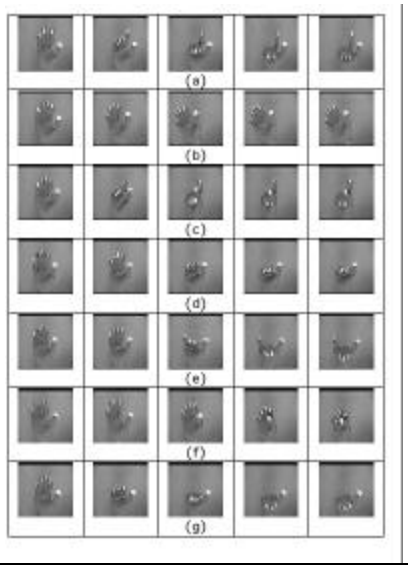
Hand Gesture Recognition

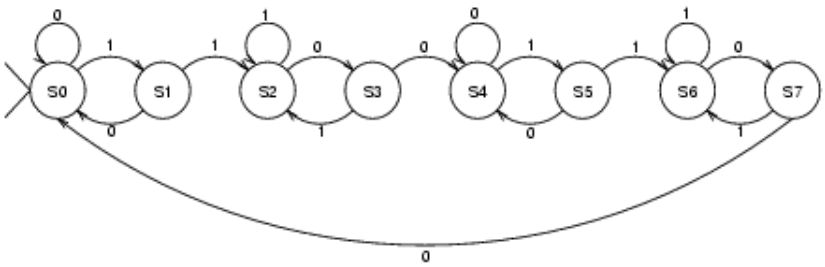
Seven Gestures



Gesture Phases

- Hand fixed in the start position.
- Fingers or hand move smoothly to gesture position.
- Hand fixed in gesture position.
- Fingers or hand return smoothly to start position.

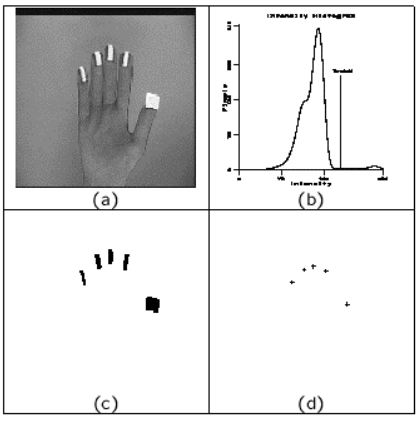
Finite State Machine



Main Steps

- Detect fingertips.
- Create fingertip trajectories using motion correspondence of fingertip points.
- Fit vectors and assign motion code to unknown gesture.
- Match

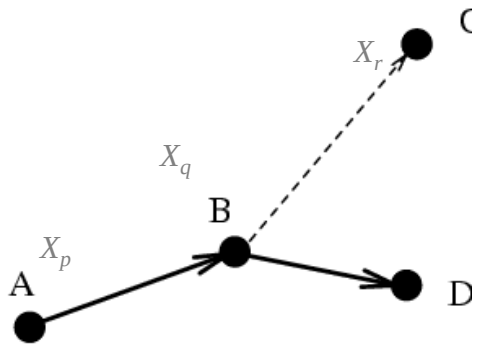
Detecting Fingertips



Proximal Uniformity Constraint

- Most objects in the real world follow smooth paths and cover small distance in a small time.
  - Given a location of point in a frame, its location in the next fame lies in the proximity of its previous location.
  - The resulting trajectories are smooth and uniform.

Proximal Uniformity Constraint



Proximal Uniformity Constraint

Establish correspondence by minimizing:

$$d(X_p^{k-1}, X_q^k, X_r^{k+1}) = \frac{\| \overline{X_p^{k-1} X_q^k} - \overline{X_q^k X_r^{k+1}} \|}{\sum_{x=1}^{m}\sum_{z=1}^{m} | \overline{X_x^{k-1} X_y^k} - \overline{X_y^k X_z^{k+1}} \|} + \frac{\| \overline{X_q^k X_r^{k+1}} \|}{\sum_{x=1}^{m}\sum_{z=1}^{m} \| \overline{X_y^k X_z^{k+1}} \|}$$

Greedy Algorithm

- For k=2 to n-1 do
- Construct M, an mxm matrix, with the points from k-th frame along the rows and points from (k+1)-th frame along the columns. Let

$$M[i, j] = d(X_p^{k-1}, X_q^k, X_r^{k+1})$$

when

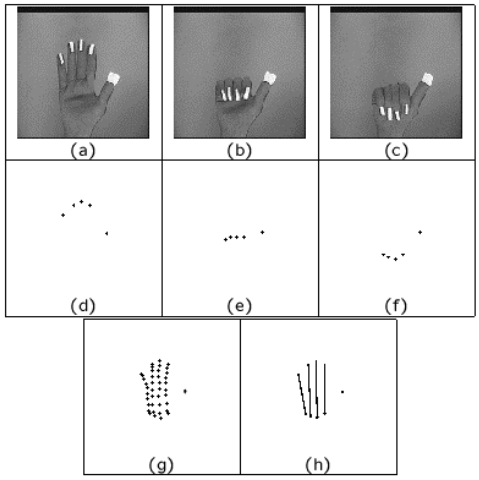
- for a=1 to m do
  - Identify the minimum element $[i, l_i]$ in each row i of M
  - Compute *priority matrix*, B, such that $B[i, l_i] = \sum_{j=1, j \neq l_i}^{m} M[i, j] + \sum_{k=1, k \neq i}^{m} M[k, l_i]$ for each *i*.
  - Select $[i, l_i]$ pair with highest *priority* value and make $f^k(i) = l_i$
  - Mask row *i* and column l_i from *M*.

Example

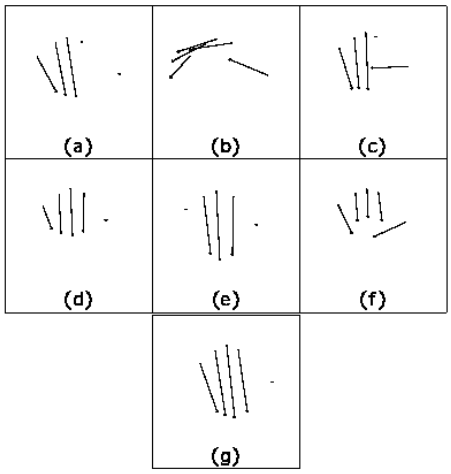
$$M = \begin{bmatrix} .6 & .3 \\ .7 & .2 \end{bmatrix}$$

$$B = \begin{bmatrix} .8 \\ 1 \end{bmatrix}$$

Vector Extraction



Vector Representation of Gestures



Results

Results

Run	Frames	L	R	U	D	T	G	S
1	200	√	√	√	√	√	√	√
2	250	√	√	√	√	√	√	√
3	250	√	√	√	X	√	√	√
4	250	√	√	√	√	√	√	√
5	300	√	√	√	√	√	√	√
6	300	√	√	√	√	√	√	√
7	300	√	√	√	√	√	√	√
8	300	√	√	√	√	√	√	√
9	300	√	√	√	√	*	*	*
10	300	√	√	√	√	√	√	√

L = Left, R = Right, U = Up, D = Down, T = Rotate, G = Grab, S = Stop, √ - Recognized, X - Not Recognized, * - Error in Sequence.

Action Recognition Using Temporal Templates

Jim Davis and Aaron Bobick

Main Points

- Compute a sequence of difference pictures from a sequence of images.
- Compute Motion Energy Images (MEI) and Motion History Images (MHI) from difference pictures.
- Compute Hu moments of MEI and MHI.
- Perform recognition using Hu moments.

MEI and MHI

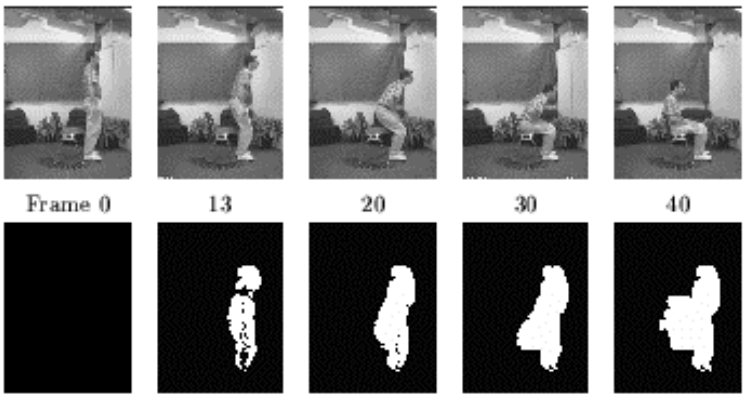
Motion-Energy Images (MEI)

$$E_t(x,y,t) = \bigcup_{i=0}^{t-1} D(x,y,t-i)$$ **Difference Pictures**

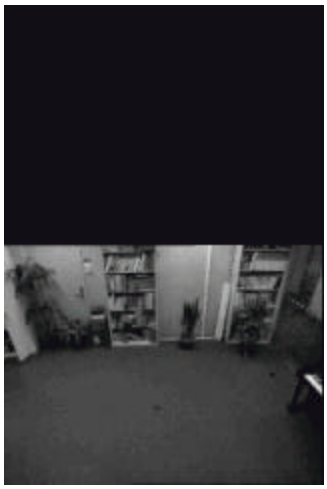
Motion History Images (MHI)

$$H_t(x,y,t) = \begin{cases} t & if\ D(x,y,t)=1 \\ \max(0, H_t(x,y,t-1)-1) & otherwise \end{cases}$$

MEIs



Color MHI Demo



Main Points

- Use seven Hu moments of MHI and MEI to recognize different exercises.
- Use seven views (-90 degrees to +90 degrees in increments of 30 degrees).
- For each exercise several samples are recorded using all seven views, and the mean and covariance matrices for the seven moments are computed as a model.
- During recognition, for an unknown exercise all seven moments are computed, and compared with all 18 exercises using Mahalanobis distance.
- The exercise with minimum distance is computed as the match.
- They present recognition results with one and two view sequences, as compared to seven view sequences used for model generation.

Moments

General Moments

$$m_{pq} = \int_{-}\int x^p y^q r(x,y)dxdy$$

Central Moments (Translation Invariant)

$$m_{pq} = \int\int (x-\bar{x})^p (y-\bar{y})^q r(x,y)\ d(x-\bar{x})d(y-\bar{y})$$

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}}$$

Central Moments

$$\begin{aligned}
m_{00} &= m_{00} \\
m_{01} &= 0 \\
m_{10} &= 0 \\
m_{20} &= m_{20} - m\bar{x}^2 \\
m_{11} &= m_{11} - m\bar{x}\bar{y} \\
m_{02} &= m_{02} - m\bar{y}^2 \\
m_{30} &= m_{30} - 3m_{20}\bar{x} + 2m\bar{x}^3 \\
m_{21} &= m_{21} - m_{20}\bar{y} - 2m_{11}\bar{x} + 2m\bar{x}^2 y \\
m_{12} &= m_{12} - m_{02}\bar{x} - 2m_{11}\bar{y} + 2m\bar{x}y^2 \\
m_{03} &= m_{03} - 3m_{02}\bar{y} + 2m\bar{y}^3
\end{aligned}$$

Moments

Hu Moments: translation, scaling and rotation invariant

$$u_1 = m_{20} + m_{02}$$

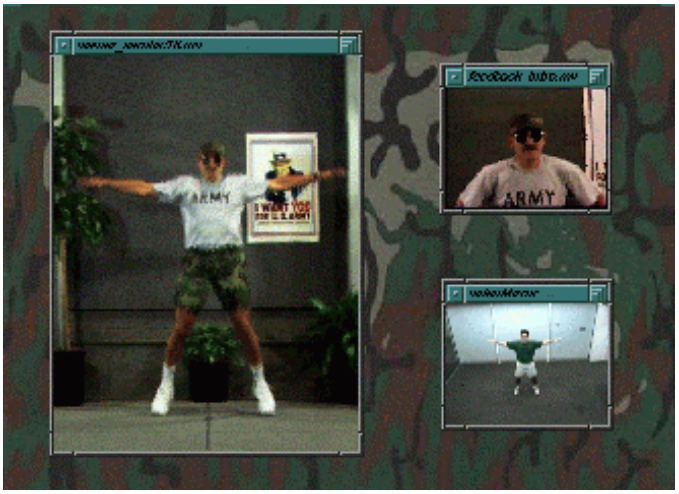
$$u2 = (m_{20} - m_{02})^2 + m_{11}^{\ 2}$$

$$u_3 = (m_{30} - 3m_{12})^2 + (3m_{12} - m_{03})^2$$

$$u_4 = (m_{30} + m_{12})^2 + (m_{21} + m_{03})^2$$

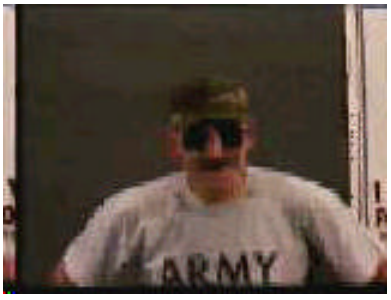
$$\vdots$$

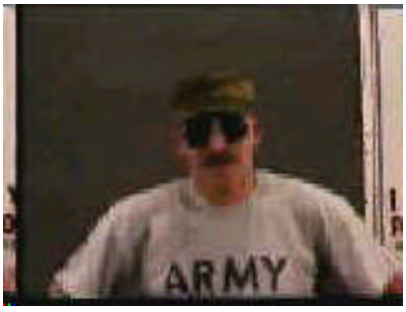
PAT (Personal Aerobic Trainer)



PAT (Personal Aerobic Trainer)

PAT (Personal Aerobic Trainer)





Webpage

- http://vismod.www.media.mit.edu/vismod/demos/actions/mhi_generation.mov
- http://www.cs.ucf.edu/~ayers/research.html
- http://www.cs.ucf.edu/~vision

Papers

- Claudette Cedras and Mubarak Shah, “Motion-Based Recognition: A survey”, Image and Vision Computing, March 1995.
- Jim Davis and Mubarak Shah, “Visual Gesture Recognition”, IEE Proc. Vis Image Signal Processing, October 1993.

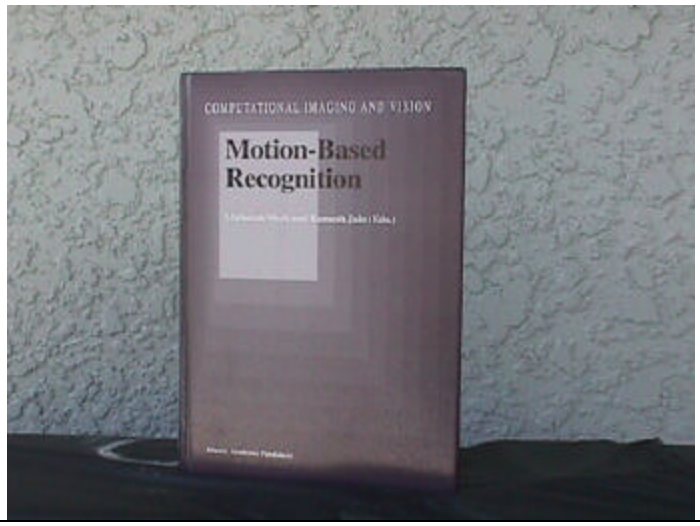
Papers

- Li Nan, Shawn Dettmer, and Mubarak Shah, "Visual Lipreading", Workshop on Face and Gesture Recognition, Zurich, 1995.
- Doug Ayers and Mubarak Shah, "Recognizing Human Activities In an Office Environment", Workshop on Applications of Computer Vision, October, 1998.

Book

- Mubarak Shah and Ramesh Jain, "**Motion-Based Recognition**", Kluwer Academic Publishers, 1997 ISBN 0-7923-4618-1.

Book



Contents

- Mubarak Shah and Ramesh Jain, "Visual Recognition of Activities, Gestures, Facial Expressions and Speech: An Introduction and a Perspective"
- Human Activity Recognition
  - Y. Yacoob and L. Davis, "Estimating Image Motion Using Temporal Multi-Scale Models of Flow and Acceleration"
  - A. Baumberg and D. Hogg, "Learning Deformable Models for Tracking the Human Body"
  - S. Seitz and C. Dyer, "Cyclic Motion Analysis Using the Period Trace"

Contents (contd.)

- R. Pollana and R. Nelson, “Temporal Texture and Activity Recognition”
- A. Bobick and J. Davis, “Action Recognition Using Temporal Templates”
- N. Goddard, “Human Activity Recognition”
- K. Rohr, “Human Movement Analysis Based on Explicit Motion Models”

Contents (contd.)

- Gesture Recognition and Facial Expression Recognition
  - A. Bobick and A. Wilson, “State-Based Recognition of Gestures”
  - T. Starner and A. Pentland, “Real-Time American Sign Language Recognition from Video Using Hidden Markov Models”
  - M. Black , Y. Yacoob and S. Ju, “Recognizing Human Motion Using Parameterized Models of Optical Flow”

Contents (contd.)

  - I. Essa and A. Pentland, “Facial Expression Recognition Using Image Motion”
- Lipreading
  - C. Bregler and S. Omohumdro, “Learning Visual Models for Lipreading”
  - A. Goldschen, O. Garcia and E. Petajan, “Continuous Automatic Speech Recognition by Lipreading”
  - N. Li, S. Dettmer and M. Shah, “Visually Recognizing Speech Using Eigensequences”

A Framework for the Design of Visual Event Detectors

Niels Haering

Motivation

- Communication between humans and computers
  - a word is worth a thousand pixels
- Image / video understanding
  - object recognition, motion analysis, scene interpretation, event detection/recognition, content abstraction
- Image / video retrieval
  - index into large image and video databases
- Compression
  - MPEG7

A Framework for the Design of Visual Event Detectors

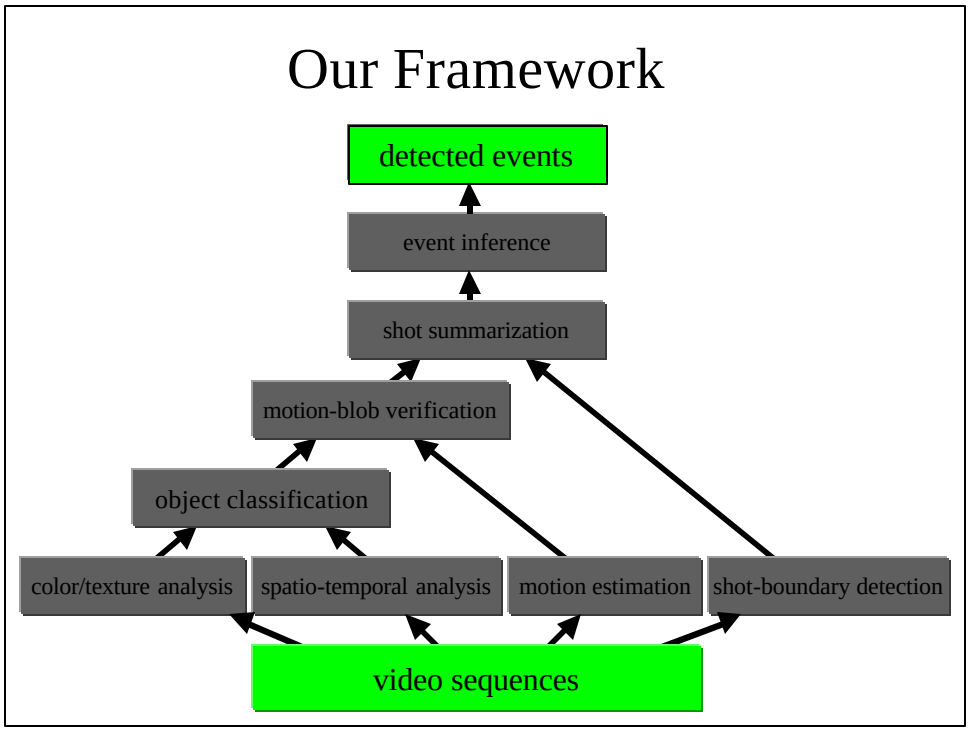
Niels Haering

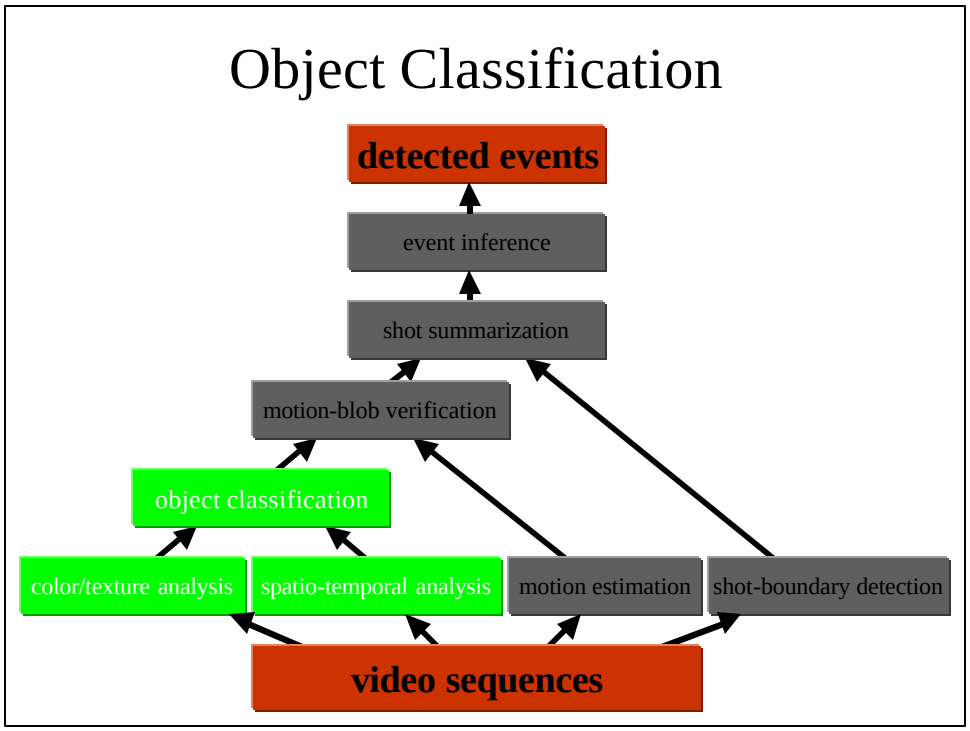
Motivation

- Communication between humans and computers
  - a word is worth a thousand pixels
- Image / video understanding
  - object recognition, motion analysis, scene interpretation, event detection/recognition, content abstraction
- Image / video retrieval
  - index into large image and video databases
- Compression
  - MPEG7

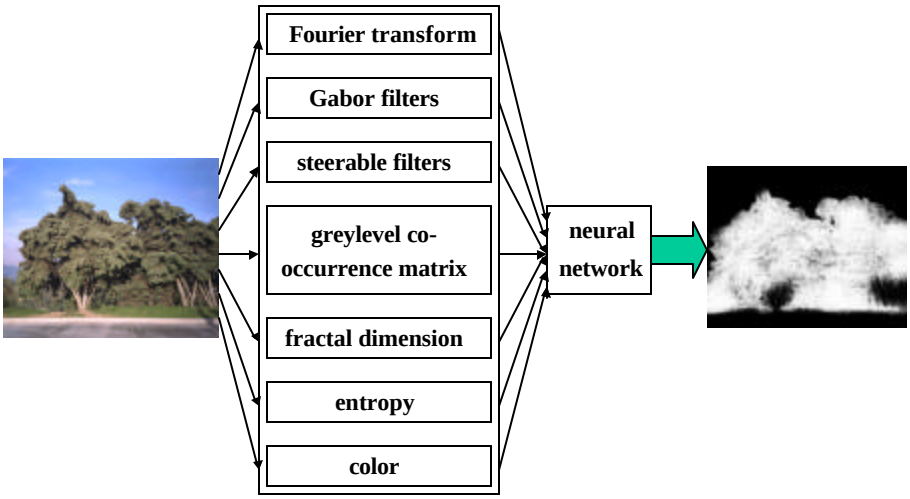
A Framework for the Design of Visual Event Detectors

- Rich internal representation of the world
- Hierarchy of abstractions
- Meaningful event summaries

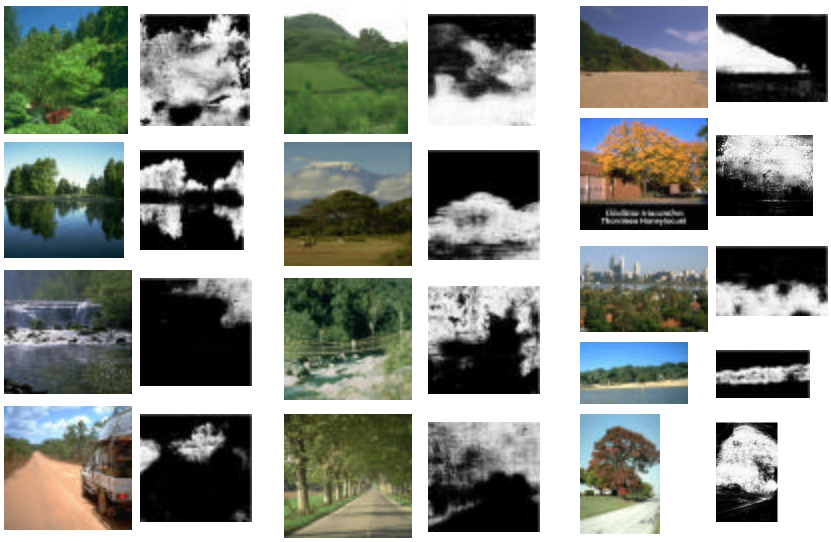

Our Framework
detected events
event inference
shot summarization
motion-blob verification
object classification
color/texture analysis
spatio-temporal analysis
motion estimation
shot-boundary detection
video sequences


Object Classification
detected events
event inference
shot summarization
motion-blob verification
object classification
color/texture analysis
spatio-temporal analysis
motion estimation
shot-boundary detection
video sequences

Object Classification

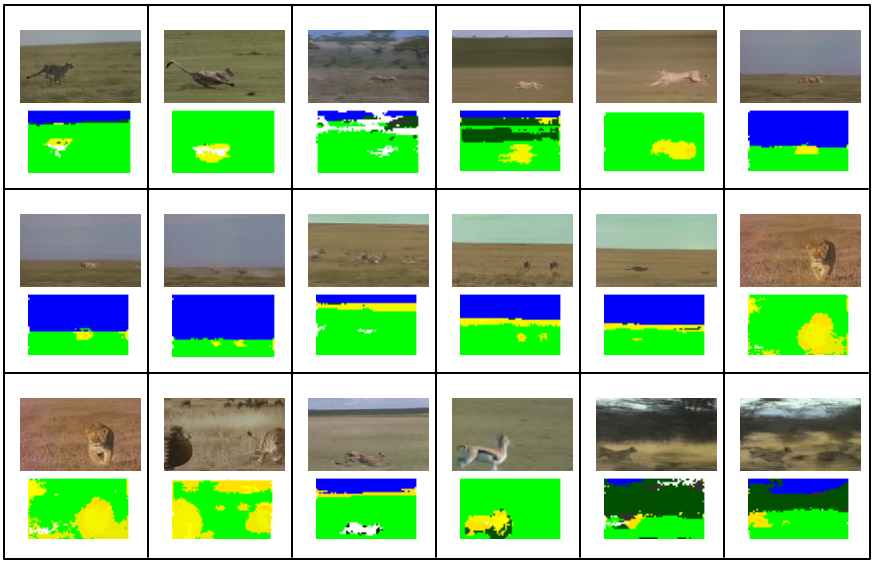


Deciduous Trees

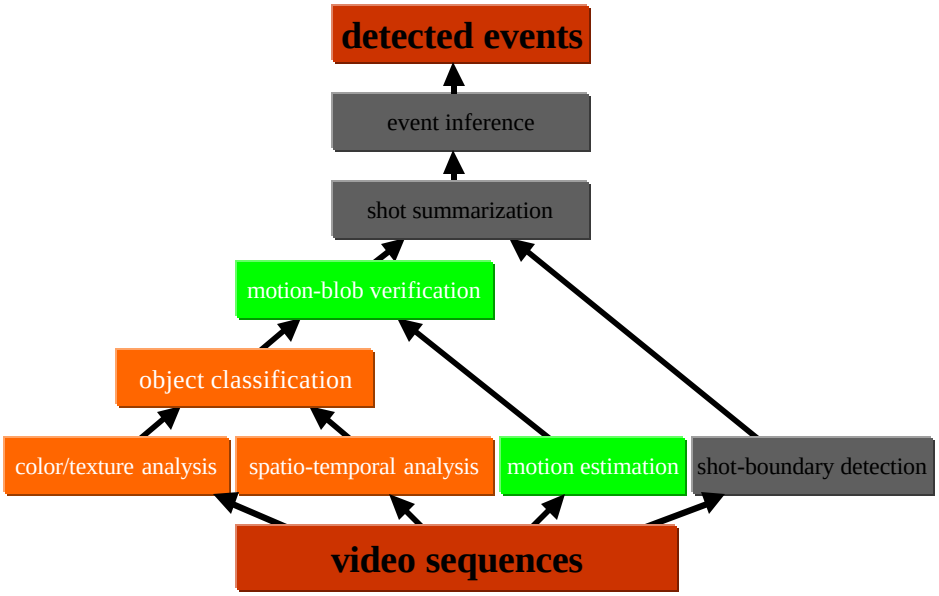


Sky

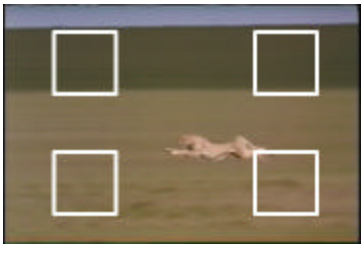
Animals, Sky, Grass, Trees, Rock

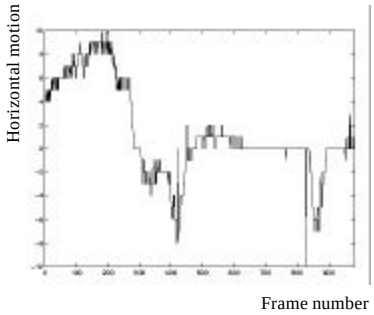


Motion-blob Verification



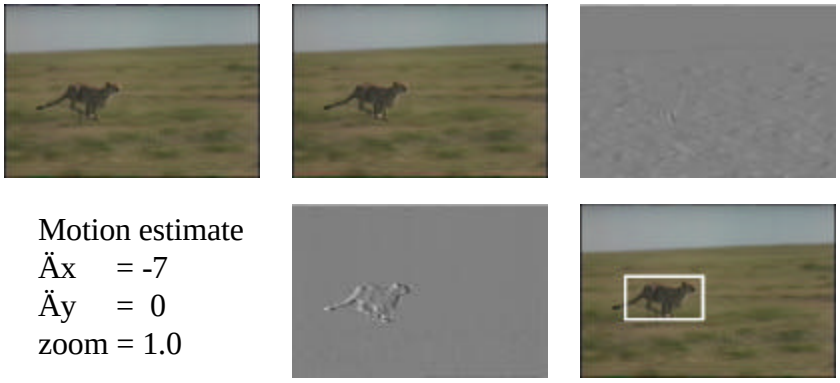
Motion Estimation





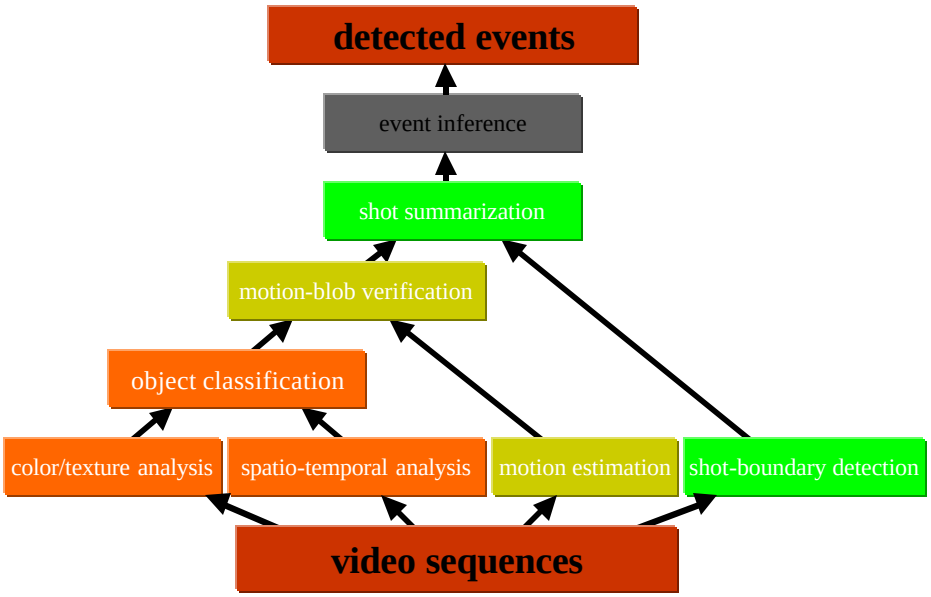
- three parameter system: x-, y-translation, and zoom,
- 4 motion estimates based on pyramid,
- 4 motion estimates based on previous best match,
- “texture” measure prevents ambiguous matches

Motion-blob detection



Motion estimate
Äx = -7
Äy = 0
zoom = 1.0

Shot Summarization



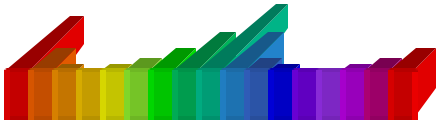
Shot Detection

Characteristics of shot boundaries:

- Change of camera/viewpoint
- Change of color characteristics

4 Bins for Value
4 Bins for Saturation
8 bins for hue

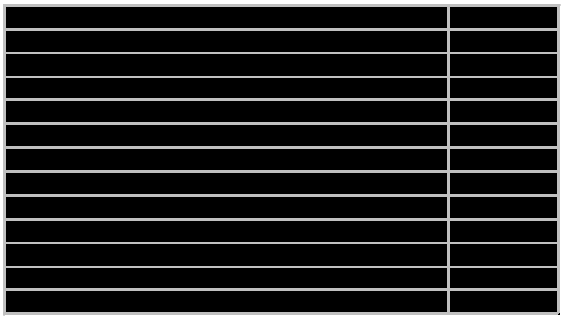
Frame 1

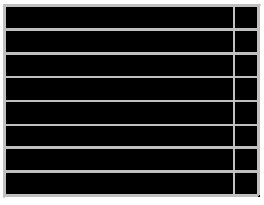


Frame 2

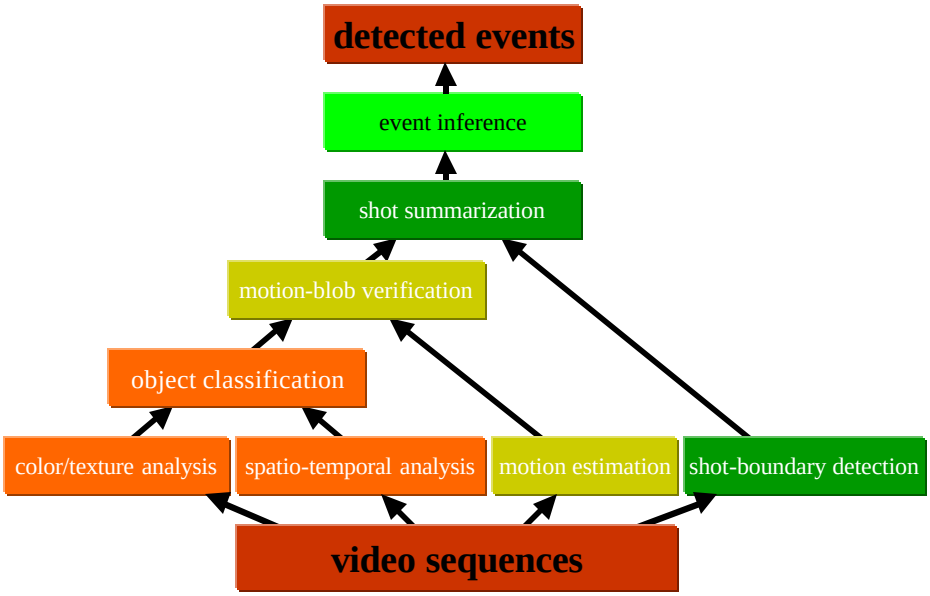
= 0.79

Shot Summaries

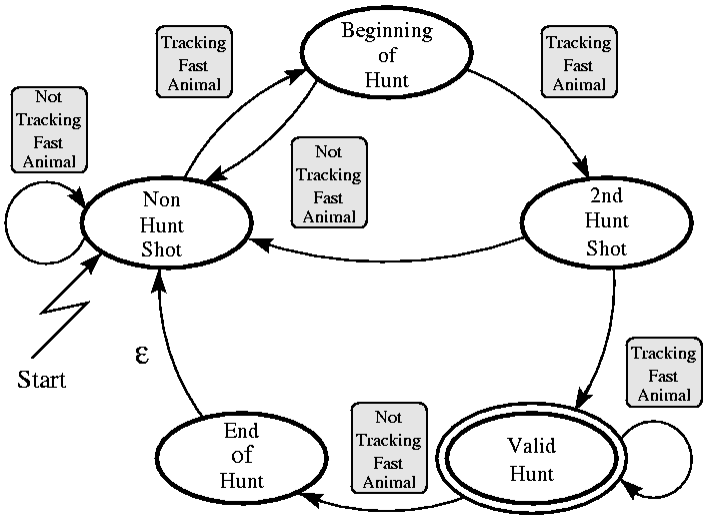




Event Inference



Hunt events



Hunts

Hunt

Non-hunt

Hunts

Non-hunt

Hunt

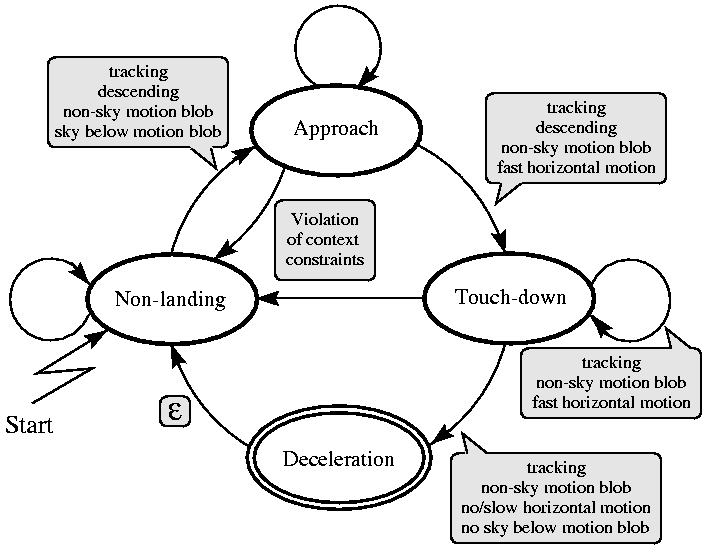
Non-hunt

Event Detection

[redacted]

Landing Events

Landing Events



Landing Events

Non-landing

Approach

Touch-down

Deceleration

Non-landing

Landing Events

Non-landing

Approach

Touch-down

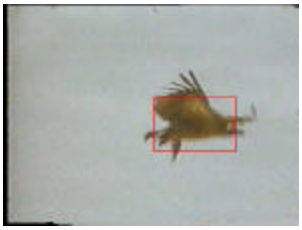
Deceleration

Non-landing

Landing Events

Non-landing

Approach



Touch-down

Deceleration

Non-landing

Conclusions

- Many natural objects are easily recognized by their color and texture signatures (shape is often not needed)
- Many events are easily detected and recognized by the classes of the comprising objects and their approximate motions
- The proposed visual event detection is robust to changes in scale, color, shape, occlusion, lighting conditions, view points and distances, and image compression