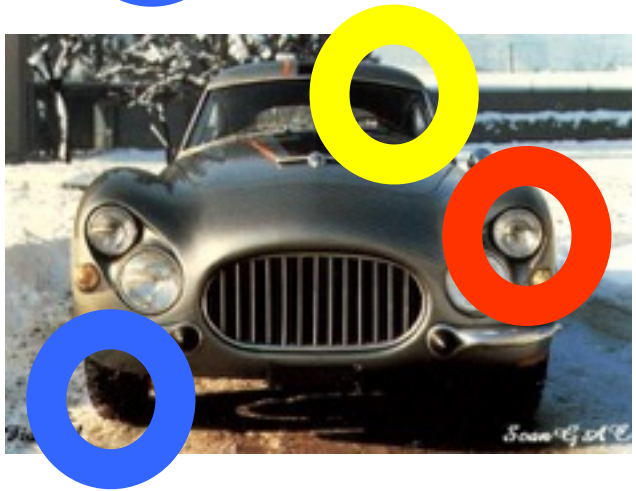
COMPUTER VISION ANALYSIS OF ACTORS AND ACTIONS

Rémi Ronfard - Montpellier – April 2015

Outline

1. Parts and Structure in Computer Vision
2. Visual Recognition of Gestures and Actions
3. Application to Theatre Performances

Parts and Structure in Computer Vision



The correspondence problem

- Model with P parts
- Image with N possible assignments for each part
- Consider mapping to be 1-1

- N^P combinations!!!

The correspondence problem

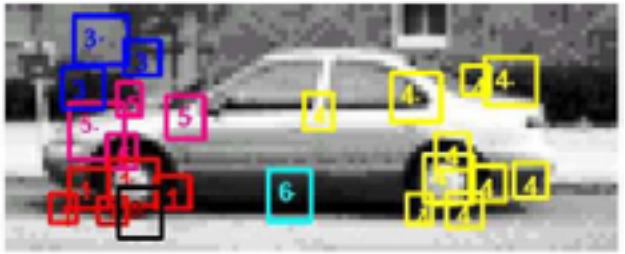
- 1 – 1 mapping
  - Each part assigned to unique feature

As opposed to:

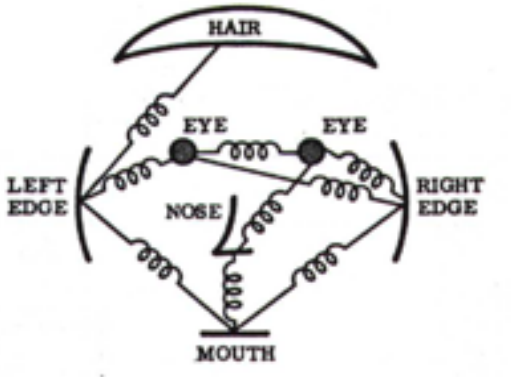
- 1 – Many

Bag of words approaches
- Sudderth, Torralba, Freeman ’05
- Loeff, Sorokin, Arora and Forsyth ‘05

Conditional Random Field
- Quattoni, Collins and Darrell, 04



History of Parts and Structure approaches



- Fischler & Elschlager 1973

- Yuille ‘91
- Brunelli & Poggio ‘93
- Lades, v.d. Malsburg et al. ‘93
- Cootes, Lanitis, Taylor et al. ‘95
- Amit & Geman ‘95, ‘99
- Perona et al. ‘95, ‘96, ’98, ’00, ’03, ‘04, ‘05
- Felzenszwalb & Huttenlocher ’00, ’04, ‘08
- Crandall & Huttenlocher ’05, ’06
- Leibe & Schiele ’03, ’04

- Many papers since 2000

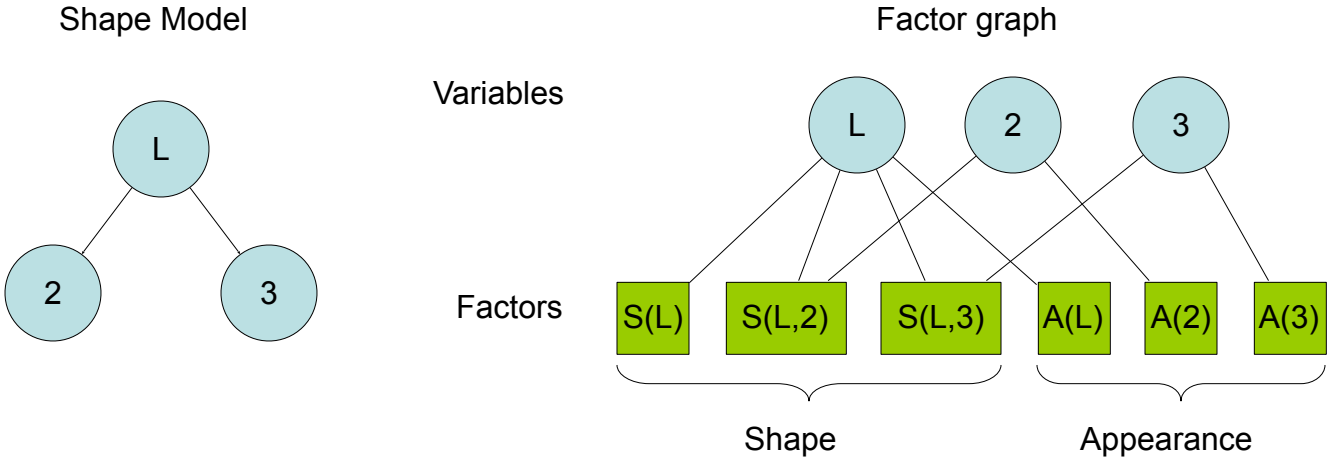
Sparse representation

+ Computationally tractable (10^5 pixels → 10^1 -- 10^2 parts)
+ Generative representation of class
+ Avoid modeling global variability
+ Success in specific object recognition

- Throw away most image information
- Parts need to be distinctive to separate from other classes

Connectivity of parts

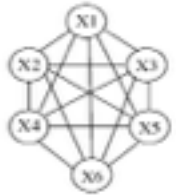
- Complexity is given by size of maximal clique in graph
- Consider a 3 part model
  - Each part has set of N possible locations in image
  - Location of parts 2 & 3 is independent, given location of L
  - Each part has an appearance term, independent between parts.



Different connectivity structures

Fergus et al. ’03
Fei-Fei et al. ‘03

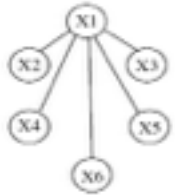
$O(N^6)$



a) Constellation [13]

Crandall et al. ‘05
Fergus et al. ’05

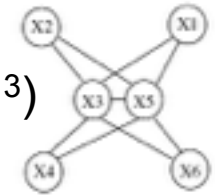
$O(N^2)$



b) Star shape [9, 14]

Crandall et al. ‘05

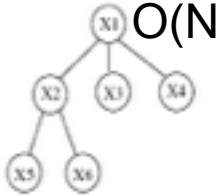
$O(N^3)$



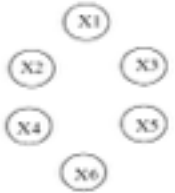
c) k-fan ($k = 2$) [9]

Felzenszwalb &
Huttenlocher ‘00

$O(N^2)$

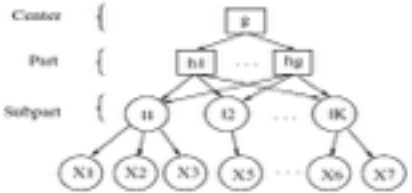


d) Tree [12]



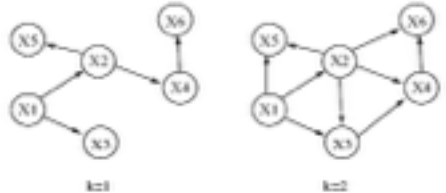
e) Bag of features [10, 21]

Csurka ’04
Vasconcelos ‘00



f) Hierarchy [4]

Bouchard & Triggs ‘05



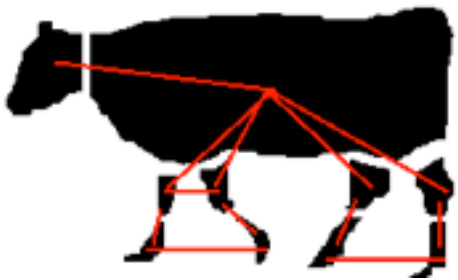
g) Sparse flexible model

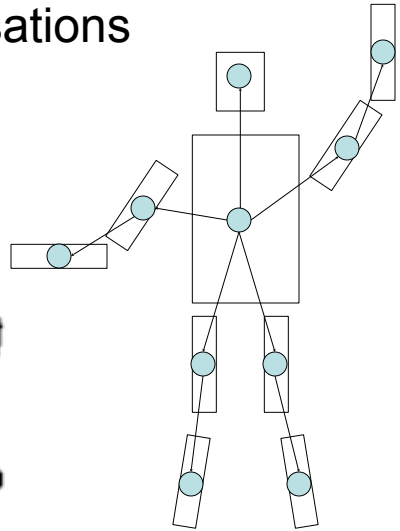
Carneiro & Lowe ‘06

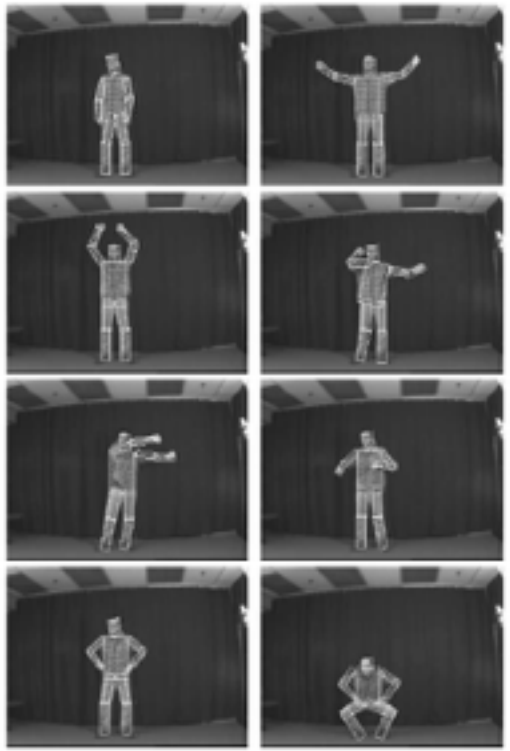
from Sparse Flexible Models of Local Features
Gustavo Carneiro and David Lowe, ECCV 2006

Some class-specific graphs

- Articulated motion
  - People
  - Animals
- Special parameterisations
  - Limb angles



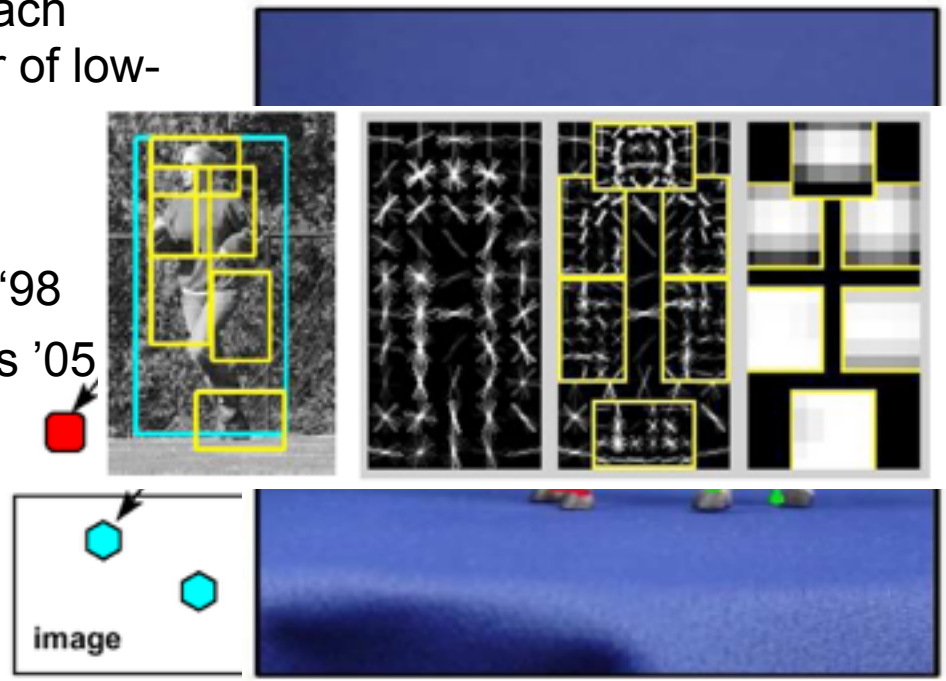




Images from [Kumar, Torr and Zisserman 05, Felzenszwalb & Huttenlocher 05]

Hierarchical representations

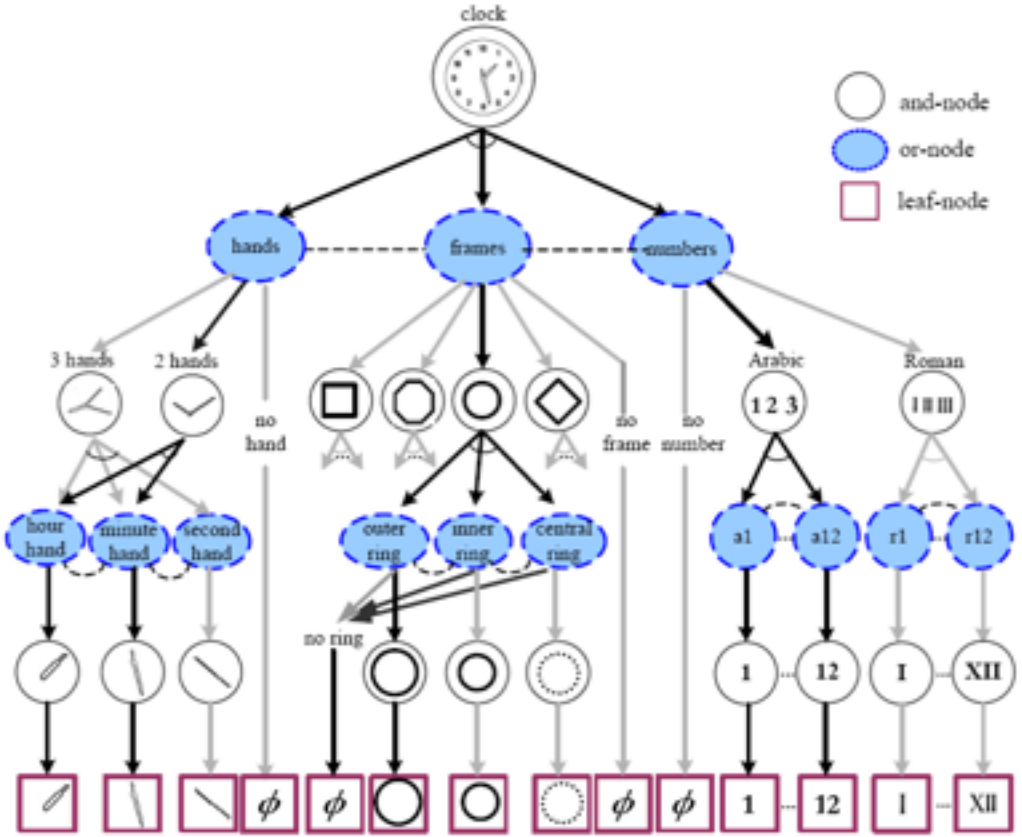
- Pixels → Pixel groupings → Parts → Object
- Multi-scale approach increases number of low-level features
- Amit and Geman ‘98
- Bouchard & Triggs ’05
- Felzenszwalb, McAllester & Ramanan ‘08



Images from [Amit98,Bouchard05,Felzenszwalb’08]

Stochastic Grammar of Images

S.C. Zhu et al. and D. Mumford



How to model location?

- Explicit: Probability density functions
- Implicit: Voting scheme

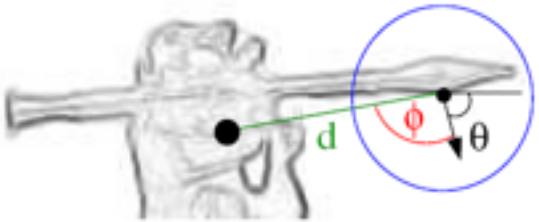
- Invariance
  - Translation
  - Scaling
  - Similarity/affine
  - Viewpoint

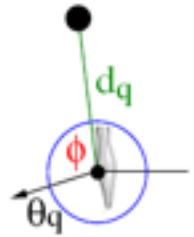
Explicit shape model

- Cartesian
  - E.g. Gaussian distribution
  - Parameters of model, μ and Σ
  - Independence corresponds to zeros in Σ
  - Burl et al. ’96, Weber et al. ‘00, Fergus et al. ’03

$$\mu = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix} \quad \Sigma = \begin{pmatrix} x_1x_1 & x_1x_2 & x_1x_3 & x_1y_1 & x_1y_2 & x_1y_3 \\ x_2x_1 & x_2x_2 & x_2x_3 & x_2y_1 & x_2y_2 & x_2y_3 \\ x_3x_1 & x_3x_2 & x_3x_3 & x_3y_1 & x_3y_2 & x_3y_3 \\ y_1x_1 & y_1x_2 & y_1x_3 & y_1y_1 & y_1y_2 & y_1y_3 \\ y_2x_1 & y_2x_2 & y_2x_3 & y_2y_1 & y_2y_2 & y_2y_3 \\ y_3x_1 & y_3x_2 & y_3x_3 & y_3y_1 & y_3y_2 & y_3y_3 \end{pmatrix}$$

- Polar
  - Convenient for invarianc rotation





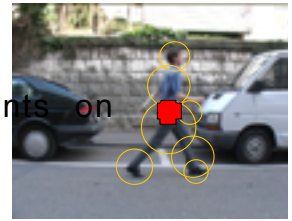
Mikolajczyk et al., CVPR ‘06

Implicit shape model

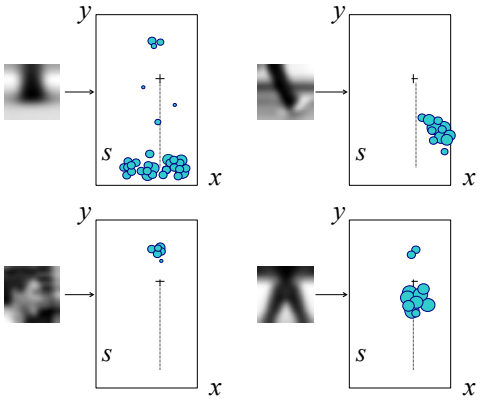
- Use Hough space voting to find object
- Leibe and Schiele ’03,’05

Learning

- Learn appearance codebook
  - Cluster over interest points on training images



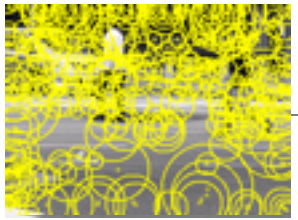
- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object
  - Centroid is given



Spatial occurrence distributions

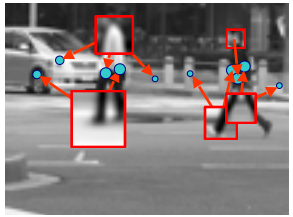
Recognition

Interest Points

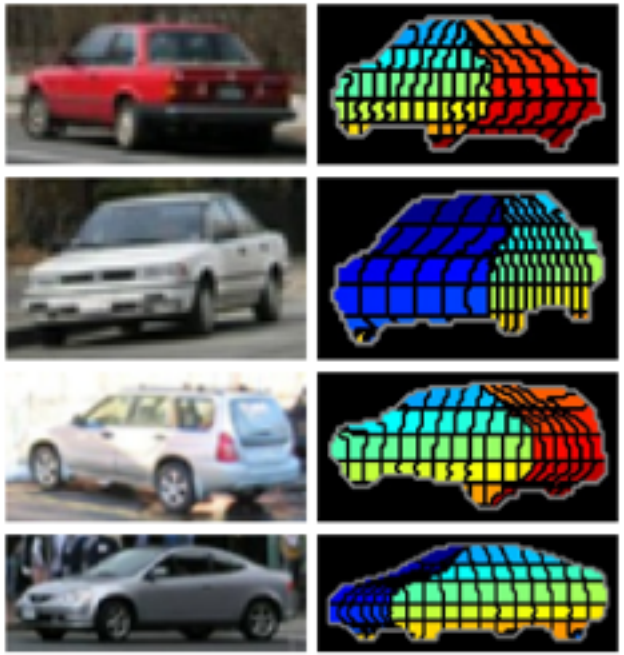


Matched Codebook Entries

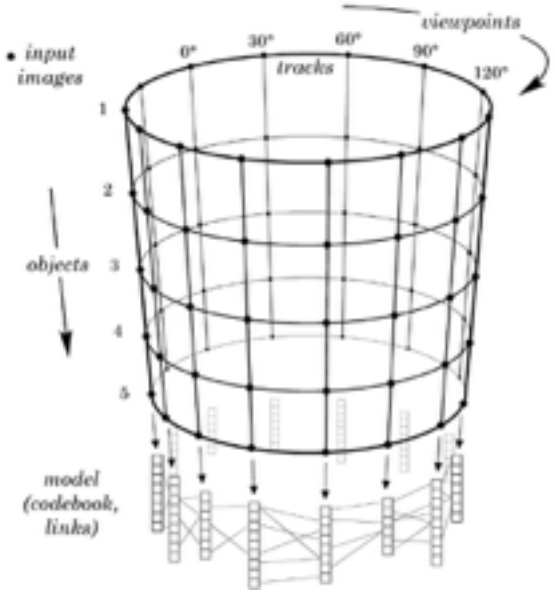
Probabilistic Voting



Multiple view points



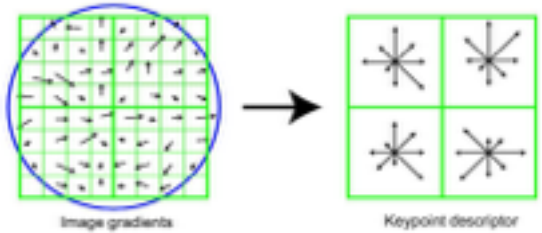
Hoiem, Rother, Winn, 3D LayoutCRF for Multi-View Object Class Recognition and Segmentation, CVPR ‘07



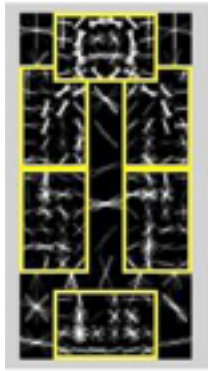
Thomas, Ferrari, Leibe, Tuytelaars, Schiele, and L. Van Gool. Towards Multi-View Object Class Detection, CVPR 06

Appearance representation

- SIFT



- HoG detectors



- Decision trees

[Lepetit and Fua CVPR 2005]

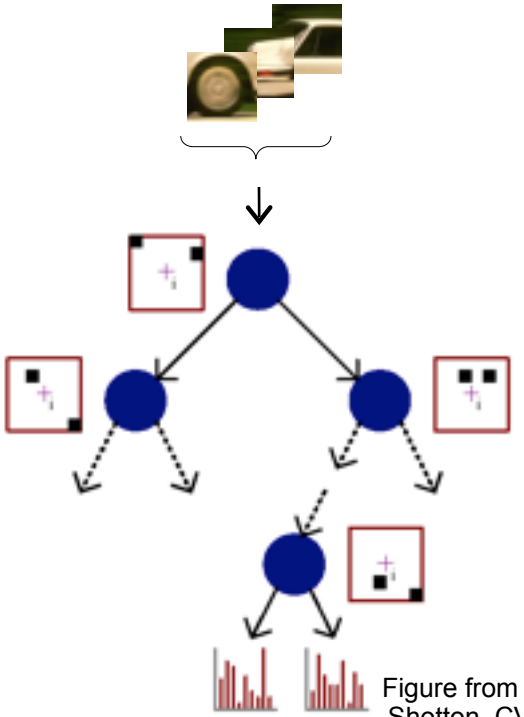
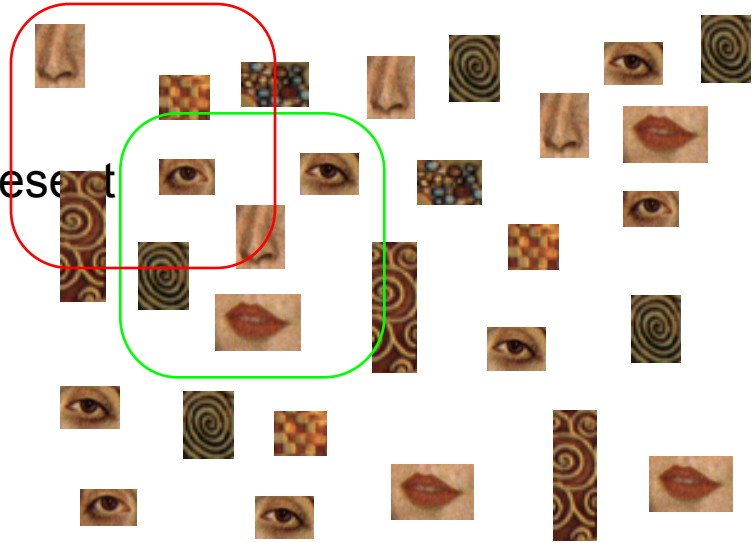


Figure from Winn & Shotton, CVPR '06

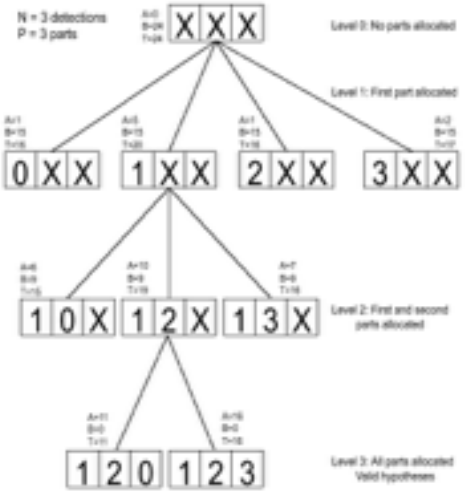
Background clutter

- Explicit model
  - Generative model for clutter as well as foreground object
- Use a sub-window
  - At correct position, no clutter is present



Efficient search methods

- Interpretation tree (Grimson ’87)
  - Condition on assigned parts to give search regions for remaining ones
  - Branch & bound, A*



Parts and Structure Summary

- Correspondence problem
- Efficient methods for large # parts and # positions in image
- Challenge to get representation with desired invariance

Future directions:

- Multiple views
- Approaches to learning
- Multiple category training

Questions ?

Visual Recognition of Gestures and Actions

Rémi Ronfard

INRIA Rhone-Alpes

remi.ronfard@inria.fr

Montpellier, April 2015

Intro

Survey the most useful and promising techniques for real-life applications of face, gesture and full-body action recognition.

Examine how those different classes of methods can be adapted to achieve invariance with respect to viewing directions and performing styles.

OUTLINE

- History and applications
  - Gesture recognition - a machine learning problem
  - What is a gesture ? What does it look like ?
- Spatial structure of gestures and actions
  - Face, hands and feet, full body
  - Actor invariance
  - View invariance
- Temporal structure of gestures and actions
  - Templates, grammars and histograms
  - Tracking, segmentation and recognition

History and applications

- What is a gesture ?
- How can the computer use gesture ?
- How can the computer recognize gesture ?

Different classes of gestures

- Gestures are movements of the human body
  - With a communicative goal
  - Or part of goal-directed action

- Action changes the world
- Gesture communicates

- Visual analysis of gesture and action
  - Detect and track human body parts
  - Recognize their gestures and actions

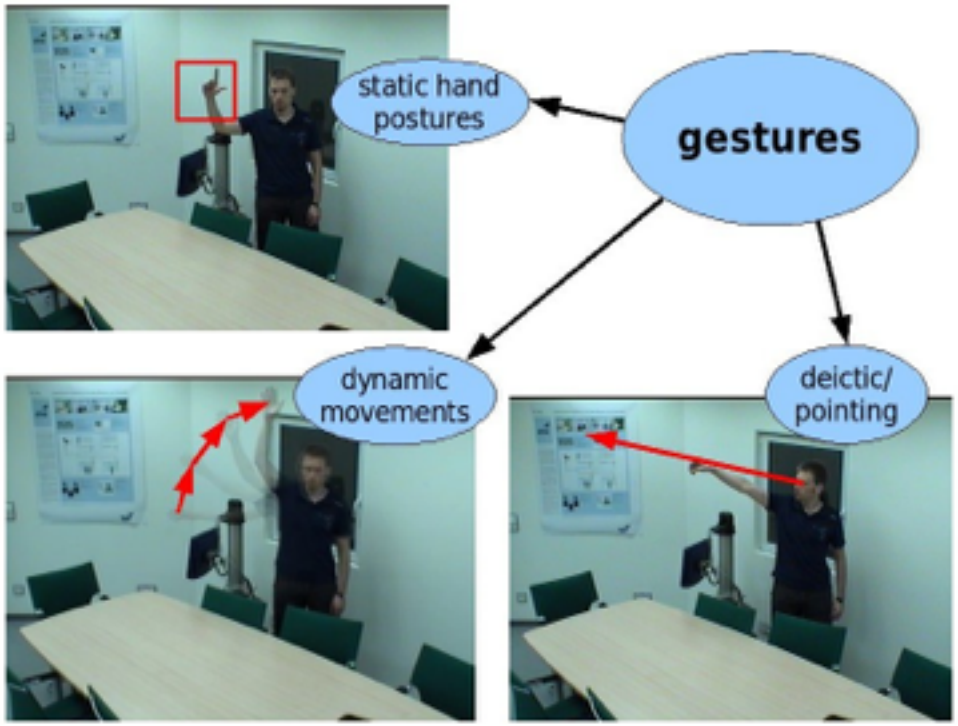
Different classes of Gestures

- Gestures as poses
  - Counting with fingers
  - Making faces
  - Pointing
- Gestures as movements
  - Yes, no
  - Come here, go away
  - From here to there
- Single body part
  - Hand gestures
  - Facial expressions
- Multiple body parts
  - Hand gestures
  - Clapping hands
  - Scratching head
  - Blowing a kiss
  - Are you crazy ?

Examples

- Source: Real-time Detection and Interpretation of 3D Deictic Gestures

- J. Richarz, T. Plötz and G. A. Fink, ICPR 2008.



Gesture Measurement

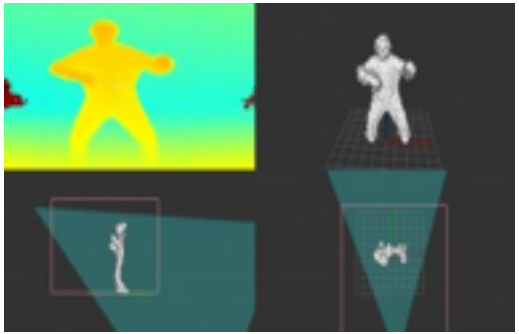
- Visual measurements
  - Markers
  - Local features
  - Body parts
  - Shape
  - Color and texture
  - Motion
- From one or more cameras (binocular, trinocular and multiview stereo)
- Physical measurements
  - Accelerometers
  - Gravitometers
  - RFID tags
  - Laser scanners
  - Radars, Lidars
  - Time-Of-Flight (TOF) cameras

Applications –reading sign language

- Source: Buehler et al. Oxford University, 2010

Gesture-based remote controls

- First proposed by Bill Freeman
- Now using Time-of-Flight Camera, allowing people to turn up the volume by moving their hand in a circle, switch the channel by swiping to the right, pause by extending their hands in a "stop" gesture, and so on.



- Source: Softkinetic-Optrima

Teaching robots by demonstration

- Gesture recognition and imitation
- Visual gesture to motor commands

Minority Report gesture Interface

- John Underkoffler helped Steven Spielberg design a futuristic human computer interface for a movie that takes place in the year 2054.

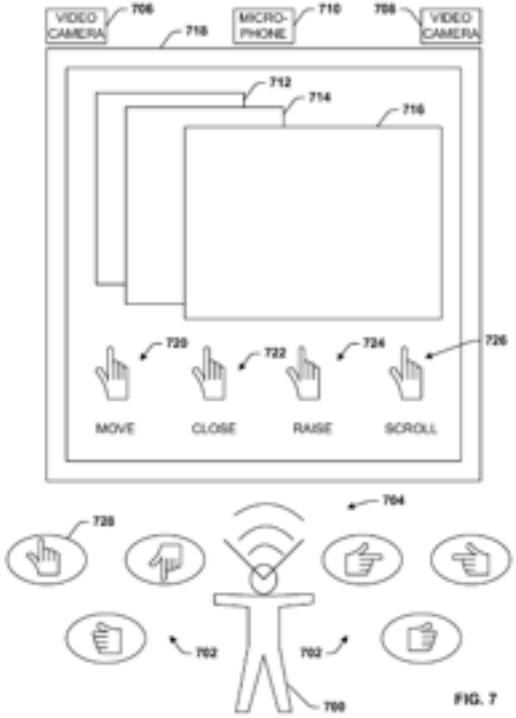
- Now introducing G-Speak by Oblong

Gesture generation by imitation

- Gesture modeling and animation, ACM TOG 2008
- Michael Neff, U. California, Davis
- Michael Kipp, DFKI
- Irene Albrech and Hans-Peter Seidel, MPI Informatik

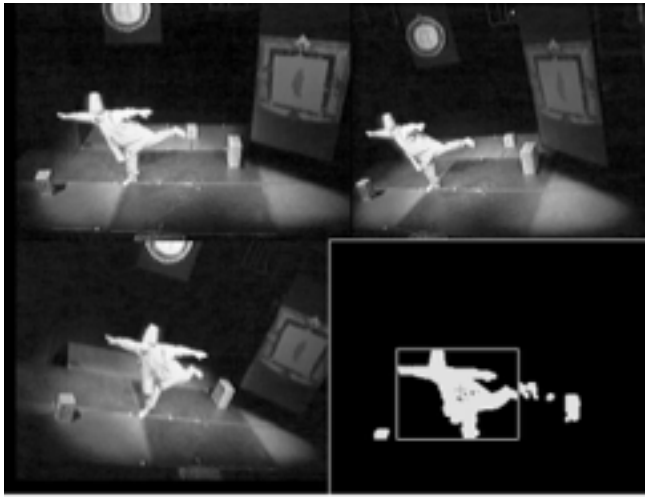
Other sensors and applications

- Gesture-Based Interface to Windows by **Andrew Wilson and Nuria Oliver, Microsoft Research**
- Microsoft Kinect (NATAL)



Binocular and trinocular stereo

PANASONIC 3D CAMERA



COMPUTER
THEATRE at MIT

Speech vs gesture

- Phonemes
- Phonetics
  - Mel-Cepstrum Coefficients
  - Speaker invariance
- Language models
  - Phonemes to words
  - Word to word statistics

- Visemes and Kinemes
  - Facial Action Units
  - Body Action Units
- Visual representations
  - Actor invariance
  - View-invariance
  - Temporal resolution
- Multiple contexts
  - Gesture lexicons
  - Gesture profiles

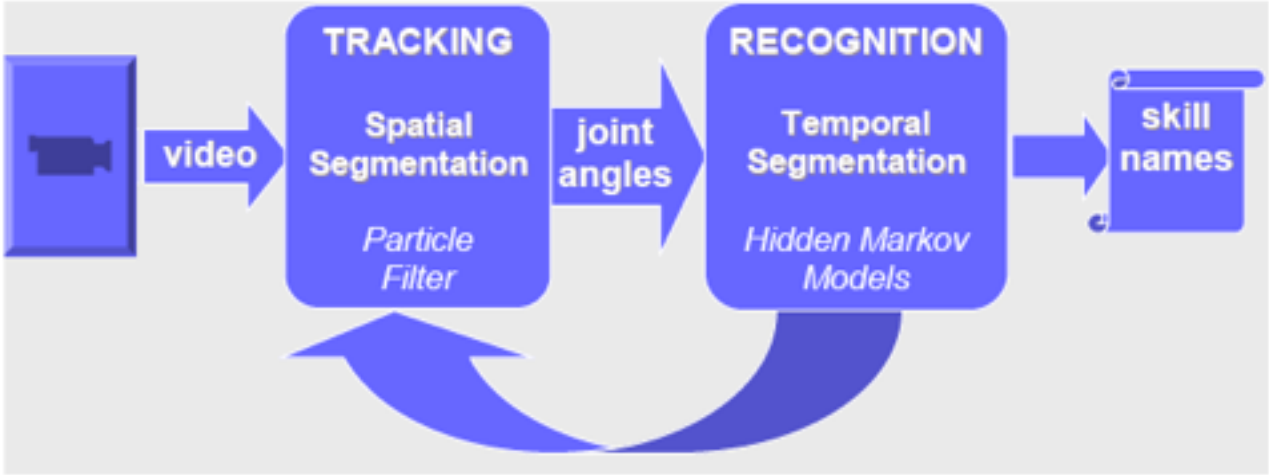
gesture in dialogue (mcneil, 1992)

- Communicative
  - EMBLEM
    - Yes, no, thumbs up
  - DEICTIC
    - pointing
  - ICONIC/METAPHORIC
    - Representing
  - BEAT
    - Stress and rythm
- Non-communicative
  - ADAPTOR
    - TOUCH SELF
    - TOUCH OBJECT
    - TOUCH OTHER
  - GOAL-DIRECTED
    - Raise arm to reach object
    - Scratch head to ease pain
    - Chase a mosquito
  - UNVOLONTARY
    - Embarassment

Body language

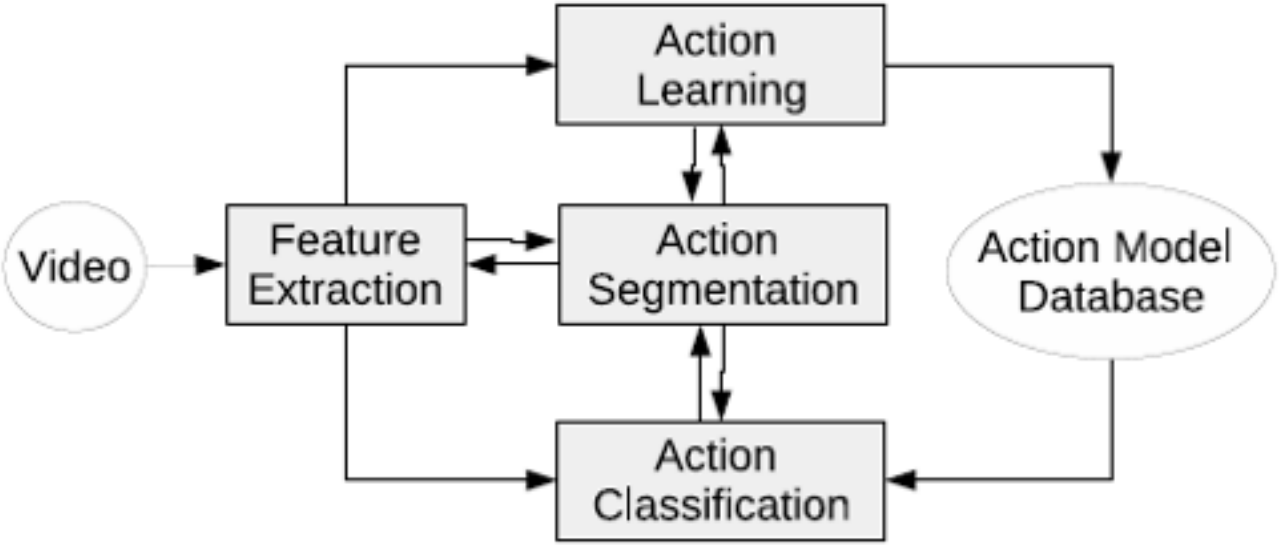
- **Hand and Mind. What Gestures Reveal about thought,** David McNeill, 1992.
- **Gesture: Visible Action as Utterance,** Adam Kendon, 2004.
- **Gesture and Thought,** David McNeill, 2007.

Tracking and recognition

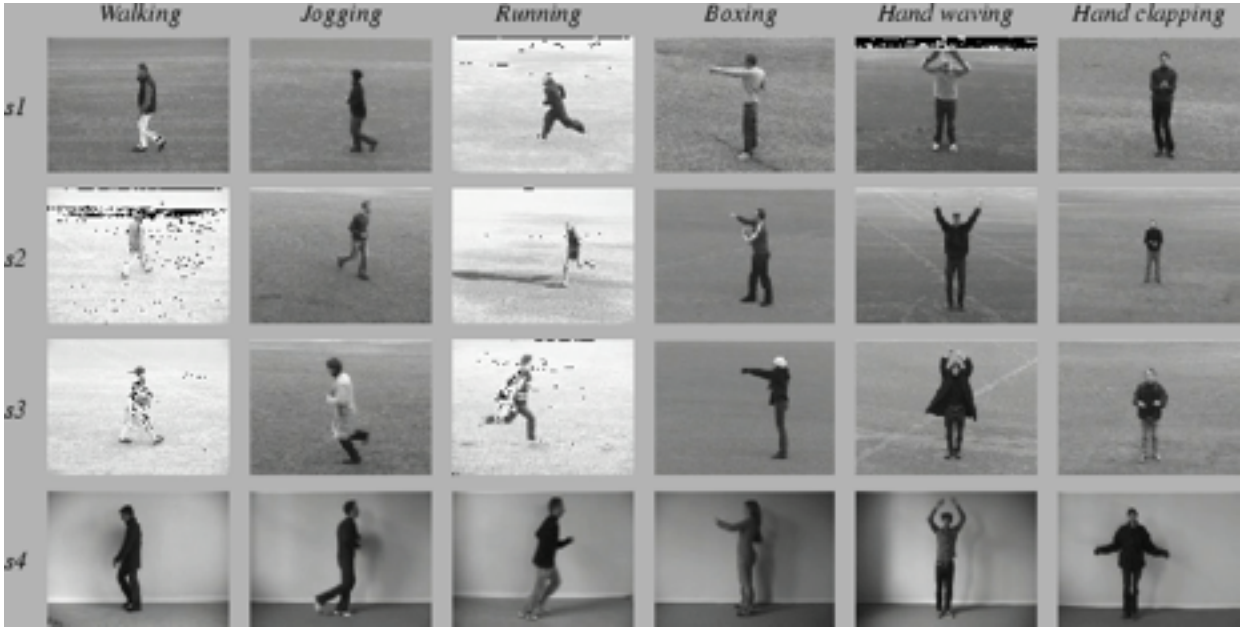


- Source: Green and Guang, Quantifying and recognizing human movement patterns form monocular video images, I*EEE Trans. on Circuits and Systems for Video Technology*.

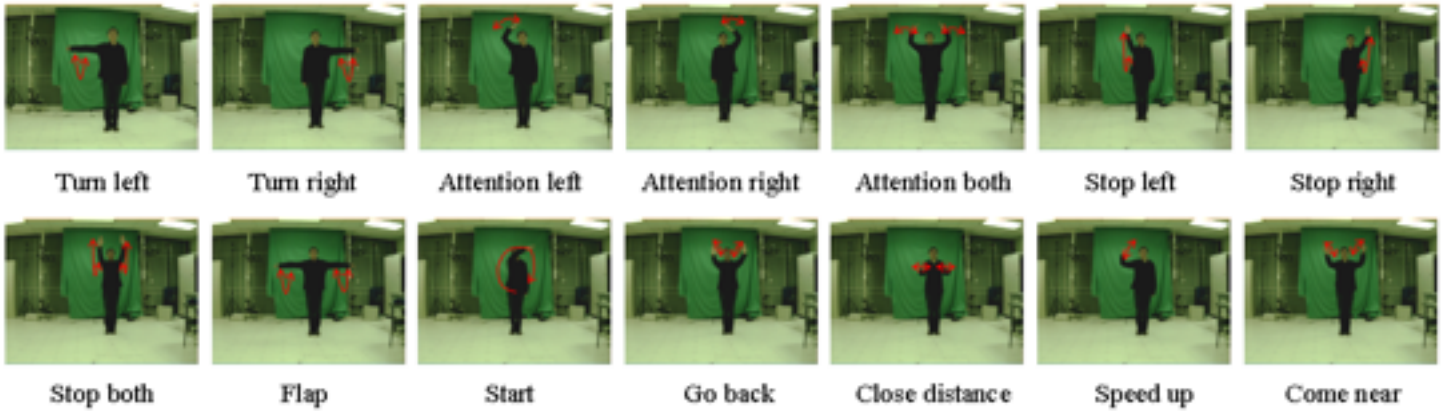
Gesture and action recognition



KTH FULL BODY GESTURE DATASET



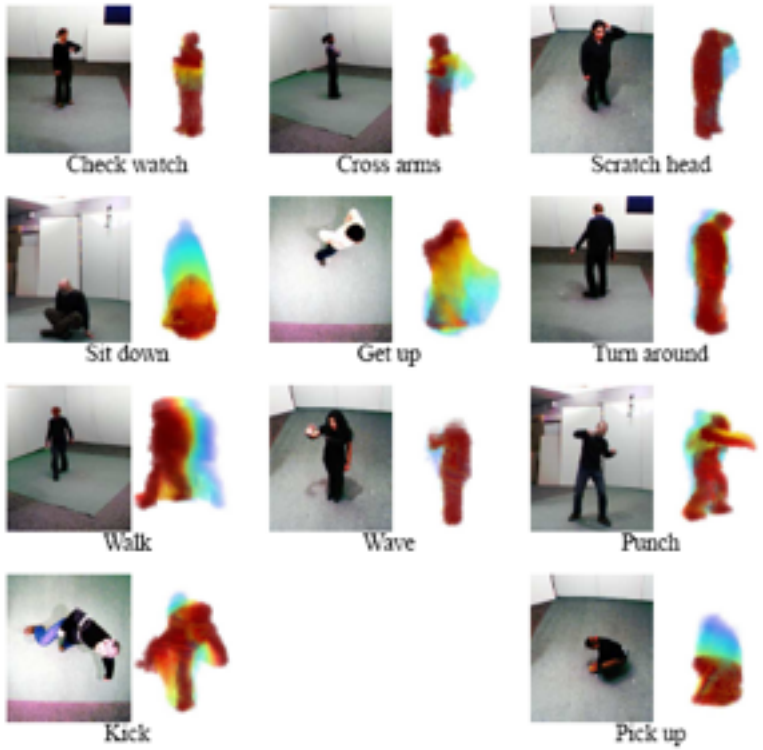
KECK DATASET (Univ MARYLAND)



Source: Lin, Jiang, Davis, Recognizing Actions by Shape-Motion Prototype Trees, ICCV, 2009.

INRIA IXMAS Full-body gesture dataset

- 11 actions by 10 actors
- 3different poses
- 5 cameras



Different classes of problems

- Recognize gesture, given body part trajectories (easiest, but body part tracking is hard)
  - Faces
  - Hands
  - Feet
- Recognize gesture, given human body trajectory (human tracking still hard)
- Recognize gesture, given image sequence (hardest)

Different classes of problems

- Single actor
  - Train on actor
  - Test on same actor
- Single view
  - Train in one view
  - Test on same view
- Multiple actors
  - Train on one actor
  - Test on other actor
- Multiple views
  - Train in one view
  - Test on other view

Machine learning approaches

- TRAINING
  - Define gestures from training examples
  - One or more actors
  - One or more views
  - Given body parts
  - Given human tracking
  - Given sequence
- RECOGNITION
  - Recognize gestures from test examples
  - Same or different actors
  - Same or different views
  - Same modalities as in training
- EVALUATION
  - Precision and recall
  - Generalization

The recognition problem

- Task - Given a test video clip, does it contain gesture or action ?

1. Learn models with training set of labeled examples.
2. Evaluate precision and recall with testing set of labeled examples.

- Many approaches for encoding spatial and temporal structure of action

- Variations

1. Given a test detection box for an actor, is actor performing gesture or action ?
2. Given a test detection boxes for a body part, is body part performing gesture or action?

The segmentation problem

- Given video clip of gesture or action, where and when is it happening ?
- Given video clip, how many different gestures and actions ? Where and when are they happening ?
- Fewer references
  - Motion maxima (Marr and Vaina)
  - Sliding window and Dynamic Time-Warping
  - Semi-Markov Models (HMMs and CRFs) provide a segmentation as a result of recognition

The viewpoint recognition problem

- Recognize gestures and actions from different viewpoints
  - How many training viewpoints ?
  - How many testing viewpoints ?
- Few approches
  - Exhaustive search
  - View-invariant features (Perez)
  - Transformed Grammars (Frey & Jojic)
- Few databases – CMU, IXMAS

Spatiotemporal features

- Image histograms
  - Space Bag of Words
  - Color and texture (SIFT)
  - Oriented Gradient (HOG)
- Image templates
  - Face
  - Hand
  - Body
- Recognition by parts
  - Pictorial structures

- Temporal histograms
  - Time Bag of Words
  - Oriented Flow (HOF)
  - Spatio-temporal Interest Points (STIPS)
- Temporal templates
  - Motion History Images
- Recognition by parts
  - Markov States and Transitions (HMM)
  - Grammars

Taxonomy of learning methods

Space / Time	*Body model*	*Image model*	Spatial Bag-of-words
Template	***Body template***	***Image template***	***Bag of trajectories***
Grammar	***Body grammar***	***Image grammar***	***Bag grammar***
Temporal Bag-of-words	***Bag of body parts***	***Bag of keyframes***	***Bag of events***

Spatial structure of gesture

- **Body coordinates = Body models**
- **Image coordinates = Image models**
- **Unstructured space = Histograms**
  - **spatial bag-of-words (BOW)**

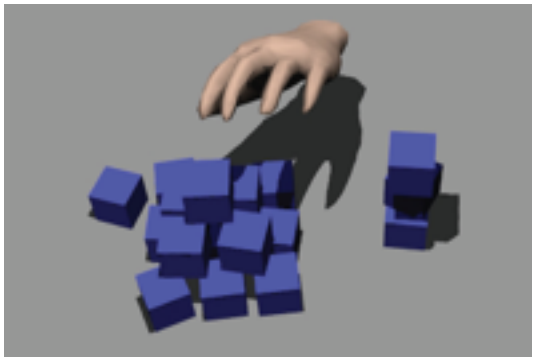
Body models

- Detection and/or tracking of body parts
- Labeling of body parts (and actors)
- Movement and pose of body parts

- Pros: gestures are naturally defined in terms of body parts
- Cons: recognition cannot recover from tracking or detection errors

Hand tracking

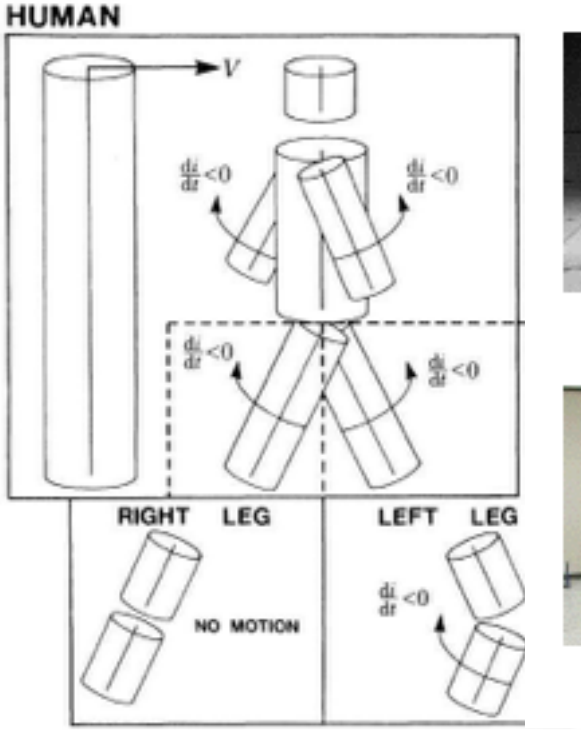
- Color-to-Finger indexing
- Wang and Popovic, MIT Media Lab, 2010



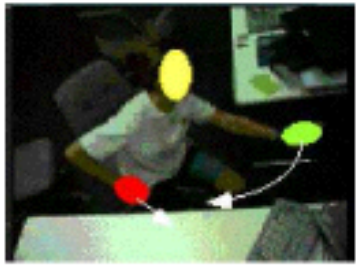
Upper body pose estimation

- Source: Buehler, Oxford Univ. 2010

Full body gesture

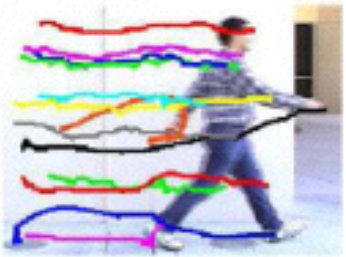


(b)



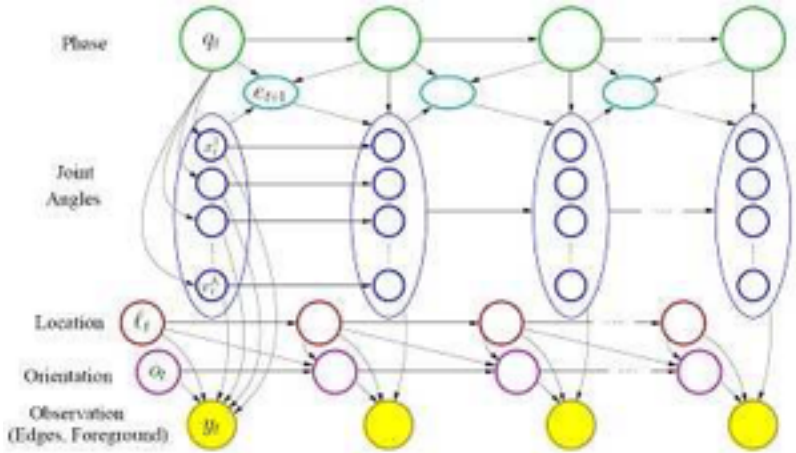
(d)

(c)



(e)

Full body gesture



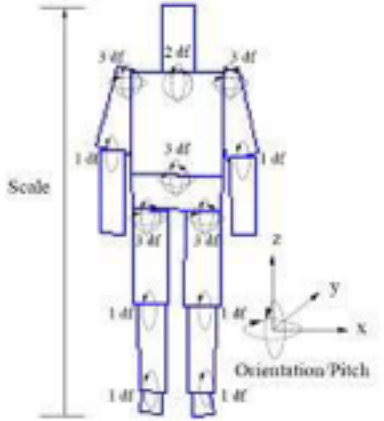


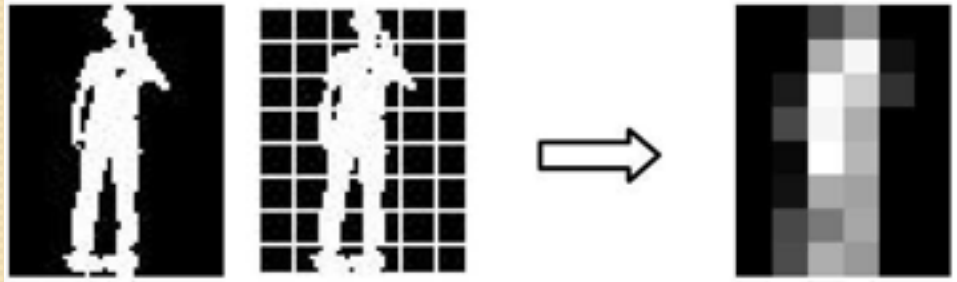
Image models

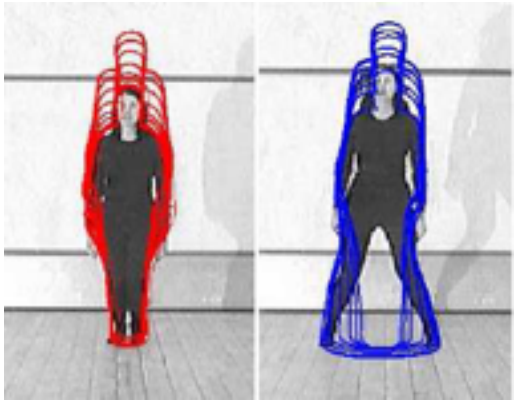
- Do not assume body parts can be recognized
- Learn global model of image appearance and motion
- Pros: Gestures can be learned from unsegmented examples; no body part labeling necessary; fast.
- Cons: Requires multiple models for different viewpoints, distances and body sizes.

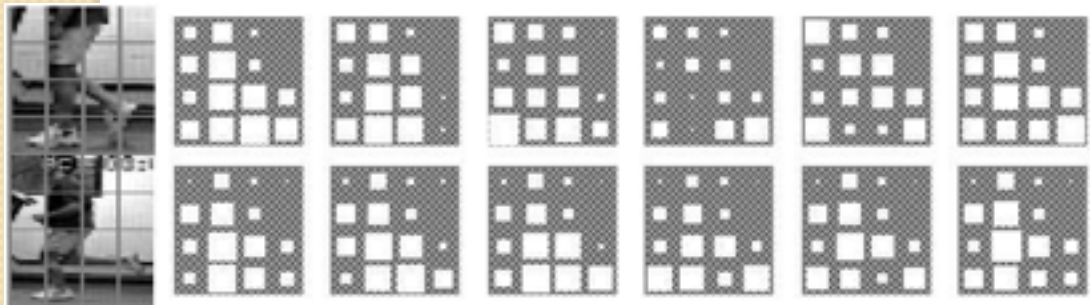
Image Descriptors

- Silhouettes
  - Background-subtraction
  - Image difference
- Colors and textures
  - SIFT images (dim=128)
- Oriented Gradients and Flows
  - Dense optical flow
  - Image of HOG-HOF coefficients (dim=16)
- Self-similarity

Image models for full body gesture

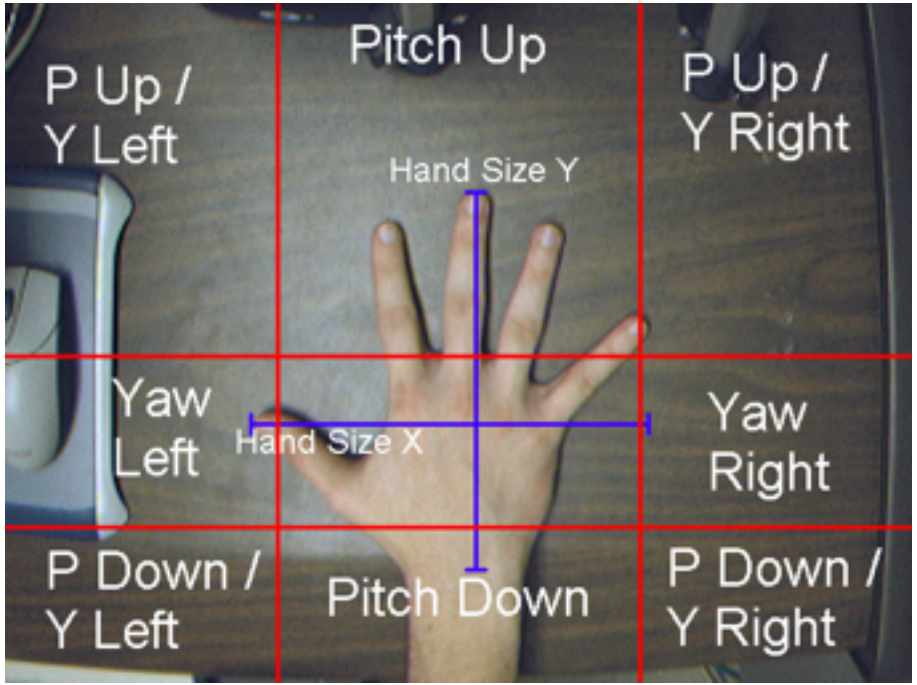






Recognizing hand movements

- A Method for Temporal Hand Gesture Recognition
- Joshua R. New
- Knowledge Systems Laboratory
- Jacksonville State University

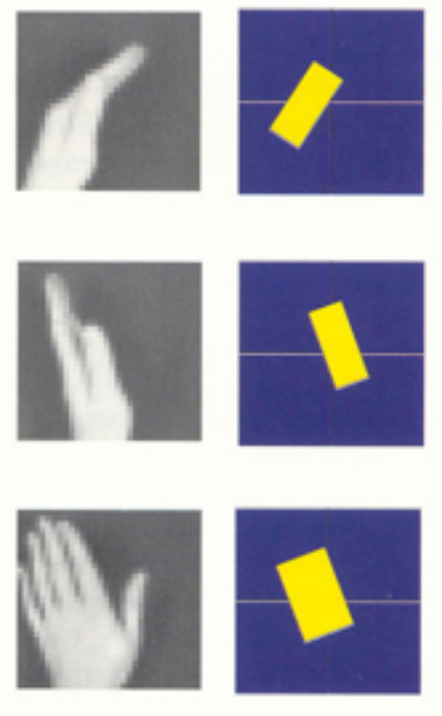


Hand gestures with image moments

$$M_{00} = \sum_x \sum_y I(x, y) \qquad M_{11} = \sum_x \sum_y xyI(x, y)$$

$$M_{10} = \sum_x \sum_y xI(x, y) \qquad M_{01} = \sum_x \sum_y yI(x, y)$$

$$M_{20} = \sum_x \sum_y x^2 I(x, y) \qquad M_{02} = \sum_x \sum_y y^2 I(x, y)$$

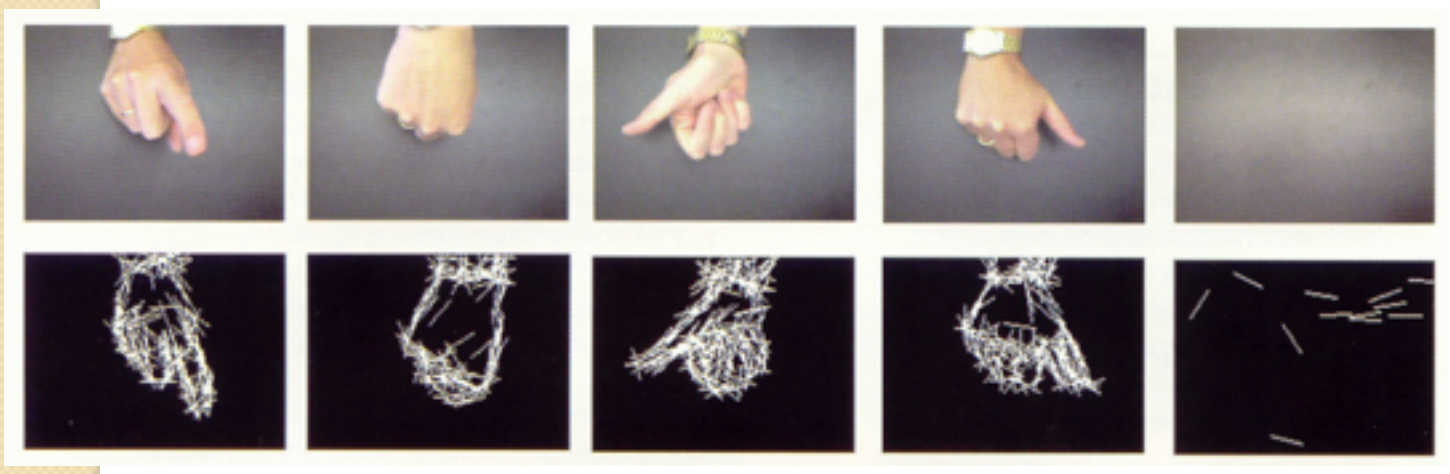


- Source: Freeman et al. Computer Vision for Interactive Computer Graphics. IEEE CGA, 1998.

Orientation histograms

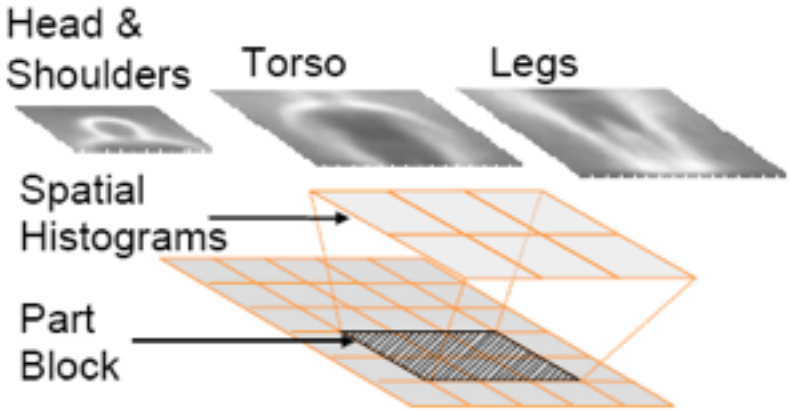
$$\sum_i \left(\Phi_1(i) - \Phi_2(i)\right)^2$$

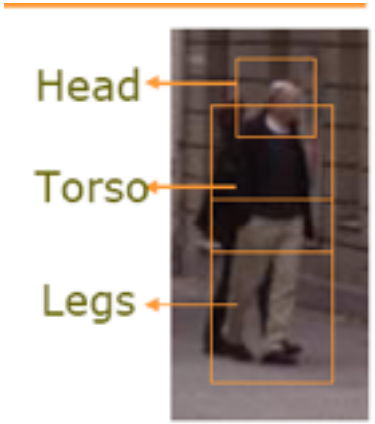
$$\Phi(i) = \sum_{x,y} \begin{cases} 1 \text{ if } \left|\theta(x,y) - \frac{360^\circ}{N} i\right| < \frac{360^\circ}{N} \\ 0 \text{ otherwise} \end{cases}$$



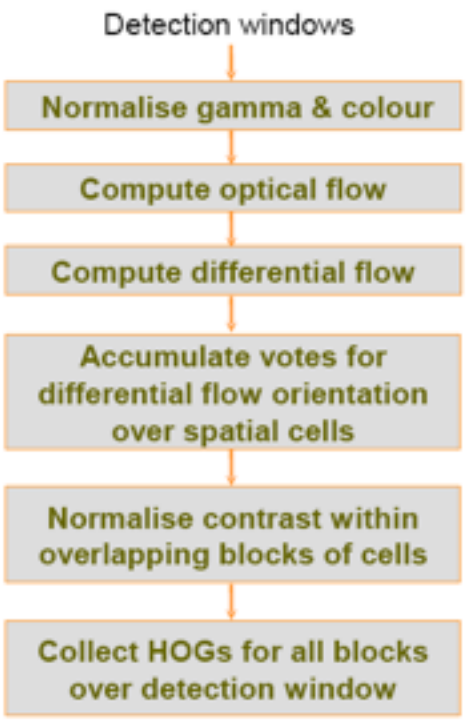
Multiple body parts

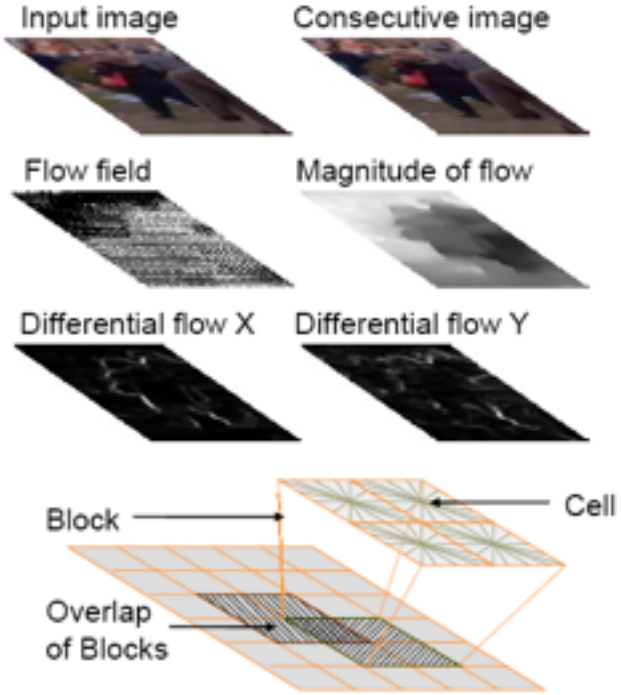
- Build templates of Gradient Orientations (HOG)
- Source: Dalal & Triggs, CVPR 2005





Hog-Hof templates





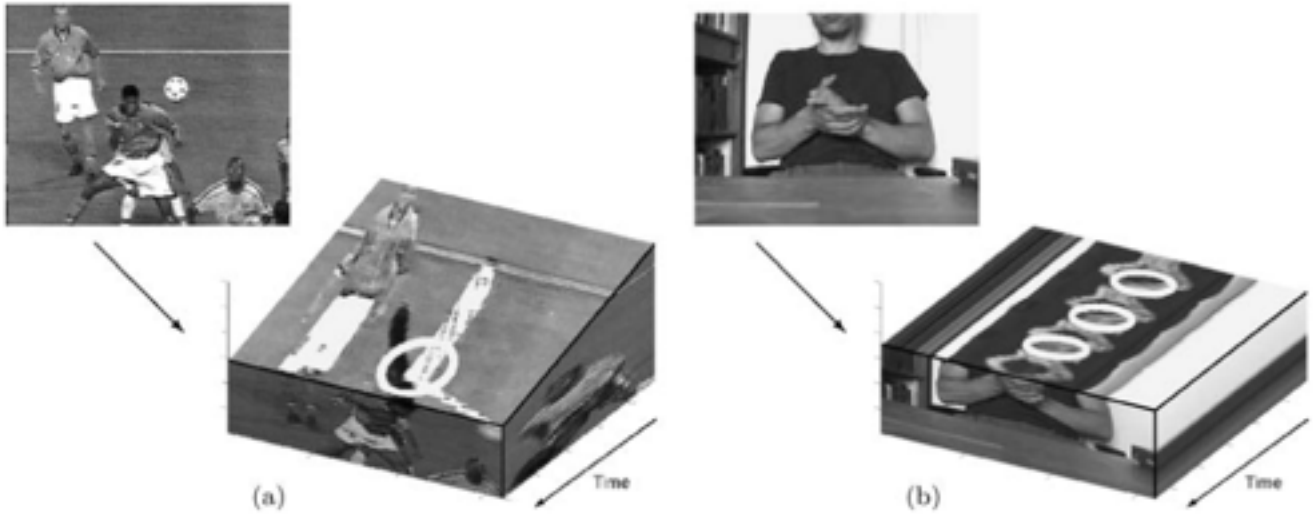
Source: Dalal, 2005

Spatial « Bag-of-Word » models

- Discretize image descriptors into a finite lexicon of « visual words »
- Compute frequency of visual words in entire image
- Object recognition: SIFT
  - 128-dimensional descriptor
  - At « interest points » or regular grid
- Gesture recognition: HOG-HOF
  - 32-dimensional descriptor
  - At « spatio-temporal interest points » or regular grid

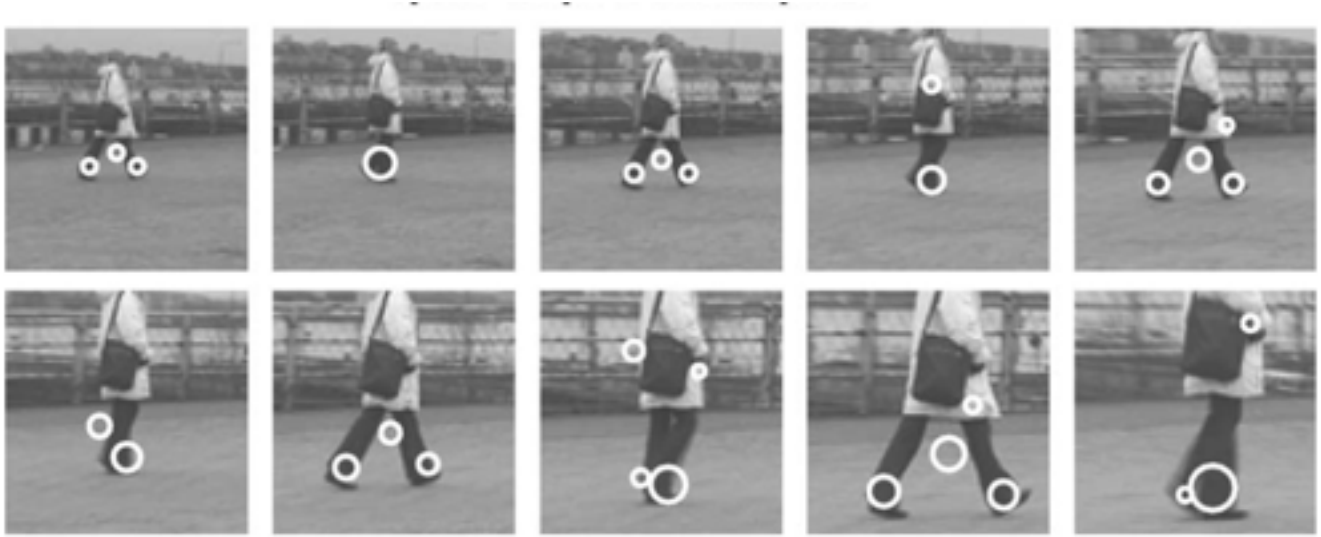
Spatio-temporal interest points

- Source: Laptev, IJCV 2005



Spatio-temporal interest points

- Source: Laptev, IJCV, 2005



Temporal structure of gesture

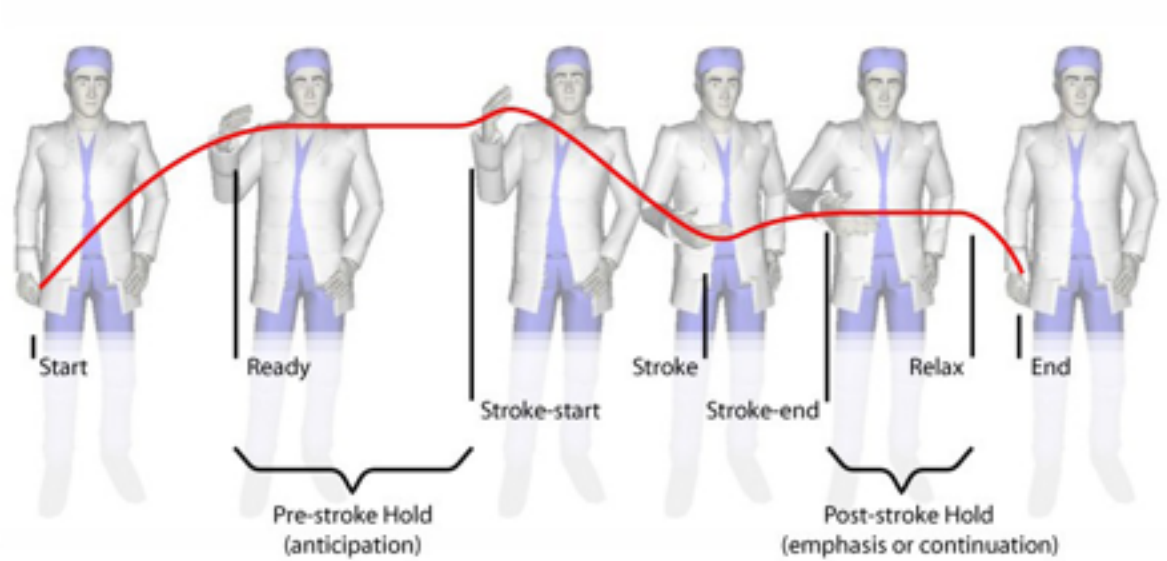
Global time-slice = Gesture Templates

Moments in time = Gesture Grammars

Unstructured time = Gesture Histograms
(Temporal Bag-Of-Words)

Temporal structure of gesture

- Source: Kipp et al. Behavior Markup Language



Temporal structure of gesture

GESTURE → [**preparation**] [**hold**] STROKE [**hold**].

STROKE → **main_stroke** (**after_stroke**)*.

GESTURE → {S-GESTURE | H-GESTURE}
S-GESTURE → [**preparation**] [**hold**] STROKE [**hold**]
H-GESTURE → [**preparation**] **hold**.

- Source: Kipp. Gesture Generation by Imitation

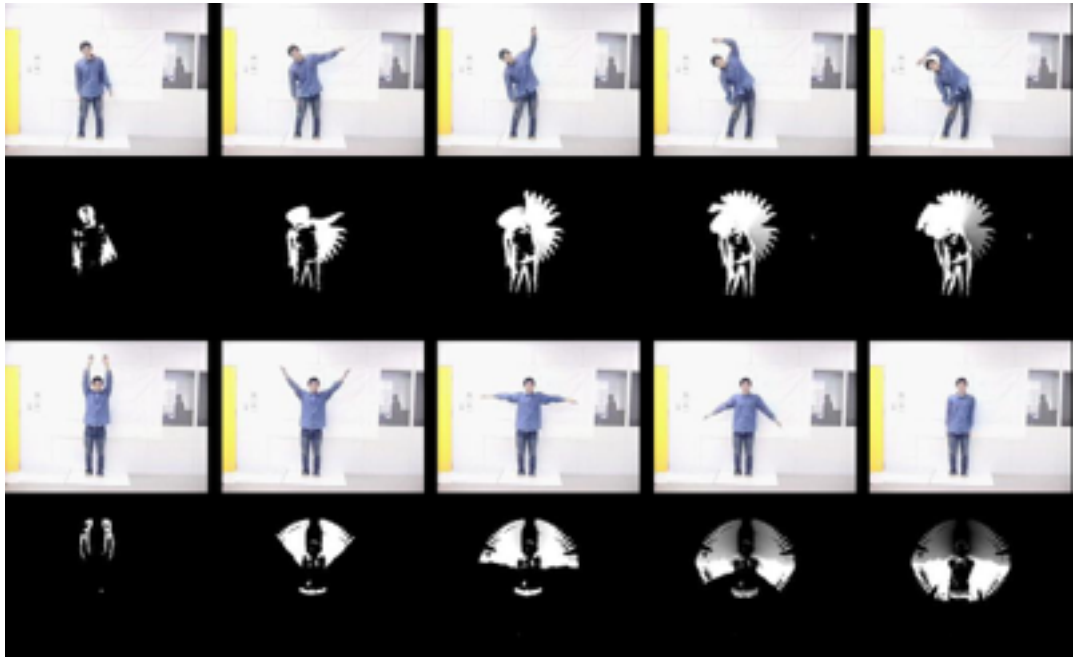
Gesture Templates

- Body is an N-dimensional articulated structure
  - recognize trajectory in N dimensions
- Template matching in N dimensions is hard !
- Dimension reduction and/or simplifications
  - Head trajectory
  - Hand trajectories
  - Head pose
- Segmentation requires sliding window (convolution)
- Multiple templates for speed, duration and intensity

Image model + Gesture Template

- Gesture is defined as a « spatio-temporal object »
  - Image x time
  - Space x time
- **Examples:**
  - Motion History Image
  - Motion History Volume

Motion history image



Representation and Recognition of Action Using Temporal Templates, Bobick and Davis, PAMI 2001

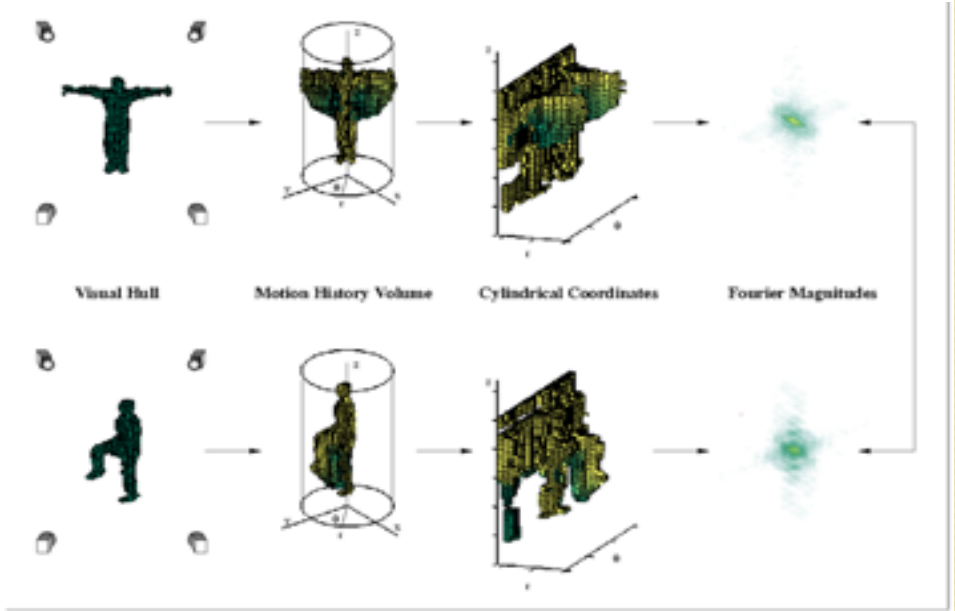
Motion History Volumes

Action Recognition in 3D using voxels

Shape from Silhouettes

+

Motion History Images



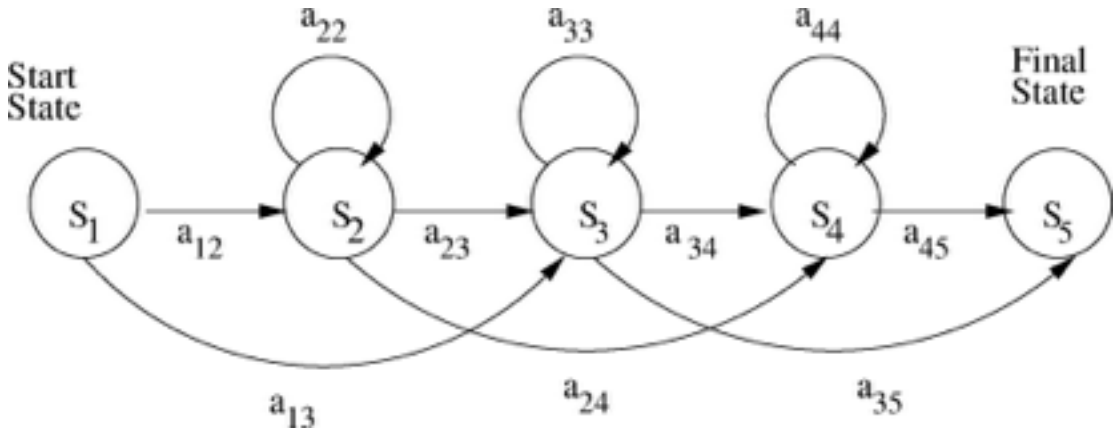
Free Viewpoint Action Recognition using Motion History Volumes. Weinland, Ronfard, Boyer, CVIU 2006

Gesture Grammars

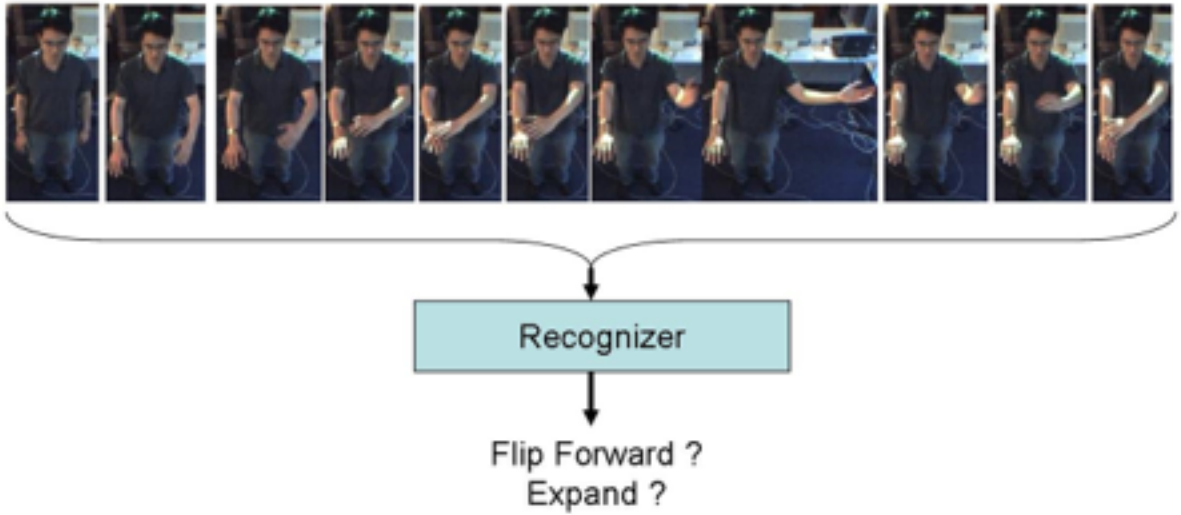
- Grammar decomposes gesture into « states» and states into observations
- Generative HMM
  - P(body part motion and pose|state)
  - P(state at t|state at t-1)
- Discriminative CRF
  - P(state|body part motion and pose)
- Probabilistic Context-free grammar
  - Repetition and grouping of states

Markov models

- Probabilistic Finite State Machines
- First-order Hidden Markov Models
- Segment Models

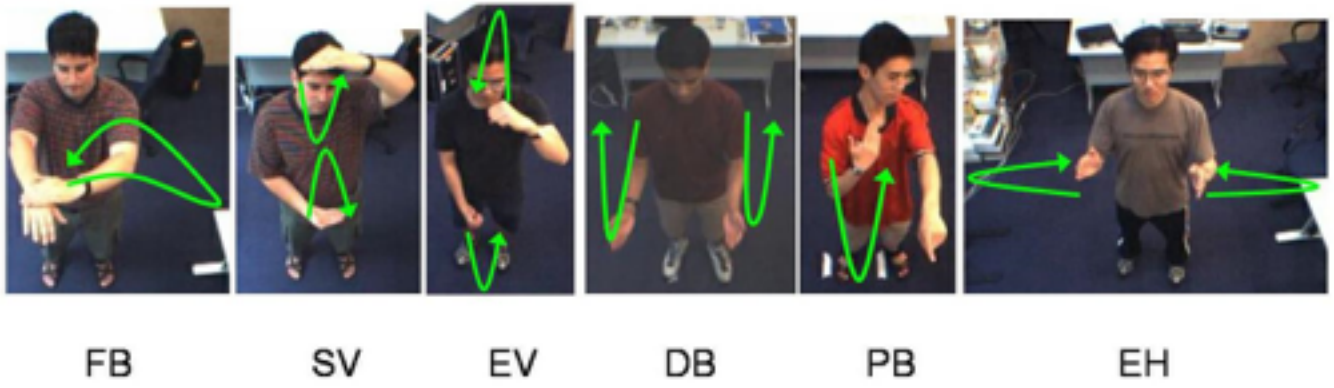


Conditional Random Fields



- **Source: Hidden Conditional Random Fields for Gesture Recognition by Wang, Quattoni, Morency, Demirdjian & Darrell, CVPR 2006**

Conditional Random Fields



FB - Flip Back, SV - Shrink Vertically, EV - Expand Vertically, DB - Double Back, PB - Point and Back, EH – Expand Horizontally.

Green arrows are the motion trajectory of the fingertip and the numbers next to the arrows symbolize the order of these arrows.

Image model + Grammar

- Grammar decomposes gesture into « states» and states into observations
- HMM is a probabilistic state machine
  - P(image motion and appearance|state)
  - P(state at t|state at t-1)
- CRF conditioned on image observations
  - P(state|image motion and appearance)
- Controled settings
- Constant viewpoint

Learning of action exemplars

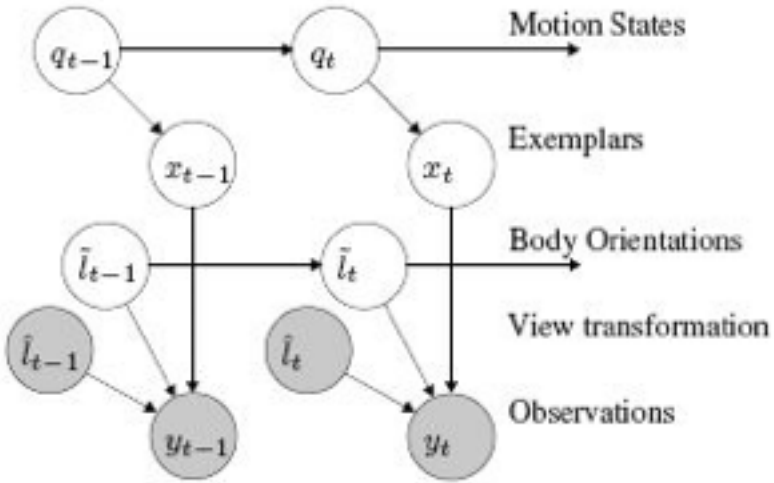
Goal – learn actions in 3D and recognize in 2D

Problems

- MHVs cannot be projected to novel views
- IXMAS actions not all simple

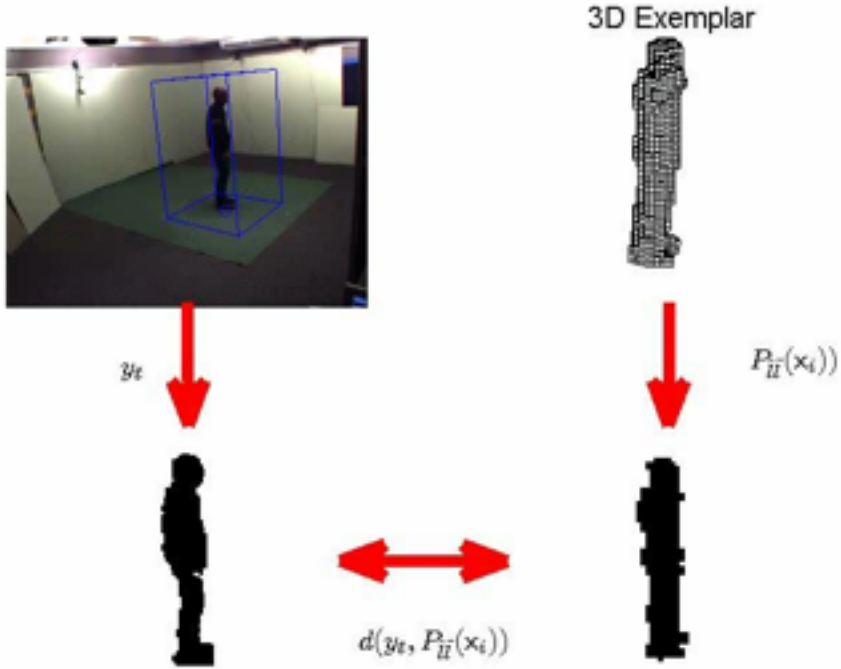
Solutions

- Metric mixture of 3D exemplars (Toyama & Blake, 2001)
- Transformed HMM for action & actor pose relative to camera (Frey & Jojic, 2001)



Action Recognition from Arbitrary Views using 3D Exemplars, Weinland, Boyer, Ronfard , 2007

Recognition with action exemplars



Recognition with action exemplars

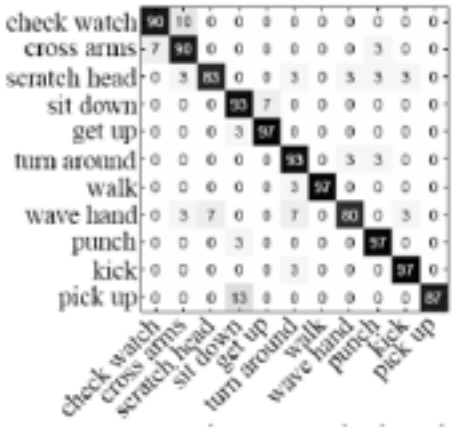
Caméra 1	Caméra 2	Caméra 3	Caméra 4	Caméra 5
65.4 %	70.0 %	54.3 %	66.0 %	33.6 %

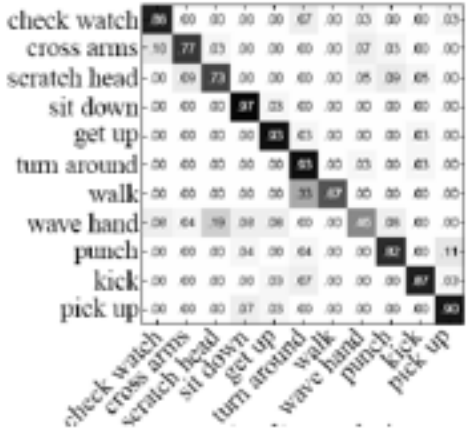
Average 2D recognition 60%

Average 3D recognition 90%

Best compromise with cameras 2+4

Outperformed by STIPS and self-similarity (Perez) on single-view recognition





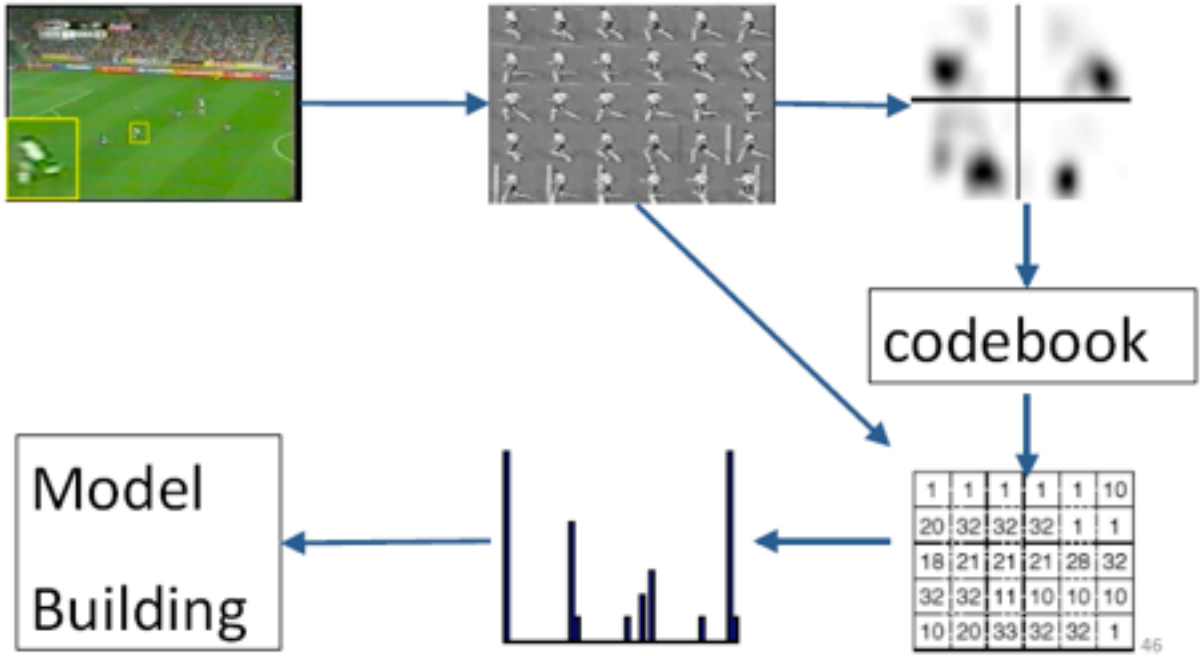
Temporal Bag of Words

- Histogram of image features over time
  - BODY MODELS
  - IMAGE MODELS
  - SPACE BAG OF WORDS
- Recognition based on distance between histograms
  - Learned histograms
  - Observed histogram

Image model + Bag of Words

- Gesture from a single or few « keyframes »
  - Still frame
  - Keyframe selection during learning and recognition
  - Same as ergodic HMM (no temporal structure)
- Gesture from many frames
  - Discretize video frames into lexicon of « visual words »
  - Compute word frequencies over time
  - P(gesture|sequence) = H(words)

Bag-of-Words Sequence Model



Source: Mori Structured Action Recognition. 2007.

Spatial + Temporal Bag-of–Words

- Build a discrete lexicon of « motion » words
  - in space and in time
- Learn the frequencies of motion words per gesture
  - P(word|gesture)
- Recognize most likely gesture from observed word frequencies over entire video segment (space and time)
  - Pros: Robust to partial observations and occlusions; invariant to sizes and durations ; very good precision in benchmarks; good generalization
  - Cons: Blind to higher-order spatial and temporal structure; hard to tune to specific cases

CONCLUSION

Gestures and actions are spatio-temporal patterns with internal structure and high complexity.

In the spatial domain, actions and gestures can be represented with body models, with image models, or with bag of isolated features.

In the temporal domain, they can be represented with templates, with grammars, or with bags of isolated features.

By combining the spatial and temporal aspects of gesture and action, one is faced with a vast number of possible combinations.

Additional references

A survey of vision-based methods for action representation, segmentation and recognition.

Daniel Weinland, Remi Ronfard, and Edmond Boyer.

Computer Vision and Image Understanding (CVIU). Vol. 115, No. 2 (February 2011), pp. 224-241.

Questions ?

Automatic Rush Generation with Application to Theatre Performances

Vineet Gandhi and Rémi Ronfard

LJK Laboratoire Jean Kuntzmann

Inria

Région RhôneAlpes

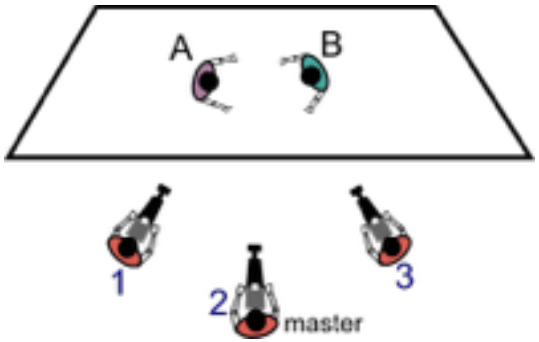
Célestins Théâtre de Lyon

Motivation

- Recordings theatre is important
  - Preserve and present cultural heritage
  - Educational, research or professional reasons
- Many organizations actively record theatre
  - The theatre on Film and Tape Archive (TOFT) - New York
  - French National Institute of Audiovisual (INA) - France
  - National Video Archive of Performance (NVAP) - UK
- Theatre for Web
  - Watch trailers and pay to watch online
  - Digital theatre, INA

Motivation

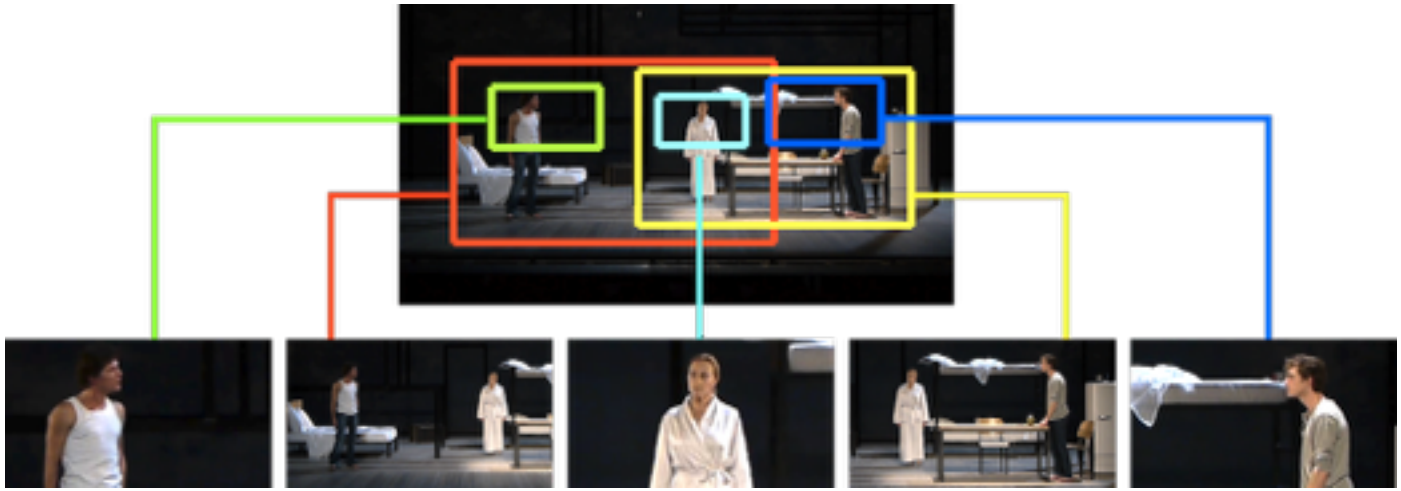
- Producing professional quality live performance videos require source video from multiple viewpoints
- These source videos (rushes) are then edited together to create the final result



High budget and Difficult!

- Multiple synchronized cameras
- Skilled cameramen
- Restricted viewpoints
- Intrusive
- No retakes

Our Approach



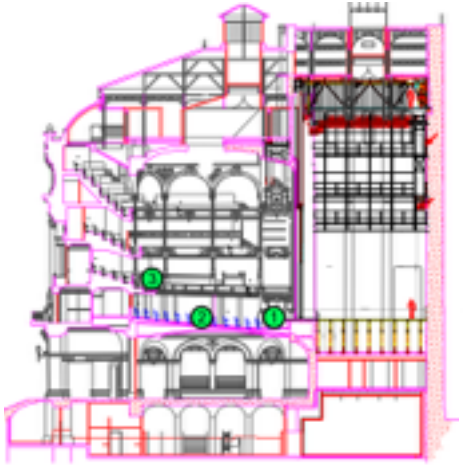
- A high resolution static camera replaces the plural camera crew
- Multiple cameras are then generated as a post-process
- Virtual pan, tilt, zoom movements (within original recording)

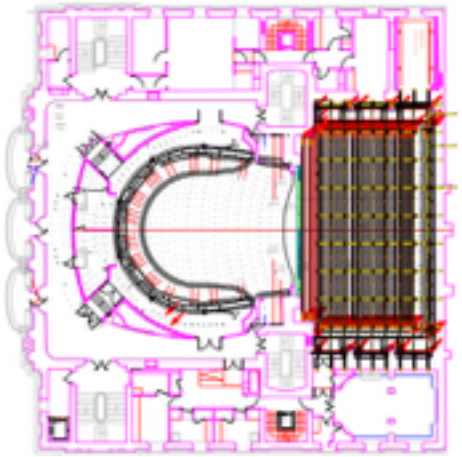
Organization

1. Theatre Database
2. Actor tracking
3. Rush Generation
4. Applications

Theatre database: Requirements

- Cover the entire stage
- Neutral angle
- Should work in varying light situations
- Good depth of field





1

2

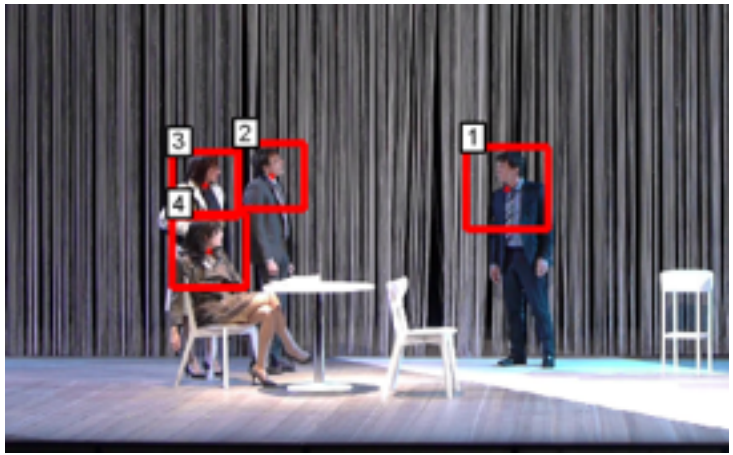
3

Organization

1. Theatre Database
2. Actor tracking
   - Related work
   - Detection
   - Tracking
   - Results
3. Rush Generation
4. Applications

Actor tracking: Goal

Goal: Identify and determine bounding box of each actor in each frame or indicate that it is not visible



Theatre [*Death of a Salesman,* Celestins 2013]

Movie [*Rope*, Alfred Hitchcock 1948]

Actor tracking: Challenges

- Occlusions
- Viewpoint and pose changes
- Scale changes
- Illumination changes
- Motion blur
- Distinguishing among actors






Actor tracking: Related work

Online approaches

- Real-time applications
- Generative [VTD Kwon et al. 2010] [IVT Ross et al. 2008]
- Discriminative [MILTrack Babenko et.al 2009] [SPT Wang et al. 2011]
- Usually fast, return smooth trajectories
- Require simple initialization
- Fail to consistently track objects over long periods
- Adapt models during tracking → drift

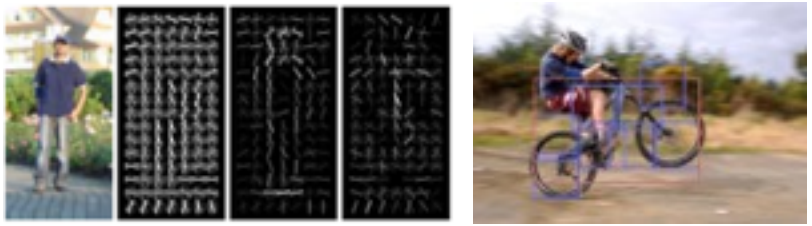
Example of drift in MILTrack

Actor tracking: Related work

Tracking using generic detectors

- Generic detectors combined using feature tracks
- Viewpoint dependent
- Does not take into account object specific features → low detection recall
- Feature tracks may drift → inaccurate localization at intermediate frames

Who are you? [Sivic CVPR 2009]

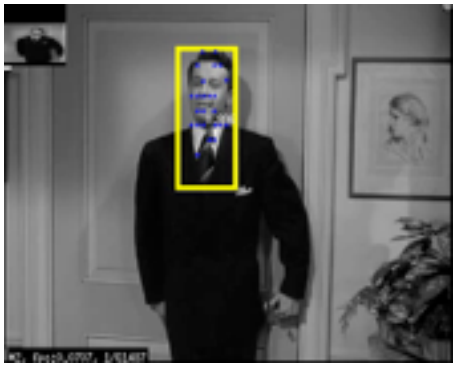


Generic detectors (Dalal CVPR 2005, Felzenszwalb CVPR 2008)

Actor tracking: Related work

TLD tracker [Kalal et al. 2012]

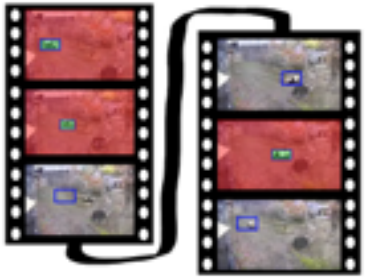
- Learns detector by gathering the weakly labelled training data during tracking process



- Addresses post failure behaviour
- Sensitive to initialization
- Biased to initial viewpoint
- Results may significantly vary based on the viewpoint chosen for initialization

Offline approaches

- Interactive offline tracking [Wei et al. 2009]
- Self paced learning [Supancic et al. 2013]
- Player tracking [Sullivan et al. 2006]



Actor tracking : Our approach

Three stage tracking framework:

- Learning → learn separate detector per actor
- Detection → perform individual detection at each frame
- Optimization → combine into smooth trajectories

Key ideas

- Offline → targeting long term tracking
- Initialized with a small set of representative training samples rather than a single initial position
- Models learnt at the beginning → non adaptive

Color blob detector: MSCR features

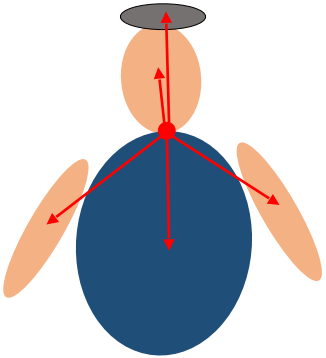
- Actor's clothing is stable in color (compared to interest point features)
- We use maximally stable color regions (MSCR) [Forssen CVPR 2007]

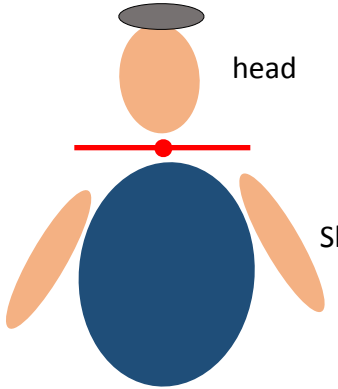
Image

MSCR features

Color blob detector: Appearance model

- Constellation model
  - Two parts: head and shoulders
  - Each actor appearance → visual vocabulary of color blobs $C = \{\gamma_i, \mu_i, \sum_i\}$





C_i → Weight (γ_i)
Mean appearance (μ_i)
Variance $(\sum_i)$

(μ_i) → 9 dimensional
(color, size, shape, position)

Color blob detector: Generative model

- Probability that an observed set of blobs B = {B_j} were generated using C = {γ_i, μ_i, $\sum_i$}

$$P_{overall} = \prod_k P_{part_k}$$ Product over parts

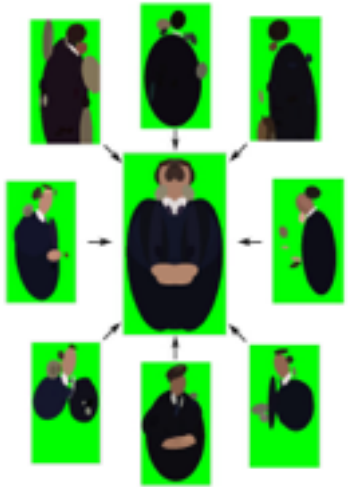
$$P_{part_k} = \sum_{ij} P(B_j, m_{ij}, C_i)$$ Sum over all blobs and clusters in the part

assignment between observed blobs and clusters

$$P(B_j, m_{ij}, C_i) = \gamma_i \cdot m_{ij} \cdot \exp\left\{-(\mu_i - B_j)^T \Sigma_i^{-1} (\mu_i - B_j)\right\}$$

Color blob detector: Model construction

- Constrained agglomerative clustering



Color blob detector: Example models

Color blob detector: Example models

(a) Brandon (b) Philip (c) Janet (d) Kenneth

(d) Mrs. Wilson (e) Mrs. Atwater (f) Rupert (g) Mr. Kentley

Color blob detector: Detection

- Sliding window search accelerated using K-nearest-neighbour (KNN) refinement


Model (C)


Image


MSCR features (B)

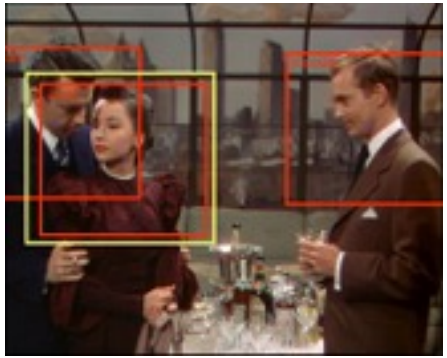
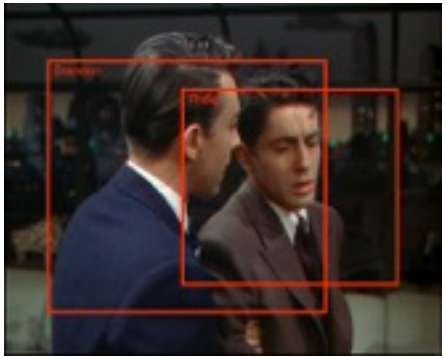
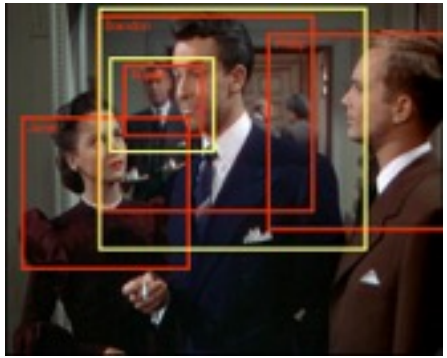

KNN Refined features (B')

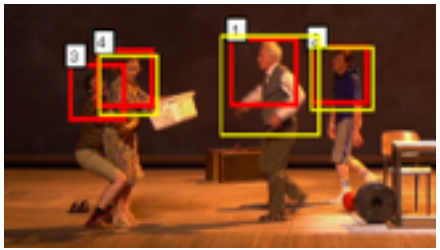

Importance map


Matched blobs


Matched clusters

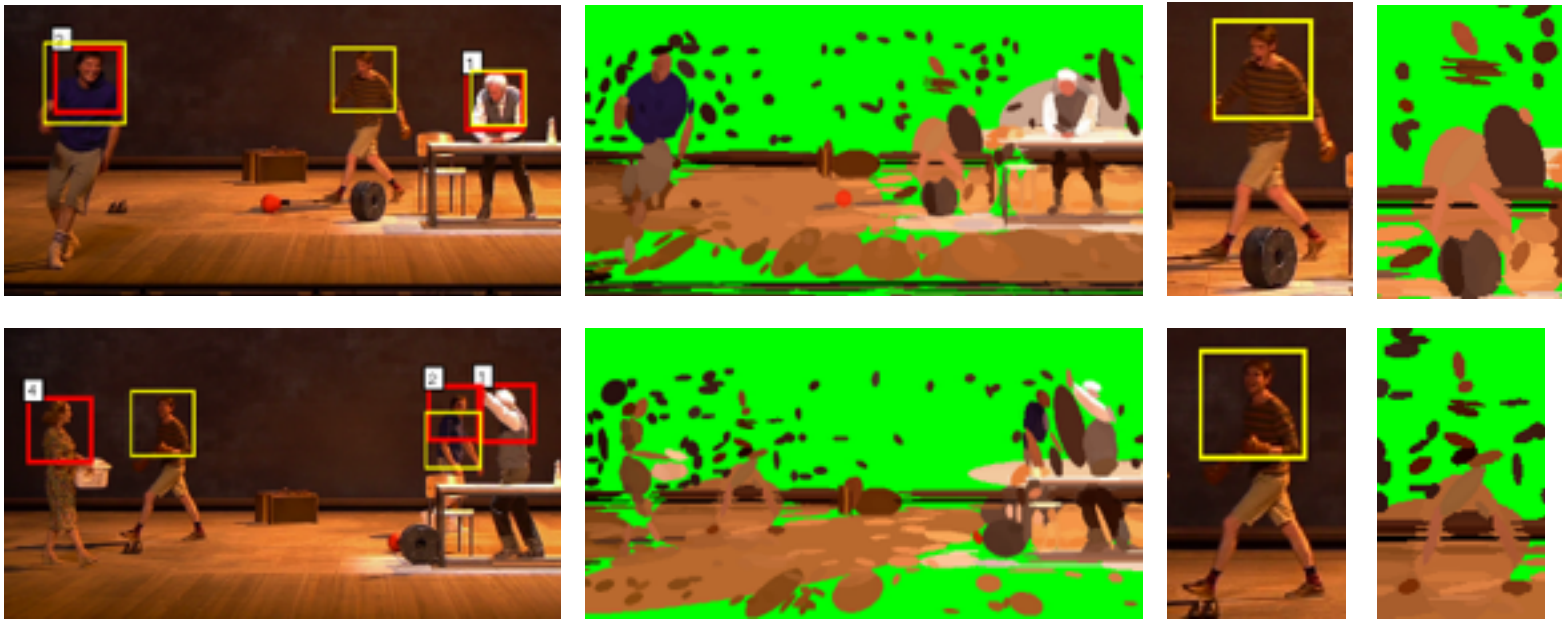
Detection results: Color blob detector

Red: Color blob detector
Yellow: Generic upper body detector [Felzenszwalb CVPR 2008]

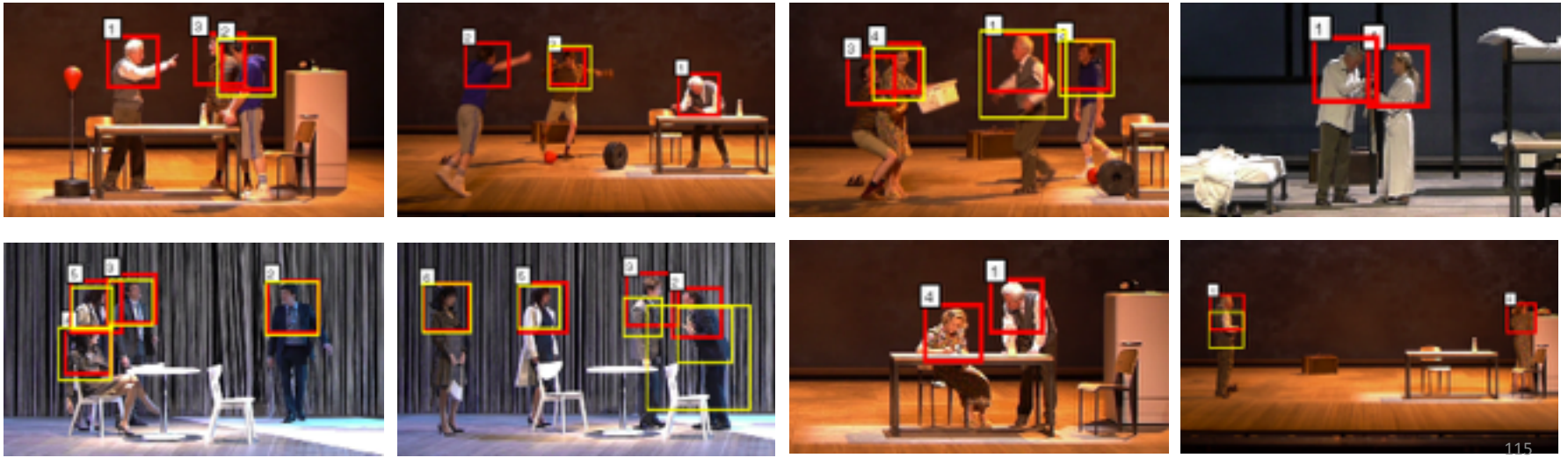
Detection results: Color blob detector



Color blob detector: Results theatre sequences

	FBD	UBD	CBD
Dos1 (4 actors, 1788 frames)	48	60	**72**
Dos2 (2 actors, 1490 frames)	76	80	**93**
Dos4 (4 actors, 1656 frames)	55	66	**84**
Dos5 (2 actors, 1094 frames)	43	54	**81**

Average recall



Actor tracking: Optimization

- We minimize following global cost function:

$$E(\xi) = \sum_{t=1}^{N} E_d(s_t) + \sum_{t=2}^{N} E_s(s_{t-1}, s_t).$$

- Data term E_d

$$E_d(s_t) = \begin{cases} -\log(P(s_t, t)) & \text{if } s_t > 0 \\ \lambda_1 & \text{if } s_t = 0 \end{cases}$$

- Smoothness term E_s

$$E_s(s_{t-1}, s_t) = \begin{cases} D_m(l(s_{t-1}, t-1), l(s_t, t)) & \text{if } s_t > 0\,, \\ \lambda_2 & \text{if } s_t = 0\,. \end{cases}$$

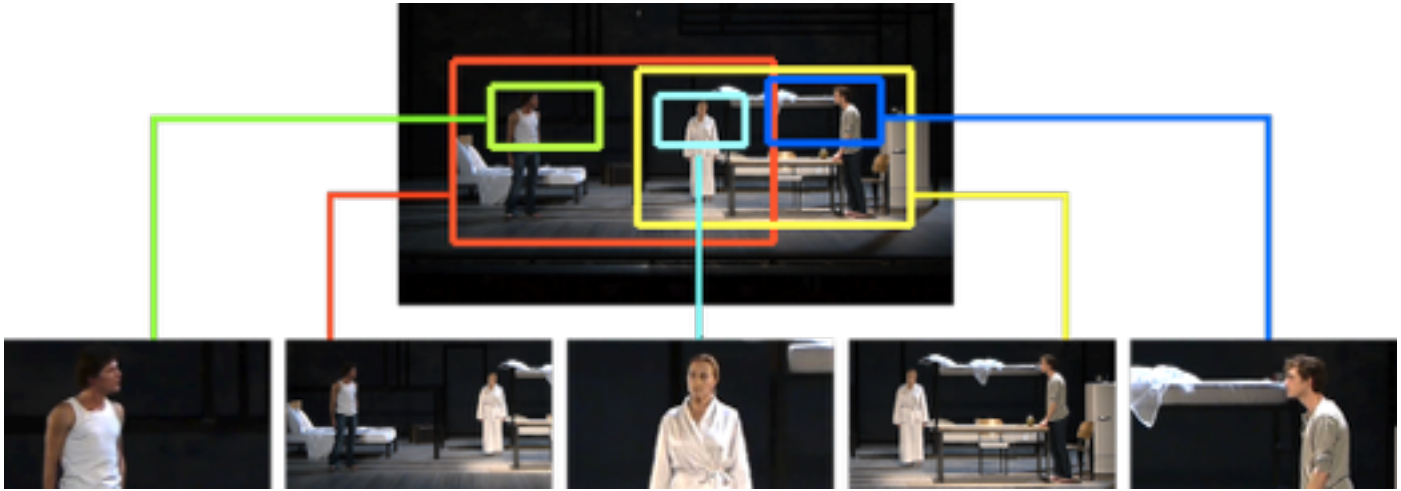
$$D_m(l_{t-1}, l_t) = \frac{(x_{t-1} - x_t)^2}{\sigma_x^2} + \frac{(y_{t-1} - y_t)^2}{\sigma_y^2} + \frac{(w_{t-1} - w_t)^2}{\sigma_w^2}.$$

Organization

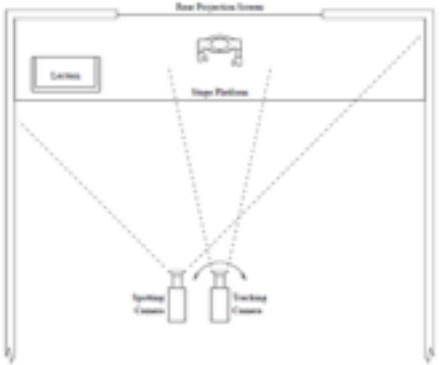
1. Theatre Database
2. Actor tracking
3. Rush Generation
   - Related Work
   - Our method
   - Results
4. Applications

Rush generation

Goal : Given actors tracks simulate virtual camera shots



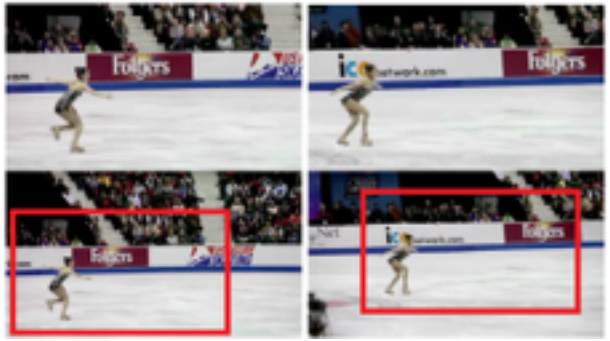
Rush generation: Related work



AutoAuditorium [Bianchi et al. 1998]

Virtual Videography [Heck et al. 2007]

Hybrid PTZ [Carr et al. 2013]



Video Stabilization [Grundmann et al. 2011]

Rush Generation: What comprises a good camera work?

- Shot composition
  - Subjects should not be cut by the image frame
  - Subjects must be given more space in the direction they travel
  - More space in the direction they look
- Cuttability (keeping the editing in mind)
  - Keep camera static
  - Screen continuity
- Camera movement
  - Camera movement should be motivated
  - Apparent actor movement should be consistent

Rush Generation: Our method

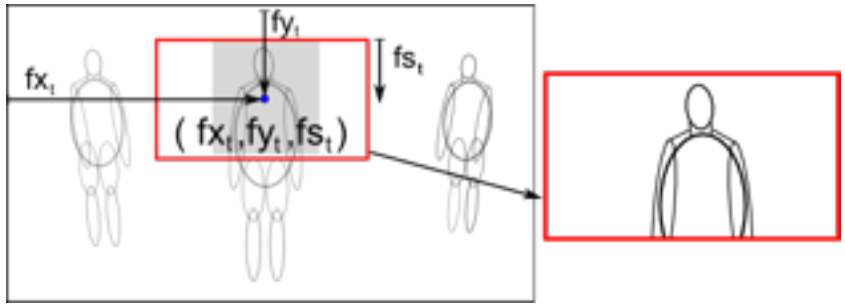
- Cinematographic principles as constraints and penalties
- We solve a constrained convex optimization problem

Input

- The master shot (single static video covering the entire field of view)
- Actor tracks
- Shot specification

Output

- The virtual camera trajectory $\xi = \{fx_t, fy_t, fs_t\}$

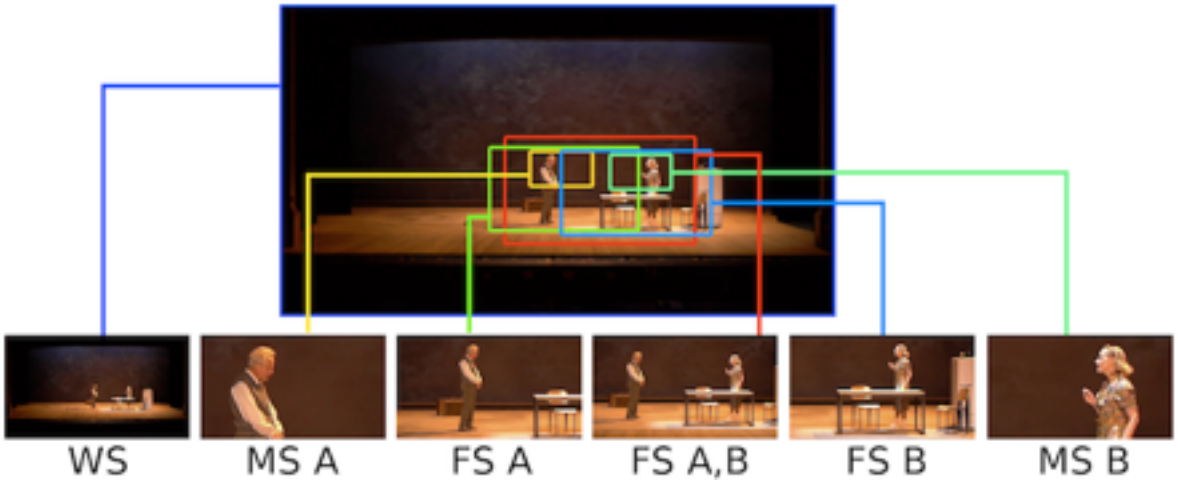


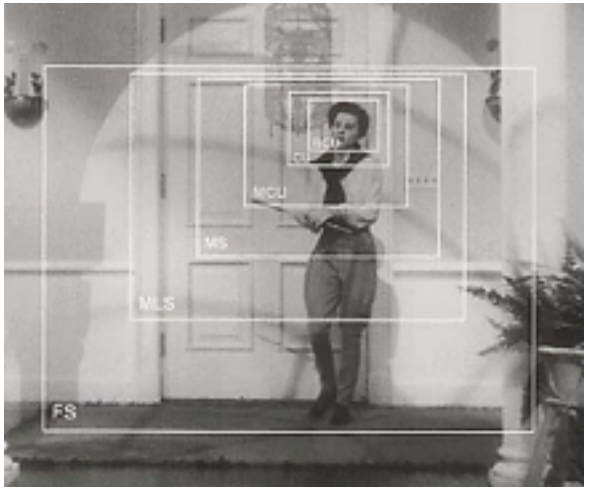
Rush Generation: Actor tracks

- $(bx_t^m, by_t^m, bs_t^m, bh_t^m)$ for each actor m at time t
  - Tracks calculated using our offline tracking algorithm
  - Use size to infer depth and ground projections

Rush Generation: Shot specification

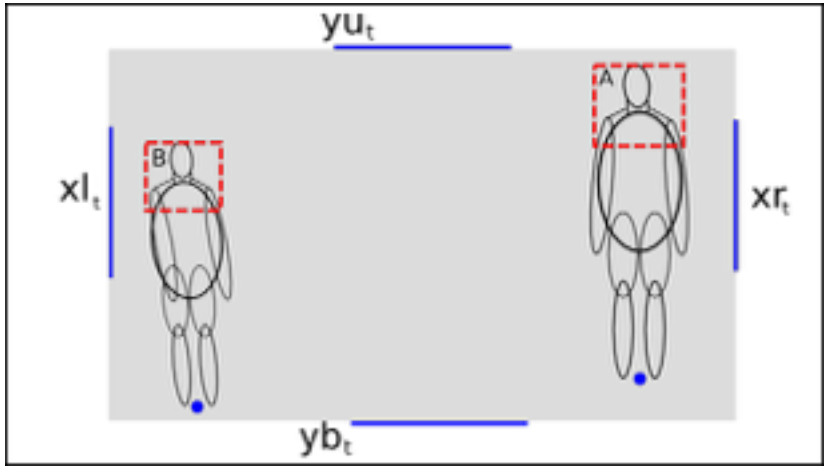
- Shot size and list of actors



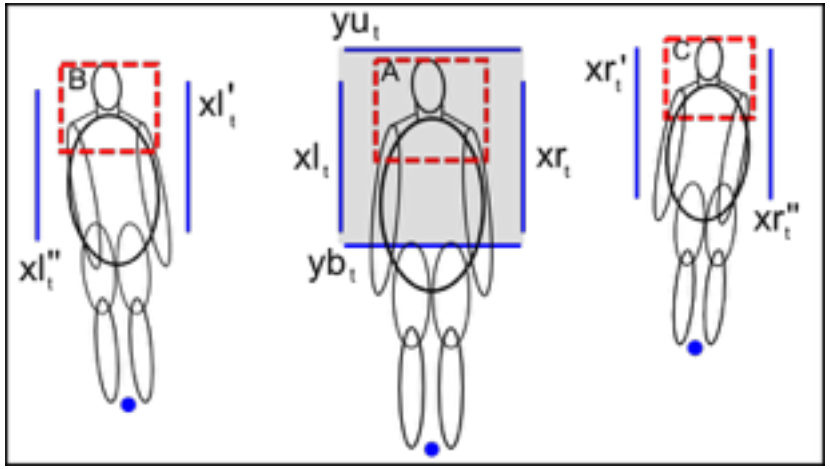


Shot sizes [B. Salt. *Moving into pictures*. 2006]

Rush Generation: Inclusion region and external actor boundaries

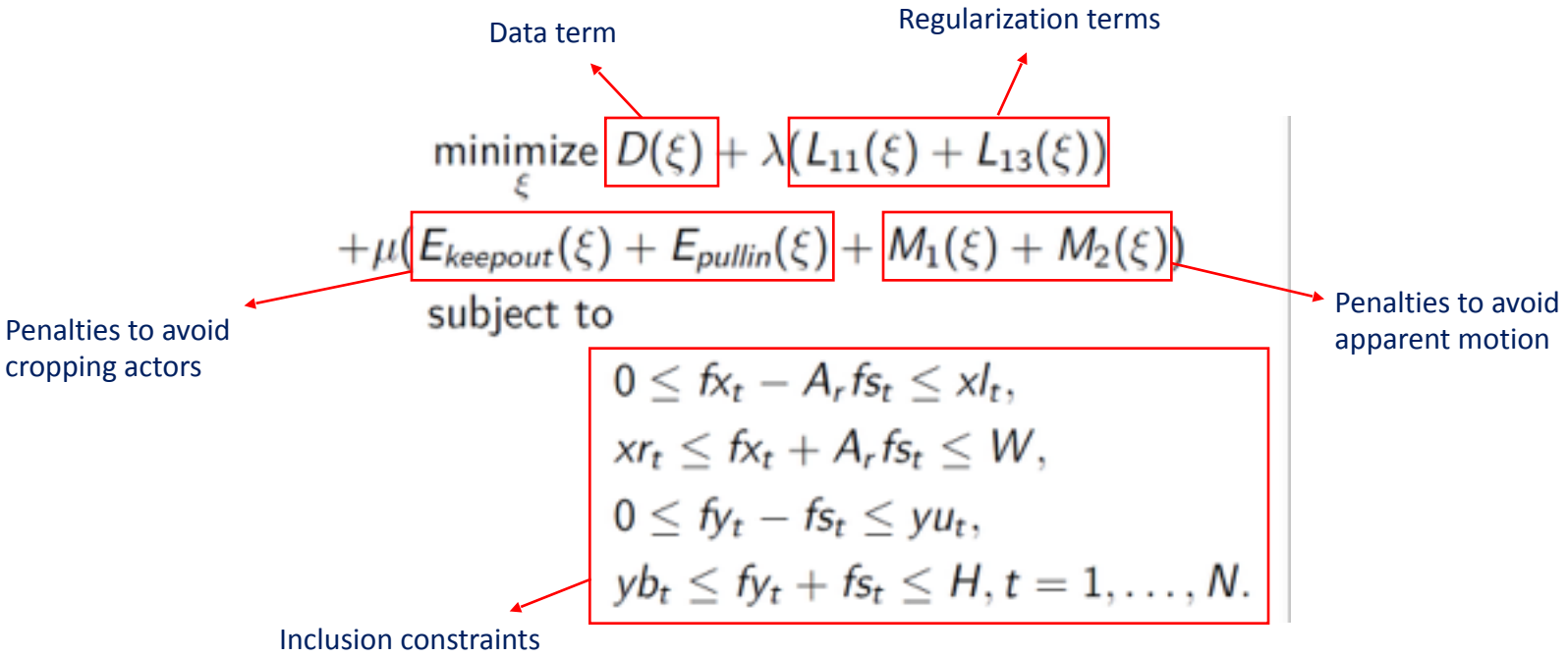


inclusion region “FS A,B”



inclusion region “MS A” and boundaries of the nearest actor on the left and the right

Rush Generation: Optimization – Overall



- The camera frame should always lie within the master shot
- The inclusion region should be enclosed within the camera frame

Rush Generation: Optimization – Shot size penalty

- The camera frame should remain close to the inclusion region

$$D(\xi) = \frac{1}{2}\sum_{t=1}^{N}((fx_t - x_t)^2 + (fy_t - y_t)^2 + (fs_t - s_t)^2),$$

$$\text{where } x_t = \tfrac{1}{2}(xl_t + xr_t),\ y_t = \tfrac{1}{2}(yu_t + yb_t) \text{ and } s_t = \tfrac{1}{2}(yb_t - yu_t)$$

- To impose screen continuity, we add an offset ($0.17A_r fs_t h_t$)



Rush Generation: Optimization – First order L1-norm regularization

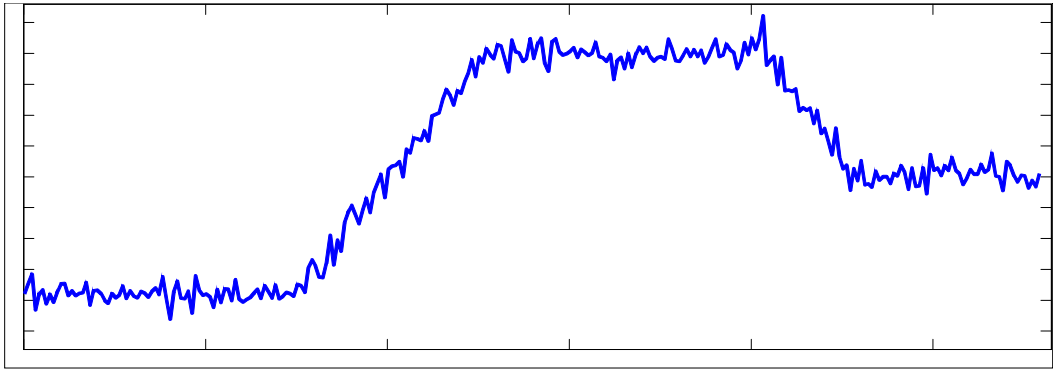
- Avoid non motivated small camera movements
- Long static camera segments → favourable for cutting

We regularize with the L1-norm of the camera velocity

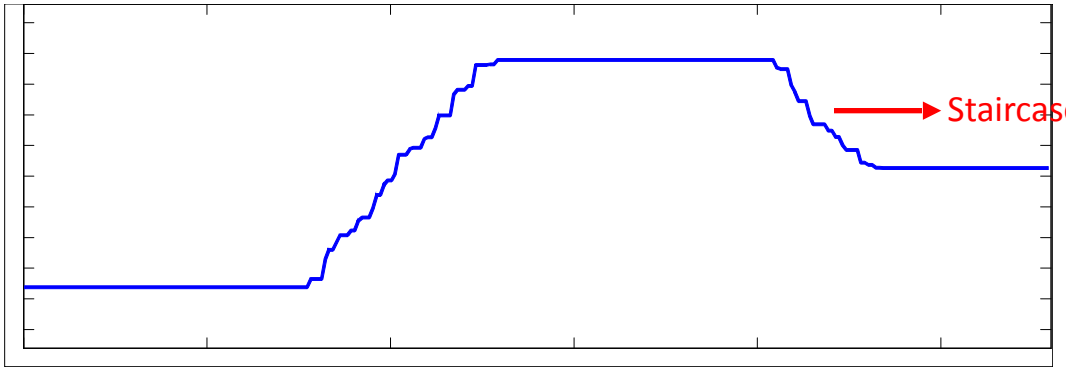
$$L_{11}(\xi) = \sum_{t=1}^{N-1}(|fx_{t+1} - fx_t| + |fy_{t+1} - fy_t| + |fs_{t+1} - fs_t|).$$

Rush Generation: Optimization – First order L1-norm regularization

Original signal x_t, t = [1 : N]



$$\sum_{t=1}^{n}(x_t - r_t)^2 + \lambda \sum_{t=1}^{n-1} |r_{t+1} - r_t|$$



Rush Generation: Optimization – Third order L1-norm regularization

- When the camera moves it should move smoothly
- Start with a segment of constant acceleration (ease in) and end with a segment of constant deceleration (ease out)

We regularize with the L1-norm of the camera jerk

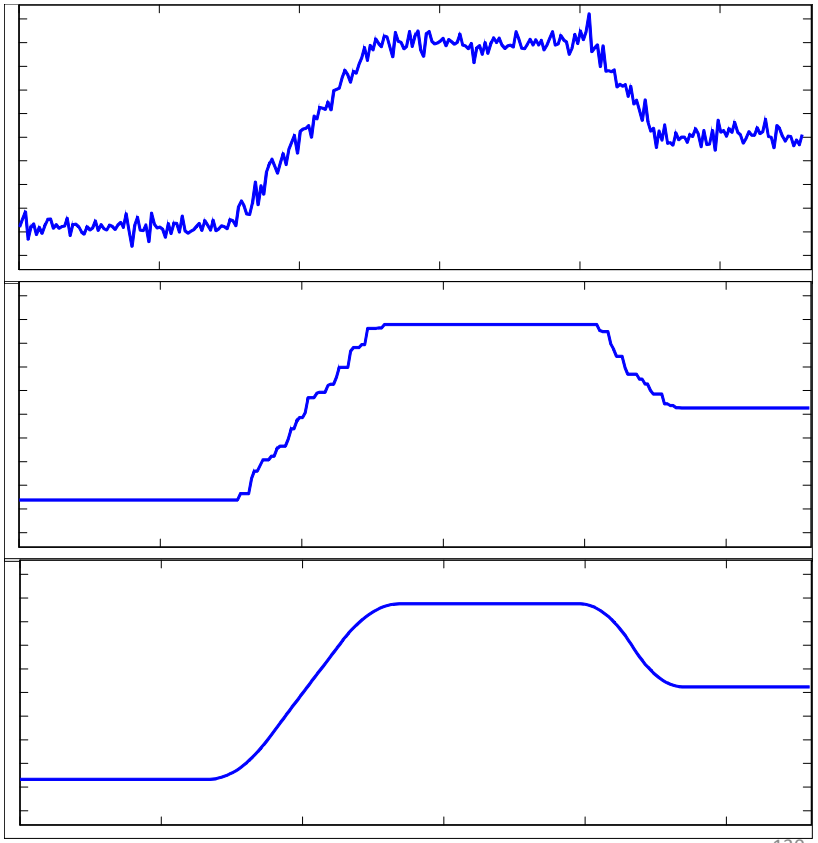
$$L_{13}(\xi) = \sum_{t=1}^{N-3} (|fx_{t+3} - 3fx_{t+2} + 3fx_{t+1} - fx_t| + |fy_{t+3} - 3fy_{t+2} + 3fy_{t+1} - fy_t| + |fs_{t+3} - 3fs_{t+2} + 3fs_{t+1} - fs_t|).$$

Rush Generation: Optimization – First order L1-norm regularization

Original signal x_t, t = [1 : N]

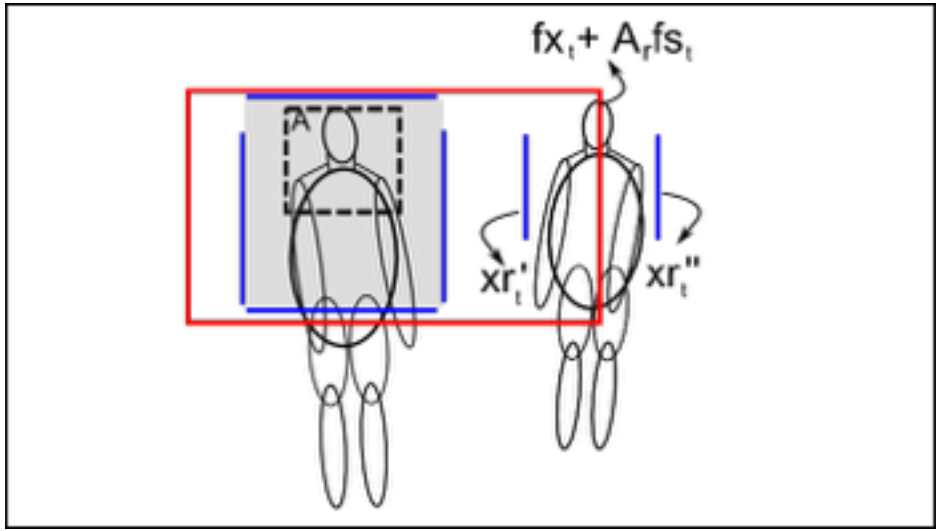
$$\sum_{t=1}^{n}(x_t - r_t)^2 + \lambda \sum_{t=1}^{n-1} |r_{t+1} - r_t|$$

$$\sum_{t=1}^{n}(x_t - r_t)^2 + \lambda \sum_{t=1}^{n-1} |r_{t+1} - r_t| \; + + \lambda \sum_{t=1}^{n-3} |r_{t+3} - 3r_{t+2} + 3r_{t+1} - r_t|$$



Rush Generation: Optimization – Pull in or keep out penalty

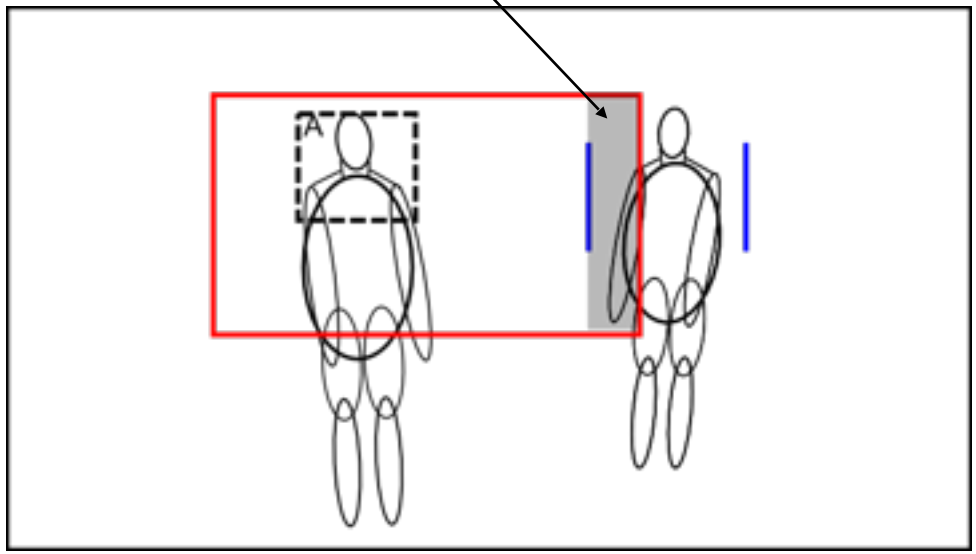
- Avoid chopping actors
- Each actor must either be in or out of the virtual camera window
- Actor included in shot specification → hard constraints
- Other actors may still come in contact with the virtual camera frame



Rush Generation: Optimization – Keep out penalty

- When the external actor is not touching the inclusion region

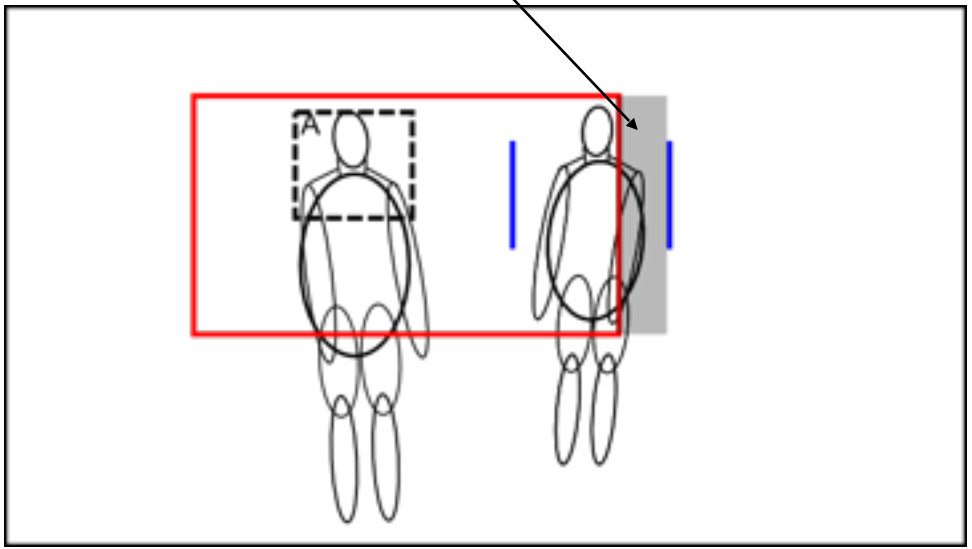
We add a penalty $E_{keepout} = |actor \cap frame|$



Rush Generation: Optimization – Pull in penalty

- When a touch event occurs

We add a penalty $E_{pullin} = |actor \setminus frame|$



Rush Generation: Optimization – Apparent motion penalty

- The apparent motion should be consistent
  - First term penalizes camera motion when actors are not moving

$$M_1(\xi) = \sum_m \sum_{t=1}^{N-1} (cx_t^m \, |fx_{t+1} - fx_t| + cy_t^m \, |fy_{t+1} - fy_t| + cs_t^m \, |fs_{t+1} - fs_t|).$$

cx_t^m, cy_t^m, cs_t^m are binary vectors → 1 if the actor is static in horizontal direction, vertical direction or size respectively

Rush Generation: Optimization – Apparent motion penalty

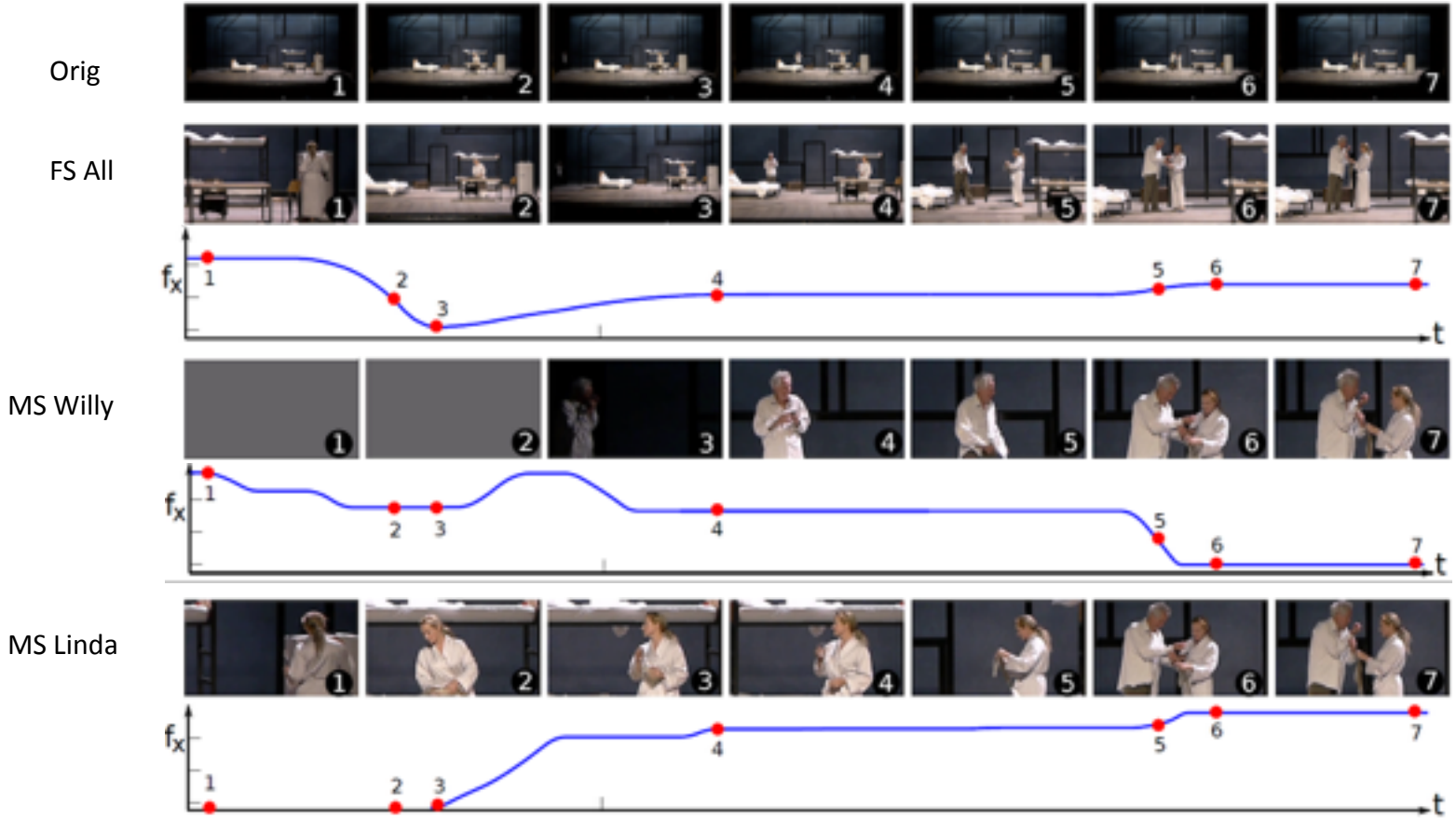
- Motion direction should be consistent

$\dot{bx}_t^m = (bx_{t+1}^m - bx_t^m)$ → actual horizontal actor motion

$(\dot{bx}_t^m - \dot{fx}_t^m)$ → apparent actor motion

Second term *M*2 penalizes negative values of $(\dot{bx}_t^m - \dot{fx}_t^m)\ \dot{bx}_t^m$

Rush Generation: Results



Rush Generation: Summary

- Rush generation as a convex minimization problem
- Considering composition, cutting and movement

Limitations and Future work

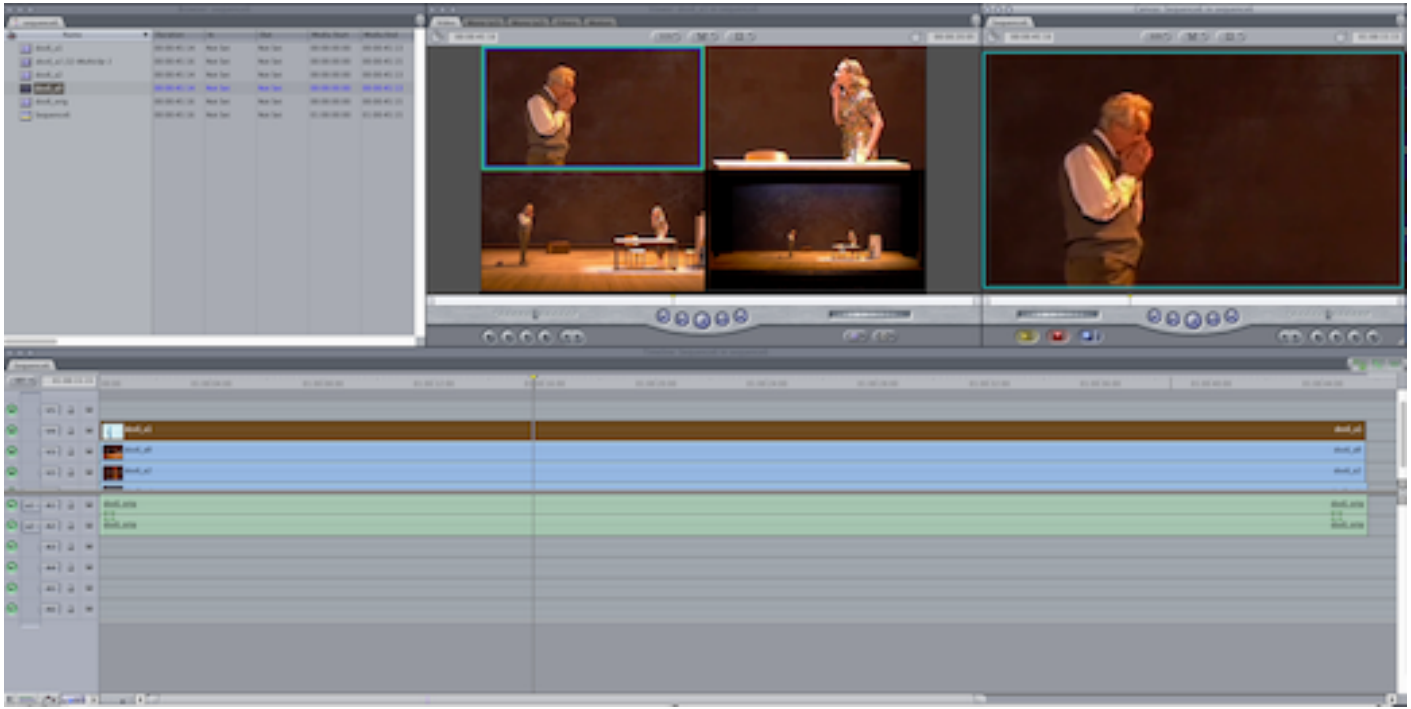
- Separate optimization for each shot specification, may lead to jump cuts
- Future work should explore joint optimization of multiple shots
- Higher resolution master shots (4K or 6K)

Organization

1. Theatre Database
2. Actor tracking
3. Rush Generation
4. Applications

Applications: multi-clip editing

- Generated rushes directly editable as multi-clip



THANK YOU!