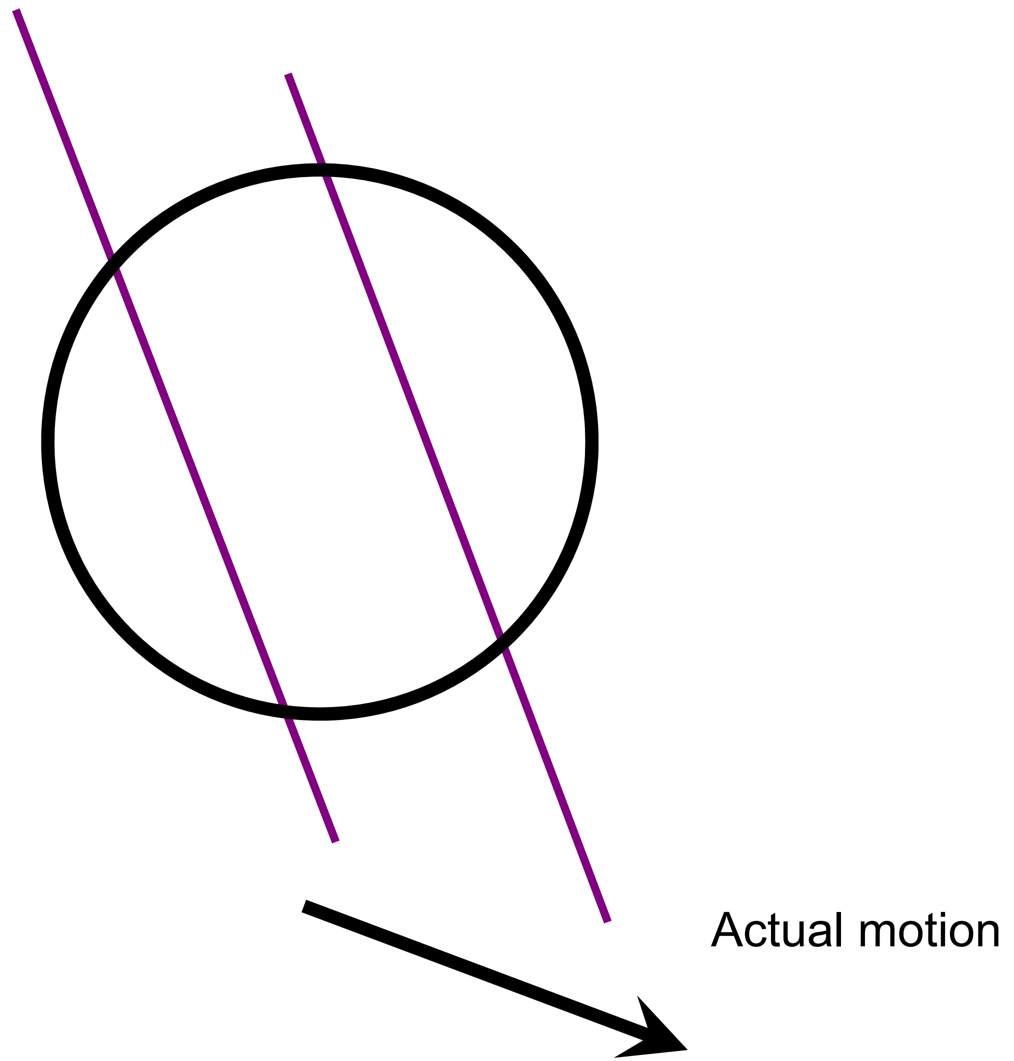
If (u, v) satisfies the equation,
so does (u+u', v+v') if $\nabla I \cdot (u', v') = 0$



The aperture problem



The aperture problem



Solving the aperture problem

- How to get more equations for a pixel?
- **Spatial coherence constraint:** pretend the pixel's neighbors have the same (u,v)
  - E.g., if we use a 5x5 window, that gives us 25 equations per pixel

$$\begin{bmatrix} I_x(\mathbf{x}_1) & I_y(\mathbf{x}_1) \\ I_x(\mathbf{x}_2) & I_y(\mathbf{x}_2) \\ \vdots & \vdots \\ I_x(\mathbf{x}_n) & I_y(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(\mathbf{x}_1) \\ I_t(\mathbf{x}_2) \\ \vdots \\ I_t(\mathbf{x}_n) \end{bmatrix}$$

B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*,1981.

informatics mathematics

Lucas-Kanade flow

- Linear least squares problem

$$\begin{bmatrix} I_x(\mathbf{x}_1) & I_y(\mathbf{x}_1) \\ I_x(\mathbf{x}_2) & I_y(\mathbf{x}_2) \\ \vdots & \vdots \\ I_x(\mathbf{x}_n) & I_y(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(\mathbf{x}_1) \\ I_t(\mathbf{x}_2) \\ \vdots \\ I_t(\mathbf{x}_n) \end{bmatrix}$$

$$\underset{n\times 2}{\mathbf{A}} \; \underset{2\times 1}{\mathbf{d}} = \underset{n\times 1}{\mathbf{b}}$$

Solution given by $(\mathbf{A}^T\mathbf{A})\mathbf{d} = \mathbf{A}^T\mathbf{b}$

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

The summations are over all pixels in the window

Lucas-Kanade flow

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

- Recall the Harris corner detector: M = $A^T A$ is the *second moment matrix*
- When is the system solvable?
  - By looking at the eigenvalues of the second moment matrix
  - The eigenvectors and eigenvalues of M relate to edge direction and magnitude
  - The eigenvector associated with the larger eigenvalue points in the direction of fastest intensity change, and the other eigenvector is orthogonal to it

Uniform region

– gradients have small magnitude
– small λ_1, small λ_2
– system is ill-conditioned

informatics mathematics Inria

Edge

– gradients have one dominant direction
– large λ_1, small λ_2
– system is ill-conditioned

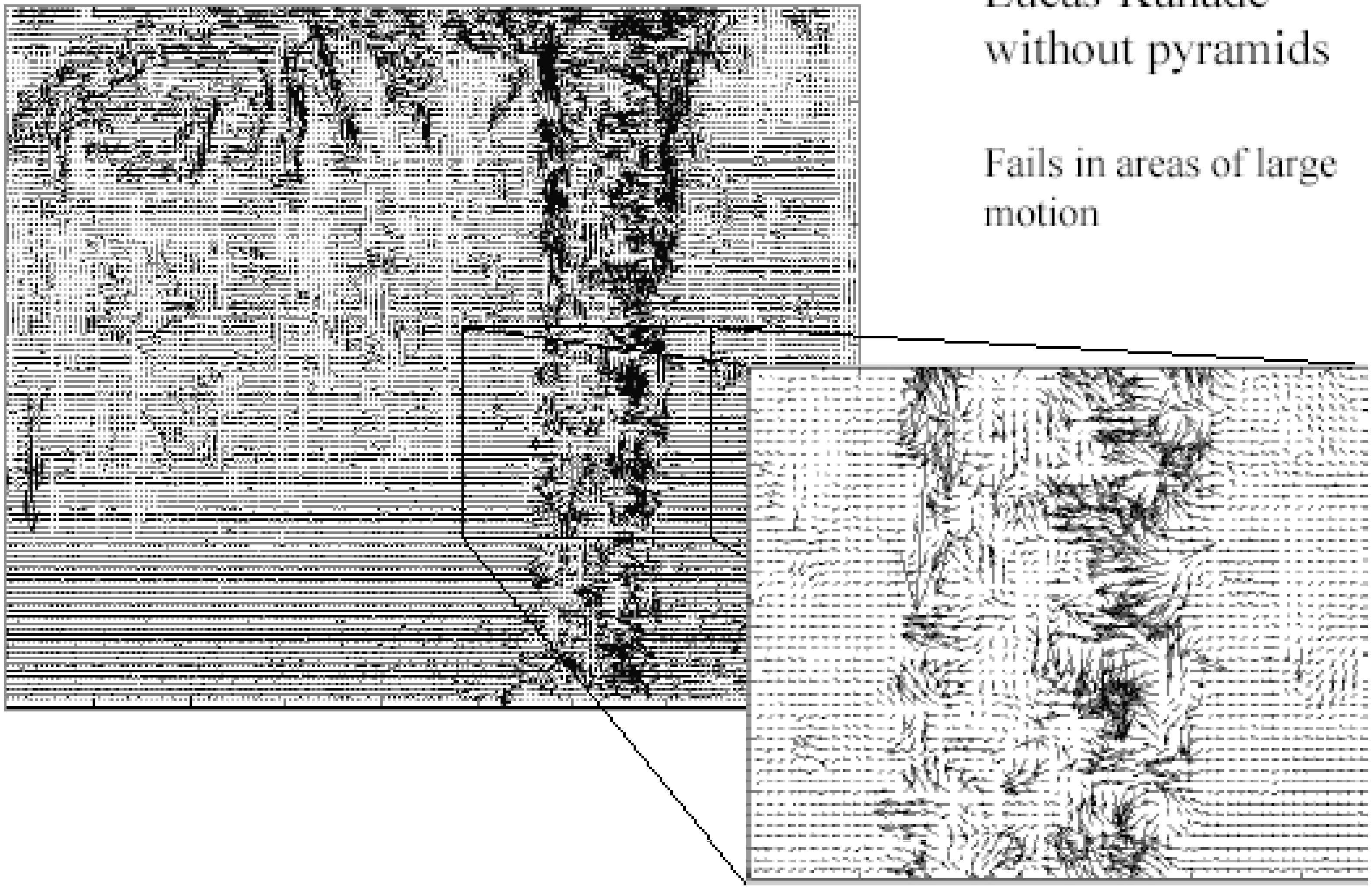
High-texture or corner region

– gradients have different directions, large magnitudes

– large λ_1, large λ_2

– system is well-conditioned

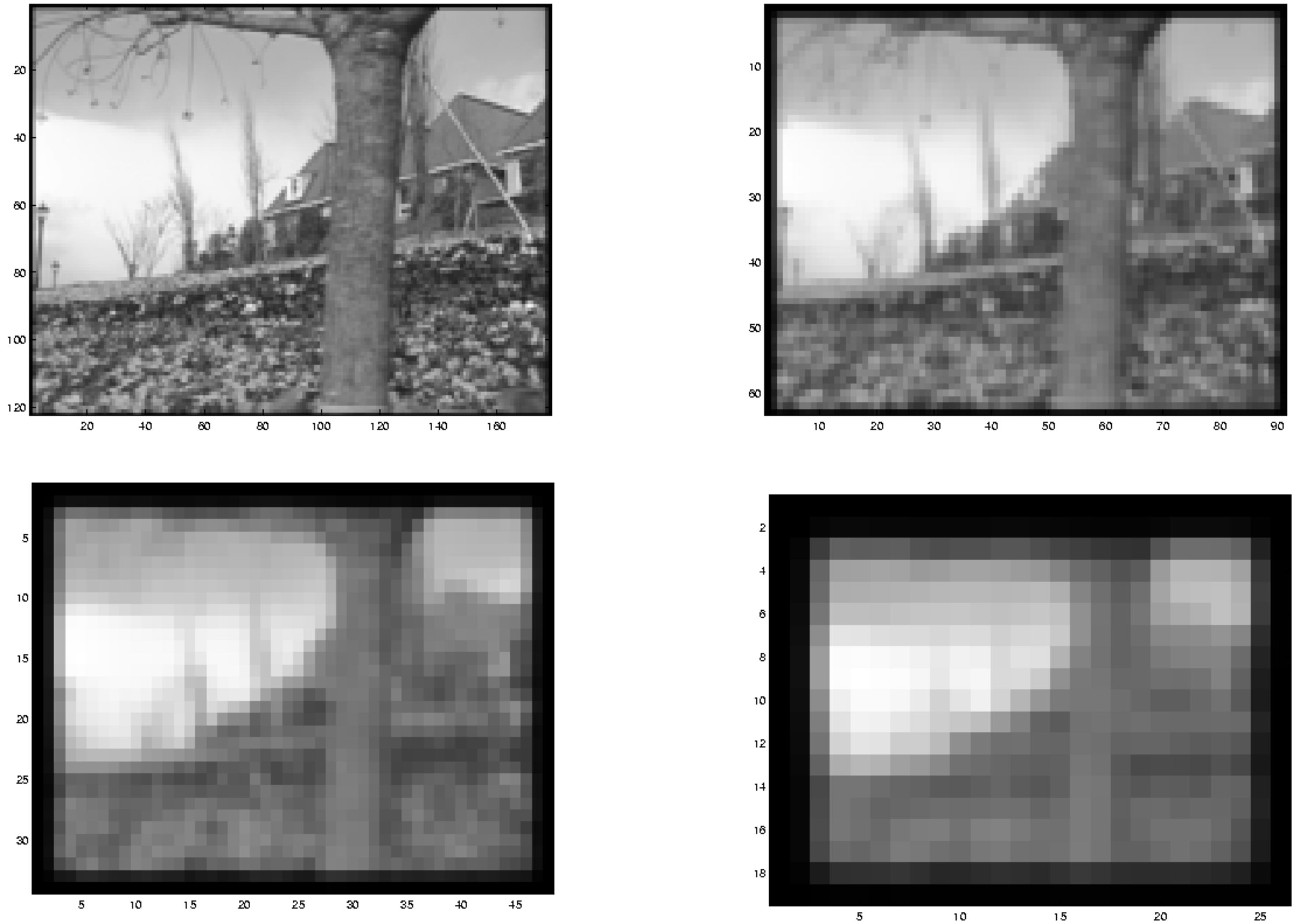
Optical Flow Results



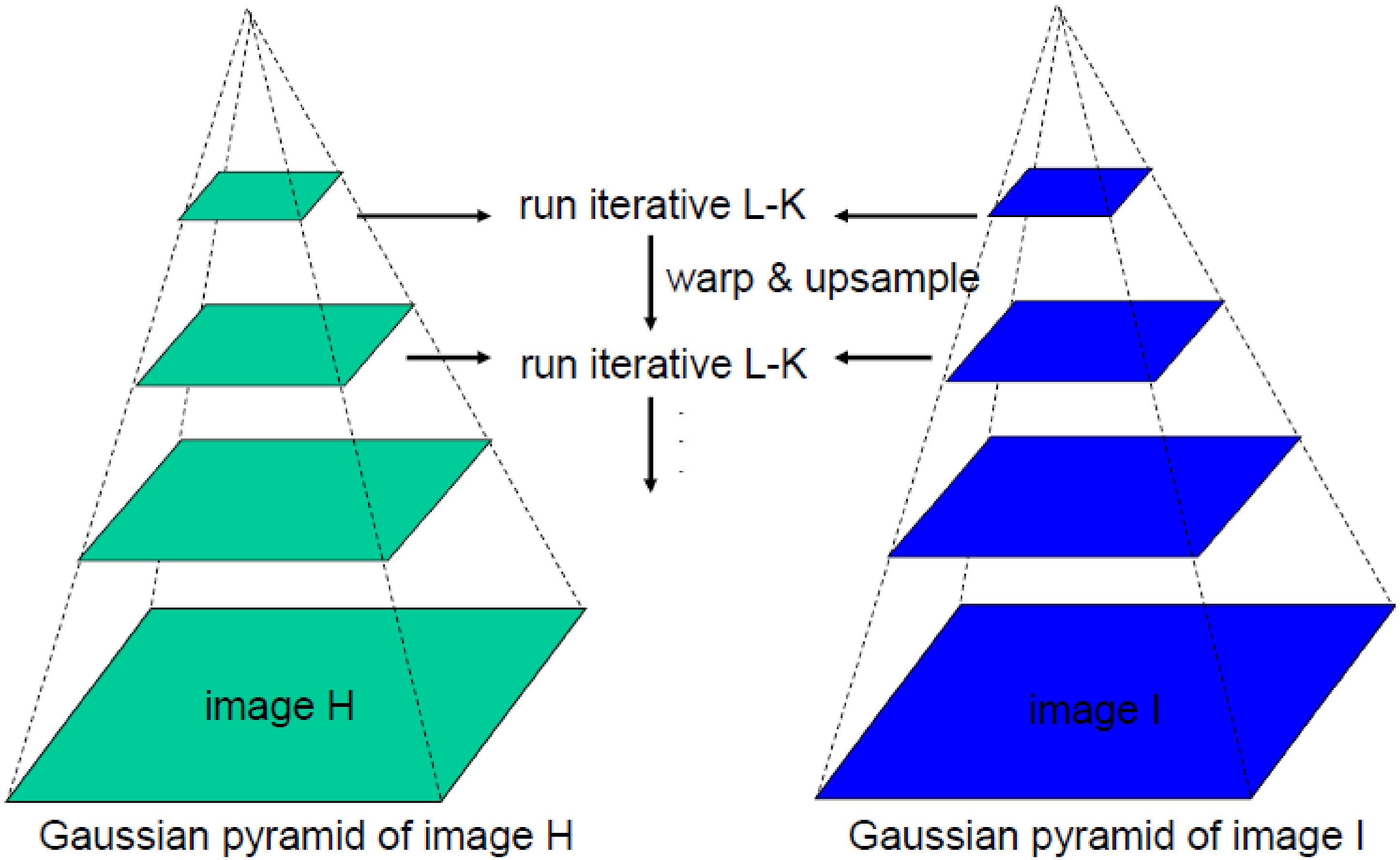
Lucas-Kanade without pyramids

Fails in areas of large motion

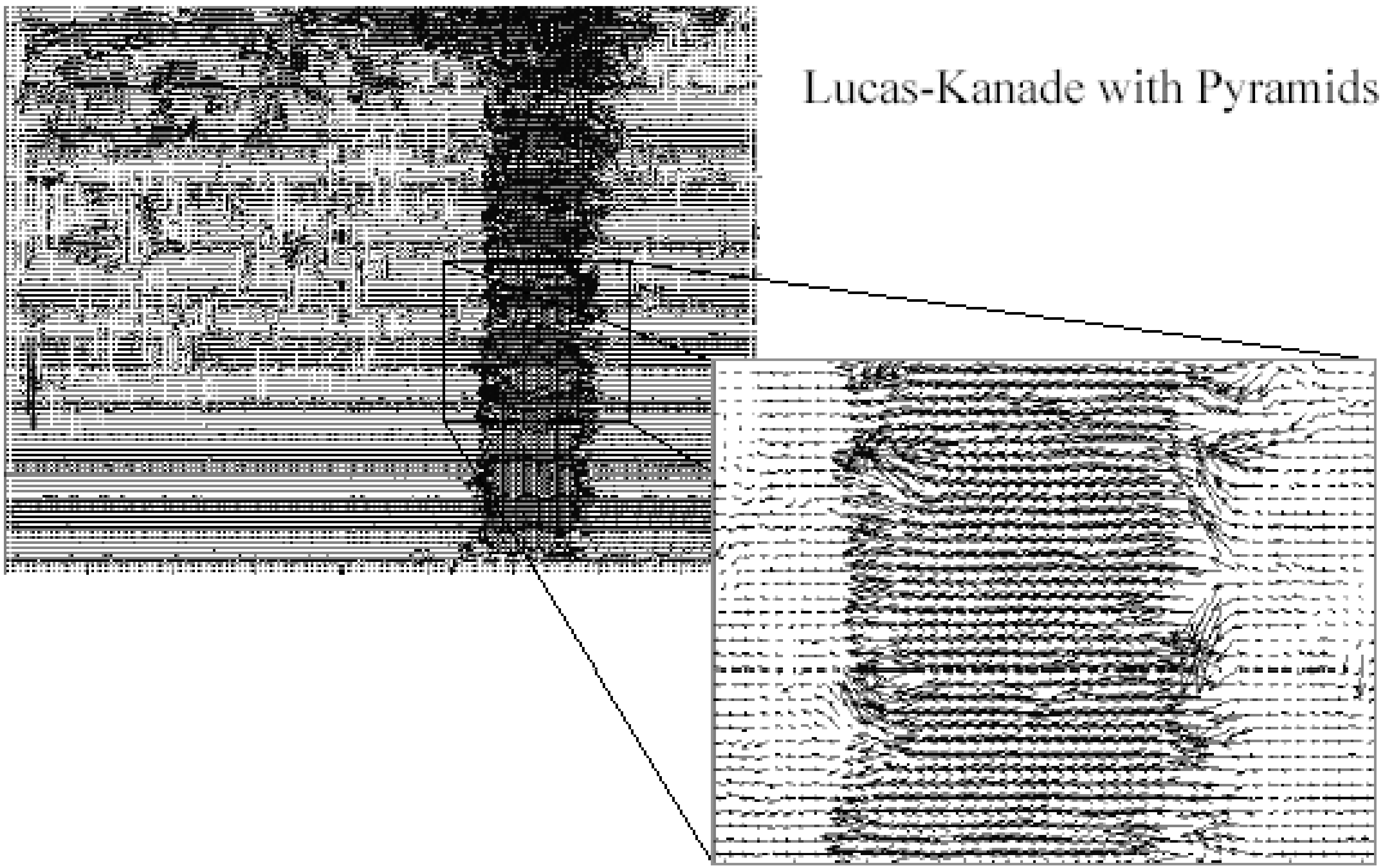
Multi-resolution registration



Coarse to fine optical flow estimation



Optical Flow Results



Horn & Schunck algorithm

Additional smoothness constraint :

$$e_s = \iint ((u_x^2 + u_y^2) + (v_x^2 + v_y^2))\,dxdy,$$

besides OF constraint equation term

$$e_c = \iint (I_x u + I_y v + I_t)^2\,dxdy,$$

minimize $e_s + \lambda e_c$ λ regularization parameter

B.K.P. Horn and B.G. Schunck, "Determining optical flow." *Artificial Intelligence*,1981

Horn & Schunck algorithm

$$E(u(x,y),v(x,y)) = \iint \underbrace{(I_x u + I_y v + I_t)^2}_{\text{Data term brightness constancy}} + \alpha \underbrace{((u_x^2 + u_y^2) + (v_x^2 + v_y^2))}_{\text{Smoothness term}} dxdy$$

$$E(u,v) = \int_{\Omega} F(x,y,u,v,u_x,u_y,v_x,v_y)dxdy$$

Euler-Lagrange equations

$$F_u - \frac{\partial}{\partial x} F_{u_x} - \frac{\partial}{\partial y} F_{u_y} = 0 \qquad F_v - \frac{\partial}{\partial x} F_{v_x} - \frac{\partial}{\partial y} F_{v_y} = 0$$

According to the calculus of variations, a minimizer of E must fulfill the Euler-Lagrange equations

Horn & Schunck

The Euler-Lagrange equations :

$$F_u - \frac{\partial}{\partial x} F_{u_x} - \frac{\partial}{\partial y} F_{u_y} = 0$$

$$F_v - \frac{\partial}{\partial x} F_{v_x} - \frac{\partial}{\partial y} F_{v_y} = 0$$

In our case ,

$$F = (u_x^2 + u_y^2) + (v_x^2 + v_y^2) + \lambda (I_x u + I_y v + I_t)^2,$$

so the Euler-Lagrange equations are

$$\Delta u = \lambda (I_x u + I_y v + I_t) I_x,$$

$$\Delta v = \lambda (I_x u + I_y v + I_t) I_y,$$

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$ is the Laplacian operator

Horn & Schunck

Remarks :

1. Coupled PDEs solved using iterative methods and finite differences

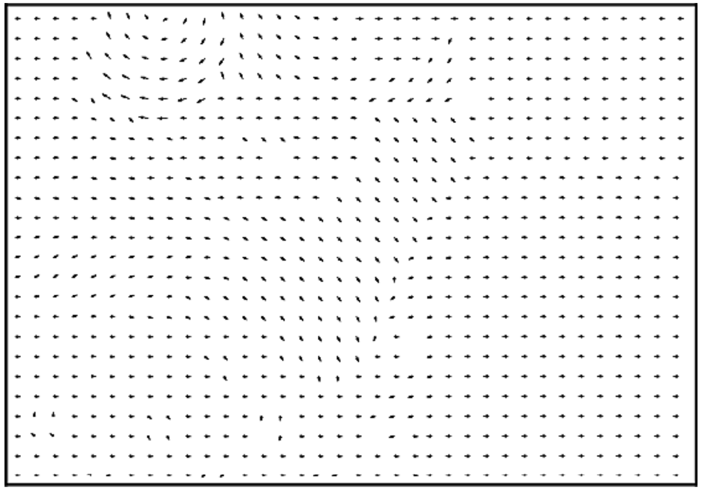
$$\frac{\partial u}{\partial t} = \Delta u - \lambda(I_x u + I_y v + I_t) I_x ,$$

$$\frac{\partial v}{\partial t} = \Delta v - \lambda(I_x u + I_y v + I_t) I_y ,$$

2. Information spreads from corner-type patterns

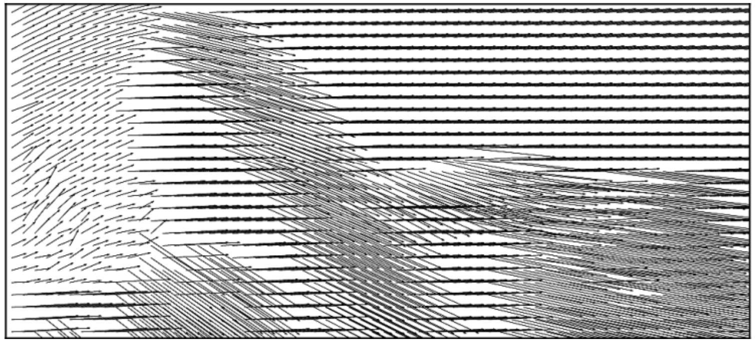
Horn & Schunck

- Works well for small displacements
  - For example Middlebury sequence



Large displacement estimation in optical flow

- Large displacement is still an open problem in optical flow estimation

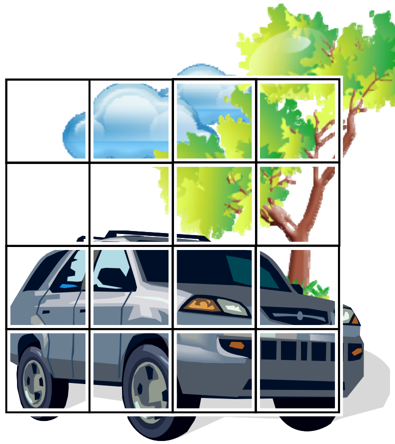


MPI Sintel dataset

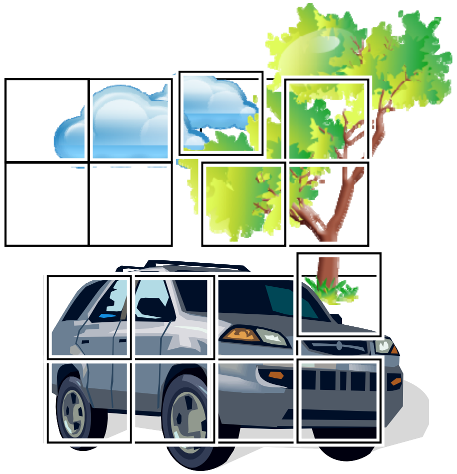
Large displacement optical flow

- Classical optical flow [Horn and Schunck 1981]
  - energy: $$E(\mathbf{w}) = \iint \underbrace{E_{data}}_{\text{color/gradient constancy}} + \alpha \underbrace{E_{smooth}}_{\text{smoothness constraint}} \mathbf{dx}$$
  - minimization using a coarse-to-fine scheme
- Large displacement approaches:
  - LDOF [Brox and Malik 2011]
    a matching term, penalizing the difference between flow and HOG matches
    $$E(\mathbf{w}) = \iint E_{data} + \alpha E_{smooth} + \beta E_{match} \mathbf{dx}$$
  - MDP-Flow2 [Xu *et al.* 2012]
    expensive fusion of matches (SIFT + PatchMatch) and estimated flow at each level
  - DeepFlow [Weinzaepfel *et al.* 2013]
    deep matching + flow refinement with variational approach

Deep Matching: main idea



First image

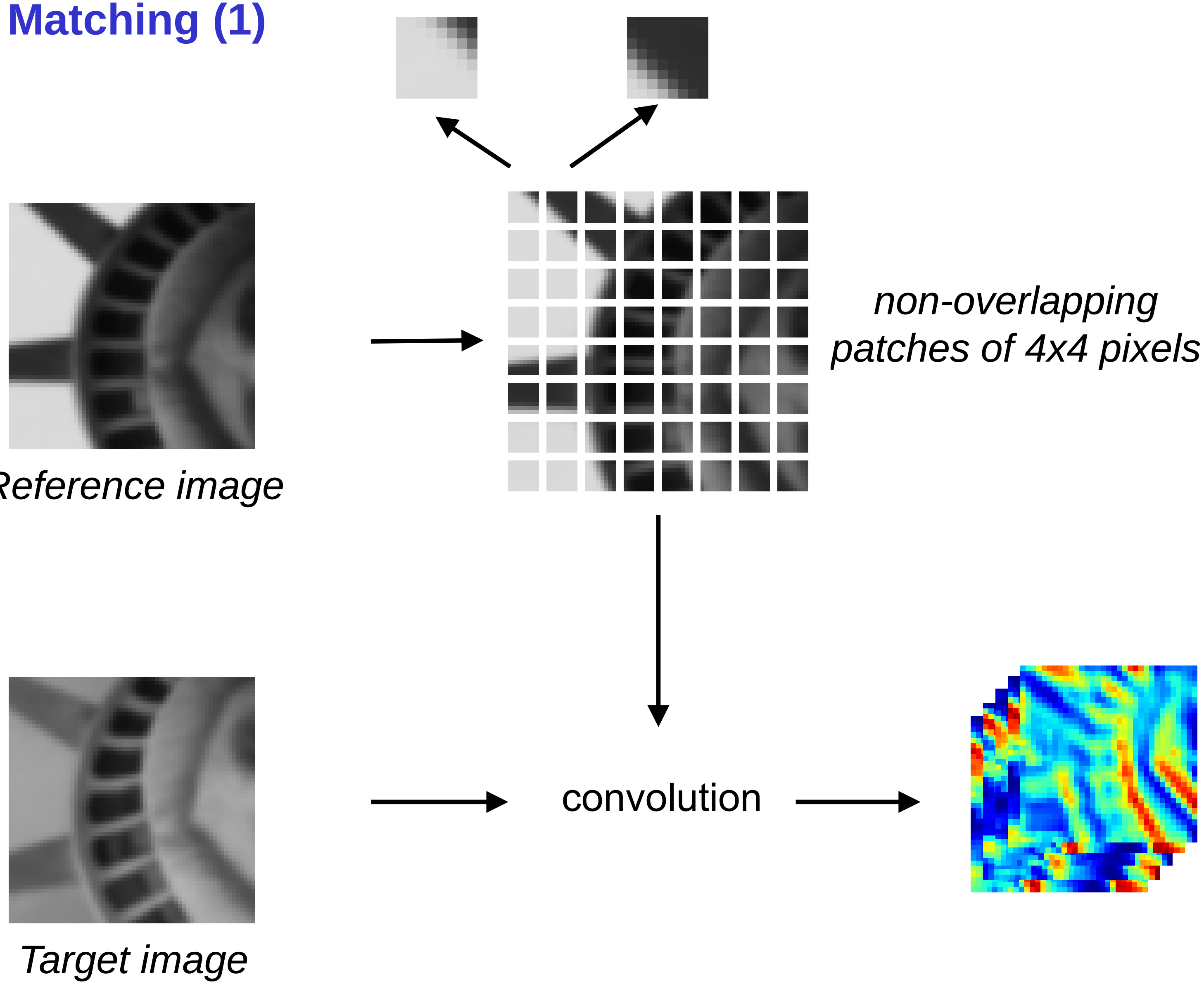


Second image

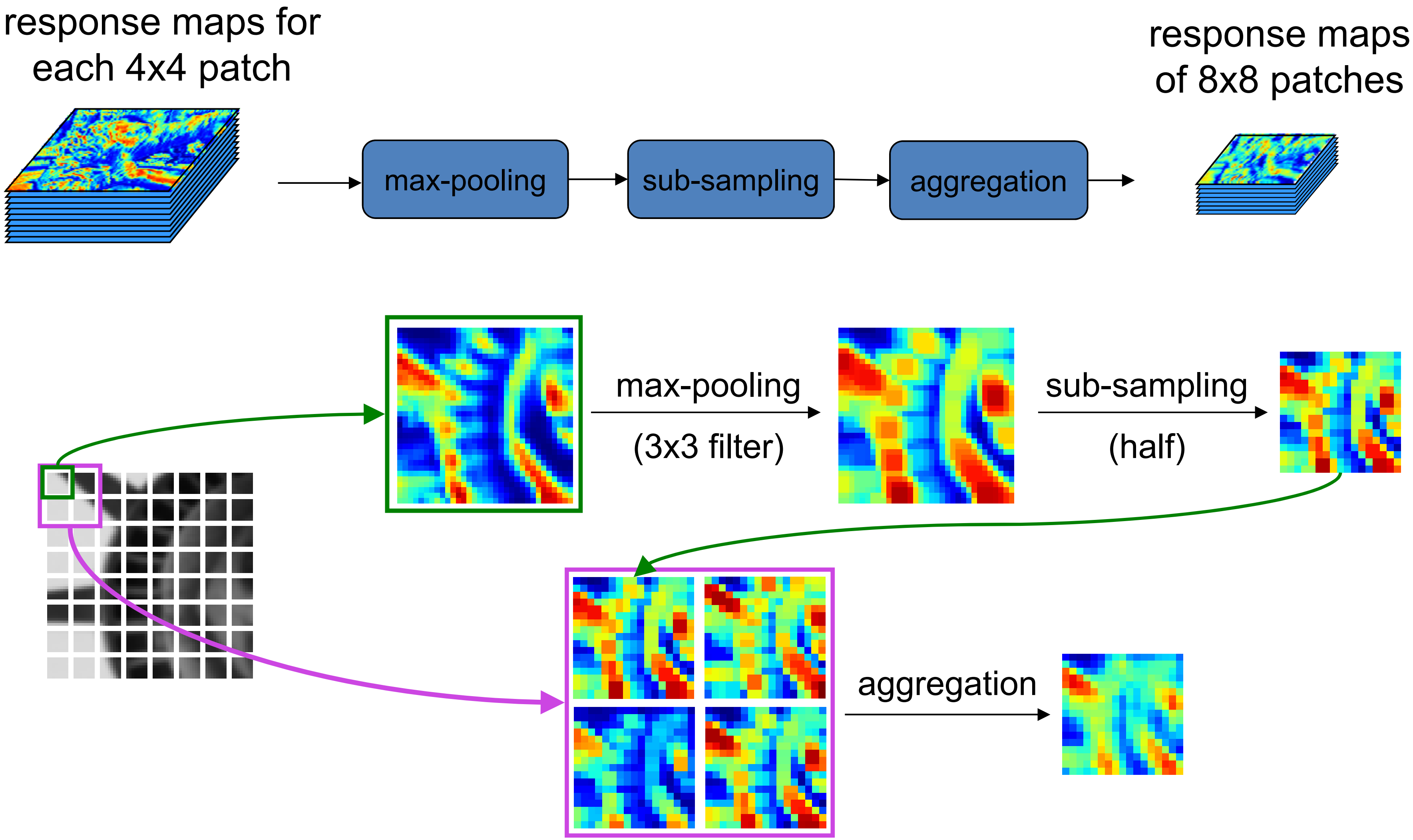
- Each subpatch is allowed to move:
  - independently
  - in a limited range depending on its size
- The approach is fast to compute using convolution and max-pooling
- The idea is applied recursively

informatics mathematics

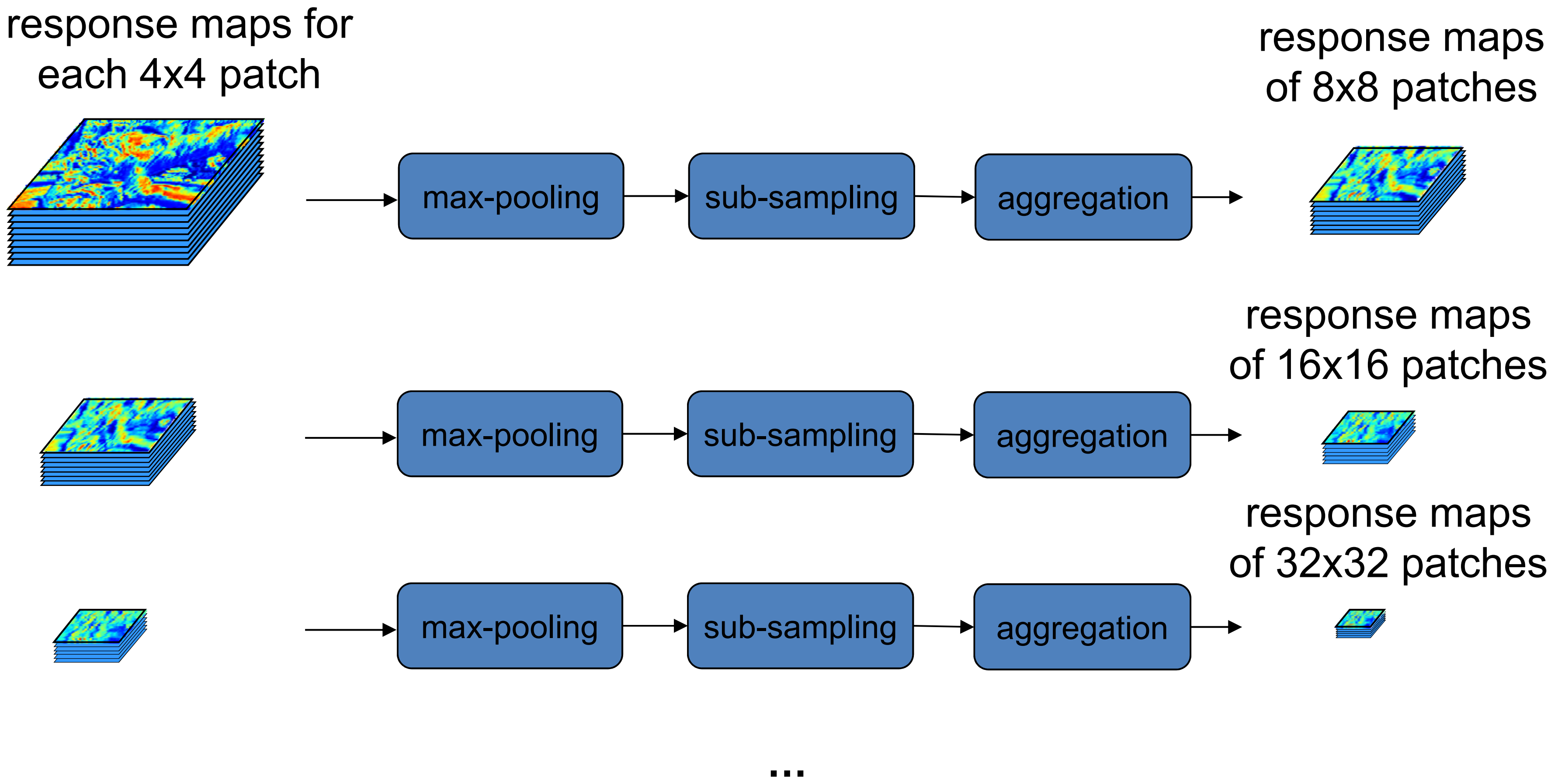
Deep Matching (1)



Deep Matching (2)

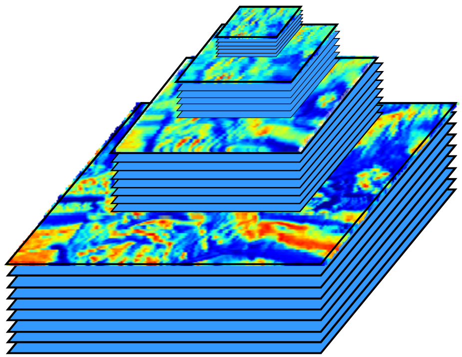


Deep Matching (2)



Pipeline similar in spirit to **deep** convolutional nets [Lecun *et al.* 1998]

Deep Matching (3)



Multi-scale response pyramid

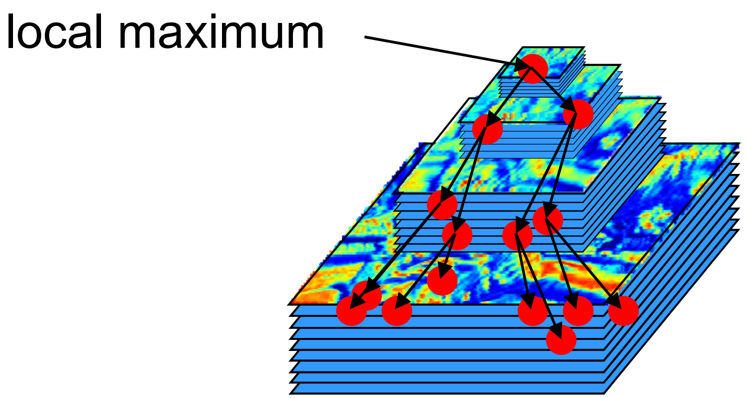
Extract scale-space local maxima

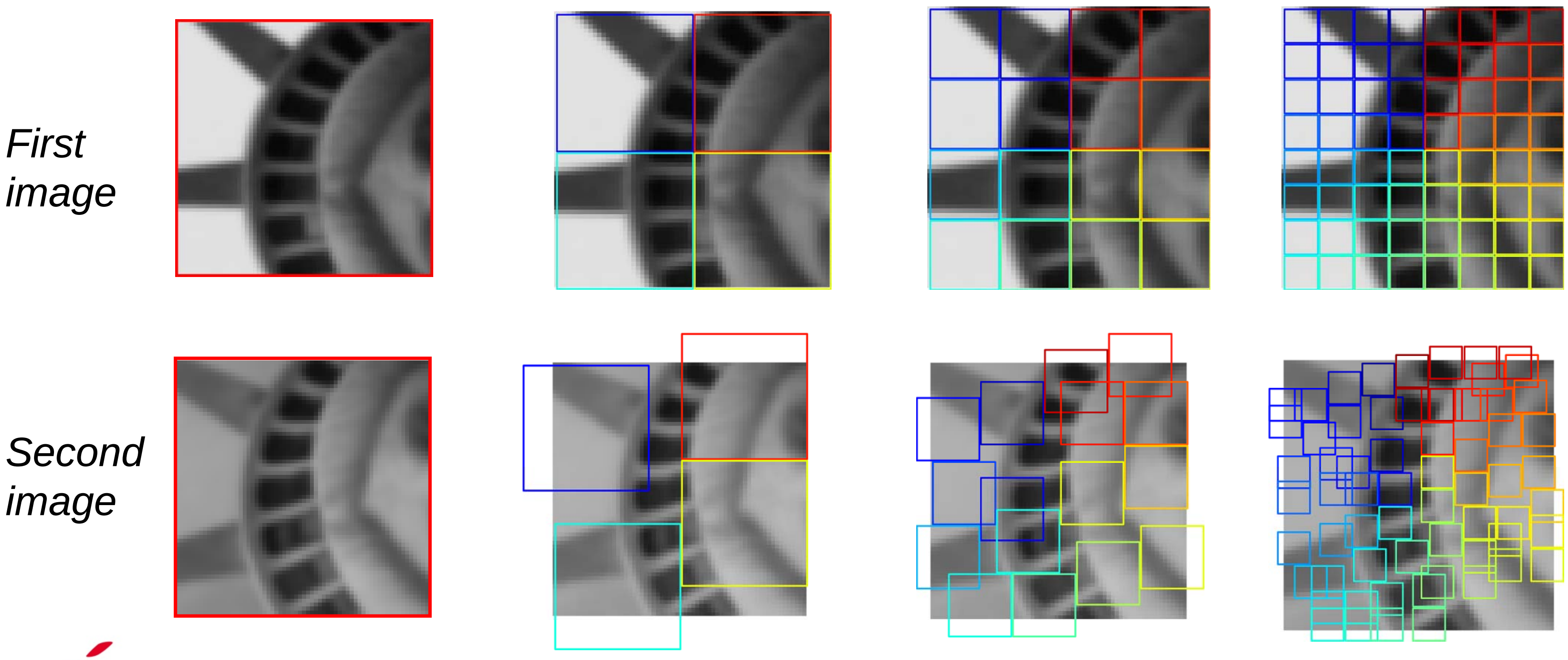
Backtrack quasi-dense correspondences

Bottom-up

Top-down

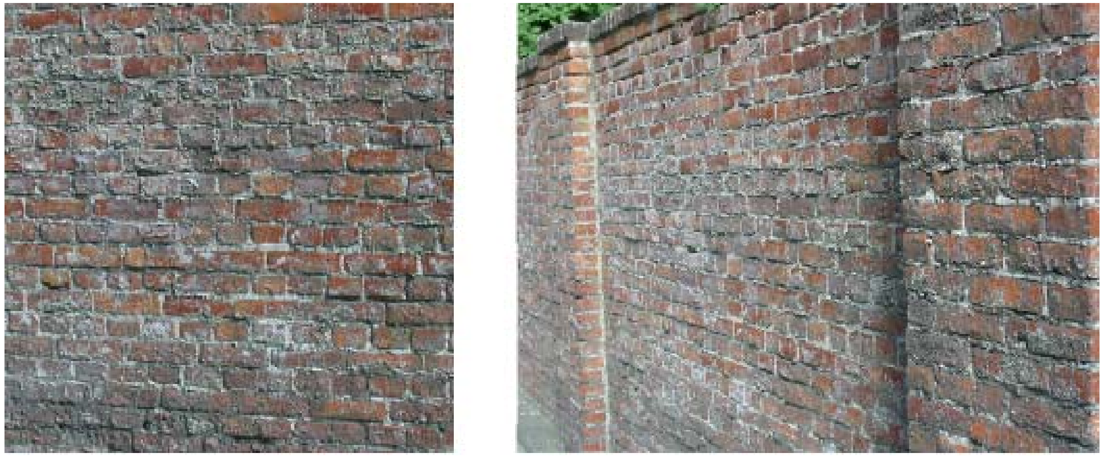
Deep Matching (3)



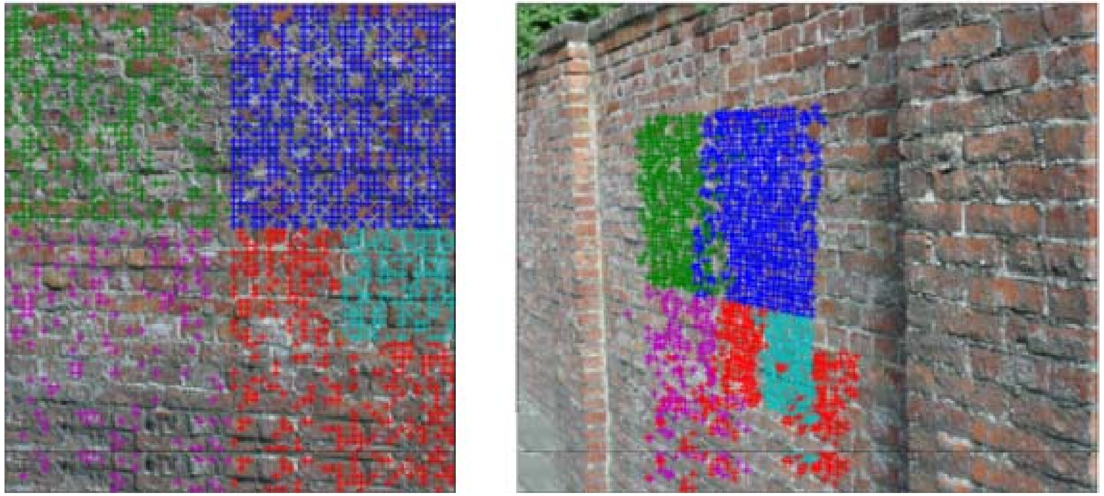


Deep Matching: example results

- Repetitive textures



First image *Second image*

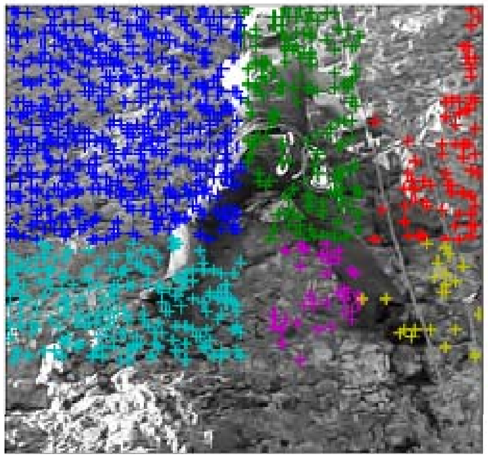
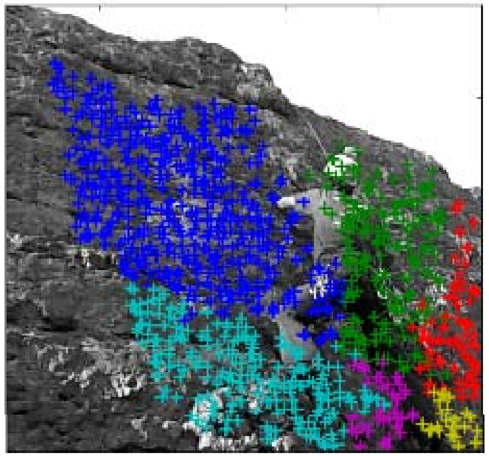


Deep Matching: example results

- Non-rigid deformation

First image *Second image*

DeepFlow

- Classical optical flow [Horn and Schunck 1981]
  - energy $E(\mathbf{w}) = \iint E_{data} + \alpha E_{smooth} \mathbf{dx}$
- Integration of Deep Matching
  - energy $E(\mathbf{w}) = \iint E_{data} + \alpha E_{smooth} + \beta E_{match} \mathbf{dx}$
  - matches guide the flow
  - similar to [Brox and Malik 2011]
- Minimization using:
  - coarse-to-fine strategy
  - fixed point iterations
  - Successive Over Relaxation (SOR)

informatics mathematics

Experimental results: datasets

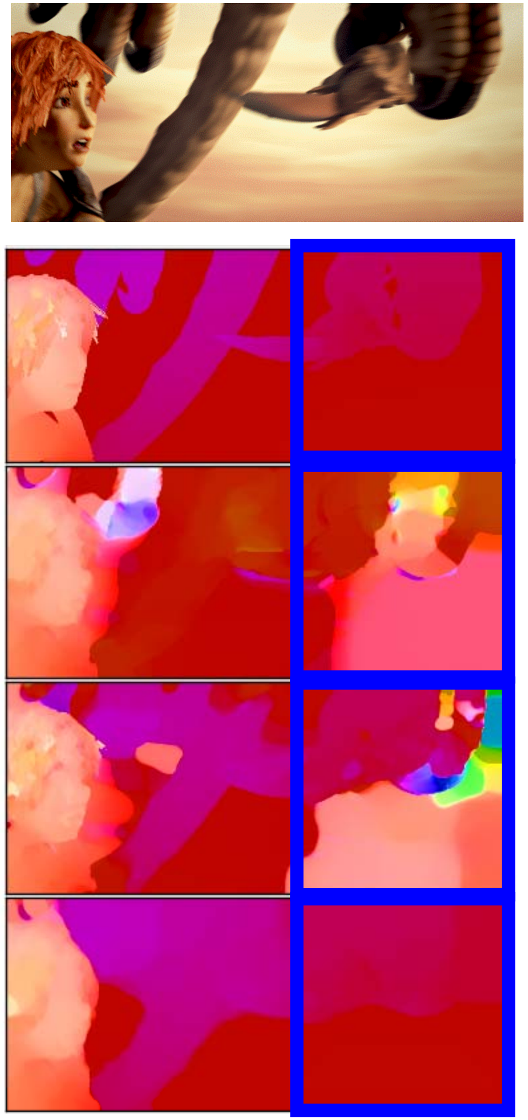
- MPI-Sintel [Butler *et al.* 2012]
  - ► sequences from a realistic animated movie
  - ► large displacements (>20px for 17.5% of pixels)
  - ► atmospheric effects and motion blur

Experimental results: datasets

- KITTI [Geiger *et al.* 2013]
  - ► sequences captured from a driving platform
  - ► large displacements (>20px for 16% of pixels)
  - ► real-world: lightings, surfaces, materials




Experimental results: sample results



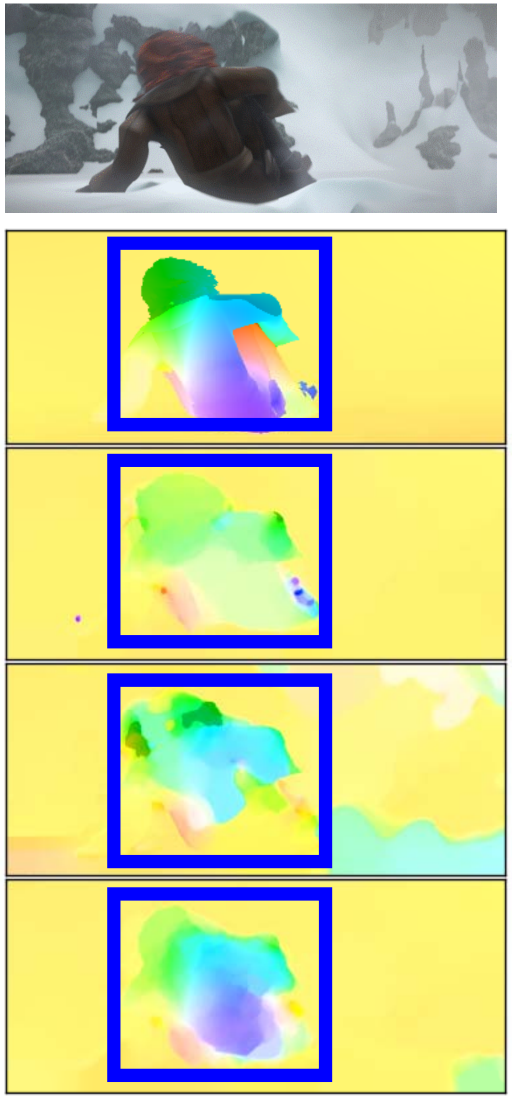
Ground-truth

LDOF [Brox & Malik 2011]

MDP-Flow2 [Xu *et al.* 2012]

DeepFlow

Experimental results: sample results



Ground-truth

LDOF [Brox & Malik 2011]

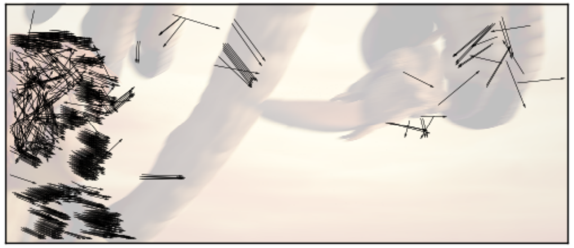
MDP-Flow2 [Xu *et al.* 2012]

DeepFlow

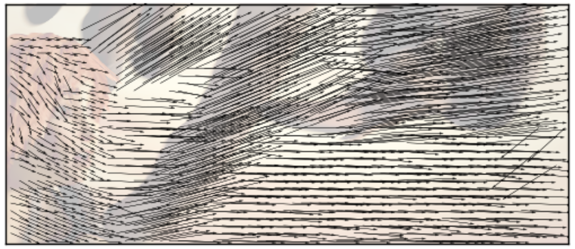
Experimental results: improvements due to Deep Matching

- Comparison on MPI-Sintel training set
  - AEE: average endpoint error
  - s40+: only on large displacements

Matching	Flow evaluation	
	AEE	s40+
No match	5.54	39.86
KLT [OpenCV]	5.51	39.20
SIFT-NN	5.44	38.28
HOG-NN	5.27	37.86
Deep Matching	**4.42**	**29.23**



HOG matching



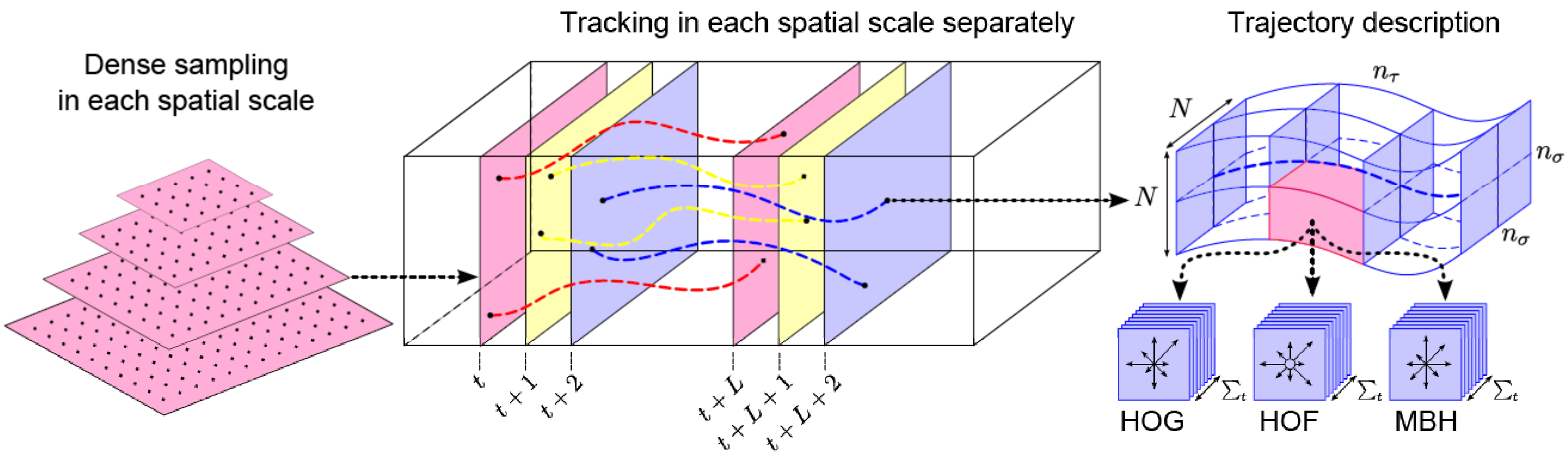
Deep Matching

Overview

- Optical flow
- *Trajectory-based low level features for action recognition*

Dense trajectories [Wang et al. IJCV'13]

- Dense sampling
- Feature tracking based on optical flow
- Trajectory-aligned descriptors



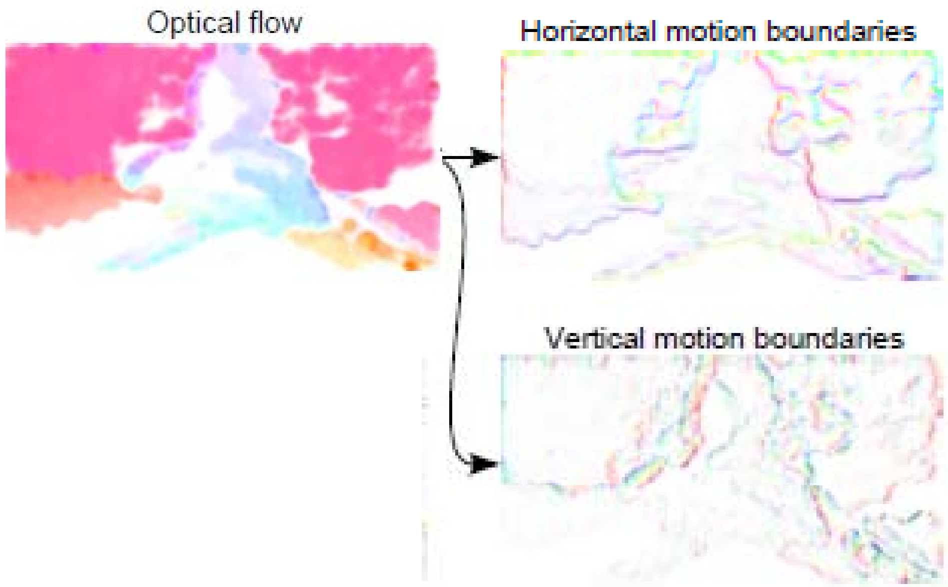
Trajectory descriptors

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



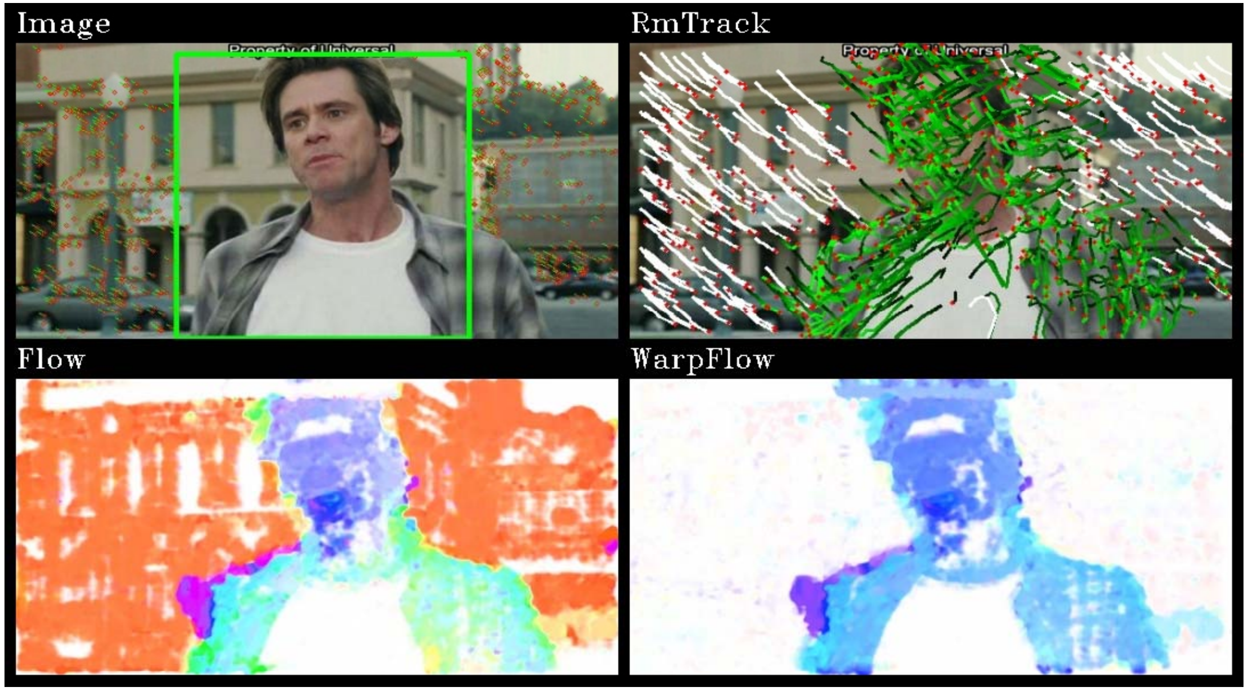
Dense trajectories

- 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, MBH

Improved dense trajectories [Wang et al. ICCV'13]

- Improve dense trajectories by explicit camera motion estimation
- Detect humans to remove outlier matches for homography estimation
- Stabilize optical flow to eliminate camera motion



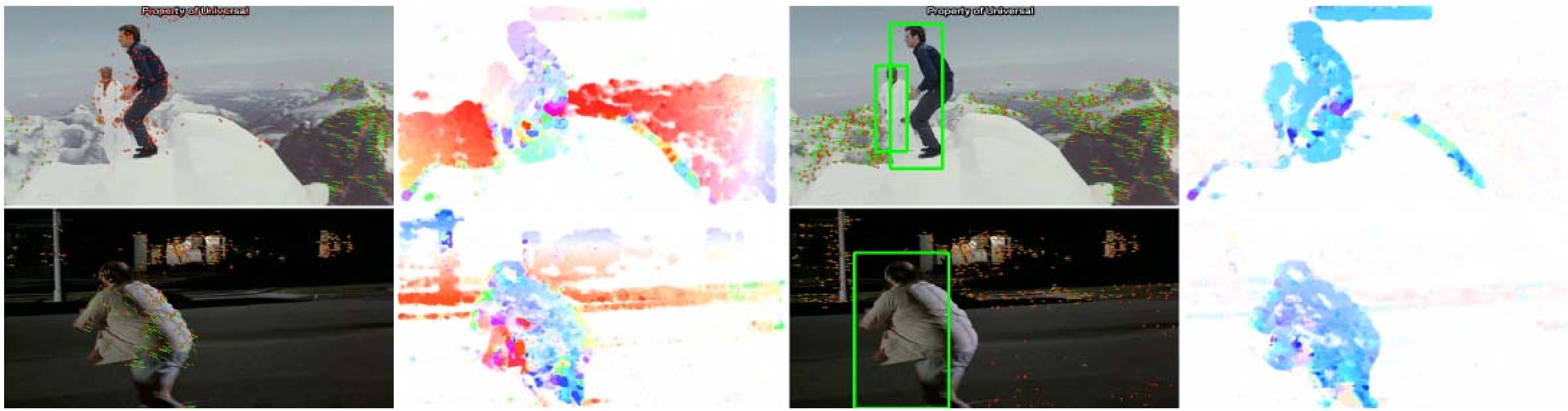
Camera motion estimation

- Find the correspondences between two consecutive frames:
  - Extract and match SURF features (robust to motion blur)
  - Use optical flow, remove uninformative points
- Combine SURF (green) and optical flow (red) results in a more balanced distribution
- Use RANSAC to estimate a homography from all feature matches

Inlier matches of the homography

Remove inconsistent matches due to humans

- Human motion is not constrained by camera motion, thus generates outlier matches
- Apply a human detector in each frame, and track the human bounding box forward and backward to join detections
- Remove feature matches inside the human bounding box during homography estimation

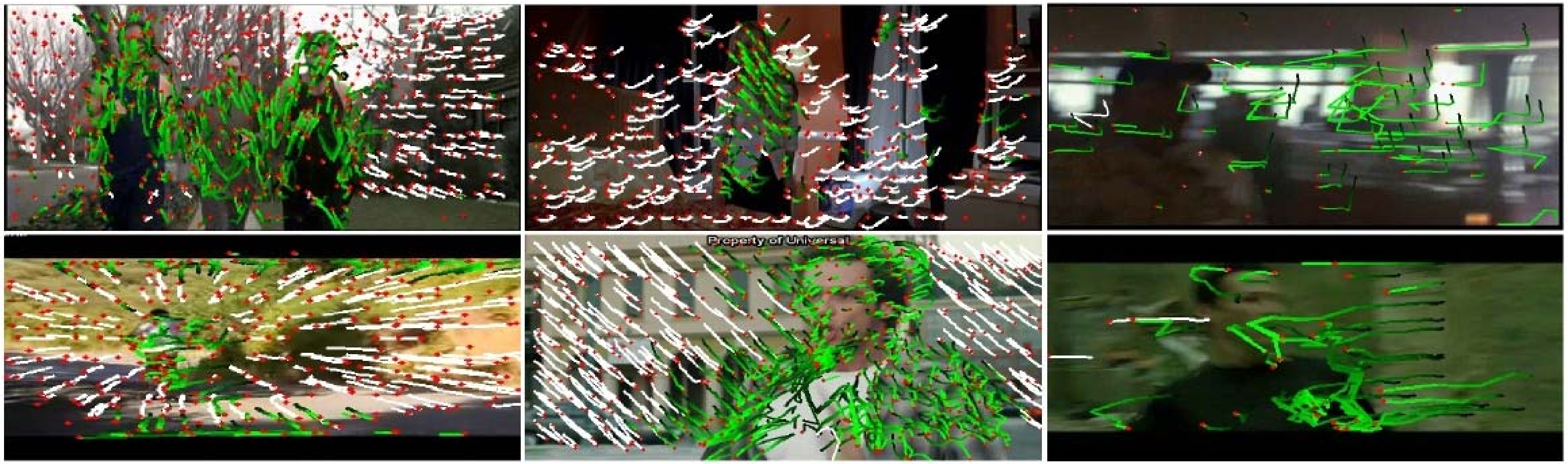


Inlier matches and warped flow, without or with HD

Remove background trajectories

- Remove trajectories by thresholding the maximal magnitude of stabilized motion vectors
- Our method works well under various camera motions, such as pan, zoom, tilt

Successful examples | Failure cases



Removed trajectories (white) and foreground ones (green)

- Failure due to severe motion blur; the homography is not correctly estimated due to unreliable feature matches

Experimental setting

- Motion stabilized trajectories and features (HOG, HOF, MBH)
- "RootSIFT" normalization for each descriptor, then PCA to reduce its dimension by a factor of two
- Use Fisher vector to encode each descriptor separately, set the number of Gaussians to K=256
- Use Power+L2 normalization for FV, and linear SVM with one-against-rest for multi-class classification

Datasets

- Hollywood2: 12 classes from 69 movies, report mAP
- HMDB51: 51 classes, report accuracy on three splits
- Olympic sports: 16 sport actions, report mAP
- UCF50: 50 classes, report accuracy over 25 groups

Evaluation of the intermediate steps

	HOG	HOF	MBH	HOF+MBH	Combined
DTF	38.4%	39.5%	49.1%	49.8%	52.2%
ITF	40.2%	48.9%	52.1%	54.7%	57.2%

Results on HMDB51 using Fisher vector

- Baseline: DTF = "dense trajectory feature"
- ITF = "improved trajectory feature”

- HOF improves significantly and MBH somewhat
- Almost no impact on HOG

- HOF and MBH are complementary, as they represent zero and first order motion information

Impact of feature encoding on improved trajectories

Datasets	Bag of features		Fisher vector	
	DTF	ITF	DTF	ITF
Hollywood2	58.5%	62.2%	60.1%	64.3%
HMDB51	47.2%	52.1%	52.2%	57.2%
Olympic Sport	75.4%	83.3%	84.7%	91.1%
UCF50	84.8%	87.2%	88.6%	91.2%

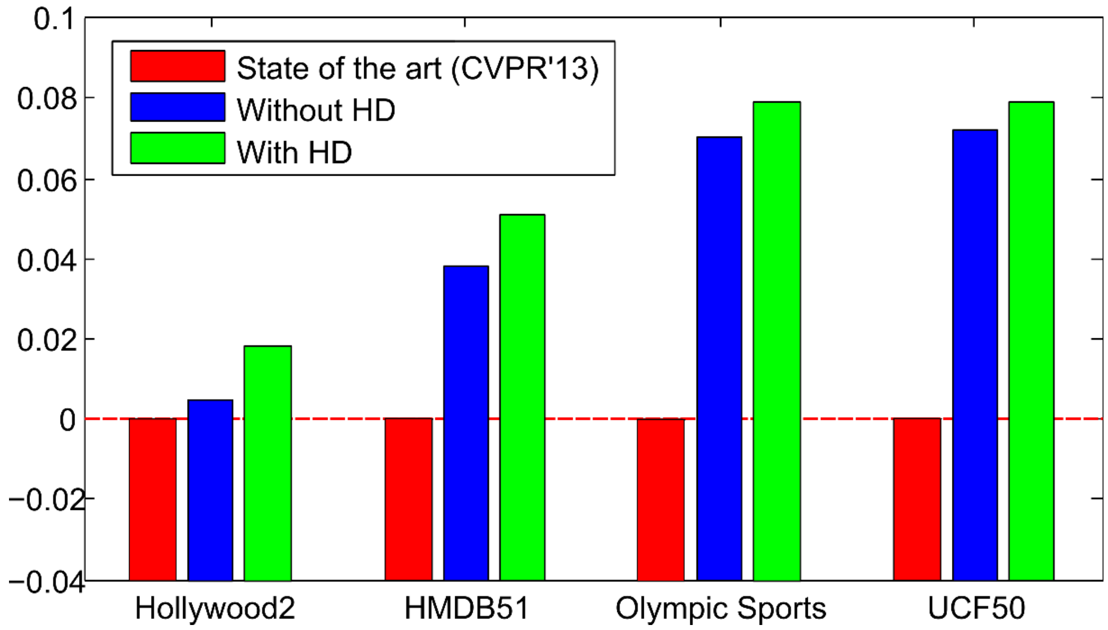
Compare DTF and ITF using different feature encoding

- Standard bag of features: train a codebook of 4000 visual words with k-means for each descriptor type; RBF-χ^2 kernel SVM for classification

- We observe a similar improvement of ITF over DTF when using BOF or FV for feature encoding

- The improvement of FV over BOF varies from 2% to 7% depending on the dataset

Impact of human detection and state of the art



HD = human detection

- Human detection always helps. For Hollywood2 and HMDB51, the difference is more significant, as there are more humans present
- Significantly outperforms the state of the art on all four datasets

Results on TrecVid MED 2013

- 100 positive video clips per event category, 5000 negative video clips
- Testing on 98000 videos clips, i.e., 4000 hours
- 20 known events, 10 adhoc events
- Videos come from publicly available, user-generated content on various Internet sites

- Descriptors: MBH, SIFT, audio, text & speech recognition

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP
Motion	44.65
Static	33.97
Audio	18.15
OCR	10.85
ASR	8.21
Visual=Motion+Static	47.22
Visual+Audio	50.41
Visual+OCR	48.97
Visual+ASR	48.28
Visual+Audio+OCR+ASR	52.28

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party
Motion	44.65	30.7
Static	33.97	25.9
Audio	18.15	33.3
OCR	10.85	10.1
ASR	8.21	3.6
Visual=Motion+Static	47.22	34.8
Visual+Audio	50.41	47.7
Visual+OCR	48.97	35.8
Visual+ASR	48.28	35.0
Visual+Audio+OCR+ASR	52.28	48.4

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party	Repair appliance
Motion	44.65	30.7	42.6
Static	33.97	25.9	43.6
Audio	18.15	33.3	43.3
OCR	10.85	10.1	32.1
ASR	8.21	3.6	39.2
Visual=Motion+Static	47.22	34.8	47.5
Visual+Audio	50.41	47.7	54.5
Visual+OCR	48.97	35.8	50.8
Visual+ASR	48.28	35.0	54.5
Visual+Audio+OCR+ASR	52.28	48.4	57.2

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party	Repair appliance	Make sandwich
Motion	44.65	30.7	42.6	22.5
Static	33.97	25.9	43.6	21.5
Audio	18.15	33.3	43.3	11.2
OCR	10.85	10.1	32.1	19.4
ASR	8.21	3.6	39.2	6.7
Visual=Motion+Static	47.22	34.8	47.5	27.8
Visual+Audio	50.41	47.7	54.5	27.3
Visual+OCR	48.97	35.8	50.8	35.7
Visual+ASR	48.28	35.0	54.5	28.8
Visual+Audio+OCR+ASR	52.28	48.4	57.2	35.4

TrecVid MED 2013 - results

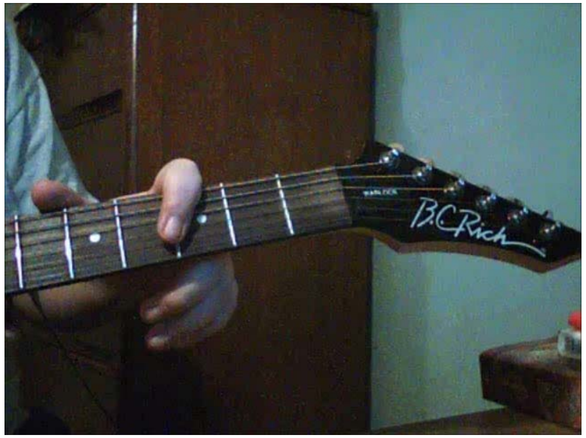
rank 1

rank 2

rank 3

Horse riding competition

TrecVid MED 2013 - results



rank 1

rank 2

rank 3

Tuning a musical instrument

informatics mathematics