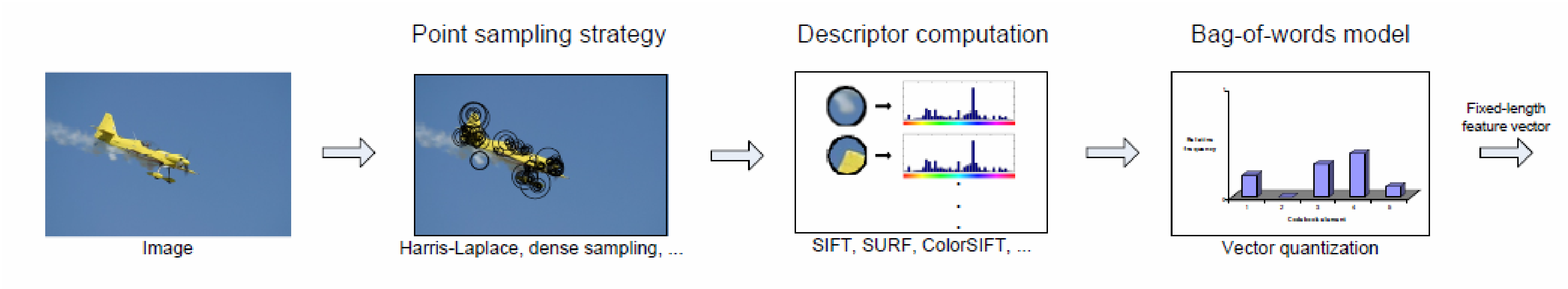
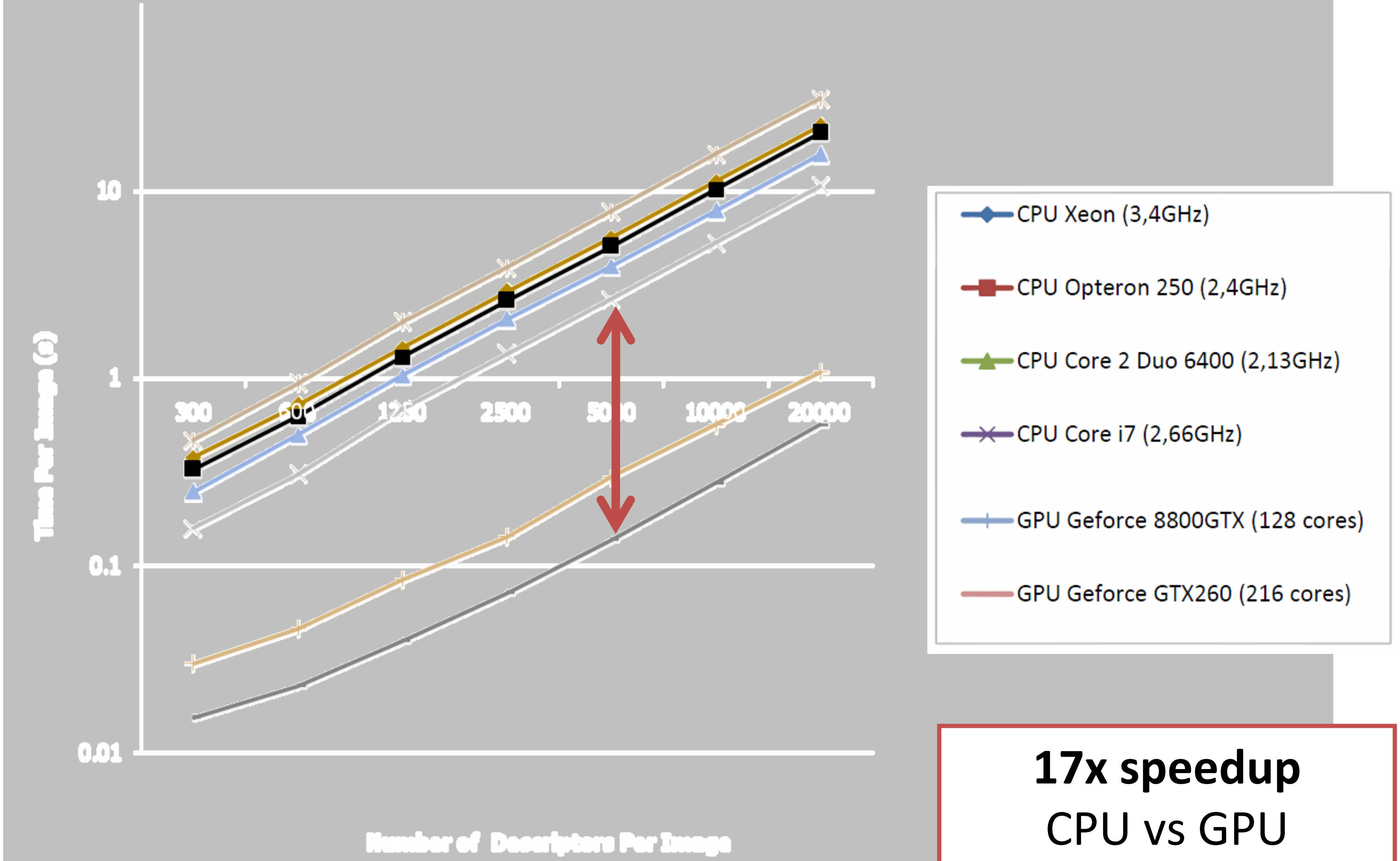
GPU-Accelerated Feature Extraction

- Single bag-of-words feature up to 15s/frame (CPU-time)
- TRECVid 2008 / PASCAL VOC 2008 consortium entries used 10 of these features
- More than 80% of time spent in vector quantization


Point sampling strategy
Descriptor computation
Bag-of-words model
Image
Harris-Laplace, dense sampling, ...
SIFT, SURF, ColorSIFT, ...
Vector quantization
Fixed-length feature vector

Vector Quantization Timings for ColorSIFT

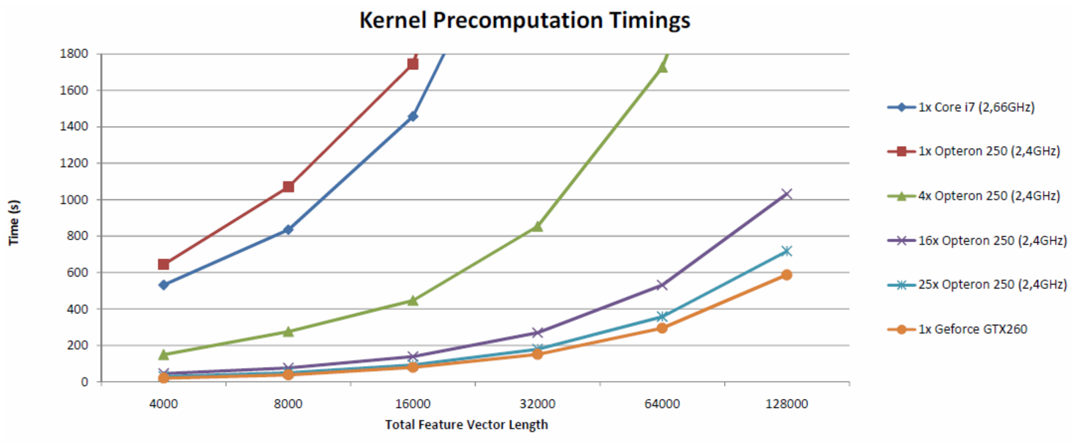


17x speedup

CPU vs GPU

Kernel Value Precomputation

- Step from image feature vectors to kernel-based classifiers from WP5 (SVM/SR-KDA)
- Computes χ^2 distance between pairs of images
- Suitable for GPU implementation: **22x speedup**
- TRECVid 2008 processing time: 800 CPU hours vs. 37 GPU hours



⇒ Process datasets order of magnitude larger

or

⇒ Single GPU replaces medium-sized cluster

Overview

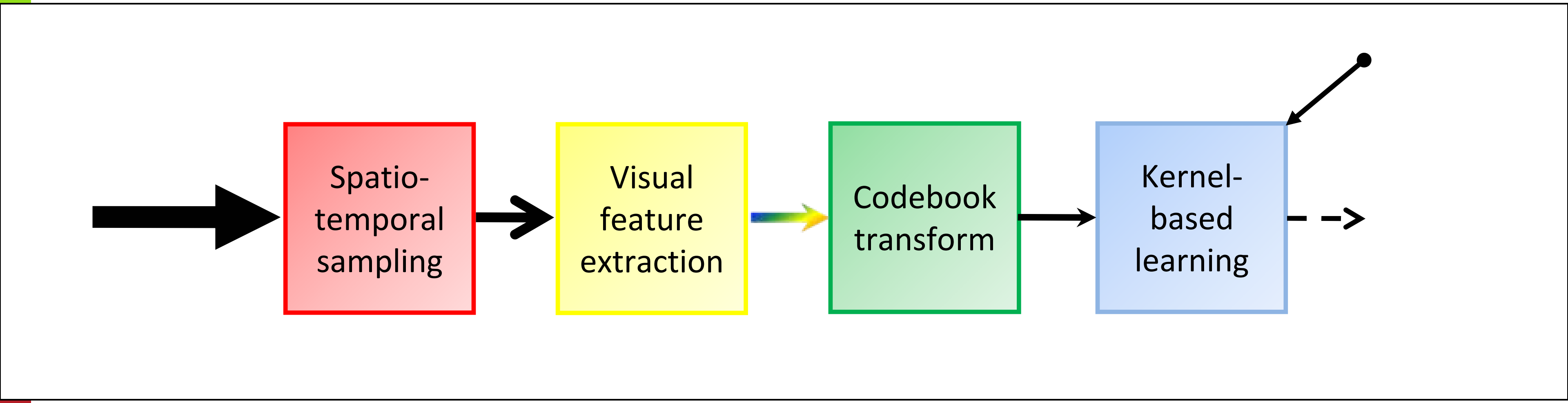
1. ***Motivation***
    - Challenges
    - Computer vision demands
2. ***Pixel-based detection***
    - Histograms
    - Density estimation
3. ***Interest point detection***
    - Local image structures (edges and corners)
    - Harris Laplace
    - Color boosted
4. ***Descriptors***
    - SIFT
    - Extension to color
5. ***Object recognition (VOC/TRECVid)***
    - Dense and point sampling
    - Code book generation
    - Results
6. ***Applications***
    - Tracking in video
    - Object replacement
    - Emotion recognition
    - Head pose estimation

TRECVid

Koen van de Sande
Cees Snoek
Jan van Gemert
Jasper Uijlings
Jan-Mark Geusebroek
Theo Gevers
Arnold Smeulders

University of Amsterdam

Concept Detection Stages



Spatio-temporal sampling

Visual feature extraction

Codebook transform

Kernel-based learning

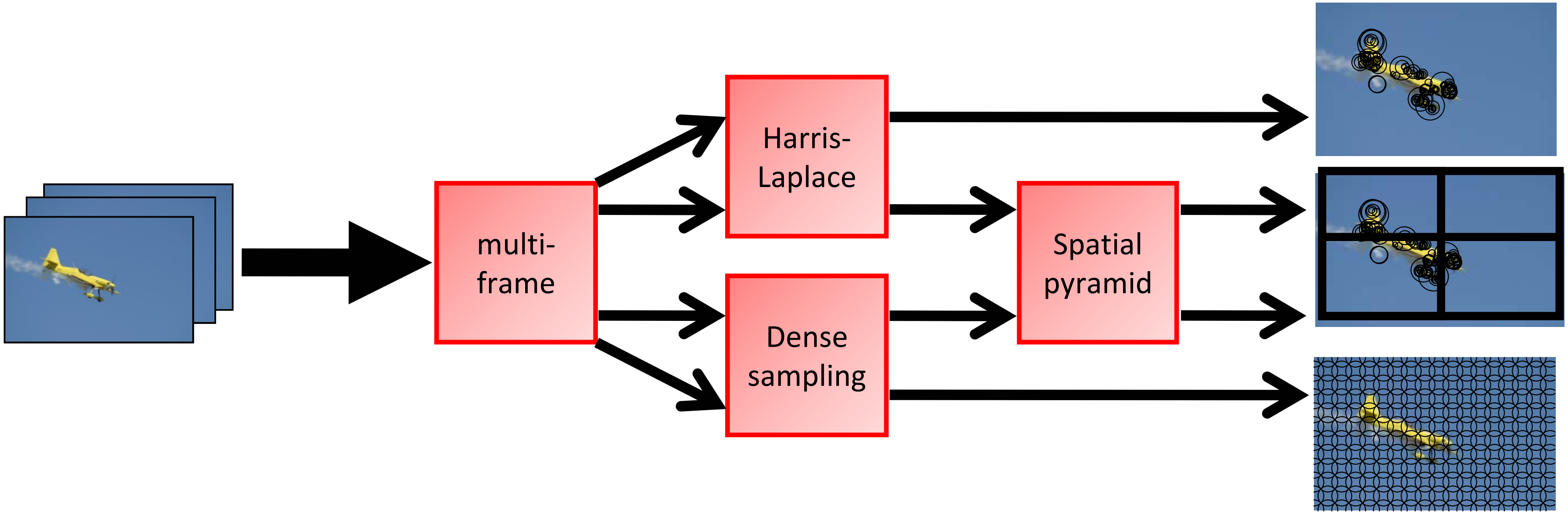
[Lazebnik, CVPR06]
[Marszalek, VOC07]

Spatio-Temporal Sampling

- Spatial pyramid

1x1	whole image
2x2	image quarters
1x3	horizontal bars

- Temporal analysis of up to 5 frames per shot



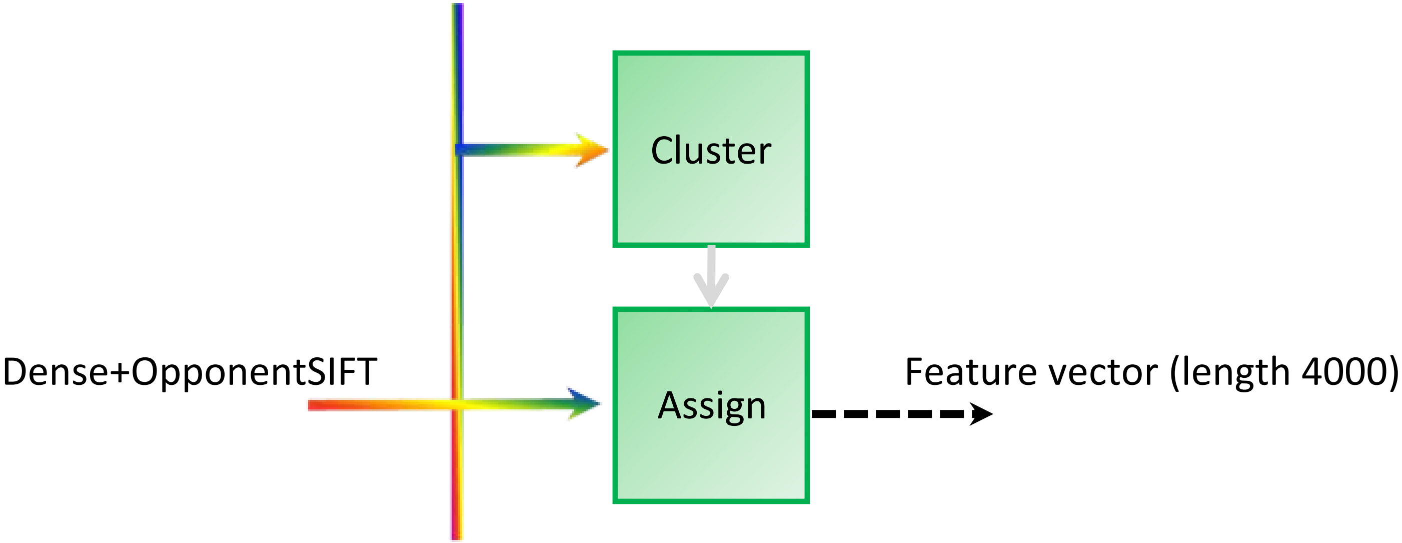
multi-frame

Harris-Laplace

Dense sampling

Spatial pyramid

Visual Codebook Model



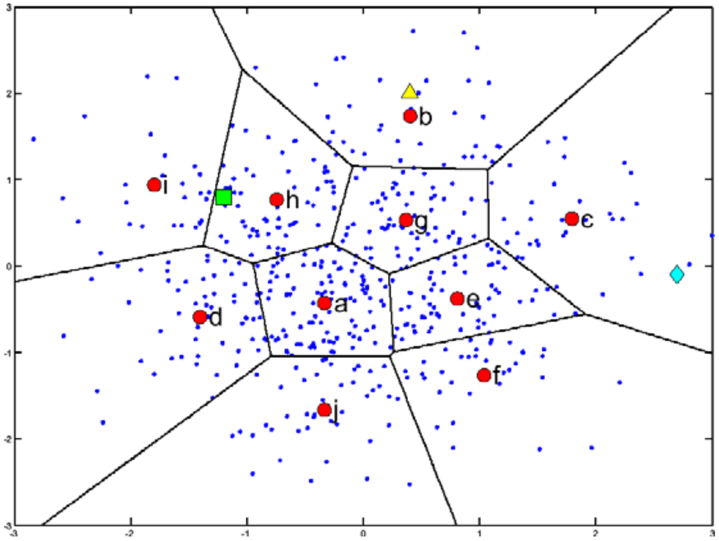
- Codebook consists of codewords
- Constructed with k-means clustering on descriptors
- We use 4,000 codewords per codebook

[van Gemert, ECCV08]

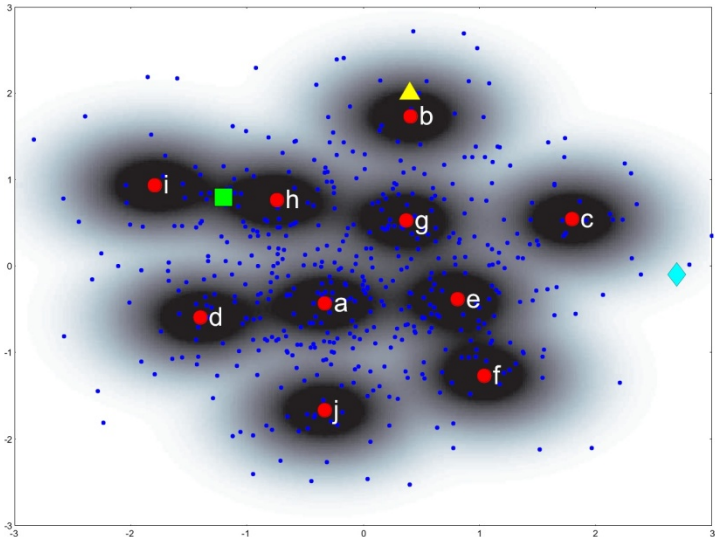
● Codeword

Codebook Assignment

Soft assignment using Gaussian kernel



Hard assignment

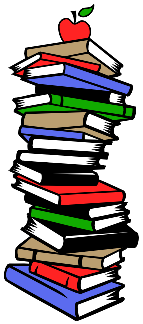


Soft assignment

Assignment	MAP on TV2007test
Hard	0,155
Soft	0,166

relative +7%

Codebook Library



Codebook	Sampling method	Descriptor	Construction	Assignment
#1	Dense	OpponentSIFT	K-means	Soft
#2	Harris-Laplace	SIFT	Radius-based	Soft
#3	Dense	*rg*SIFT	K-means	Hard
…	Dense	C-SIFT	K-means	Hard

Single codebook depends on

- Sampling method
- Descriptor
- Codebook construction method
- Codebook assignment

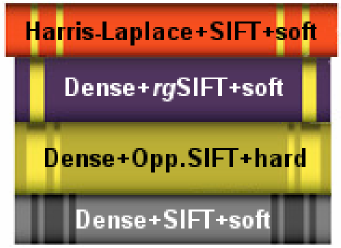
Codebook library is…

- a configuration of several codebooks

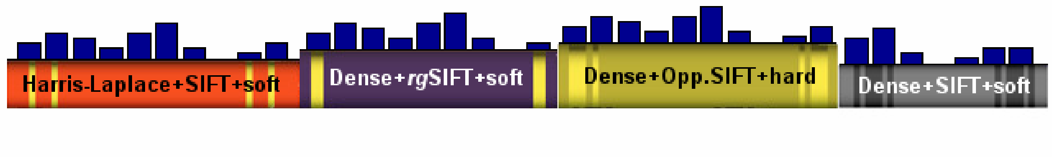
Codebook Library (cont'd)

For a frame:

- Each codebook in the library has feature vector of length 4,000



- Final feature vector is concatenation (4 books ~ length 16,000)

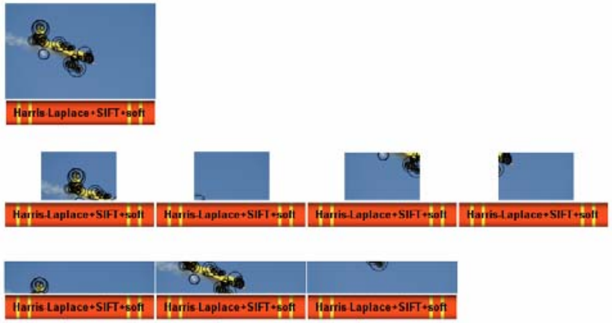


- Spatial pyramid adds more dimensions:

•1x1 4,000

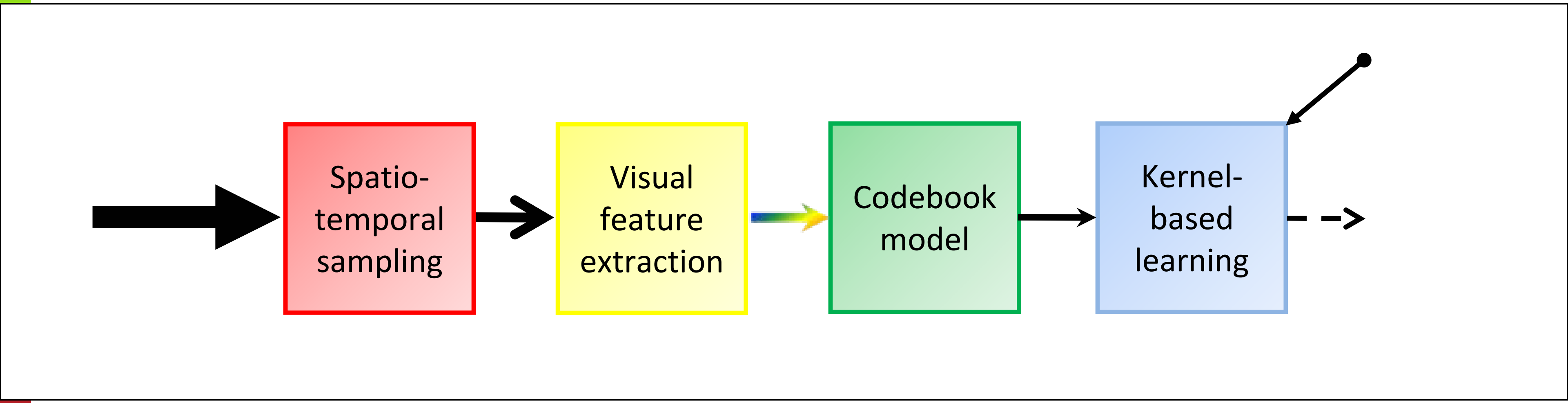
•2x2 16,000

•1x3 12,000



- Feature vector length easily >100,000…

Concept Detection Stages



[Snoek, ICME05]

Robust Temporal Approach

- No cloud computing yet: need to be efficient ☺
- Process 5 frames per shot in test set
- Linear increase in computation: x5

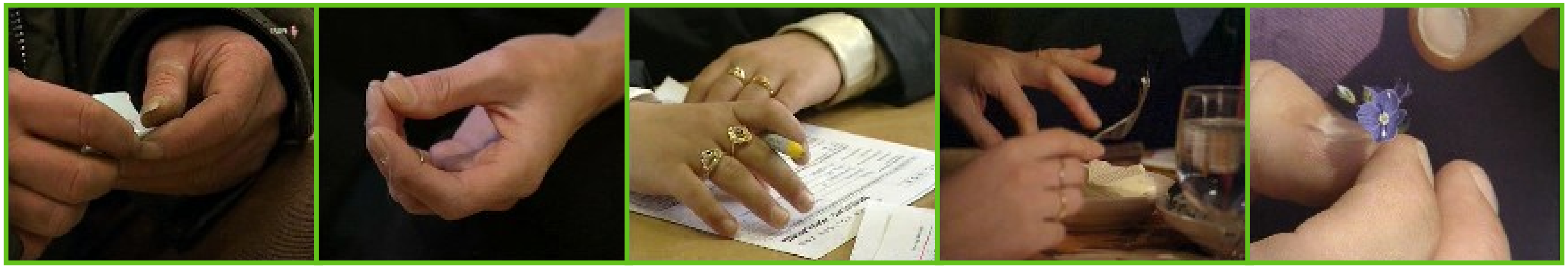
Codebook library	Frames/shot	MiAP on TV2008test
3x Color SIFT	1	0,152
3x Color SIFT	5	0,184

relative +20%

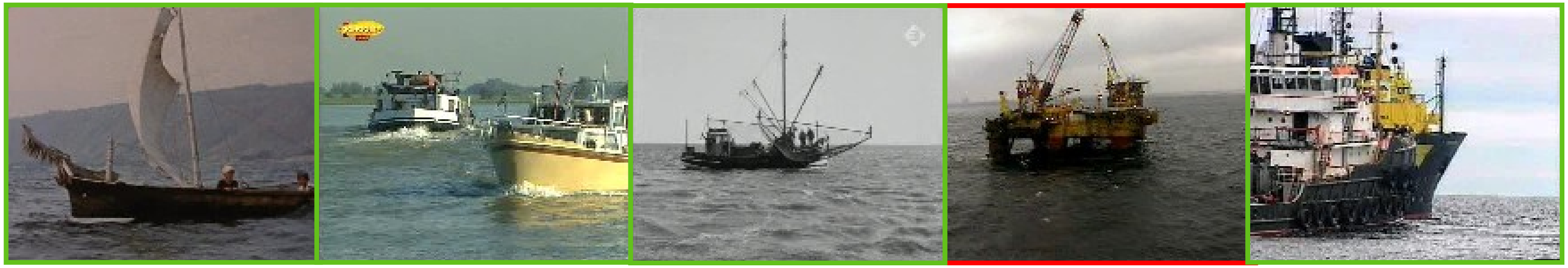
- In 2005 paper 7.5% to 38% improvement noted for multi-frame (worst-case vs. best-case using oracle)
- **Robust color SIFT *with* temporal = ~20% improvement**

The Good

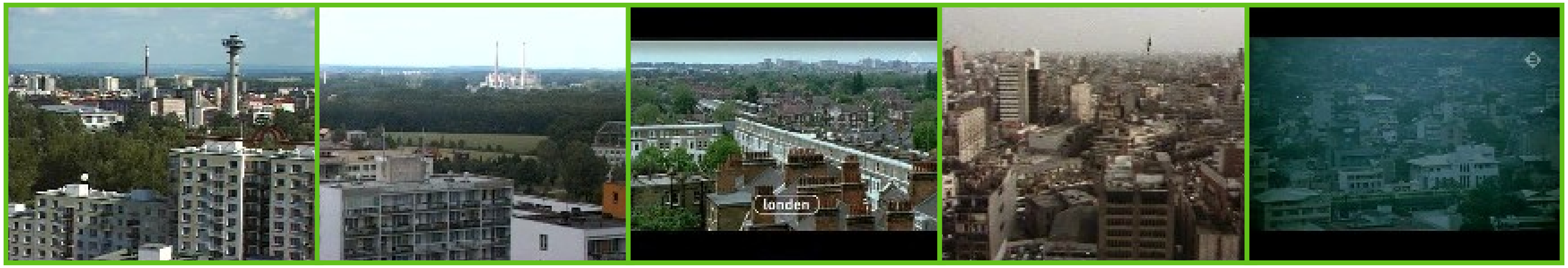
- Close-up of hands



- Boats and ships

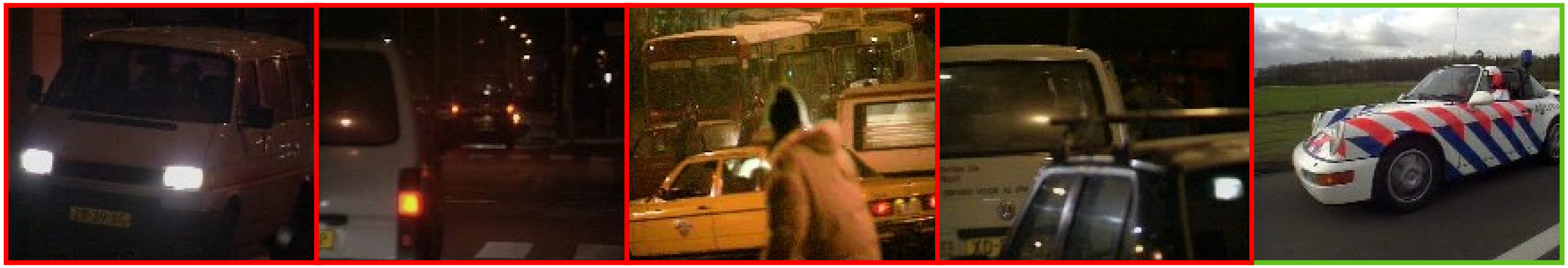


- Cityscape

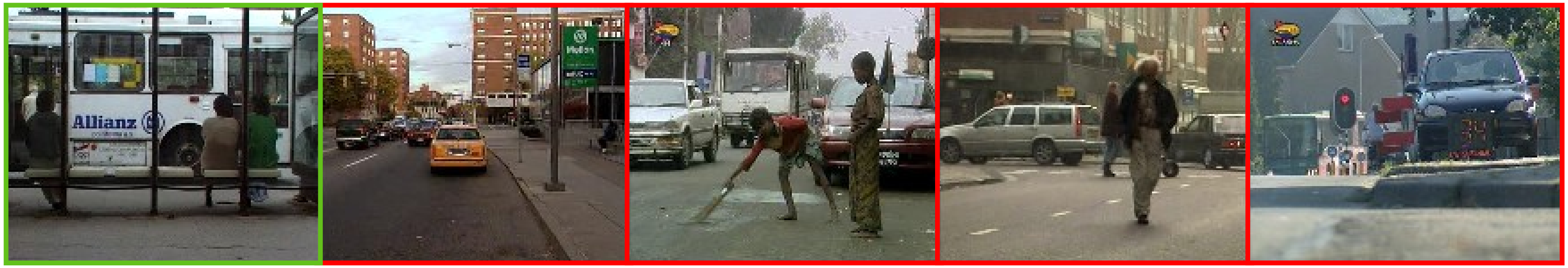


The Bad

- Emergency Vehicle (only 46 examples, many at night)

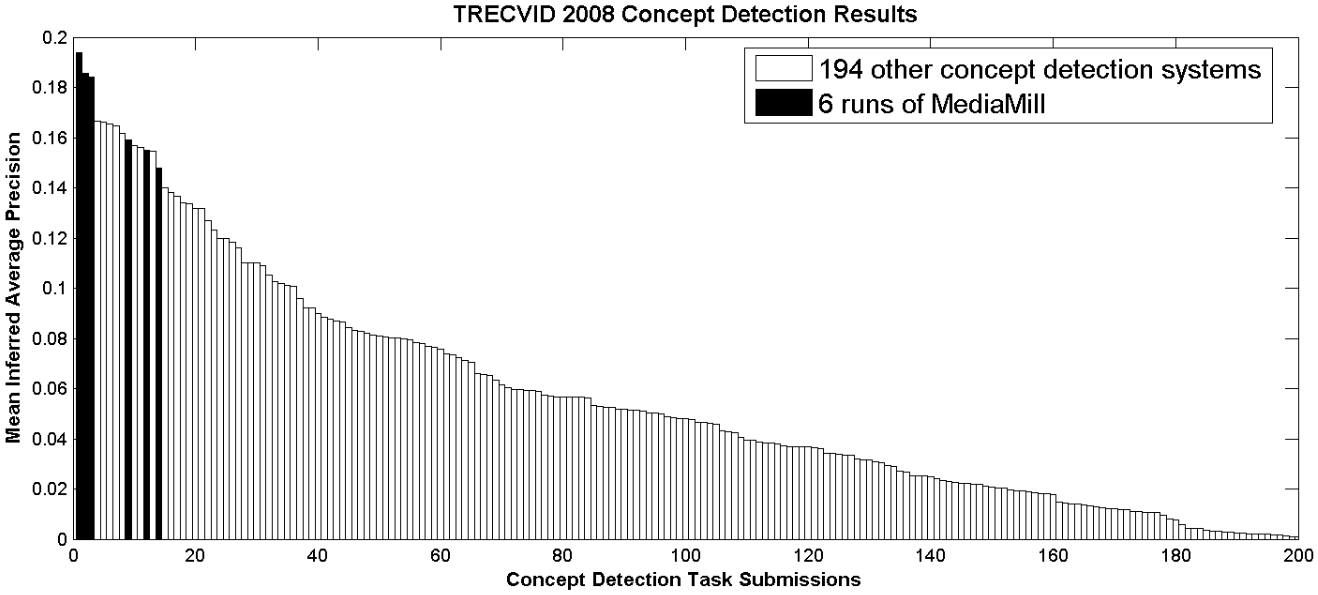


- Bus (only 64 examples)



Conclusions

- Illumination conditions affect concept detection
- SIFT+colorSIFT improves ~8%
- Soft codebook assignment improves ~7%
- Robust colorSIFT with simple multi-frame improves ~20%:
  - Room for more advanced methods in TRECVID 2009
- Precomputed kernel matrix reduces SVM computation time
- Near-duplicates from trailers hamper progress:
  - We suggest to exclude them, or count only once

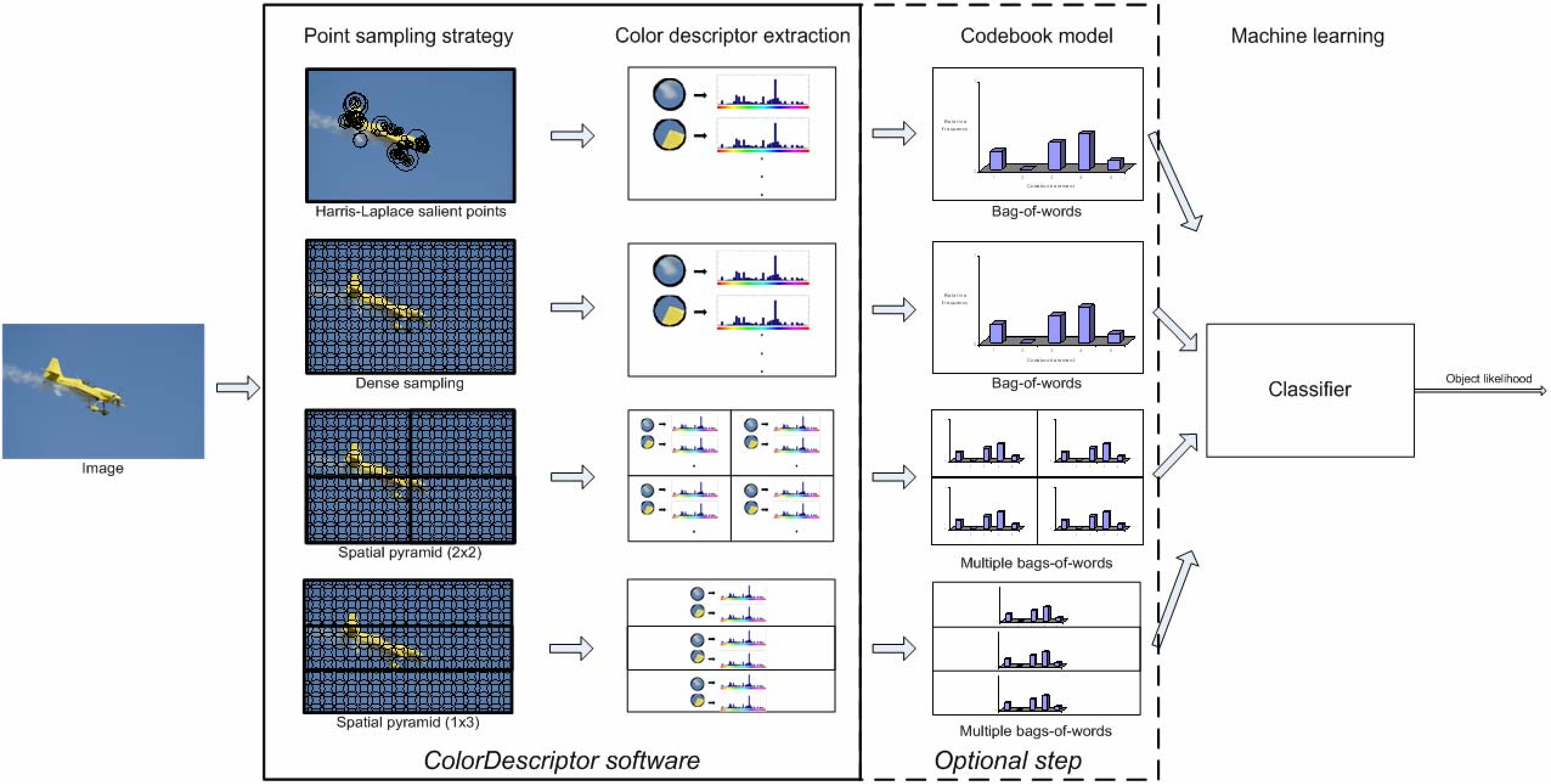


References

- K. E. A. van de Sande, T. Gevers and C. G. M. Snoek, "*Evaluation of Color Descriptors for Object and Scene Recognition*", CVPR 2008
- M. Marszalek, C. Schmid, H. Harzallah and J. van de Weijer, "*Learning Object Representations for Visual Object Class Recognition*", Visual Recognition Workshop in conjunction with ICCV 2007
- J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, A.W.M. Smeulders, "*Kernel Codebooks for Scene Categorization*", ECCV 2008
- K. Mikolajczyk and C. Schmid, "*A Performance Evaluation of Local Descriptors*", PAMI 2005
- D. G. Lowe, "*Distinctive Image Features from Scale-Invariant Keypoints*", IJCV 2004
- J. Zhang, M. Marszalek, S. Lazebnik and C. Schmid, "*Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study*", IJCV 2007
- C. G. M. Snoek et al, "*The MediaMill TRECVID 2008 Semantic Video Search Engine*", TRECVID Workshop 2008

ColorDescriptor software

for object and scene categorization



Created by Koen van de Sande
© University of Amsterdam

Visit http://colordescriptors.com for color descriptor software

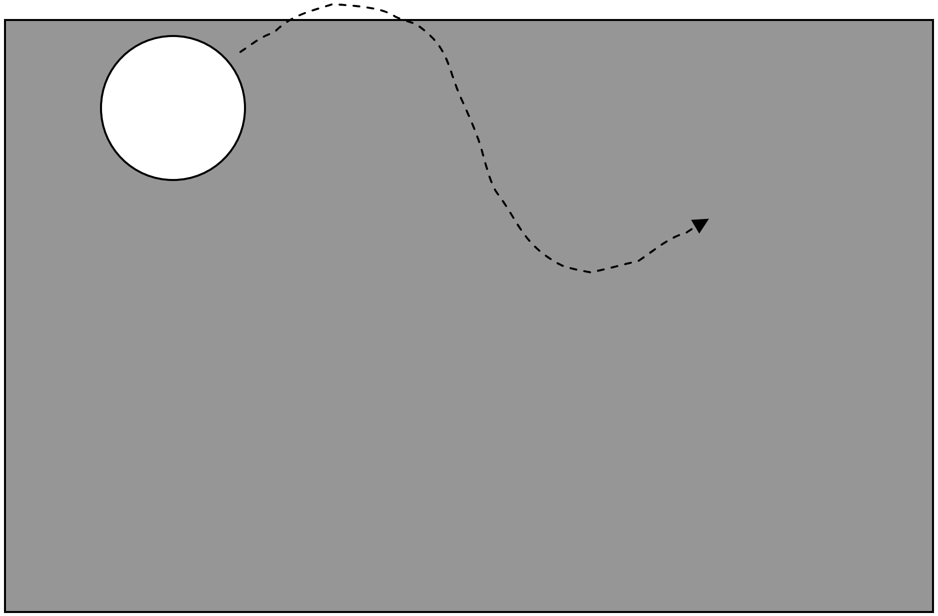
Overview

1. ***Motivation***
   - Challenges
   - Computer vision demands
2. ***Pixel-based detection***
   - Histograms
   - Density estimation
3. ***Interest point detection***
   - Local image structures (edges and corners)
   - Harris Laplace
   - Color boosted
4. ***Descriptors***
   - SIFT
   - Extension to color
5. ***Object recognition (VOC/TRECVid)***
   - Dense and point sampling
   - Code book generation
   - Results
6. ***Applications***
   - Tracking in video
   - Object replacement
   - Emotion recognition
   - Head pose estimation

Object tracking

Tracking

- Tracking can be very easy if both the target and the background are uniform in color.



Tracking

- **Background clutter**: the presence of other objects or non-informative patterns in the image complicates the detection of the right object.
- A **dynamic background**: moving camera.
- **Illumination change**: change in direction or intensity of light source, shadow…
- **Viewpoint change**: change of object pose or camera position.
- **Occlusion**: the target disappears partially or completely behind another object for a while.

Standard tracking algorithms

- **Background subtraction**.
- **Template tracking**:
  - SSD matching.
  - Correlation matching.
- **Mean-shift tracking**

Standard tracking algorithms

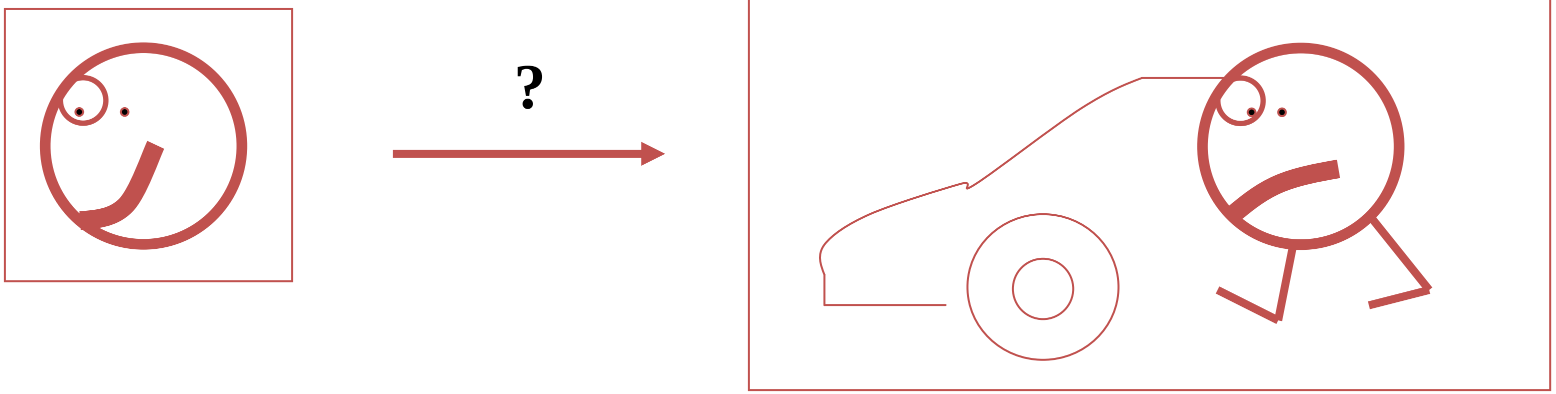
Template tracking

Mean-shift tracking

Tracking Objects based on Foreground-Background Separation

Template-based Tracking

- Tracking consists in searching for the target object in a frame by comparing with a **template** image.
- We assume that the template is fixed and given in advance.

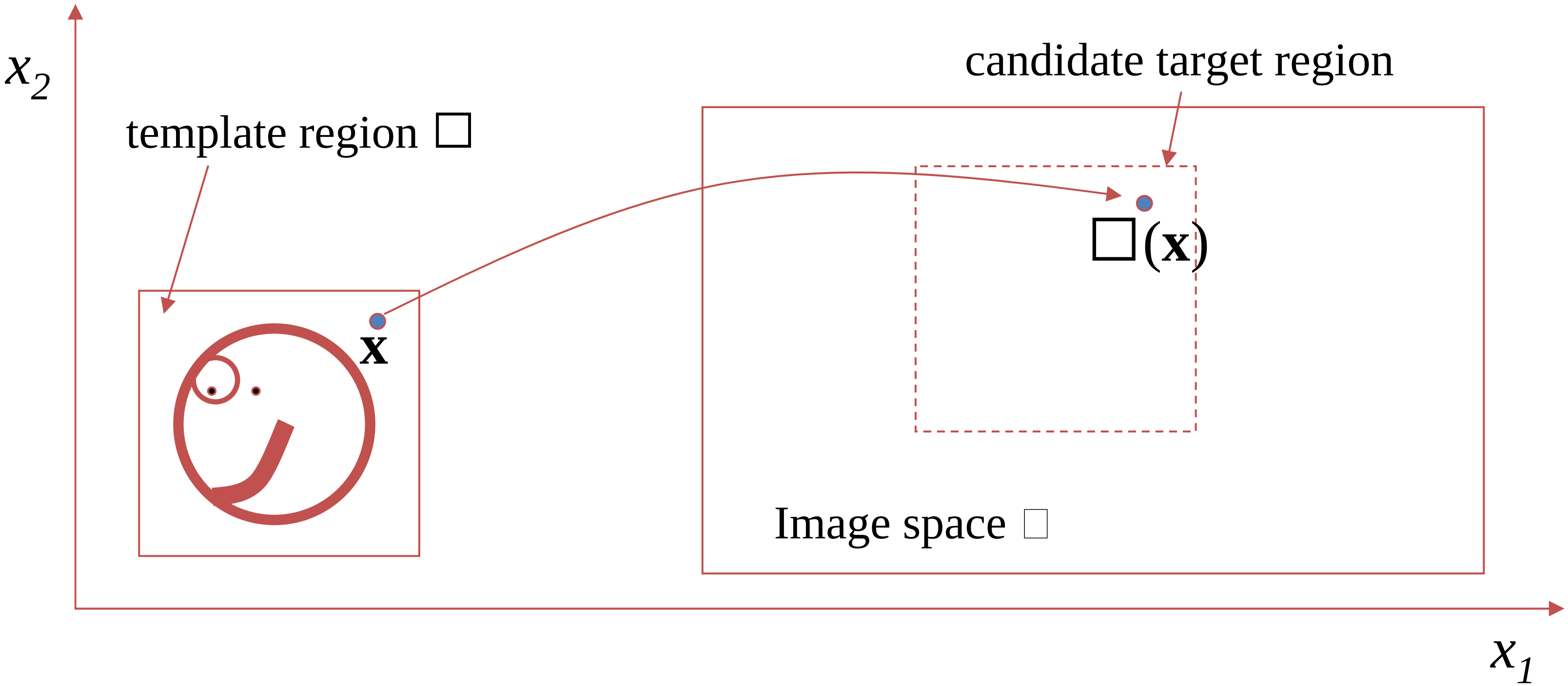


Template image
$T(\mathbf{x})$

$I(\mathbf{x},\mathbf{t})$

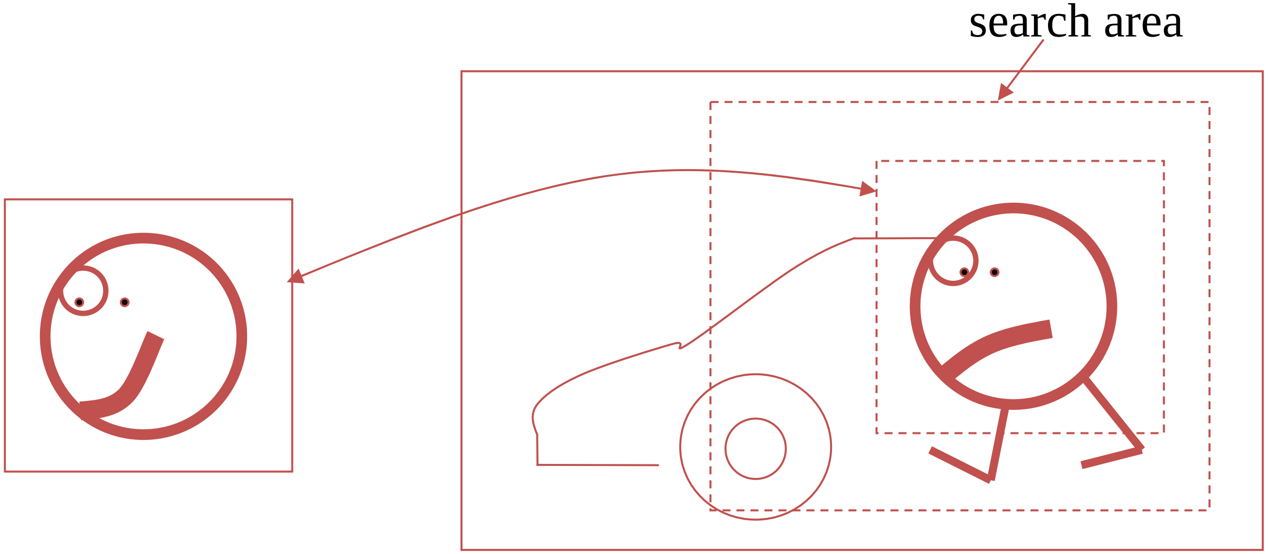
Teamplate to target transformation

- The template is mapped into a candidate target region the image using a transformation of coordinates: $\varphi(\mathbf{x})$: $\Omega \rightarrow \Sigma$. This transformation depends on a parameter vector **y**. Different candidate regions correspond to different values of **y**. So we write $\varphi(\mathbf{x}; \mathbf{y})$.



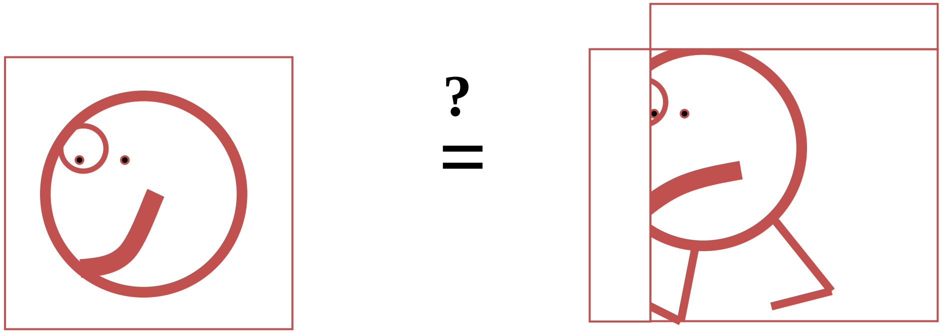
Search

- Align the template with every possible candidate region in the image, and find the most similar candidate according to a **similarity measure**.
- We search the target only in an area around the previous position exploiting general knowledge that the object won’t have moved far.



Similarity measure

- We need a measure of how similar (or far apart) the template and the candidate are.



- The similarity measure can be based on:
  - pixelwise intensity (color) difference: **SSD** and **correlation** trackers,
  - histogram difference: **mean-shift** tracker.

SSD and correlation

- SSD is short for sum-of-squared-difference:

$$D(\mathbf{y}) = \sum_{\mathbf{x}\in\Omega} [I(\mathbf{x}+\mathbf{y}) - T(\mathbf{x})]^2 \rightarrow \min_{\mathbf{y}}$$

- A simpler similarity measure is the (unnormalized) cross-correlation:

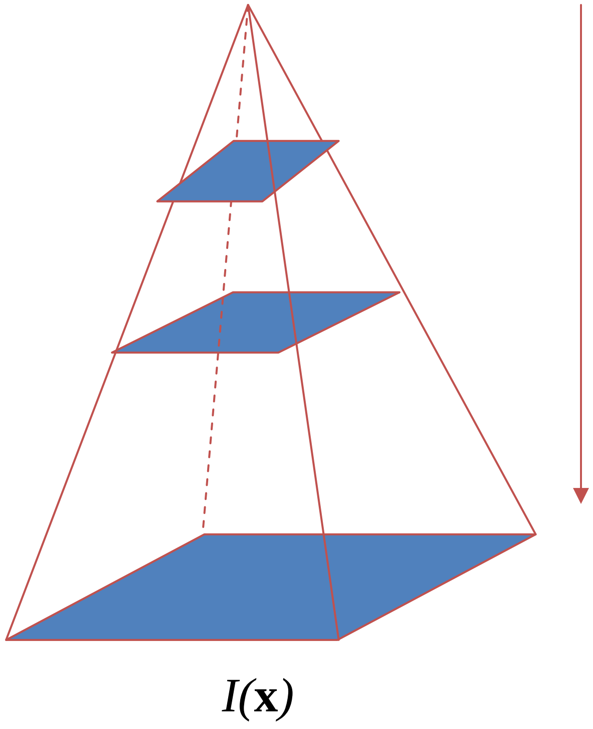
$$C(\mathbf{y}) = \sum_{\mathbf{x}\in\Omega} I(\mathbf{x}+\mathbf{y}) T(\mathbf{x}) \rightarrow \max_{\mathbf{y}}$$

Exhaustive search

- Calculate SSD for every $\mathbf{y}$ in a search window and choose the position with the least SSD.
- Strengths: robustness and simplicity in implementation.
- Weaknesses:
  - Computations could be time-consuming in case of a large search window.
  - Only suitable for translation.

Coarse-to-fine strategy

- Propagate the search results through different resolution levels using image pyramids.
- First search for the target in a low resolution and then use the result as initial point for the higher resolution.
- Able to overcome the issues of complexity and local minima:
  - Reduce complexity since images at low resolution have small sizes
  - At low resolution local minima are smoothed over.



Template tracking

Mean-shift tracking

Tracking Objects based on Foreground-Background Separation

Mean-shift tracking

- Features:
  - Target detection is performed by matching **weighted histograms**.
  - Very fast in comparison with SSD or correlation trackers,.

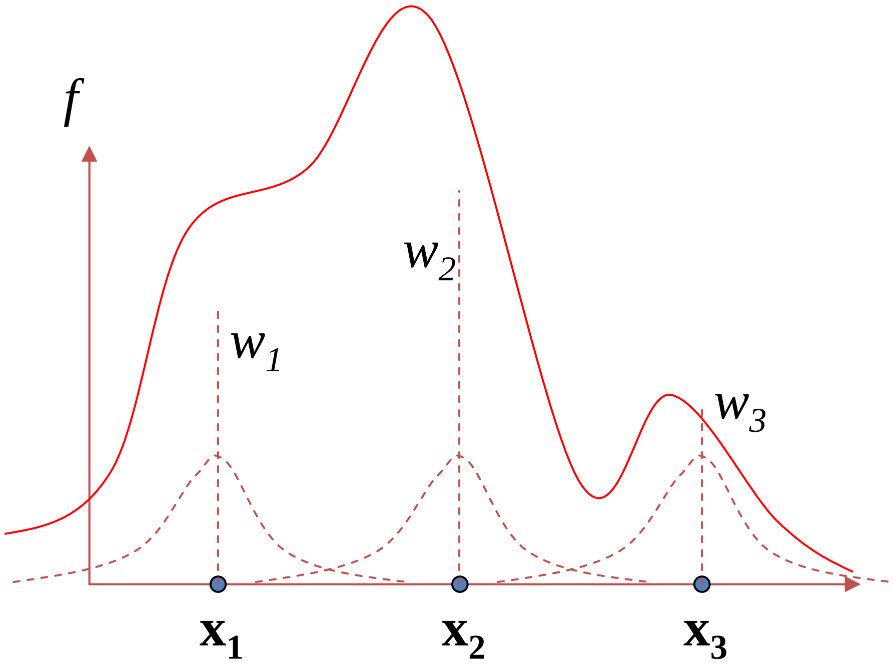
- Reference: Comaniciu et al. *Real time tracking of Non-Rigid Objects using Mean Shift*, In CVPR 2000.

Mean-shift algorithm

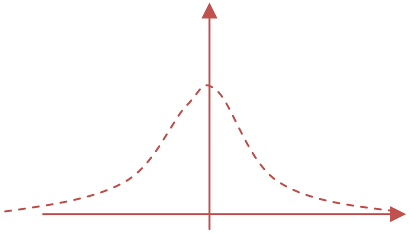
- The mean-shift algorithm finds a local maximum of a density function of the form:

$$f(\mathbf{y}) = \sum_i w_i K\left(\frac{|\mathbf{y}-\mathbf{x}_i|^2}{\sigma}\right)$$

- where K is the local kernel.



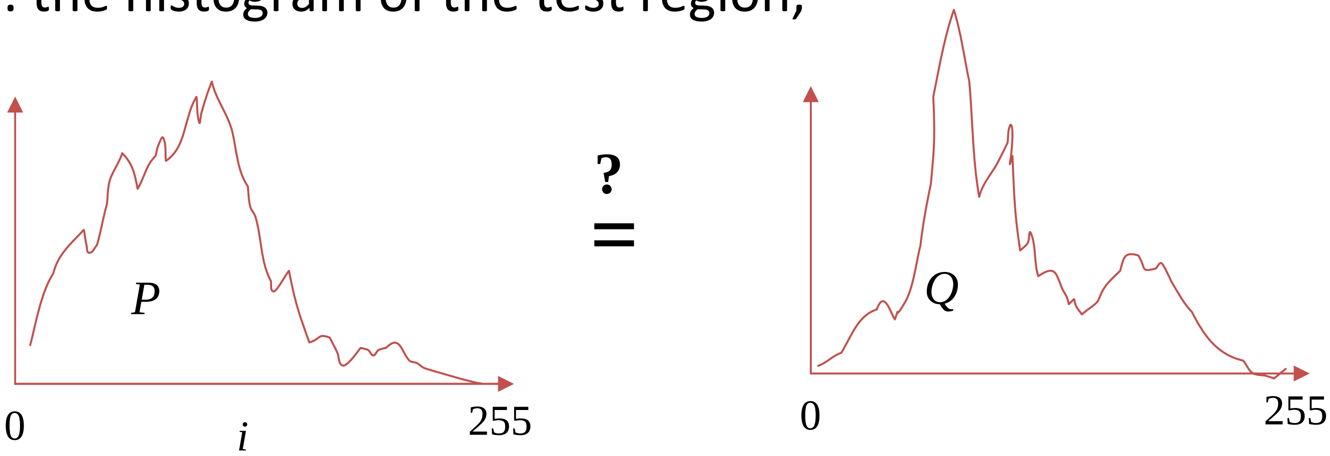
Gaussian kernel:



$$K(|\mathbf{x}|^2) = (2\pi)^{-d/2}\exp\left(-|\mathbf{x}|^2/2\right)$$

Similarity measure

- $P(i)$: the template histogram,
- $Q(i;\mathbf{y})$: the histogram of the test region,

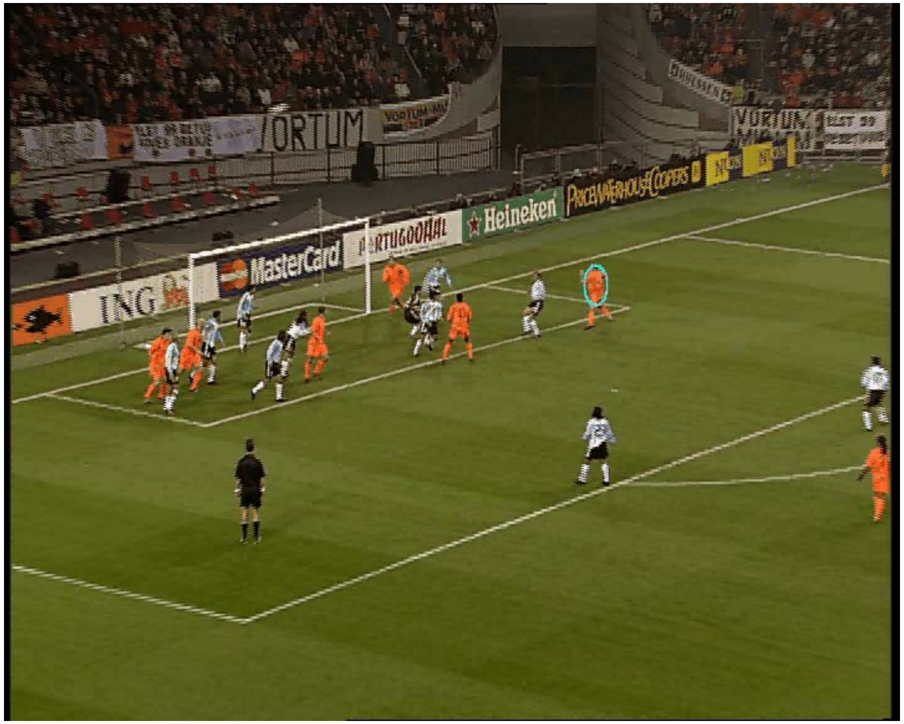


- *The Bhattacharyya coefficient* can measure the similarity between two distributions:

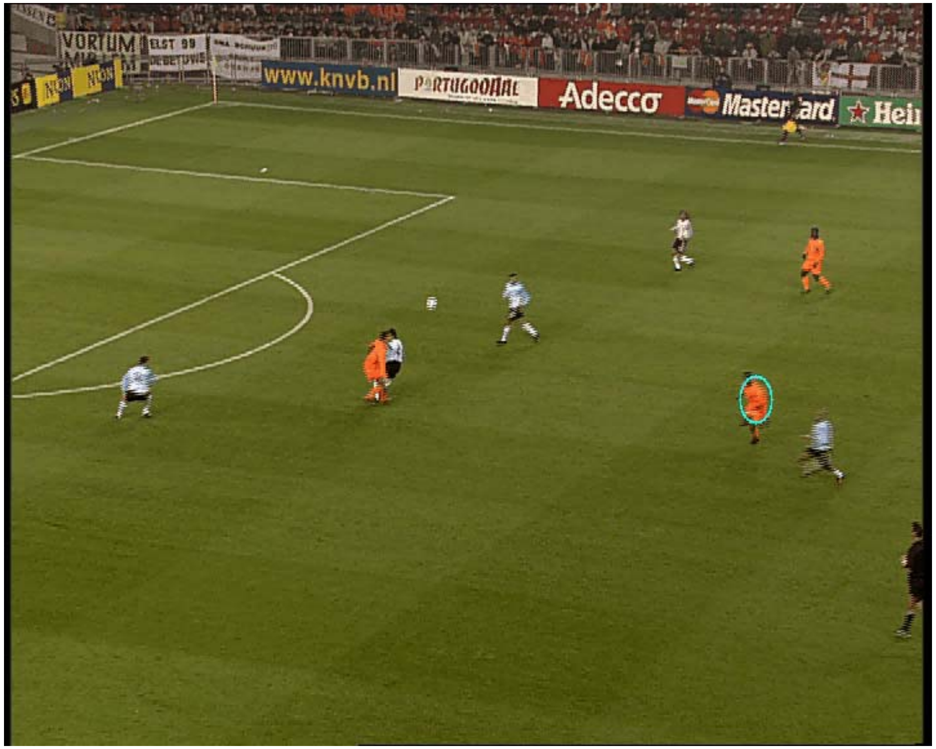
$$r(\mathbf{y}) = r(P, Q(\mathbf{y})) = \sum_{i=0}^{255} \sqrt{P(i)Q(i;\mathbf{y})} \to \max_{\mathbf{y}}$$

Color-based object tracking

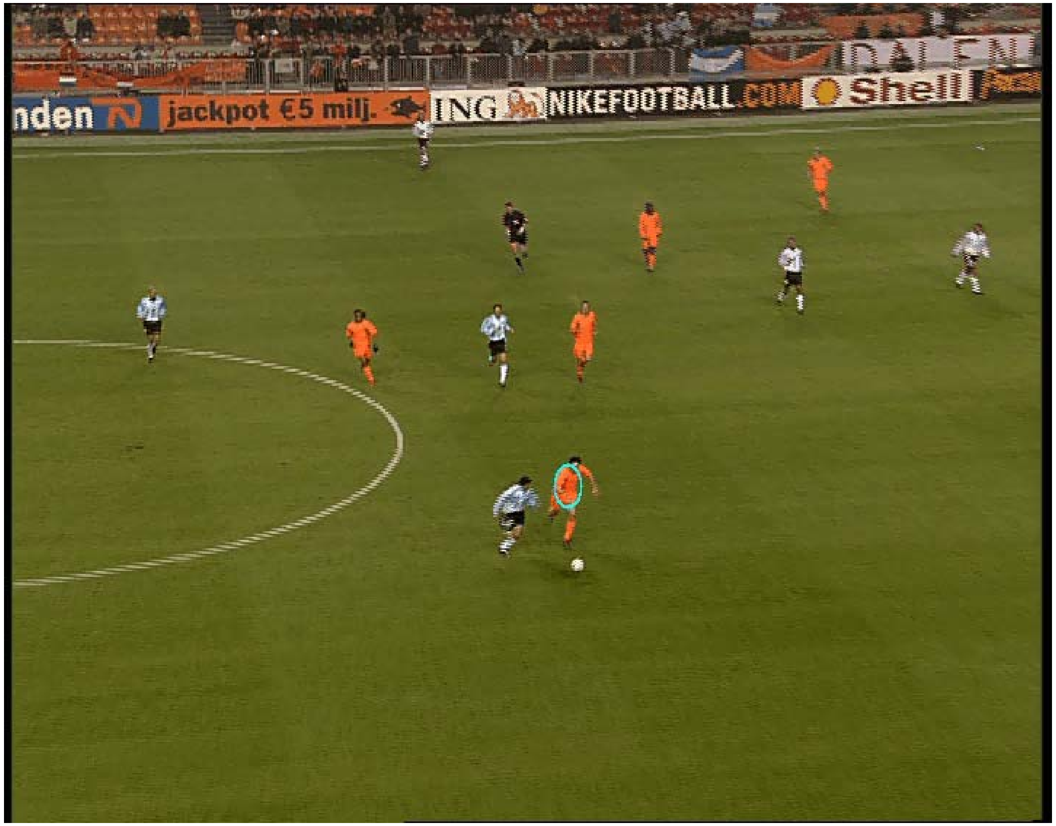
Player tracking



Player tracking with occlusion



Player tracking with occlusion



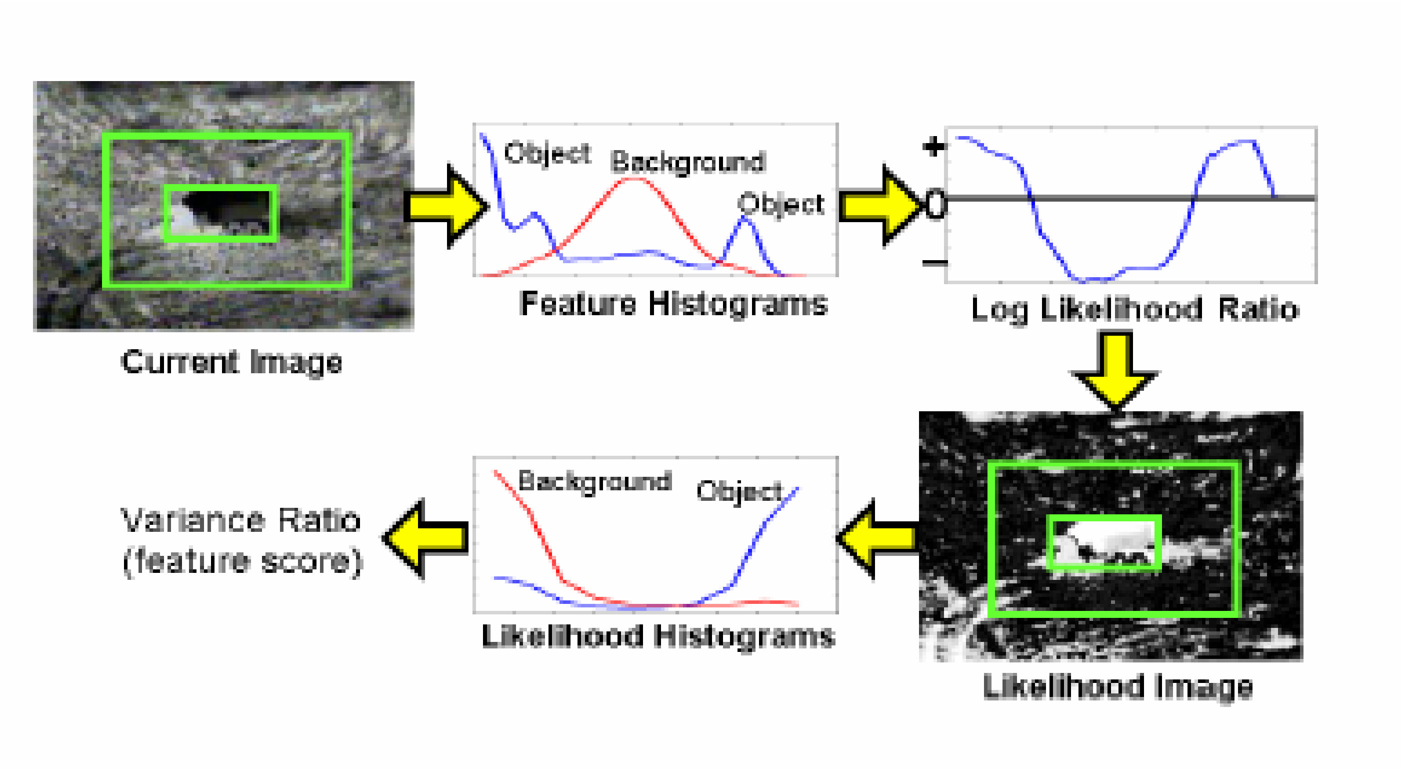
Template tracking

Mean-shift tracking

Tracking Objects based on Foreground-Background Separation

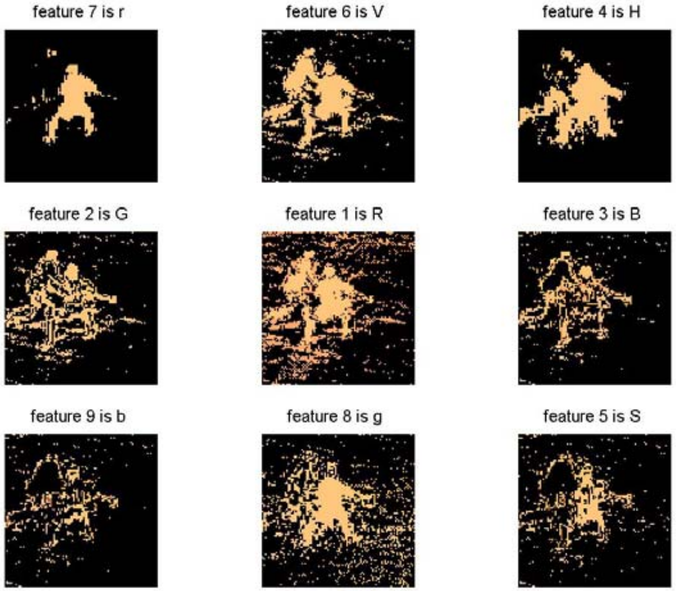
Algorithm

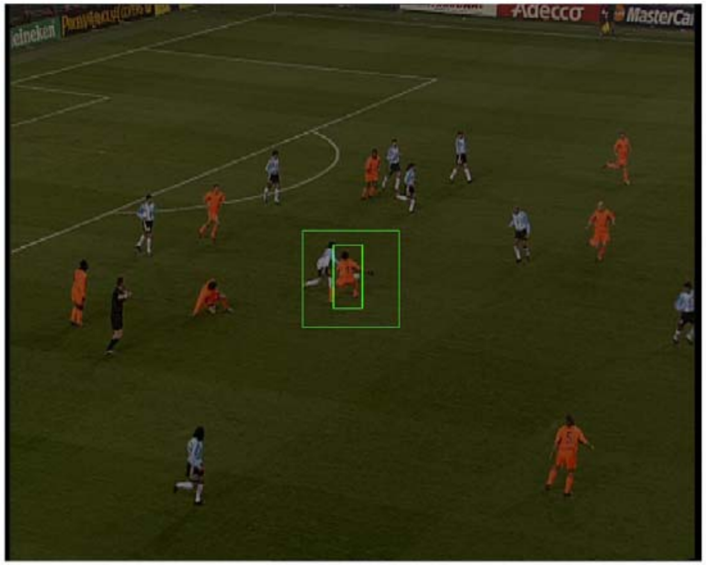
Figure taken from "Online Selection of Discriminating features" Collins and Liu



Results

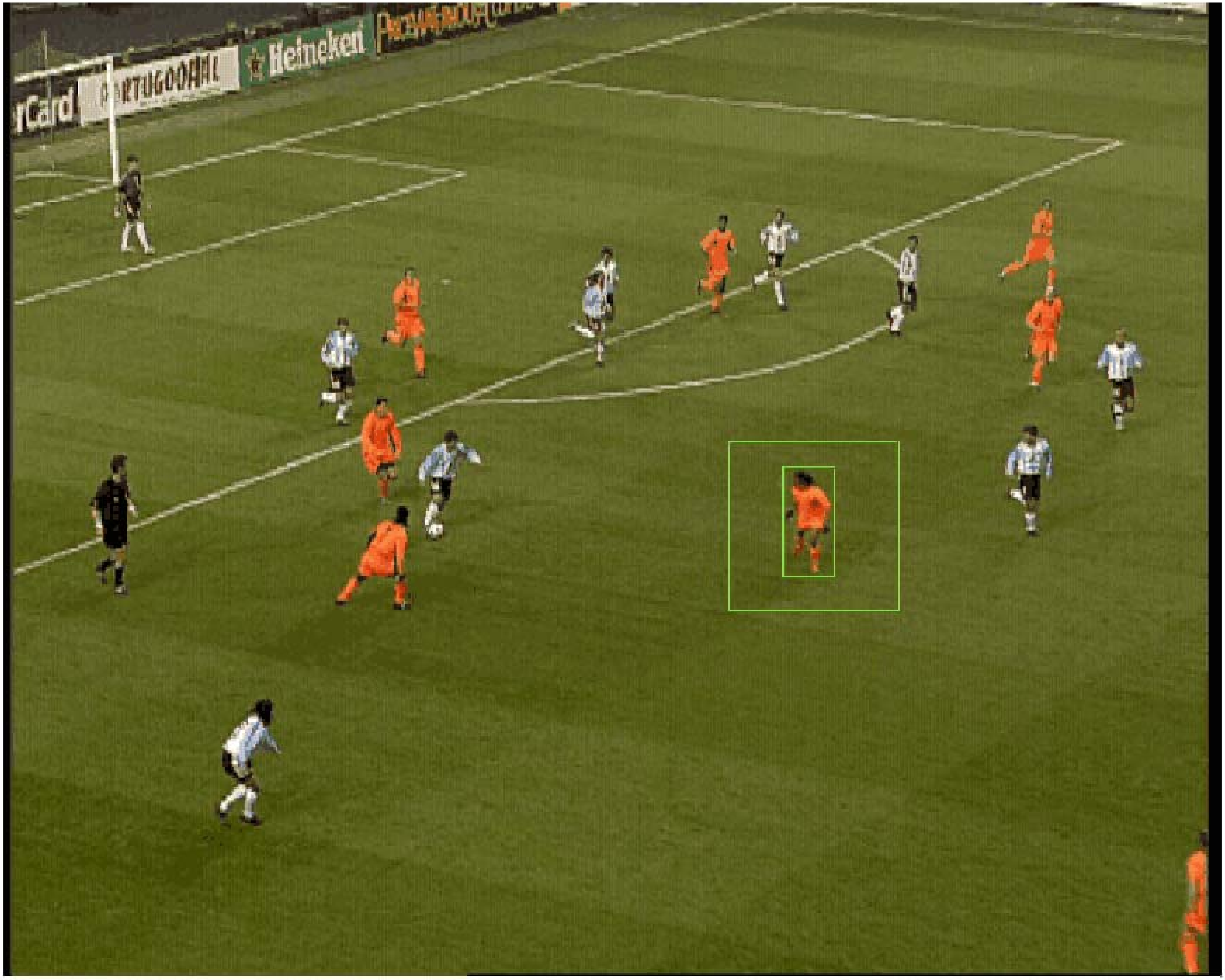
Here the same sequence with tracking target Davids is shown. Only now the intensity of a view frames is changed artificially by: I = R+G+B /2, and switched back to normal intensity.



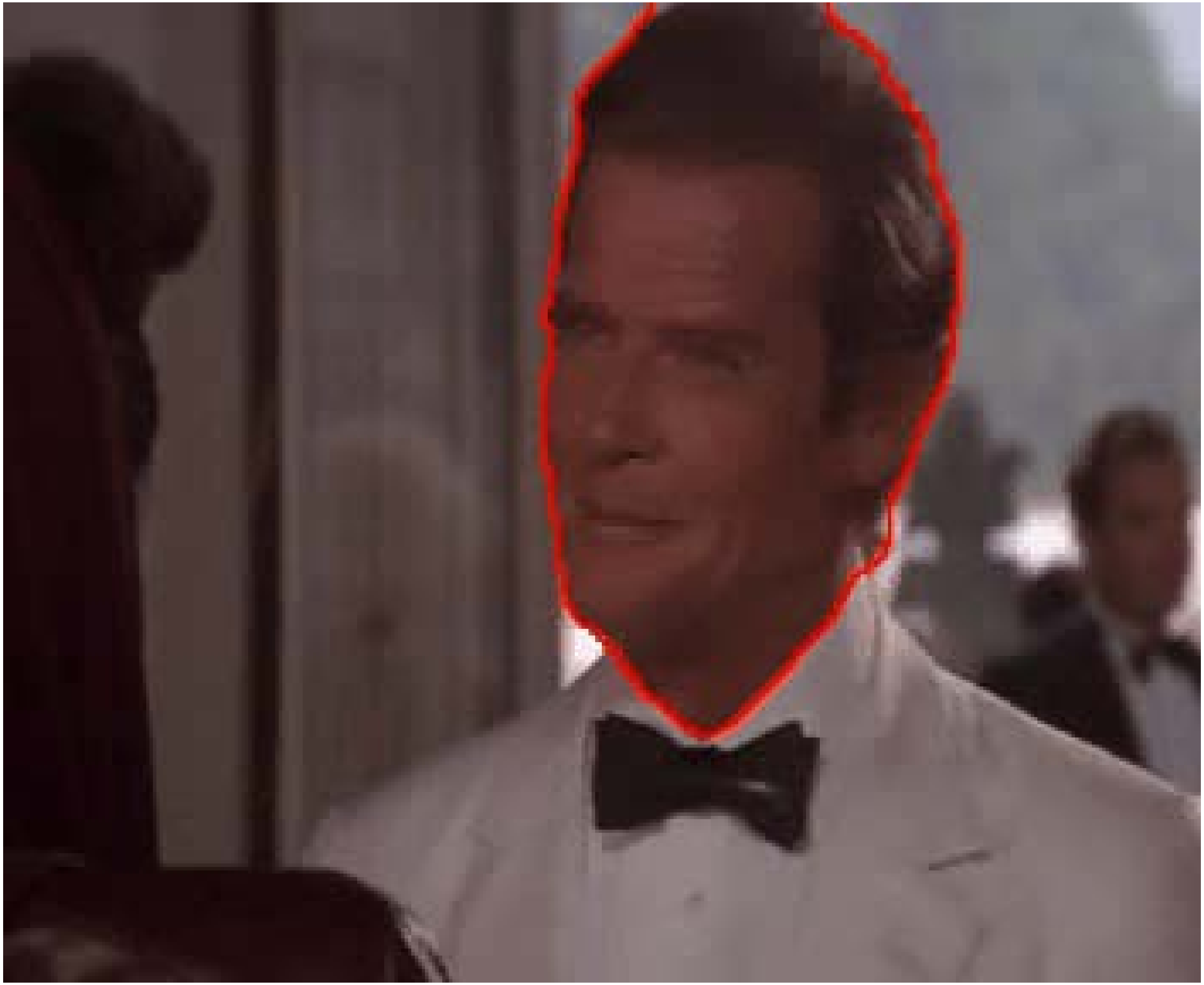


Frame of soccer game, tracking target Davids

Likelihood images associated with target Davids


rCard
Heineken

[Hieu, IEEE PAMI, 2003]

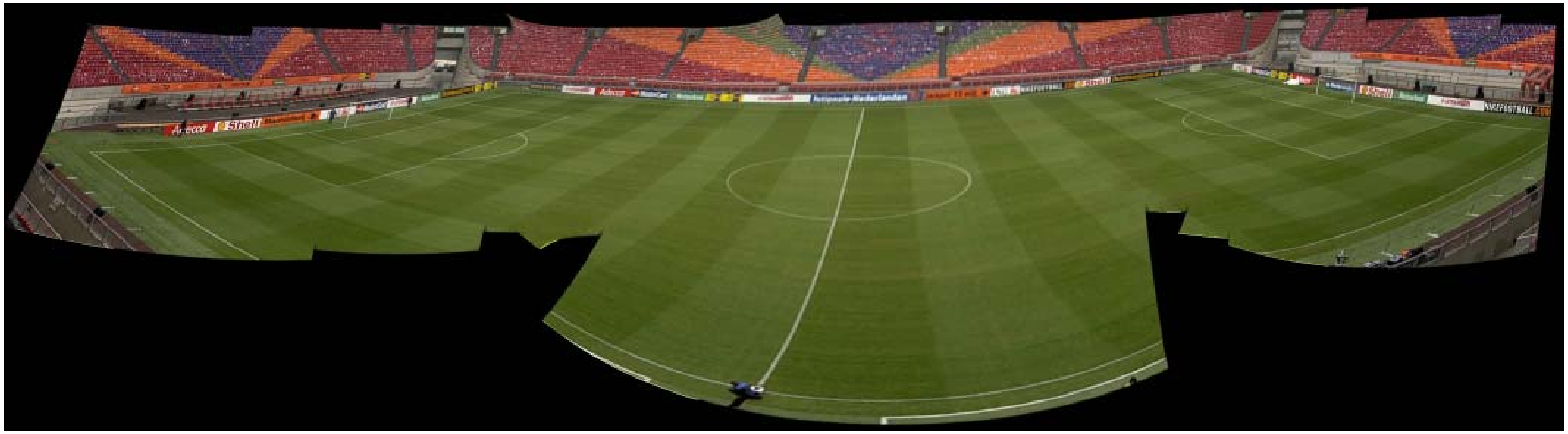


Robust to background clutter and changing object appearance

Object replacement

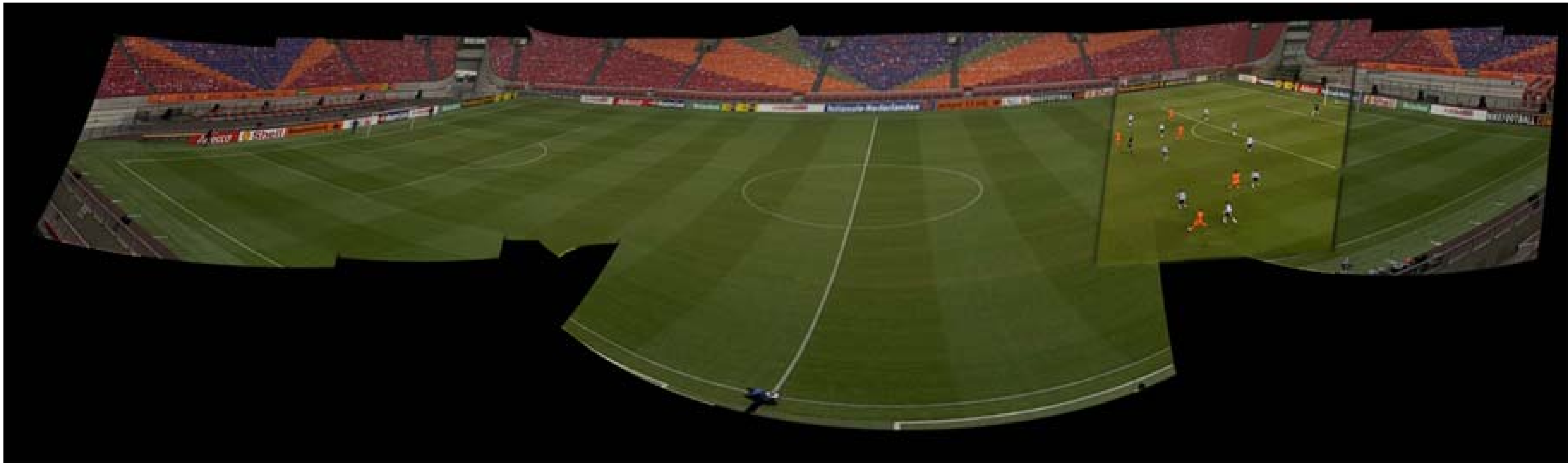
Mosaics

Mosaic created from video

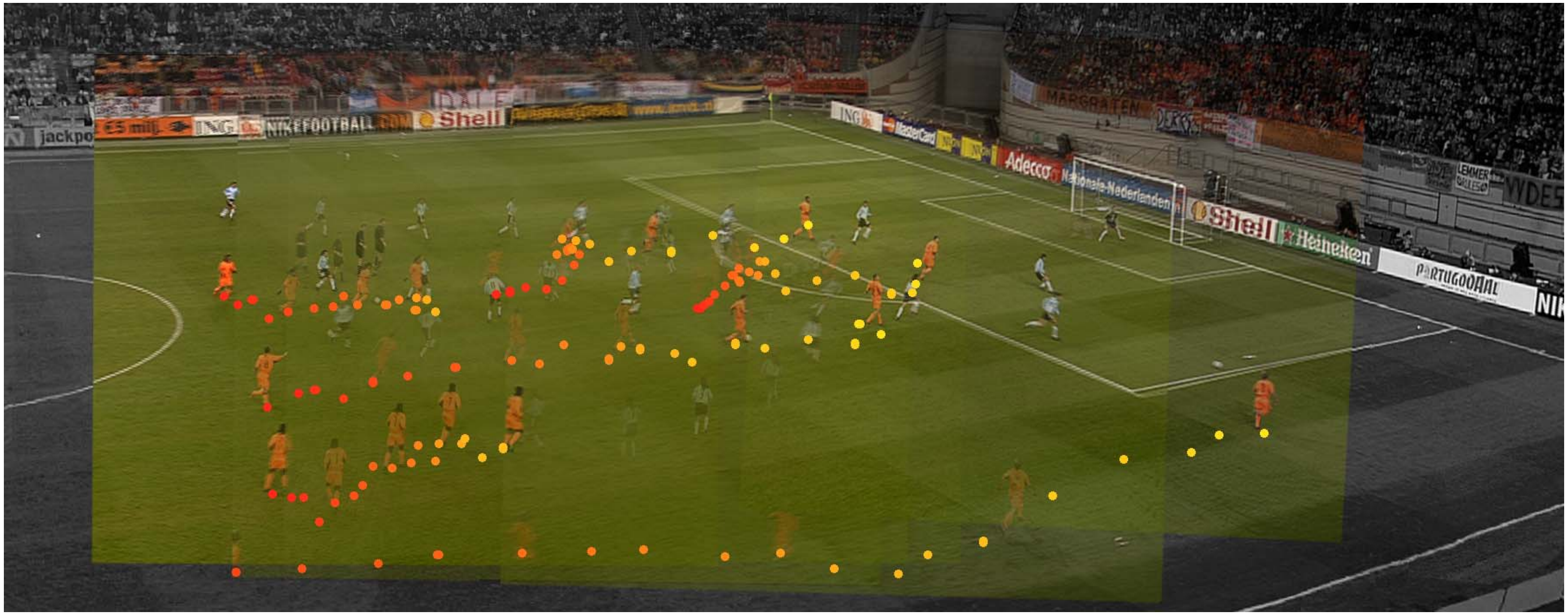


Mosaics

Using model for matching



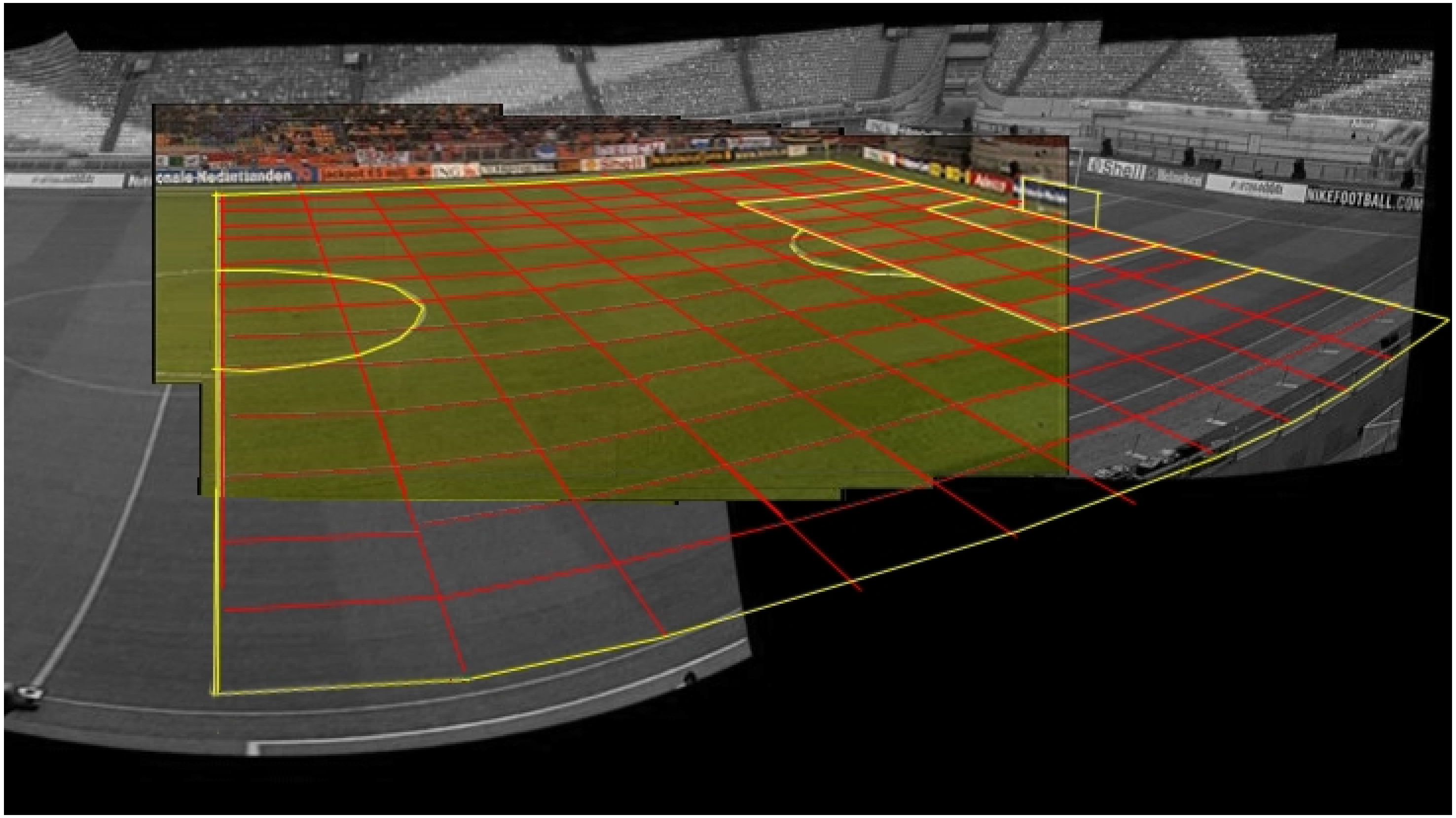
Mosaics



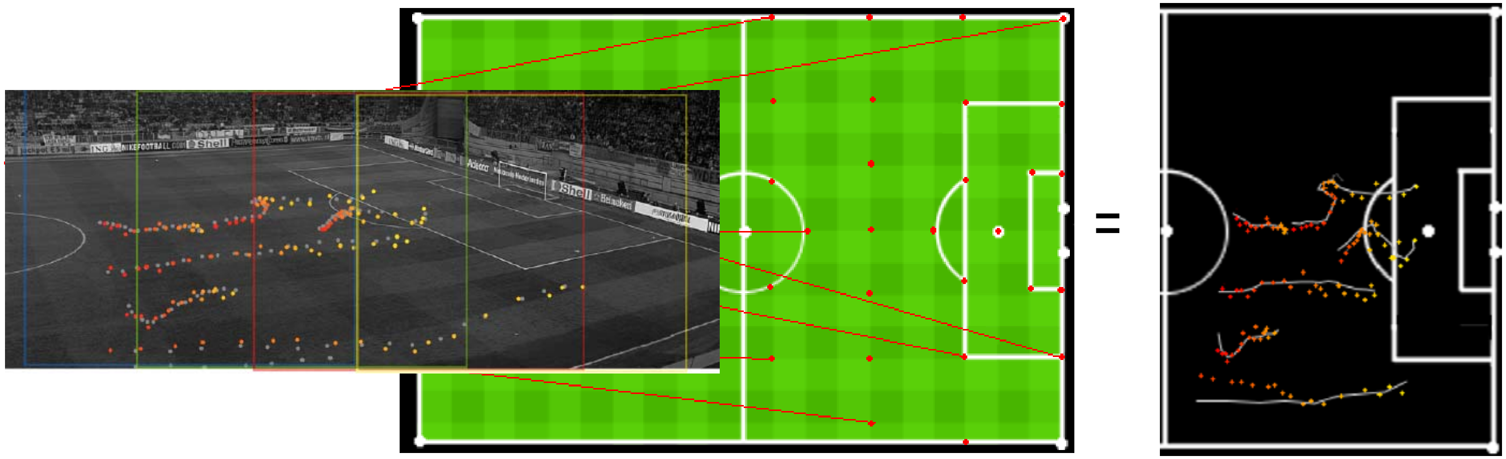
Several frames projected on the mosaic, according to their recovered registration parameters.

Showing ‘ghosts’ of players is very illustrative

Mosaics

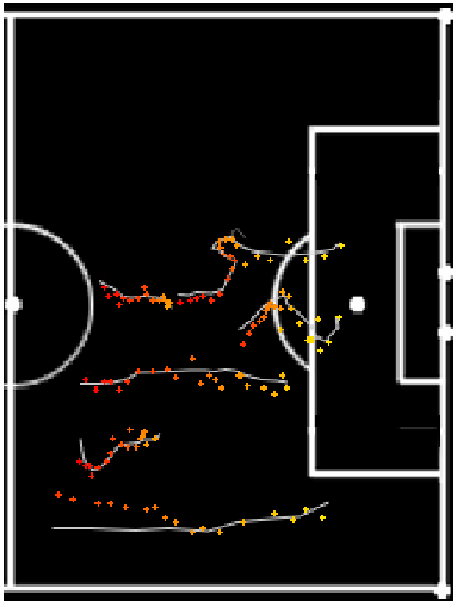


Homography Transform Phase



- After iteratively plotting the foot-positions of each frame a trajectory plot is constructed. Distinctive or salient features are selected and mapped to the geometrically correct line-model. Finally, conversion to an orthogonal perspective using a homography is performed.

Motion and Visual Tracking



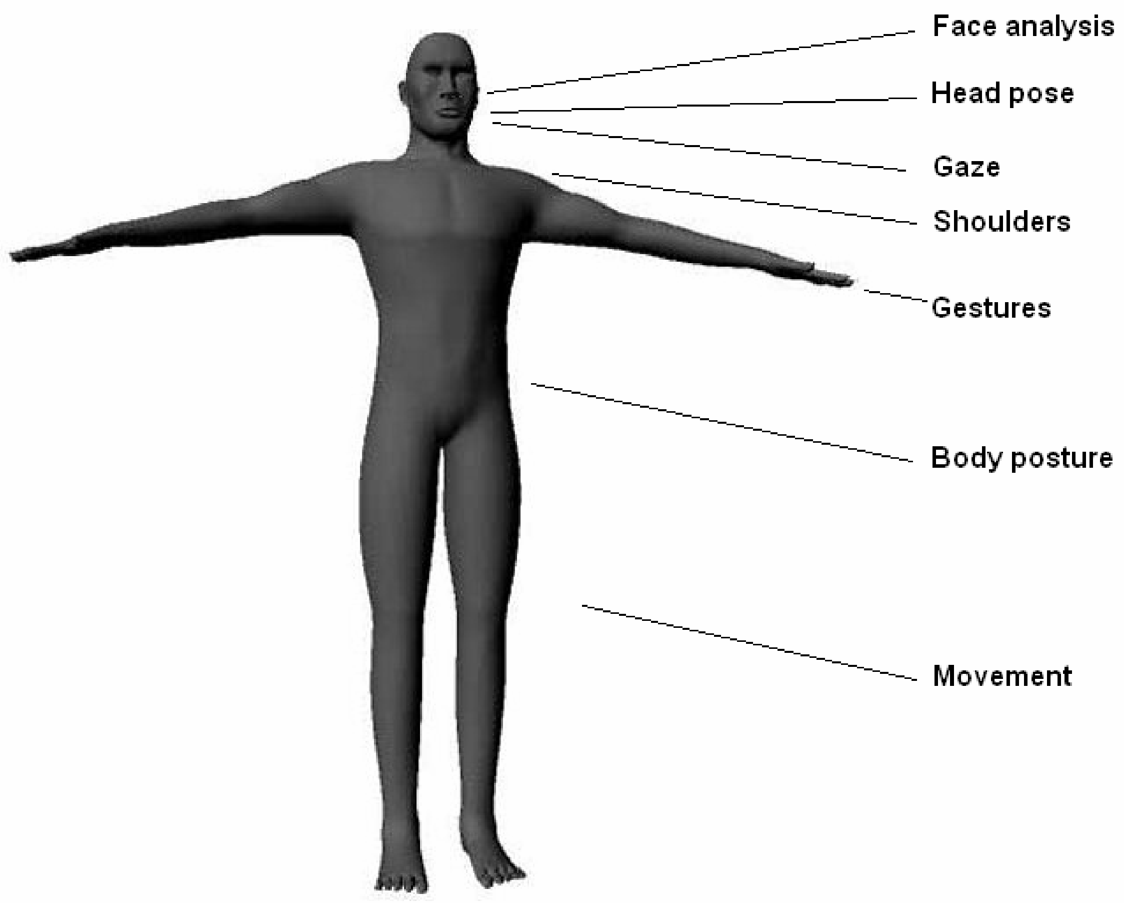
Motion and Visual Tracking

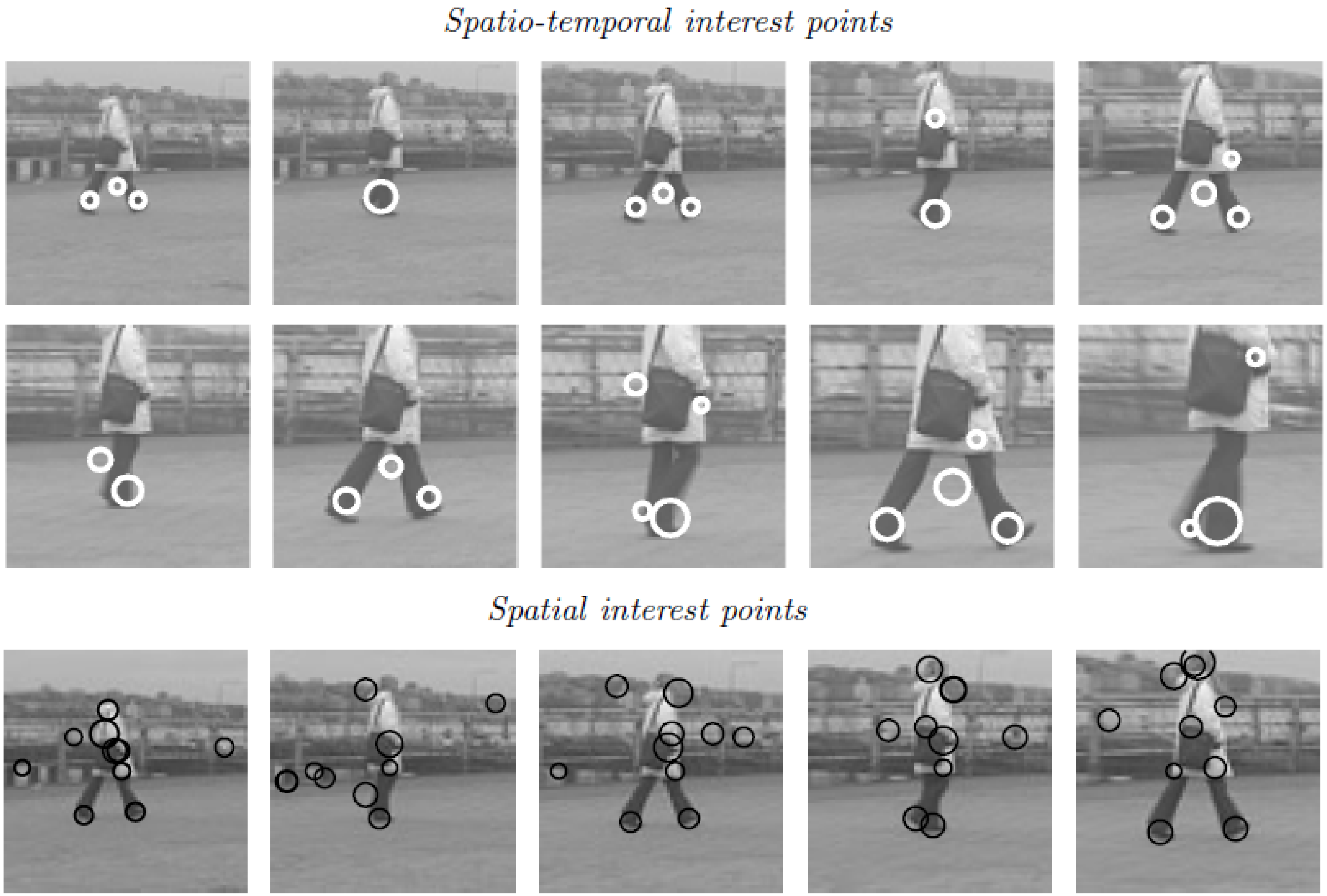
Overview

1. ***Motivation***
   - Challenges
   - Computer vision demands
2. ***Pixel-based detection***
   - Histograms
   - Density estimation
3. ***Interest point detection***
   - Local image structures (edges and corners)
   - Harris Laplace
   - Color boosted
4. ***Descriptors***
   - SIFT
   - Extension to color
5. ***Object recognition (VOC/TRECVid)***
   - Dense and point sampling
   - Code book generation
   - Results
6. ***Applications***
   - Tracking in video
   - Object replacement
   - Emotion recognition
   - Head pose estimation

Activity Recognition

Visual analysis of the human body





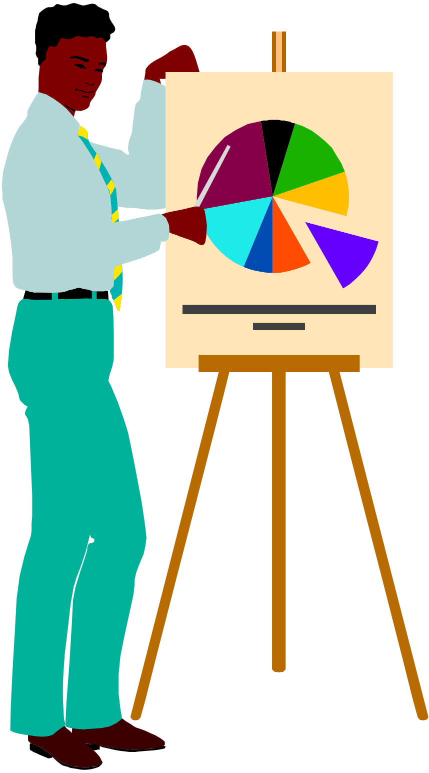
Spatio-temporal interest points

Spatial interest points

Figure 8: Top: Results of spatio-temporal interest point detection for a zoom-in sequence of a walking person. The spatial scale of the detected points (corresponding to the size of circles) matches the increasing spatial extent of the image structures and verifies the invariance of the interest points with respect to changes in spatial scale. Bottom: Pure spatial interest point detector (here, Harris-Laplace) selects both moving and stationary points and is less restrictive.

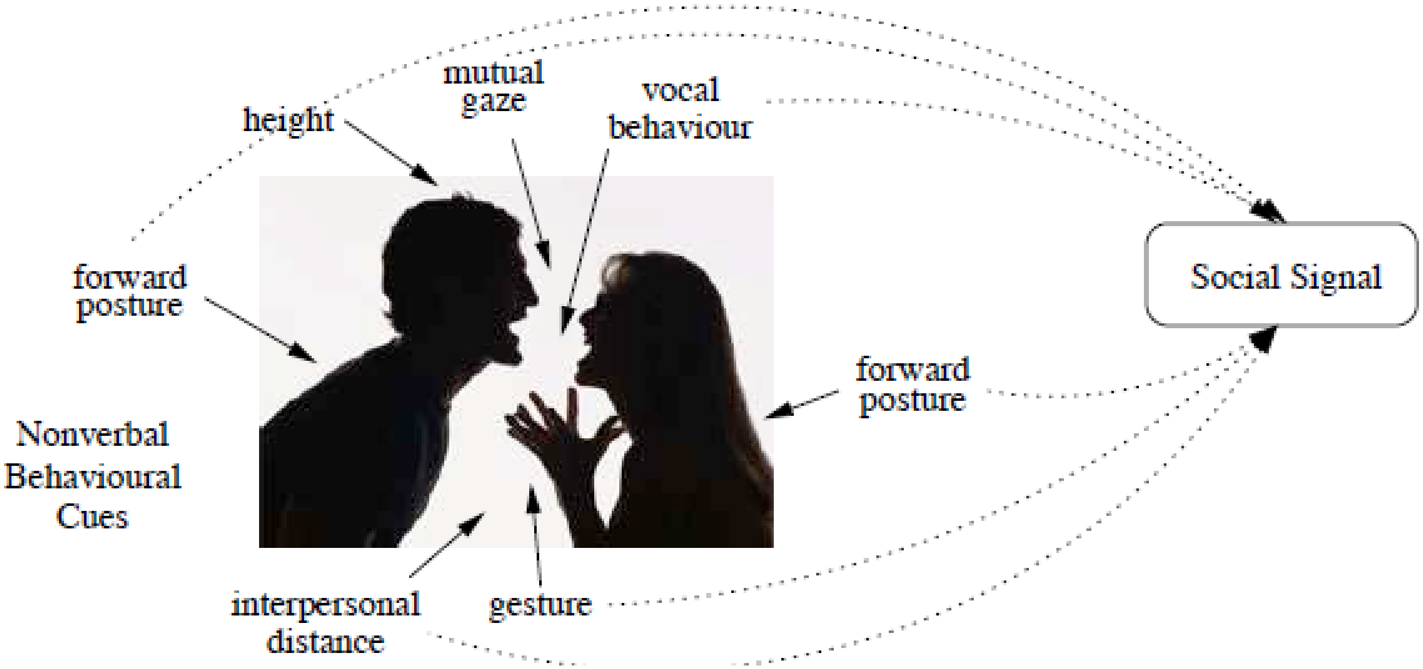
Overview

- Activity recognition
- Social signal processing



Social Signal Processing

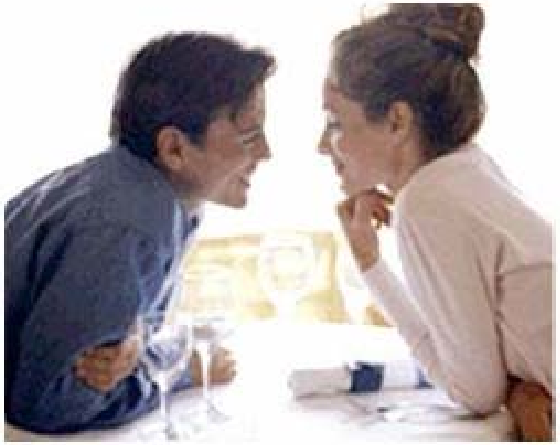
Nonverbal cues



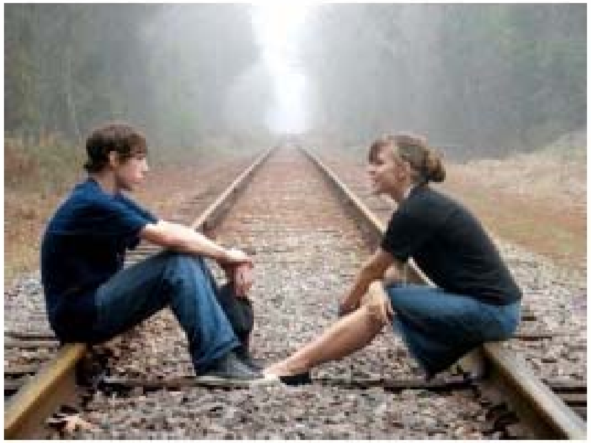
Vinciarelli, Pantic, Bourlard, 2009

Social Signal Processing

Example for posture congruence



Congruent postures

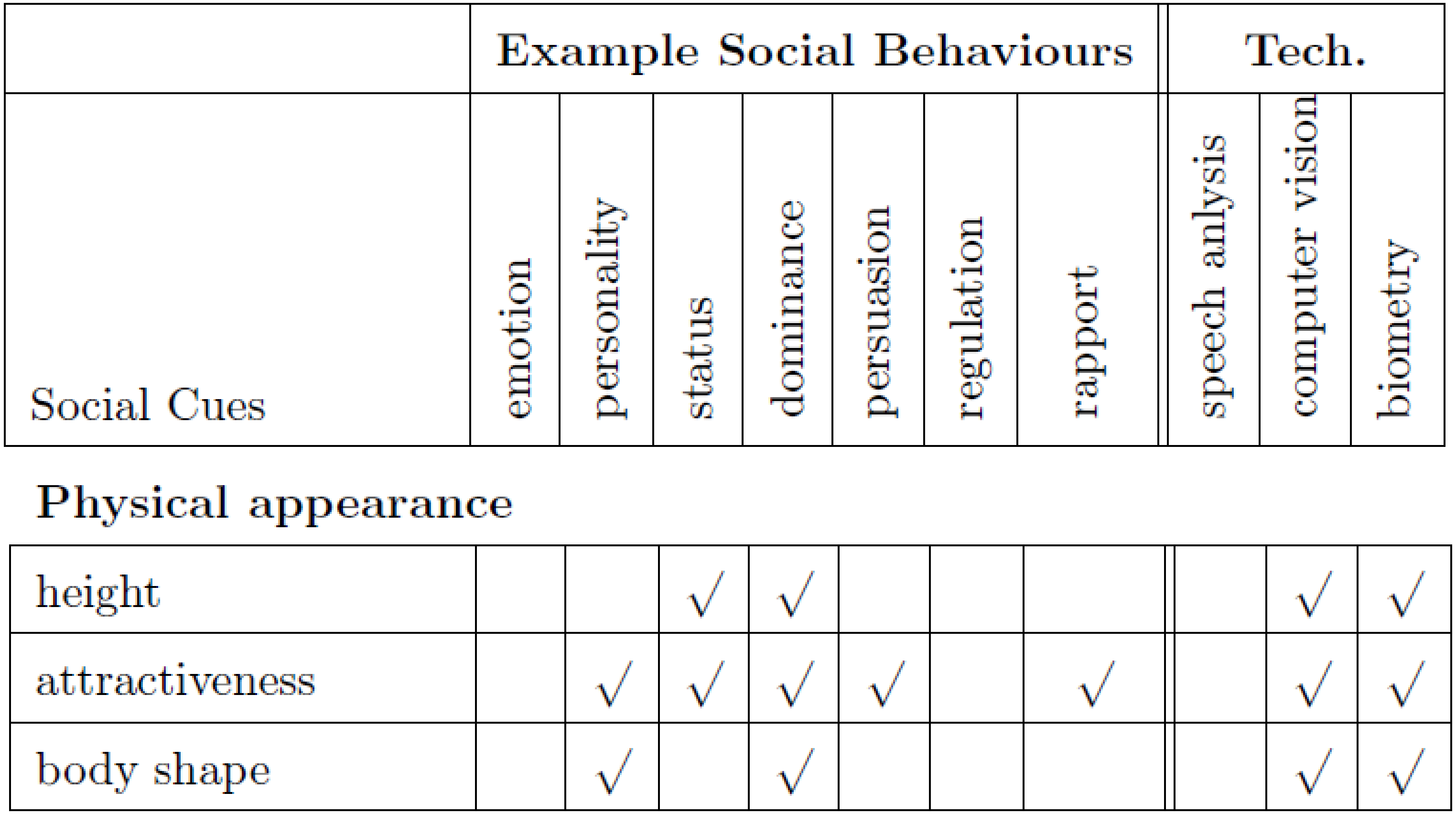


Non–congruent postures

Vinciarelli, Pantic, Bourlard, 2009

Social Signal Processing

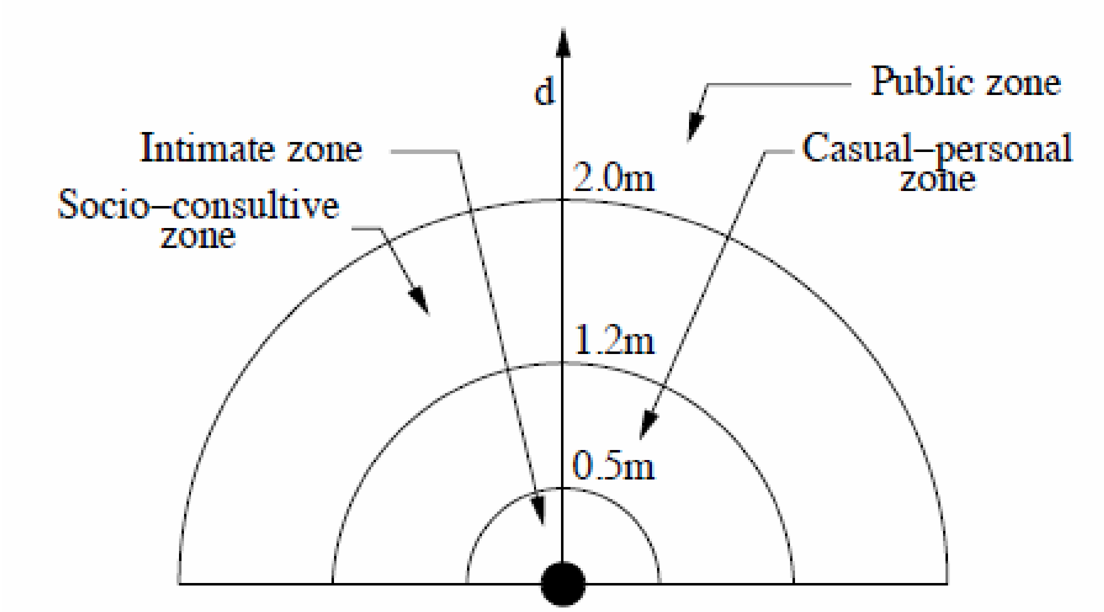
Taxonomy



	Example Social Behaviours							Tech.		
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport	speech anlysis	computer vision	biometry
Physical appearance										
height			√	√					√	√
attractiveness		√	√	√	√		√		√	√
body shape		√		√					√	√

Vinciarelli, Pantic, Bourlard, 2009

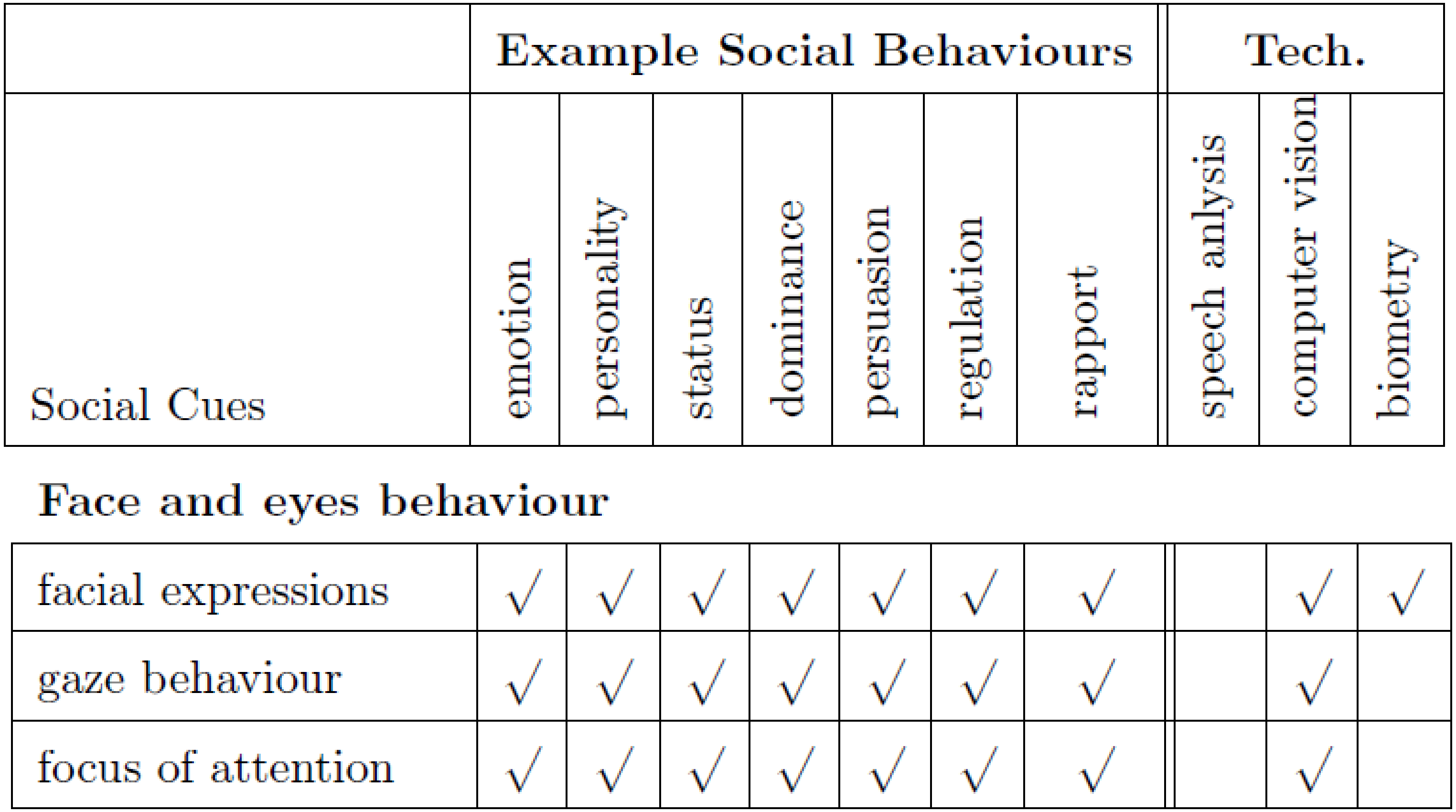
Social Signal Processing



Vinciarelli, Pantic, Bourlard, 2009

Social Signal Processing

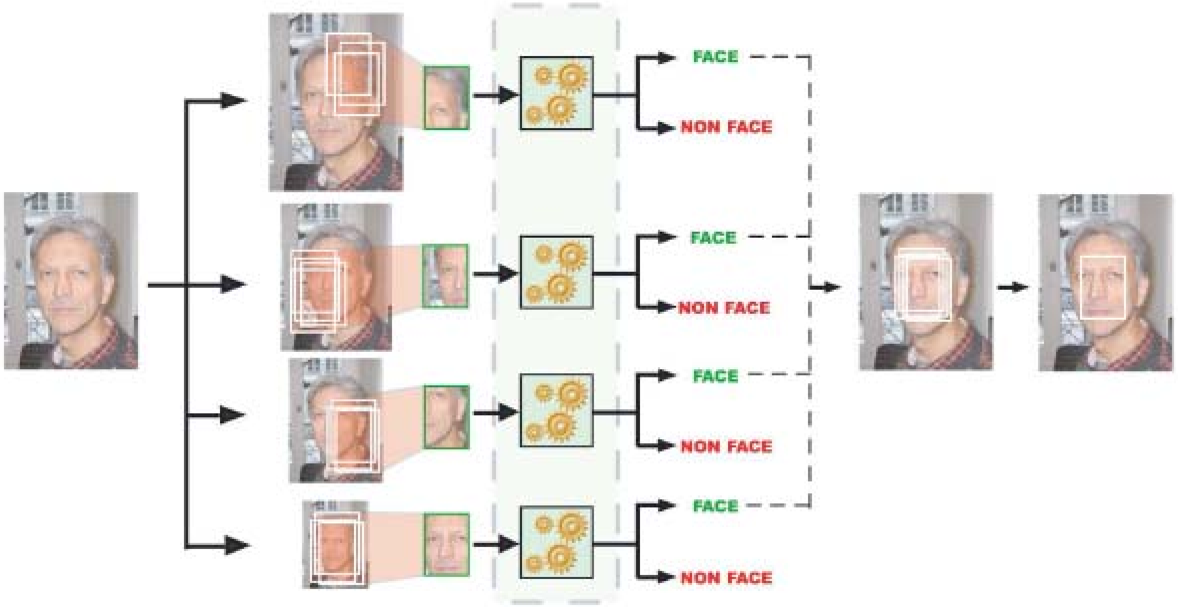
Taxonomy



	Example Social Behaviours							Tech.		
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport	speech anlysis	computer vision	biometry
Face and eyes behaviour										
facial expressions	✓	✓	✓	✓	✓	✓	✓		✓	✓
gaze behaviour	✓	✓	✓	✓	✓	✓	✓		✓	
focus of attention	✓	✓	✓	✓	✓	✓	✓		✓	

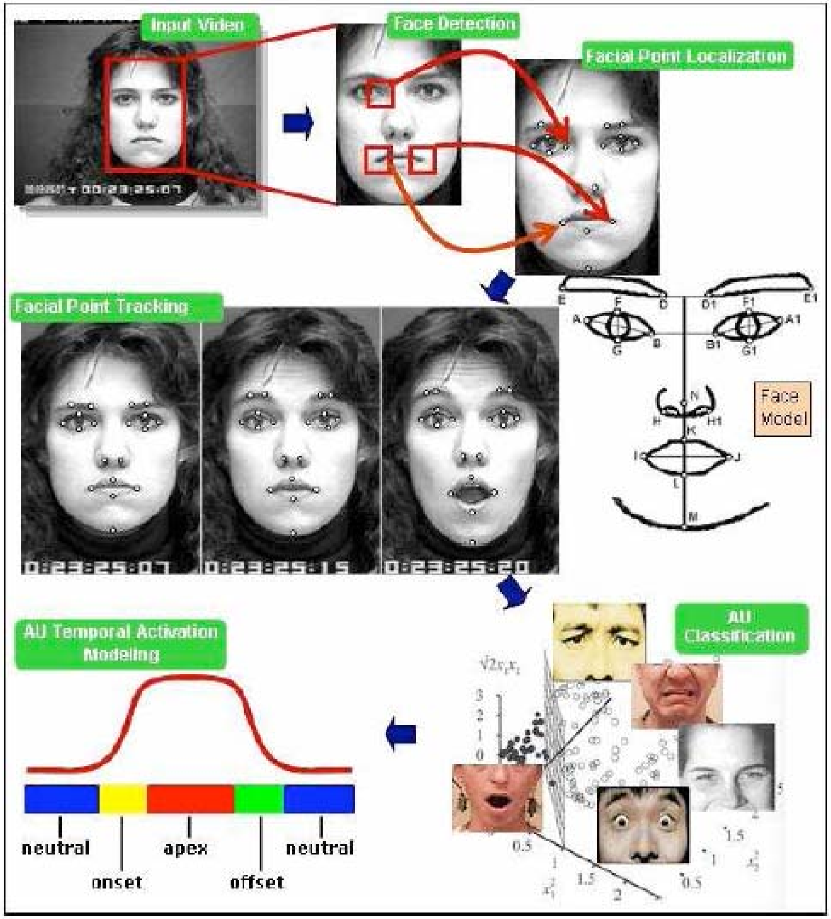
Vinciarelli, Pantic, Bourlard, 2009

Faces

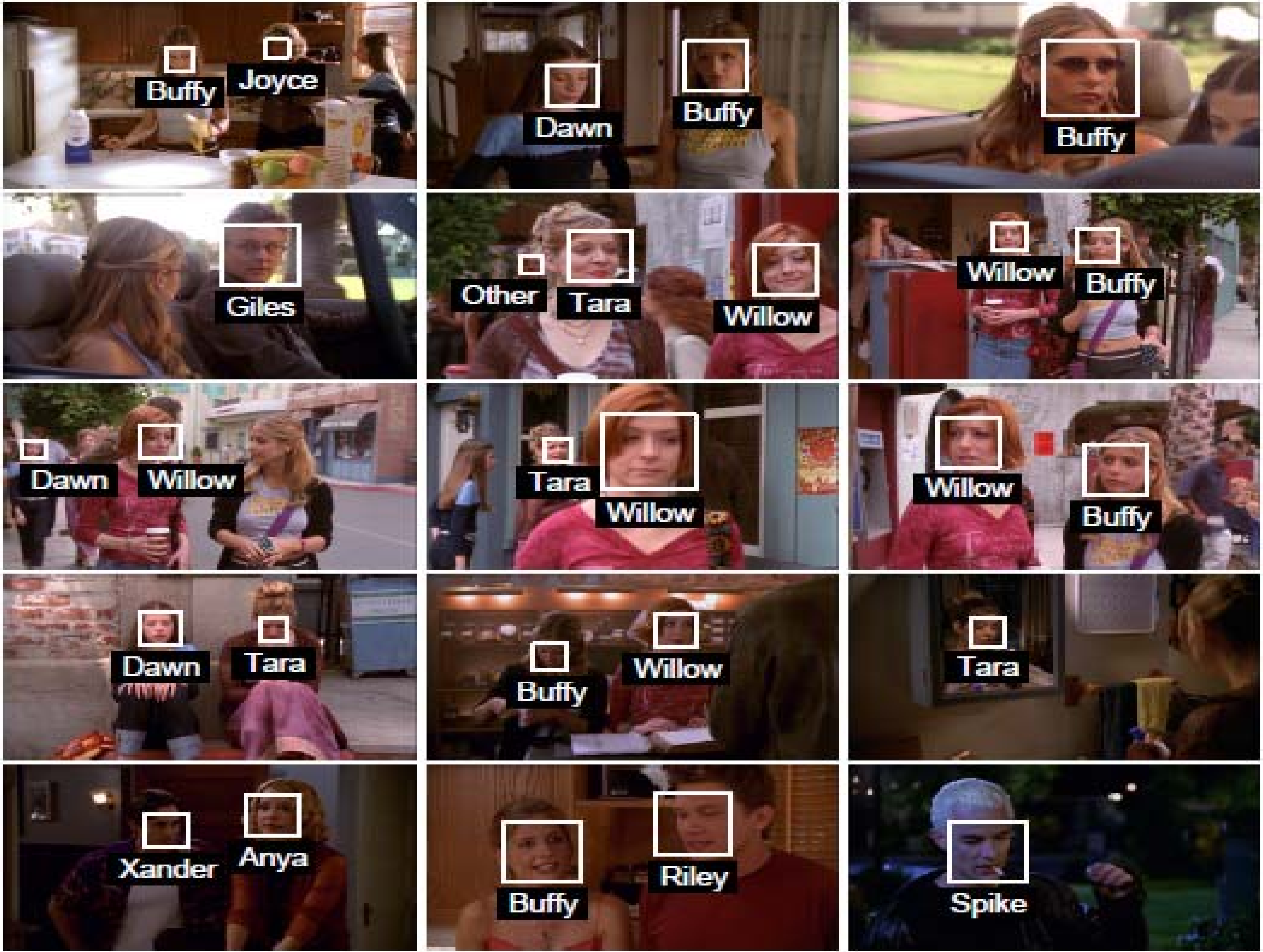


Vinciarelli, Pantic, Bourlard, 2009

Facial expression



Vinciarelli, Pantic, Bourlard, 2009

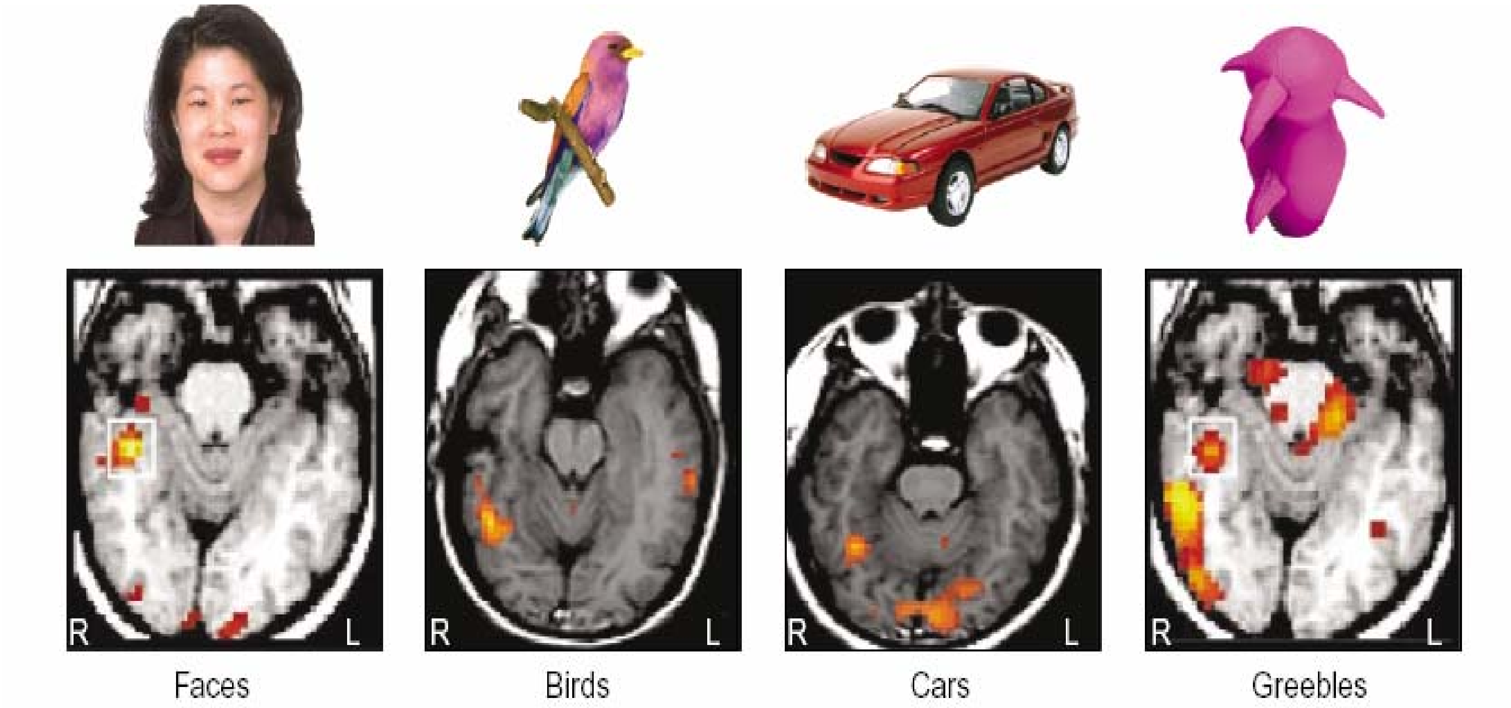

Buffy
Joyce
Dawn
Buffy
Buffy
Giles
Other
Tara
Willow
Willow
Buffy
Dawn
Willow
Tara
Willow
Willow
Buffy
Dawn
Tara
Buffy
Willow
Tara
Xander
Anya
Buffy
Riley
Spike

What is Face Recognition?

Face vs. Object Recognition

- Is face recognition different from general object recognition?
  - fMRI measurements
  - Prosopagnosia and agnosia
  - Prosopamnesia
  - Capgras syndrome
- Is there a module in the brain for face recognition?

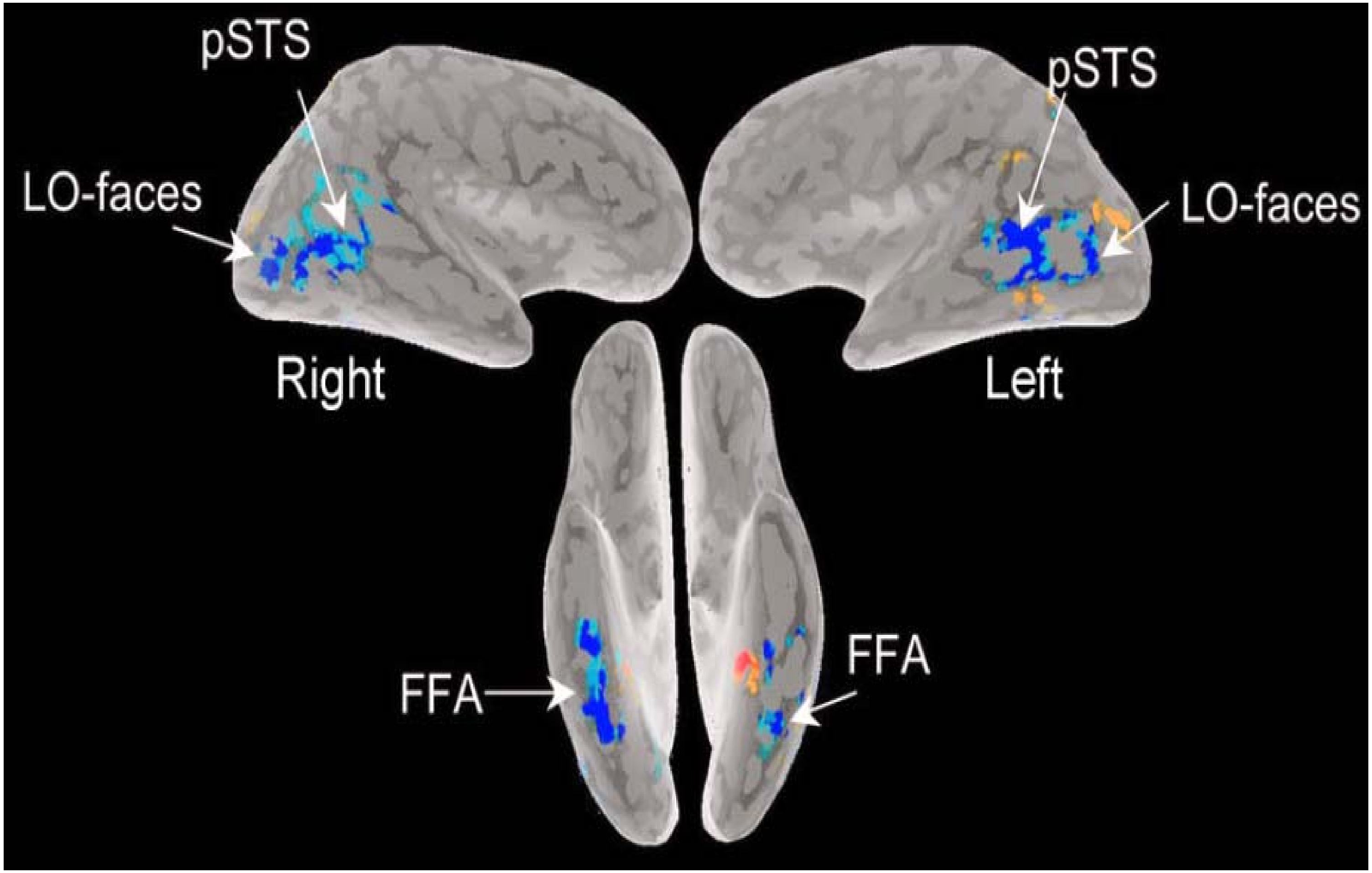
fMRI Experiments



Faces Birds Cars Greebles

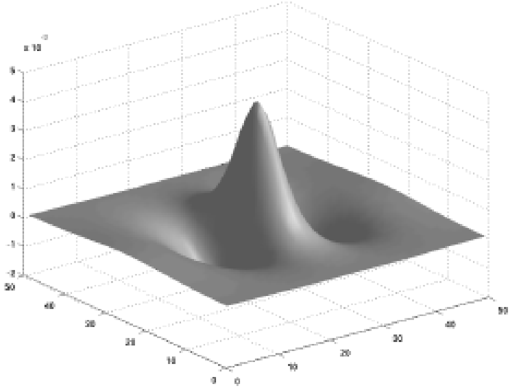
Gauthier, I., M.J. Tarr, *Vision Research* vol.37, pp.1673-1682, 1997

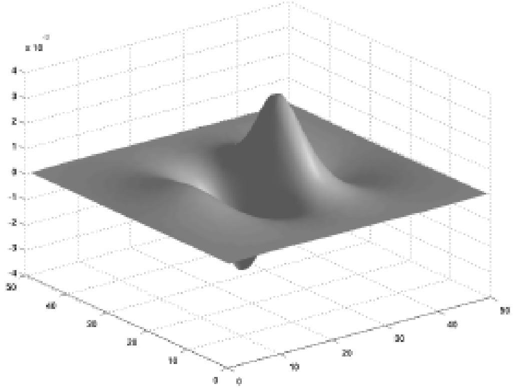
Activation for Faces

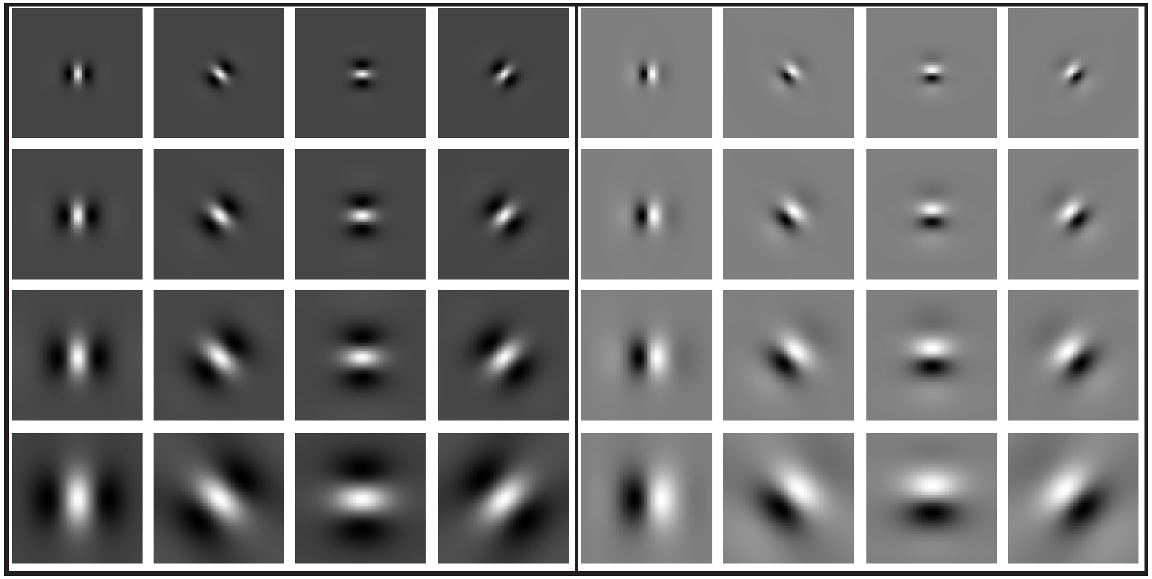


Grill-Spector, Knouf, Kanwisher, Nature Neuroscience 2004

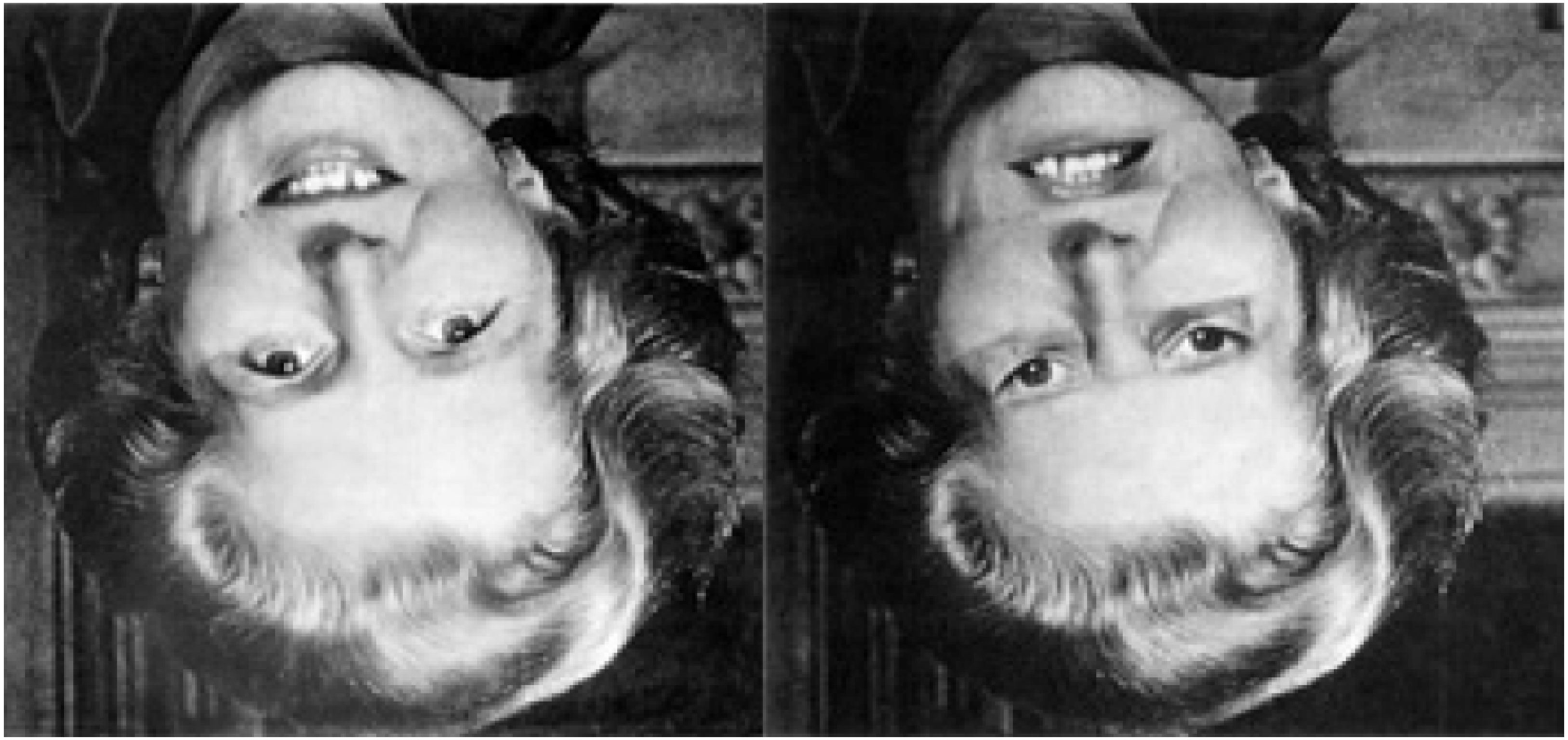
Gabor Wavelet Filters





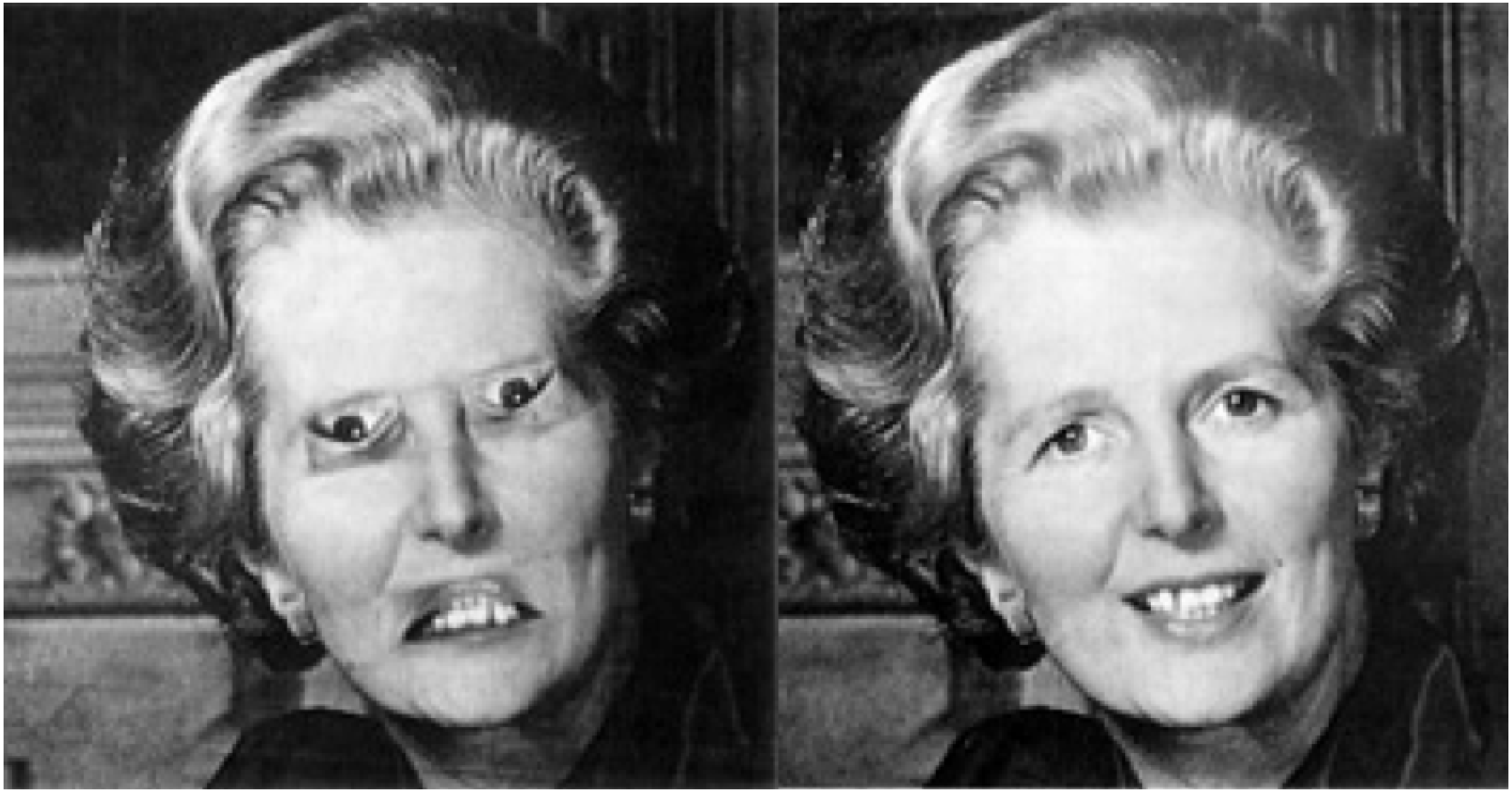


Thatcher Illusion



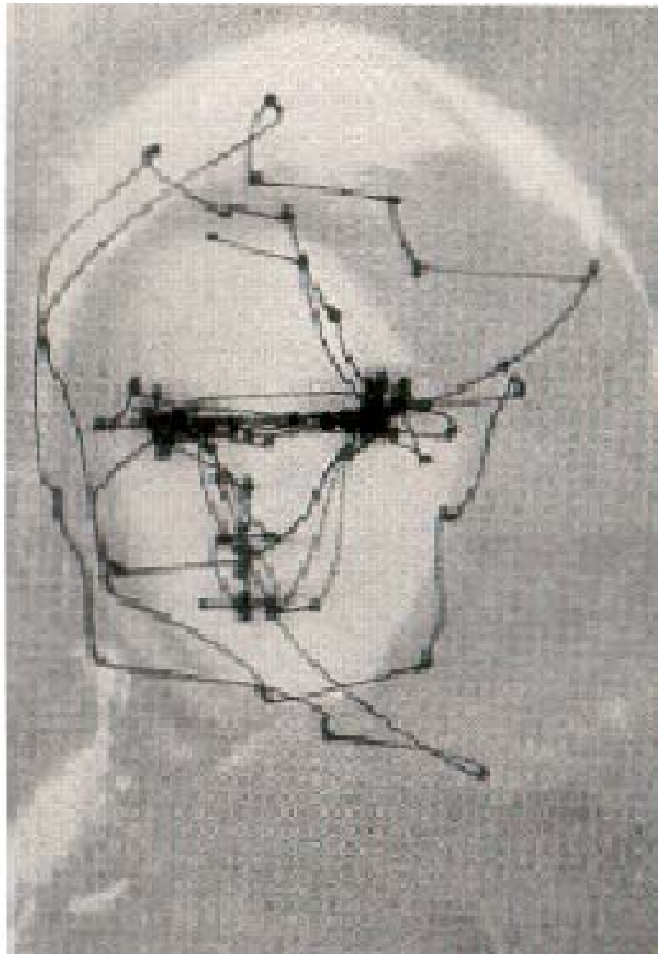
Thompson, *Perception*, vol.9, pp.483-484, 1980

Thatcher Illusion



Thompson, *Perception*, vol.9, pp.483-484, 1980

Selective Attention

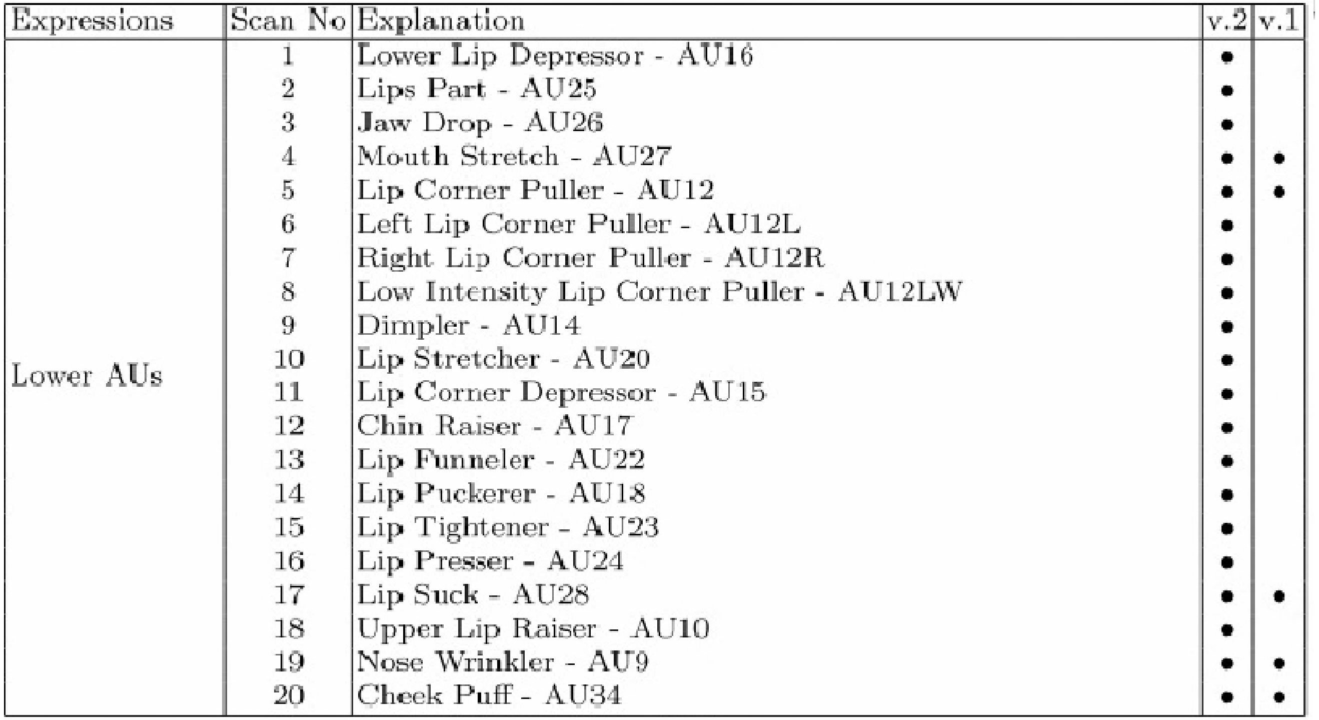


Yarbus, A.L.,Eye Movement and Vision, 1976

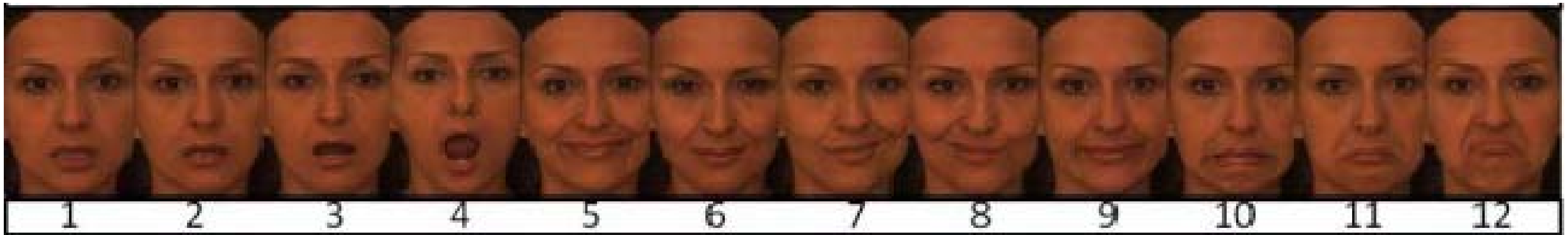
Facial expressions

Face

Lower Action Units

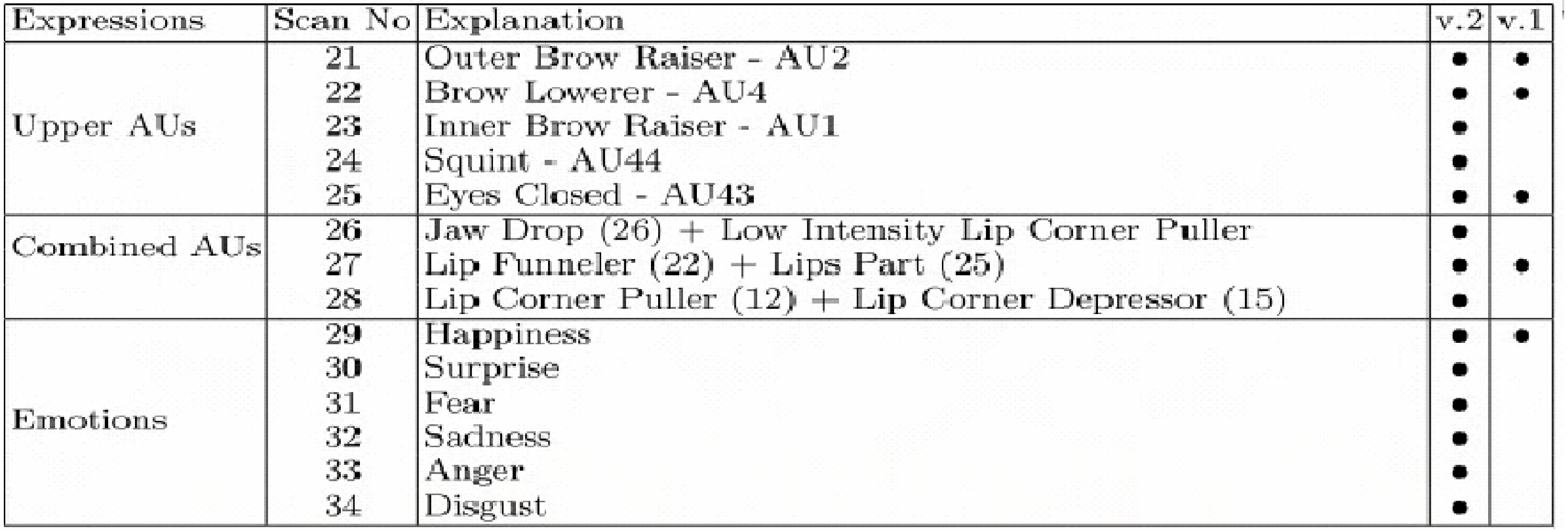


Expressions	Scan No	Explanation	v.2	v.1
Lower AUs	1	Lower Lip Depressor - AU16	●	
	2	Lips Part - AU25	●	
	3	Jaw Drop - AU26	●	
	4	Mouth Stretch - AU27	●	●
	5	Lip Corner Puller - AU12	●	●
	6	Left Lip Corner Puller - AU12L	●	
	7	Right Lip Corner Puller - AU12R	●	
	8	Low Intensity Lip Corner Puller - AU12LW	●	
	9	Dimpler - AU14	●	
	10	Lip Stretcher - AU20	●	
	11	Lip Corner Depressor - AU15	●	
	12	Chin Raiser - AU17	●	
	13	Lip Funneler - AU22	●	
	14	Lip Puckerer - AU18	●	
	15	Lip Tightener - AU23	●	
	16	Lip Presser - AU24	●	
	17	Lip Suck - AU28	●	●
	18	Upper Lip Raiser - AU10	●	
	19	Nose Wrinkler - AU9	●	●
	20	Cheek Puff - AU34	●	●

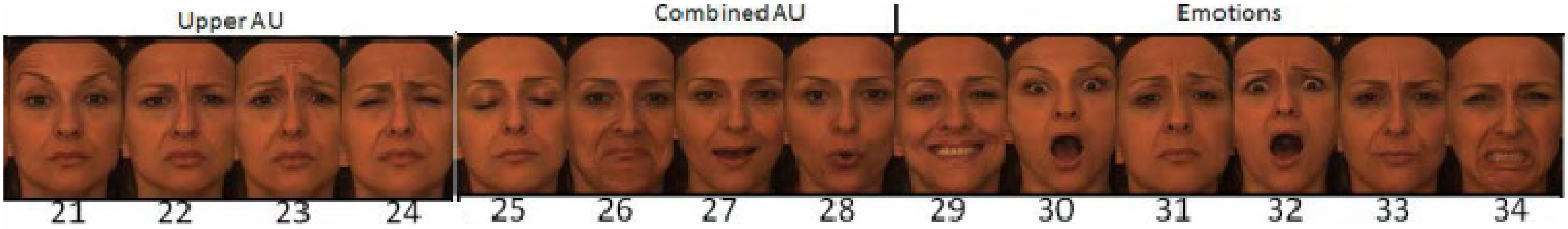


Face

Upper/Combined Action Units + Basic Expressions



Expressions	Scan No	Explanation	v.2	v.1
Upper AUs	21	Outer Brow Raiser - AU2	●	●
	22	Brow Lowerer - AU4	●	●
	23	Inner Brow Raiser - AU1	●	
	24	Squint - AU44	●	
	25	Eyes Closed - AU43	●	●
Combined AUs	26	Jaw Drop (26) + Low Intensity Lip Corner Puller	●	
	27	Lip Funneler (22) + Lips Part (25)	●	●
	28	Lip Corner Puller (12) + Lip Corner Depressor (15)	●	
Emotions	29	Happiness	●	●
	30	Surprise	●	
	31	Fear	●	
	32	Sadness	●	
	33	Anger	●	
	34	Disgust	●	



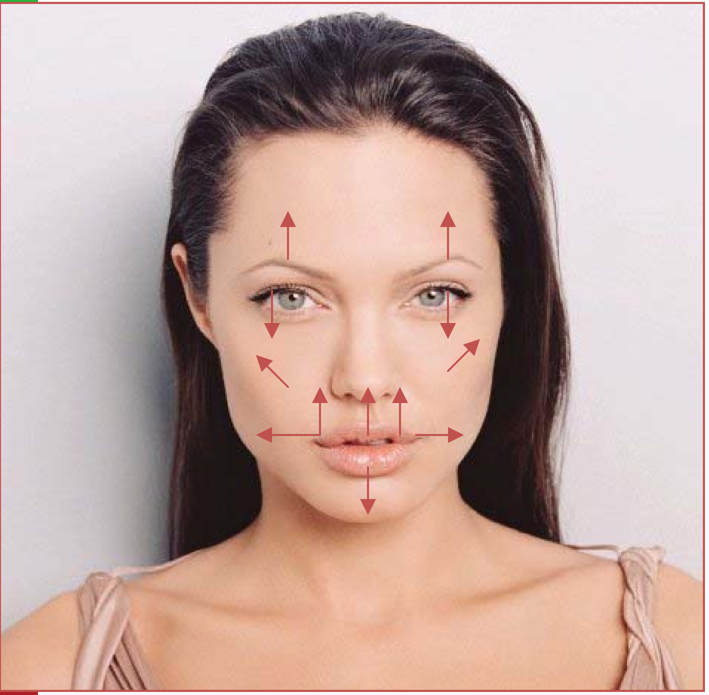
Facial Expression Recognition

With Nicu Sebe

Dept. of Information Engineering and Computer Science,
University of Trento

Beckman Institute at the University of Illinois, Urbana-Champaign, USA

Facial Expression Recognition



12 facial motion measurements

vertical movement of the lips

horizontal movement of the mouth corners

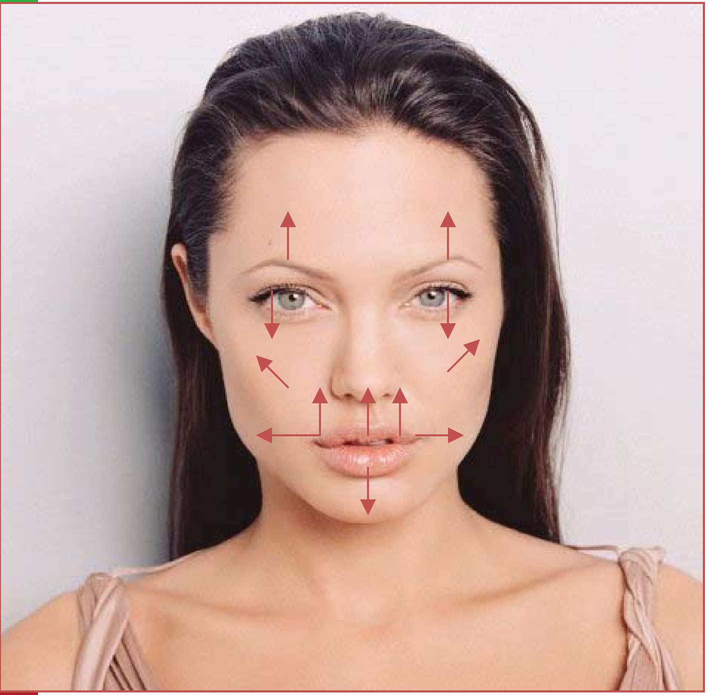
vertical movement of the mouth corners

vertical movement of the eye brows

lifting of the cheeks

blinking of the eyes

Facial Expression Recognition

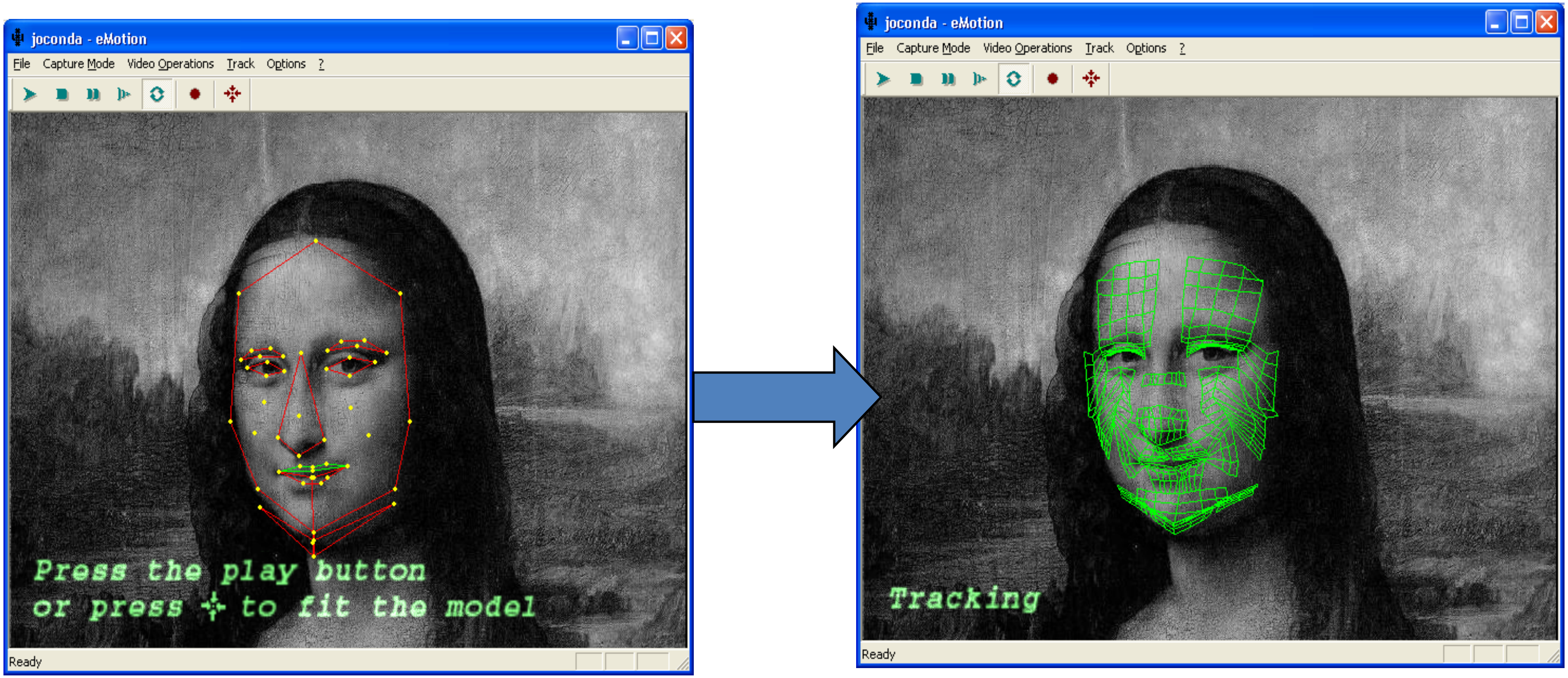


We use 12 facial features = 12 facial motion measurements

The combination of these features define the 7 basic classes of facial expression we want to classify: ***Neutral, Happy, Anger, Disgust, Fear, Sad, Surprise***

Facial Expression Recognition

Nicu Sebe

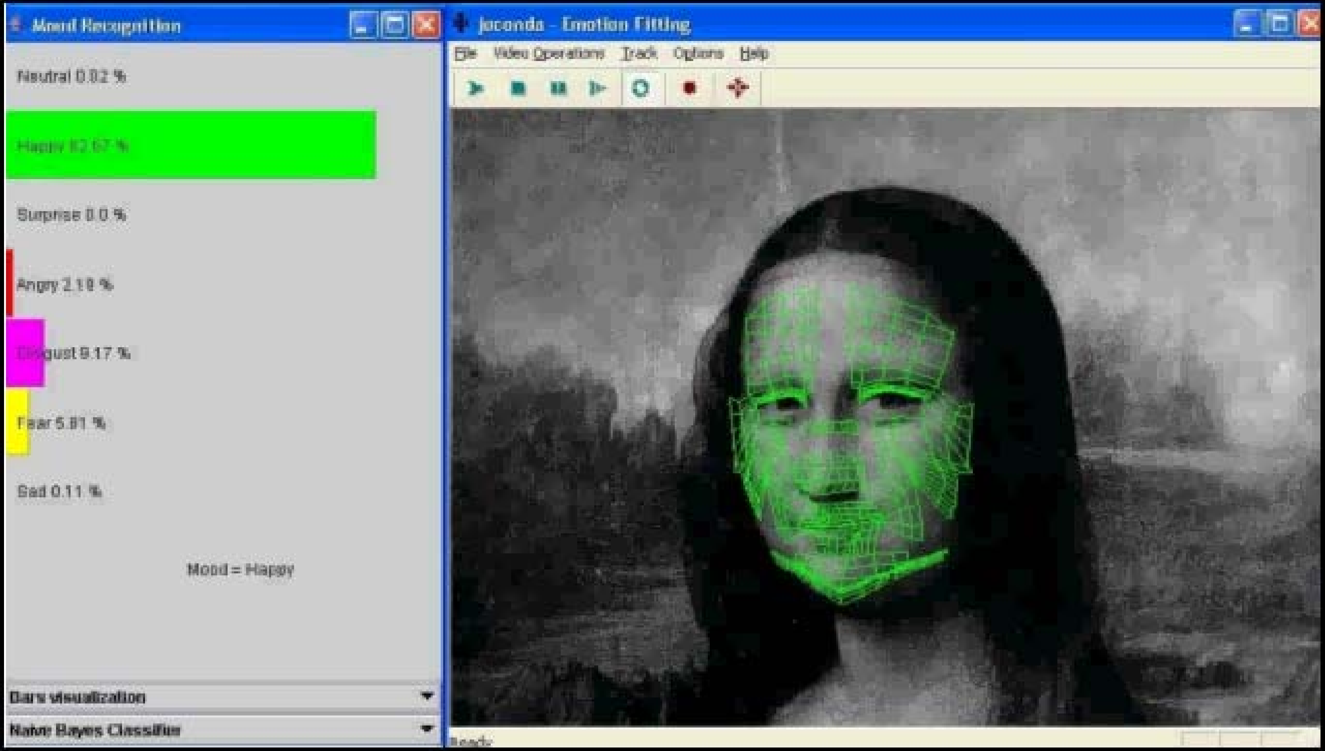


Press the play button
or press ✣ to fit the model

Ready

Tracking

Ready



Neutral 0.02 %

Happy 82.67 %

Surprise 0.0 %

Angry 2.18 %

Disgust 9.17 %

Fear 5.81 %

Sad 0.11 %

Mood = Happy

In the media...

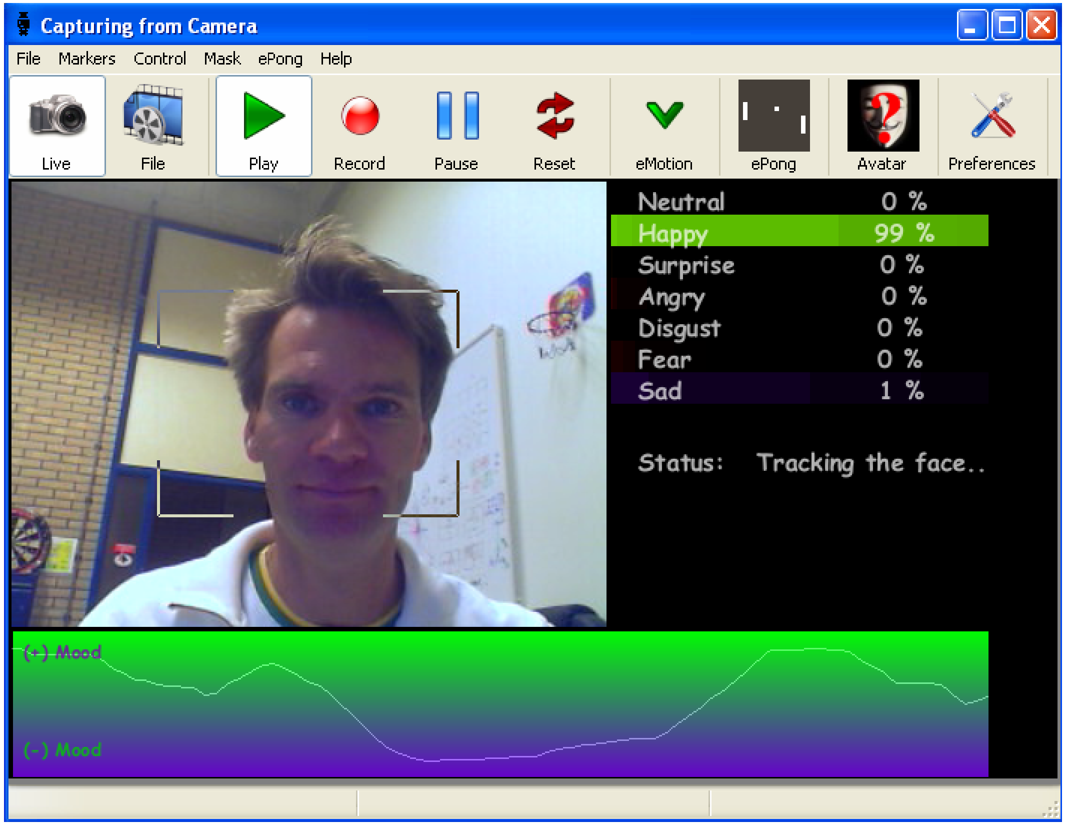
– ***TV exposure (selected)***
  - BBC news, December 15th, 2005.
  - CNN, December 16th, 2005.
  - TeleFrance 1 (TF1), December 16th, 2005.
  - RAI 1, December 15th, 2005.
  - Japan Today News, December 15th, 2005.
  - TVE, December 15th, 2008.
  - RTBF, June 16th, 2008.
  - **….**

– ***TV interviews***
  - RTL-i (Belgium) – “Ca Alors” show – January 13 th 2006.
  - Ned 3 – VARA – “De Wereld Draait Door” (live) –Jan. 2006
  - SBS6 Shownieuws – February 11th, 2006
  - Deutsche Welle (Germany) – March 1th, 2006
  - TV Tokyo (Japan) – March 22th, 2006
  - Ivanhoe Broadcast News (USA) – March 31th, 2006.

– ***Articles in Newspapers and Magazines***
  - Der Spiegel, 15th December, 2005.
  - Parool, 15th December, 2006
  - Algemeen Dagblad, 2006.
  - Time Magazine, Oct 6th, 2006.
  - The Christian Science Monitor, Boston, USA, Dec 18th, 2006.
  - Psychologie, January, 2007.
  - Metro, February, 2007.
  - Wired, July, 2007.
  - Algemeen Dagblad, 17 November, 2007.
  - Parool, 2 November, 2009.

– ***Radio***
  - Radio 1, Met het oog op morgen, January, 2006.
  - Canadian radio station CNKW, January 2006.
  - Mega stad fm, Rotterdam, March 2007.
  - Radio 1, Radio Online, August 14th, 2007.
  - Hoe?Zo!, September 2007.
  - Radio 1, Kassa, November, 2009

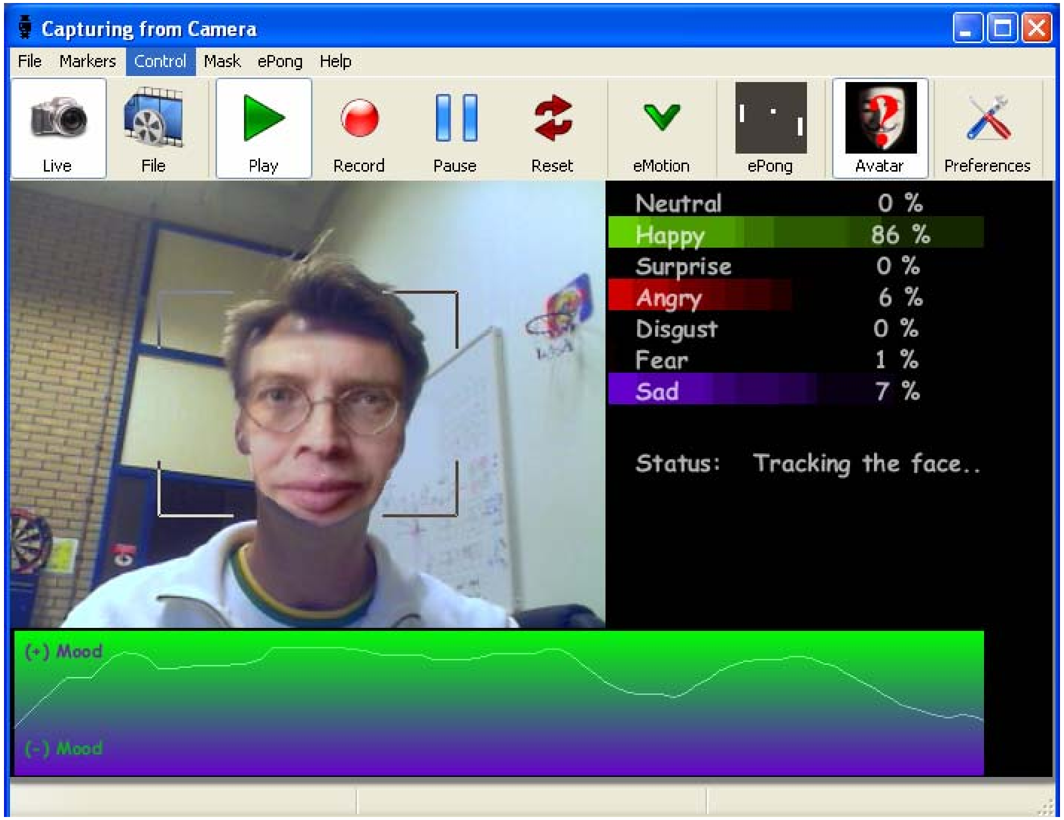
The mask



Neutral 0 %
Happy 99 %
Surprise 0 %
Angry 0 %
Disgust 0 %
Fear 0 %
Sad 1 %

Status: Tracking the face..

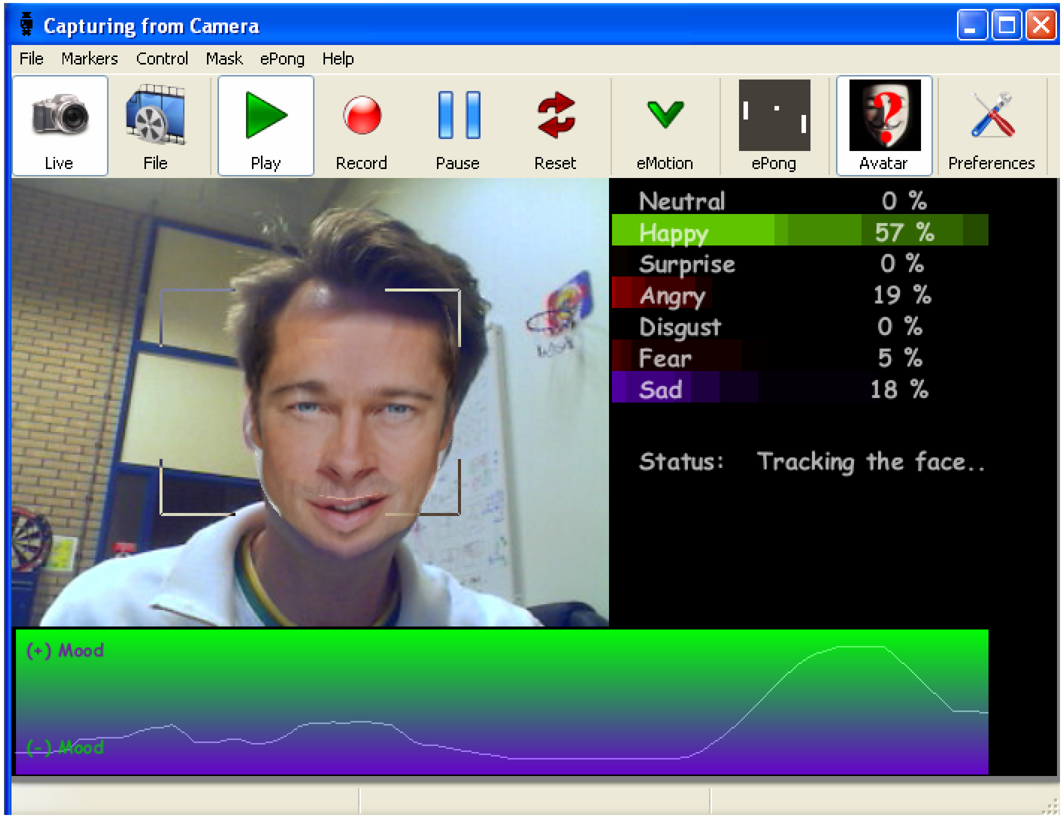
The mask



Neutral	0 %
Happy	86 %
Surprise	0 %
Angry	6 %
Disgust	0 %
Fear	1 %
Sad	7 %

Status: Tracking the face..

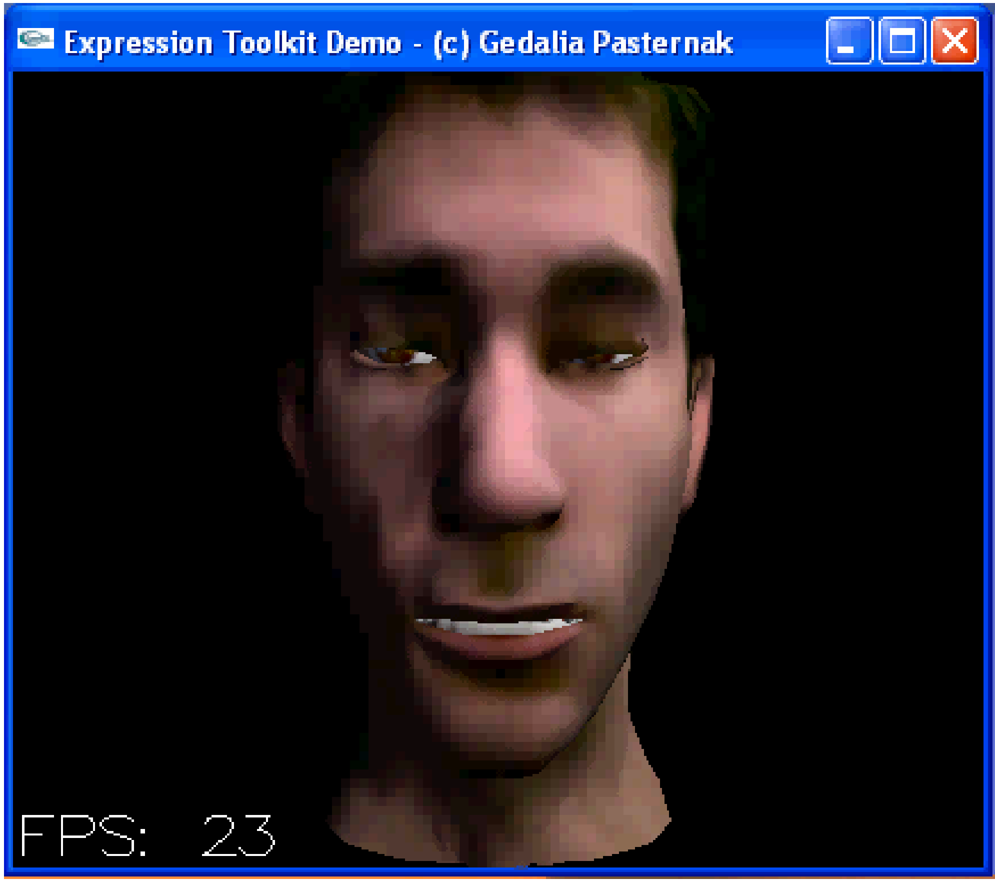
The mask



Neutral 0 %
Happy 57 %
Surprise 0 %
Angry 19 %
Disgust 0 %
Fear 5 %
Sad 18 %

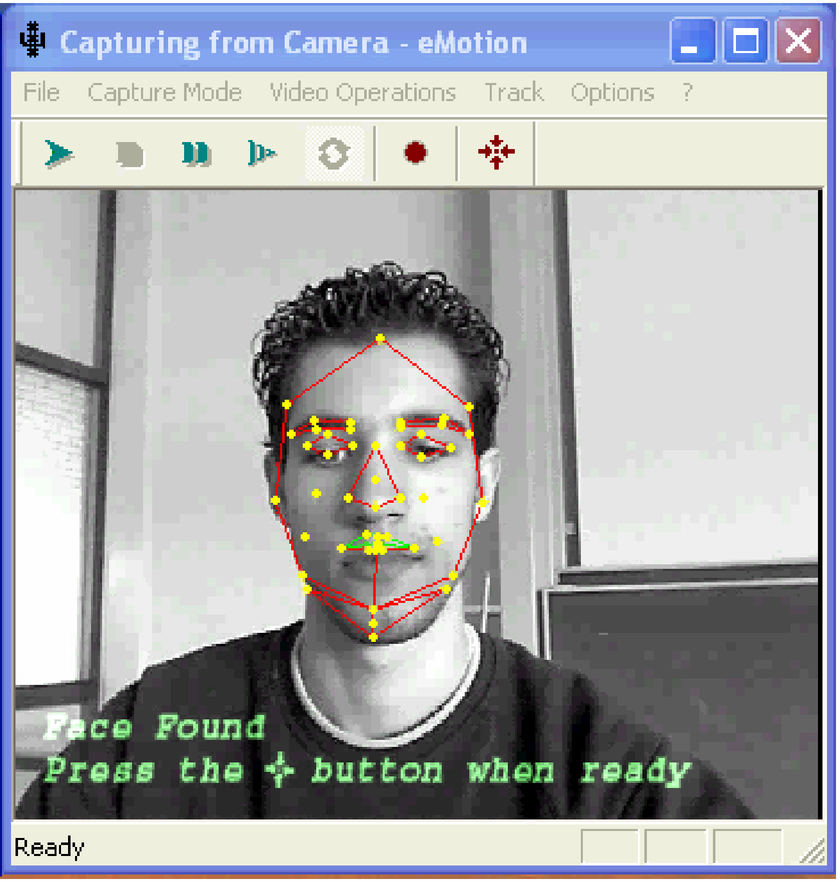
Status: Tracking the face..

Avatar



Expression Toolkit Demo - (c) Gedalia Pasternak

FPS: 23



Capturing from Camera - eMotion

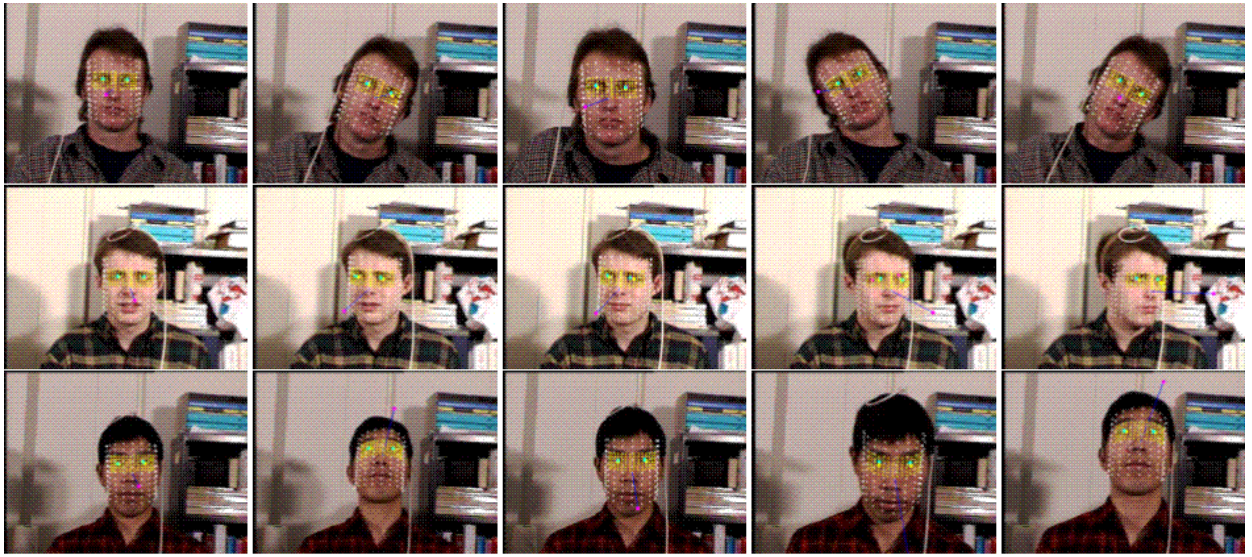
Face Found
Press the ✥ button when ready

Ready

[Valenti, cvpr08]

Human behaviour understanding

- Facial expression
- Head pose
- Eye Tracking
- Voice



Roberto Valenti
Intelligent Systems Lab Amsterdam
University of Amsterdam

Conclusions

- Large scale datasets with annotations
- Color and photometric invariance needed
- Balance between discriminative power and invariance
- Color add information to classification achieving best performance in VOC08/VOC09, TRECVid08/TRECVid09 and ImageCLEF.
- Speed up is required (e.g. GPU)
- Higher semantics like aggression, emotions etc.