

# Resynthesizing audiovisual perception with augmented reality

Parag K Mital  
Department of Computing,  
Goldsmiths, University of London  
<http://pkmital.com>



Presented for Lunch BITES, CULTURE Lab, Newcastle on  
30/06/11

## Questions

What computational processes describe audiovisual perception in the real-world?

What can augmented reality reveal about our underlying perception?

## Objectives

Build computational models of audio-visual attention using controlled experiments

Interpret these models in a real-time context situated in real-life scenarios using augmented reality and re-synthesis techniques

## Modeling

Attention Prior

Spectral/Region Segmentation

Temporal Event Segmentation

## Synthesis

Retrieval/Indexing

Scene Reconstruction

## Modeling

*Attention Prior*

Spectral/Region Segmentation

Temporal Event Segmentation

## Synthesis

Retrieval/Indexing

Scene Reconstruction

# Experimental Psychology

What processes describe human cognition?

Visual cognition

Vision research

Auditory scene analysis

Auditory attention

Psychophysics

Psychoacoustics

Multisensory/Crossmodal perception

Film cognition

# Computational Cognition

What computational models best describe human cognition?

Computer vision

Computational neuroscience

Machine learning

Speech recognition

Saliency models

# Dynamic Images and Eye Movements

John Henderson, Tim Smith, Robin Hill, Parag K Mital  
2008 – 2010 - awarded to John Henderson and funded by  
Leverhulme and ESRC

## Question

What drives human attention and eye-movement behavior during moving images?

## Objectives

Build a corpus of eye-movement data and corresponding moving images

Develop theories and tools for understanding active visual cognition

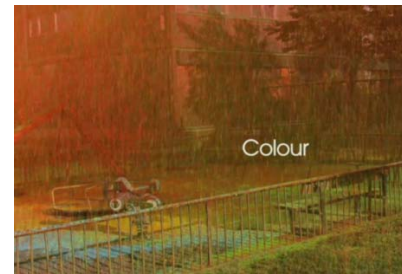
82 videos

Range between 30 seconds and  
3 minutes.

200 viewers+

Broad range of stimuli:

- adverts
- film clips
- real-world scenes
- social scenes
- film trailers
- video game trailers
- music videos
- documentaries
- news clips
- animation



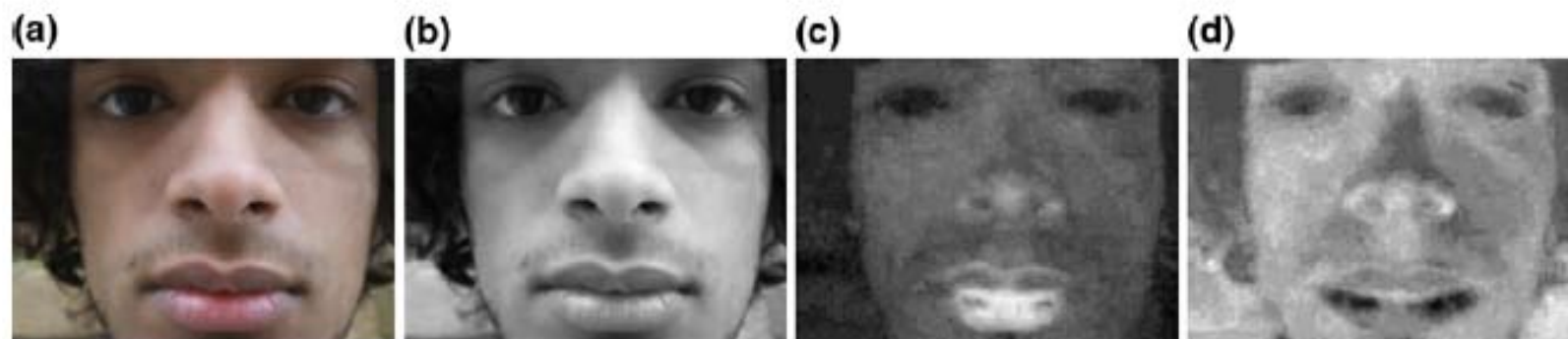
## Eye-tracking data

- X/Y coords of eyes per millisecond per eye per person, plus various eye-movement events and messages.
- >1000 lines of 8-column data per second!

## CARPE

- Gaze videos
- Gaussian Mixture Models
- Low-level feature visualizations
  - Optical flow, edges, gabors, flicker, chromaticity, luminance
- Dynamic Heatmap videos





**Fig. 1** a Original image of frame 1975 of video 24 ('Video Republic' <http://www.demos.co.uk/publications/videorepublic>); b  $L^*$  image depicting luminance (Lum); c  $a^*$  image depicting red/green opponent colors (RG); d  $b^*$  image depicting blue/yellow opponent colors (BY)

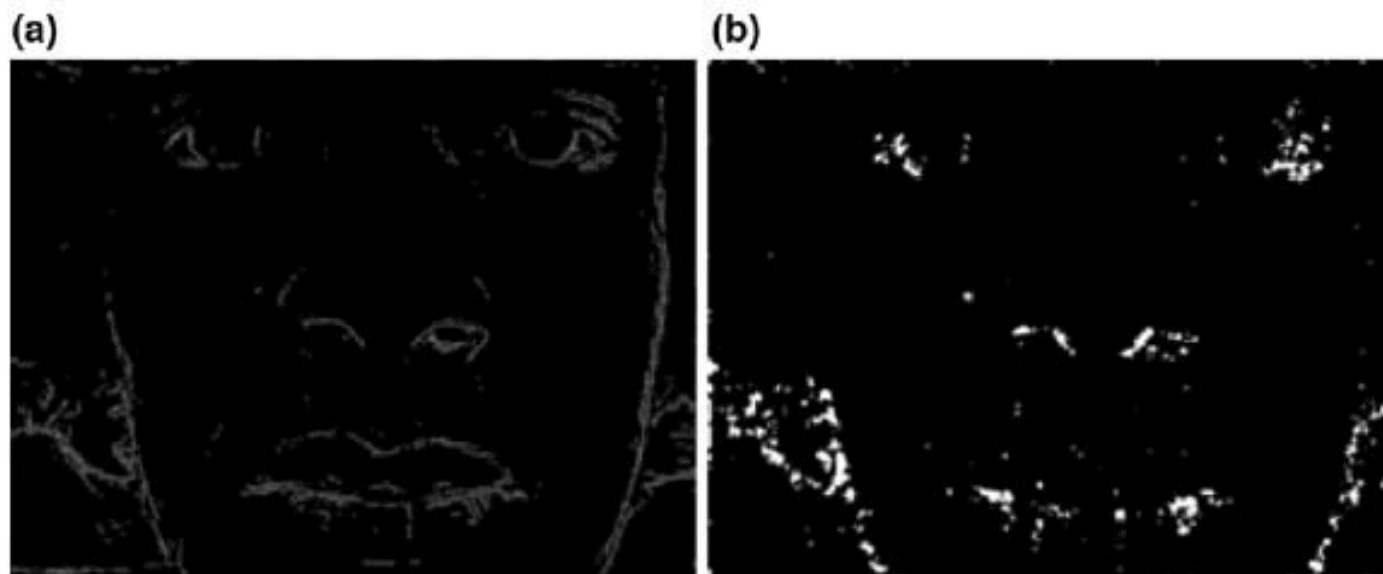
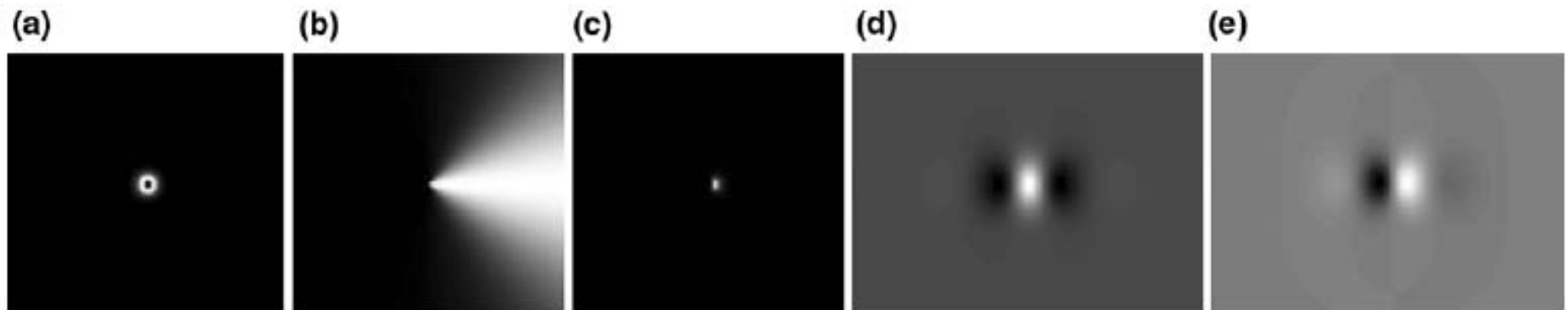
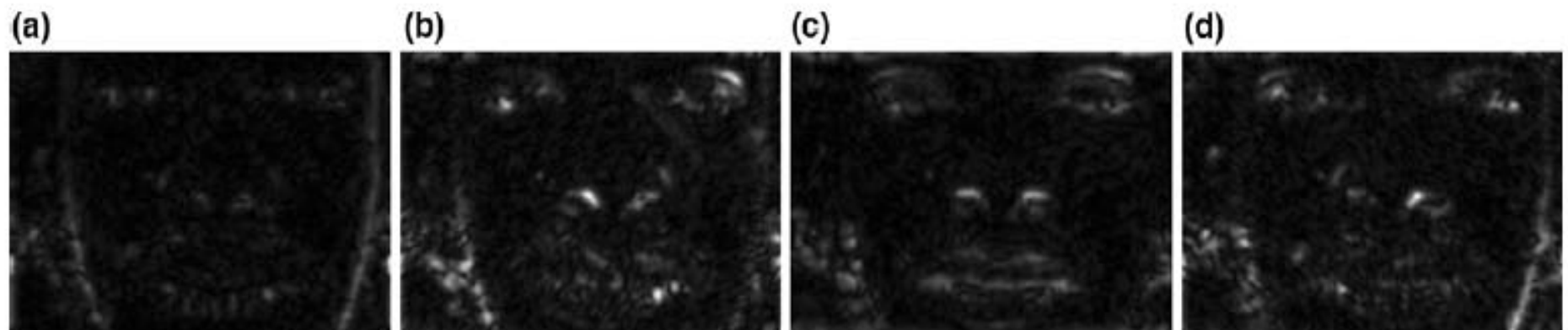


Fig. 2 a Edge map using Sobel edge detection; and b Harris corner detection map for the luminance image in Fig. 1b



**Fig. 3** The process for creating a log-Gabor kernel for  $0^\circ$  (left to right): **a** the radial map computed from multiplying a sinusoid with a Gaussian kernel; **b** the orientation of the kernel set for  $0^\circ$ ; **c** the result of multiplying the radial (**a**) and orientation (**b**) maps; **d** the even

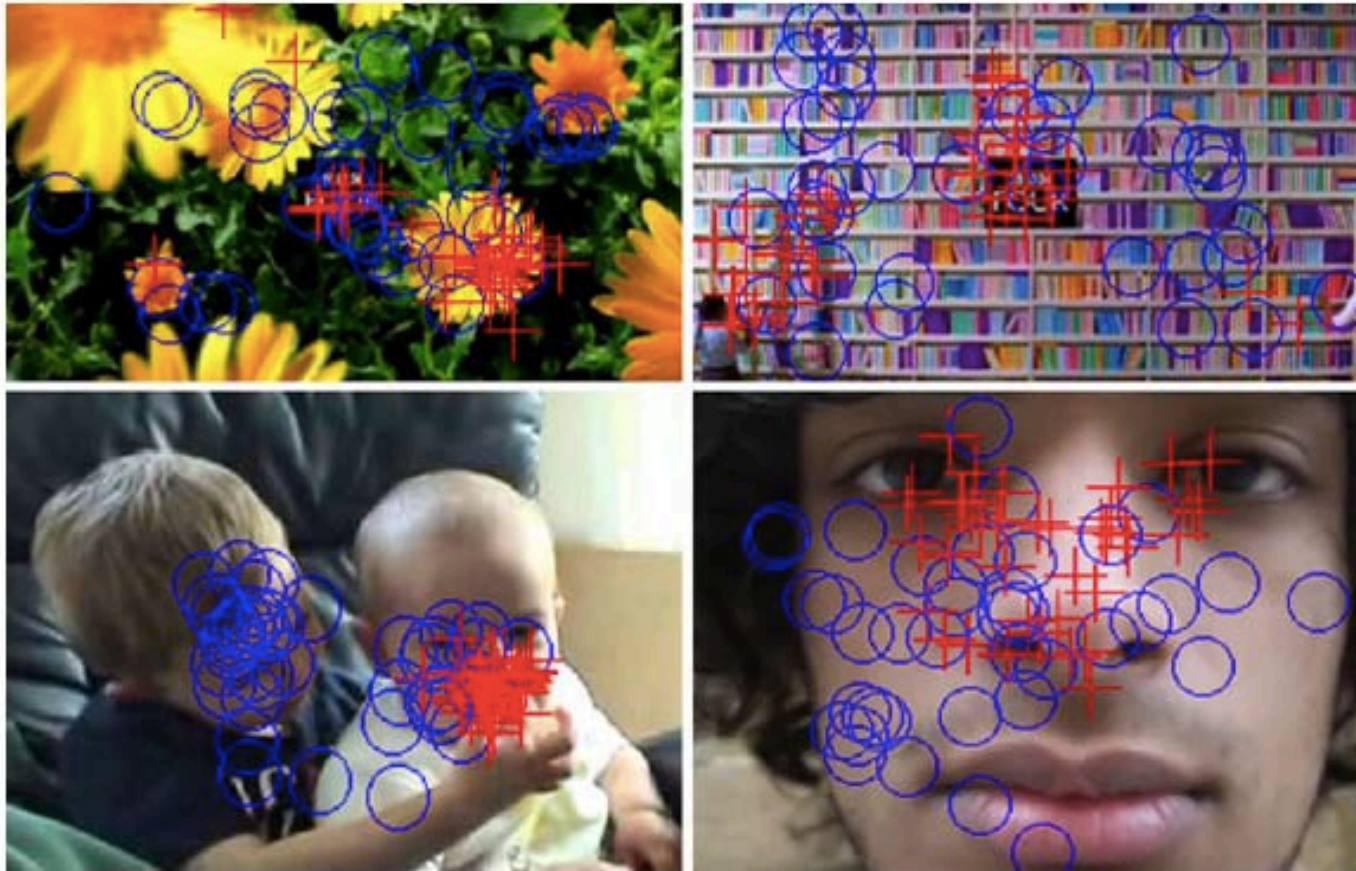
symmetric component of the log-Gabor filter taken from the real part of the inverse fourier transform of the kernel; **e** the corresponding odd symmetric component taken from the imaginary component of the kernel



**Fig. 4** Gabor-oriented maps for **a**  $0^\circ$ , **b**  $45^\circ$ , **c**  $90^\circ$ , and **d**  $135^\circ$  for the luminance image in Fig. 1b

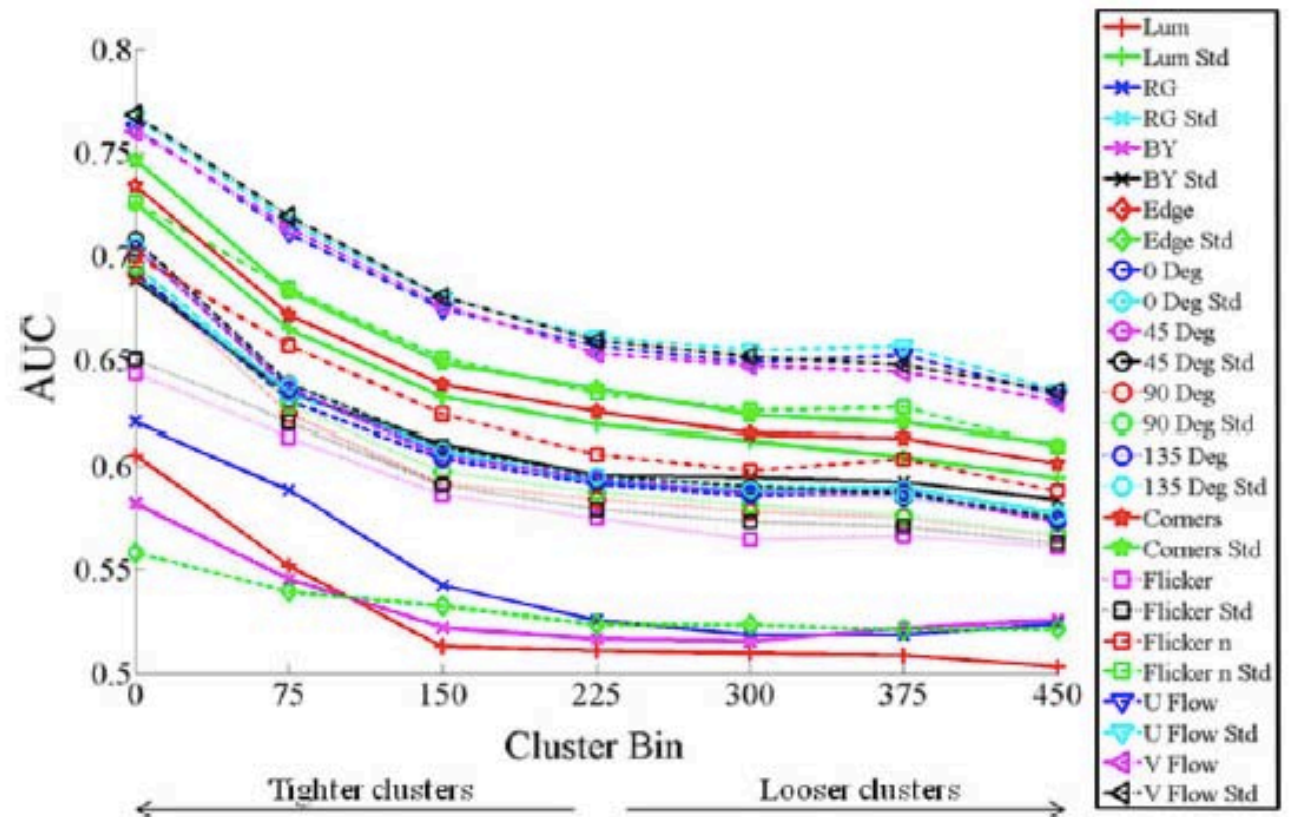


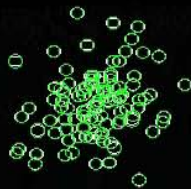
**Fig. 5** **a** High-pass flicker (Flicker); **b** low-pass flicker (Flicker-N); **c** horizontal optical flow (*U*-Flow); **d** vertical optical flow (*V*-Flow) for the frame in Fig. 1a



**Fig. 7** Example actual (*cross*) and baseline (*circle*) subject foveations for videos 1, 2, 15, and 24 (*clockwise from top left*)

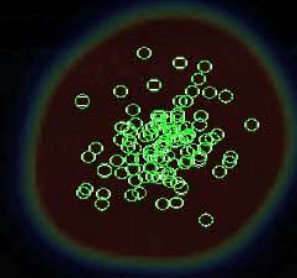
**Fig. 11** Area under ROC curves (AUC) as a function of weighted cluster covariance (bins = 75)





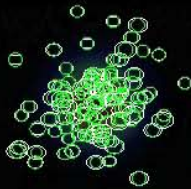
Copyright 2009 (CC-SA-NC) Henderson's Visual Cognition Lab @ Edinburgh University  
(e-mail visual.cognition@ed.ac.uk for information)  
00:00/02:46

Frame: 3



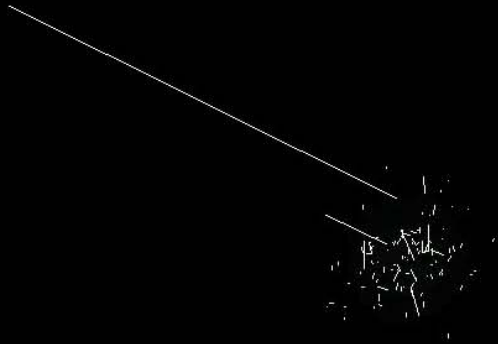
Copyright 2009 (CC-SA-NC) Henderson's Visual Cognition Lab @ Edinburgh University  
(e-mail visual.cognition@ed.ac.uk for information)  
00:00/02:46

Frame: 3



Copyright 2009 (CC-SA-NC) Henderson's Visual Cognition Lab @ Edinburgh University  
(e-mail visual.cognition@ed.ac.uk for information)  
00:00/02:46

Frame: 3



Copyright 2009 (CC-SA-NC) Henderson's Visual Cognition Lab @ Edinburgh University  
(e-mail visual.cognition@ed.ac.uk for information)  
00:00/02:46

Frame: 3

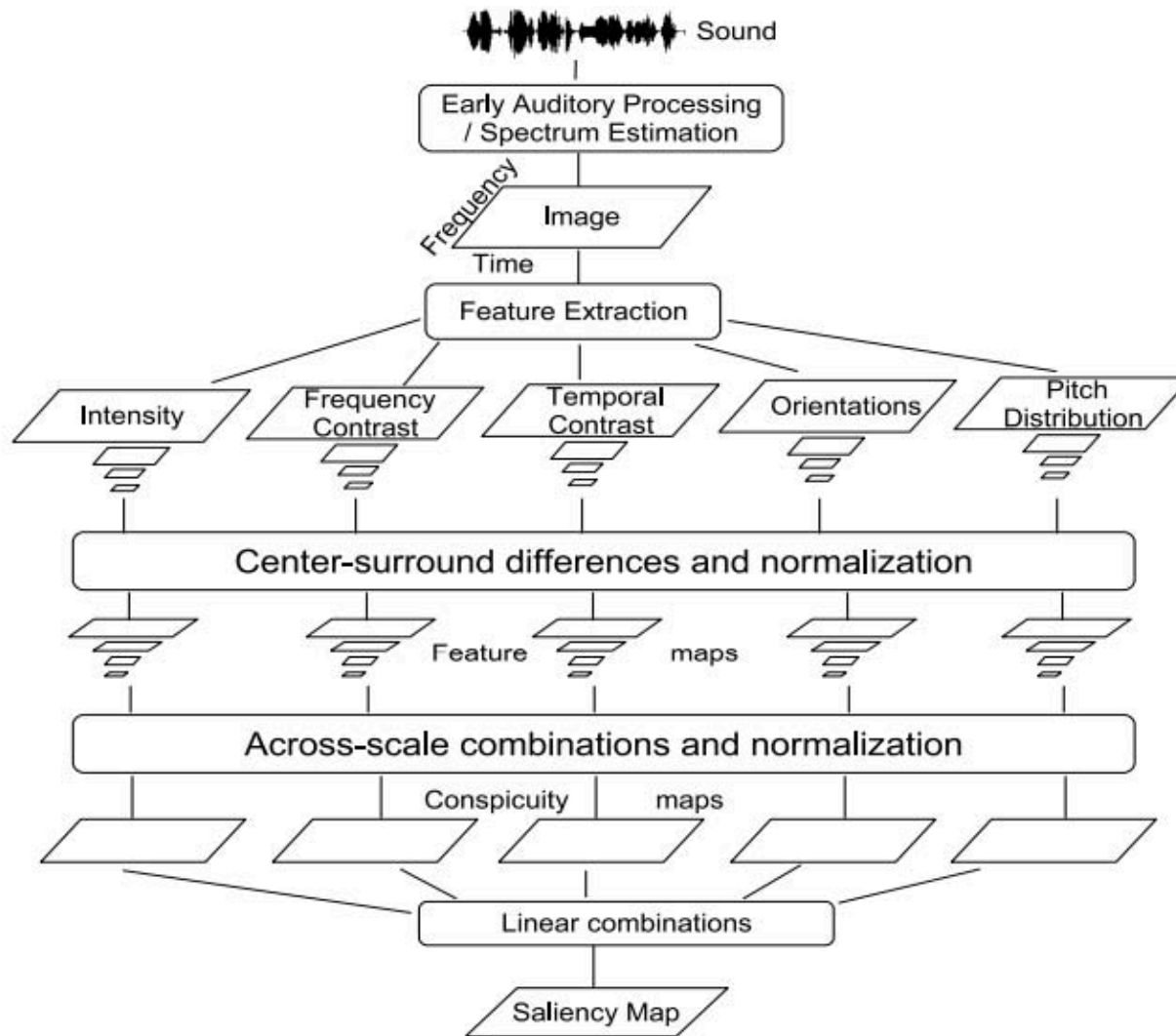
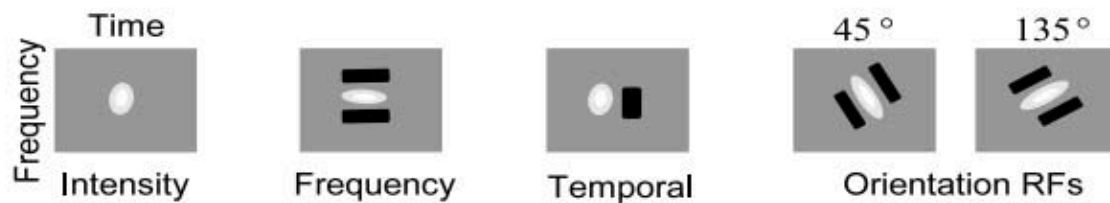


Figure 1: Auditory saliency map structure adapted from [2]



## Modeling

Attention Prior

*Spectral/Region Segmentation*

Temporal Event Segmentation

## Synthesis

Retrieval/Indexing

Scene Reconstruction

# Vision Processing

## Detection

Features (SIFT, SURF, Harris Corners)

Regions (Mean-shift, MSER)

Haar-Features (Boosted Cascades, Viola-Jones)

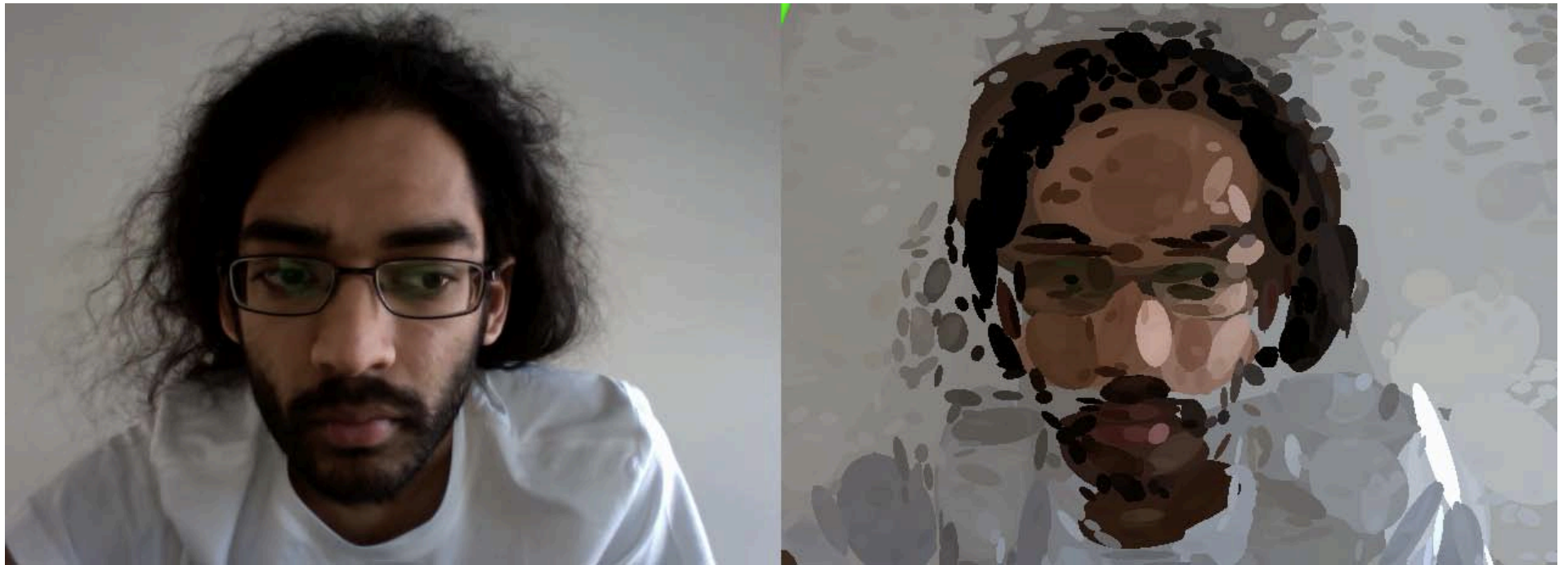
Templates (MI, SSD, Lucas-Kanade)

## Description

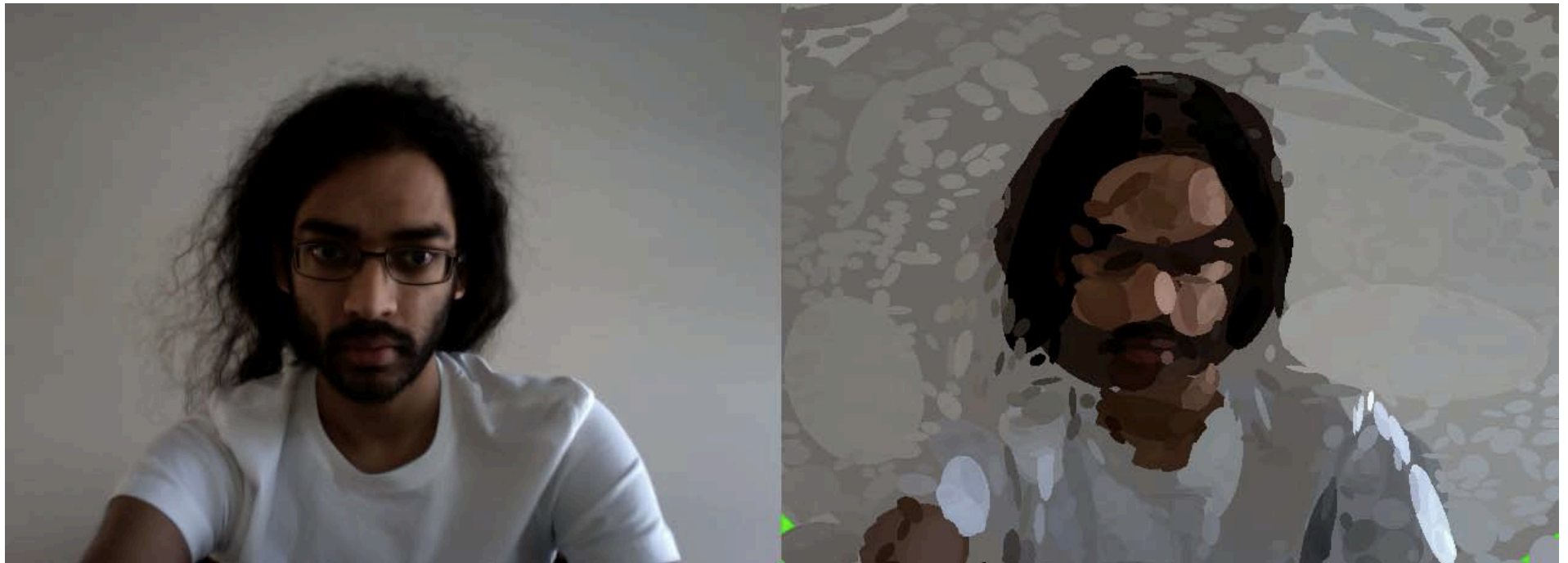
Vector codes (GIST, SIFT, SURF, BRIEF)

Trees (FIANN, LSH)

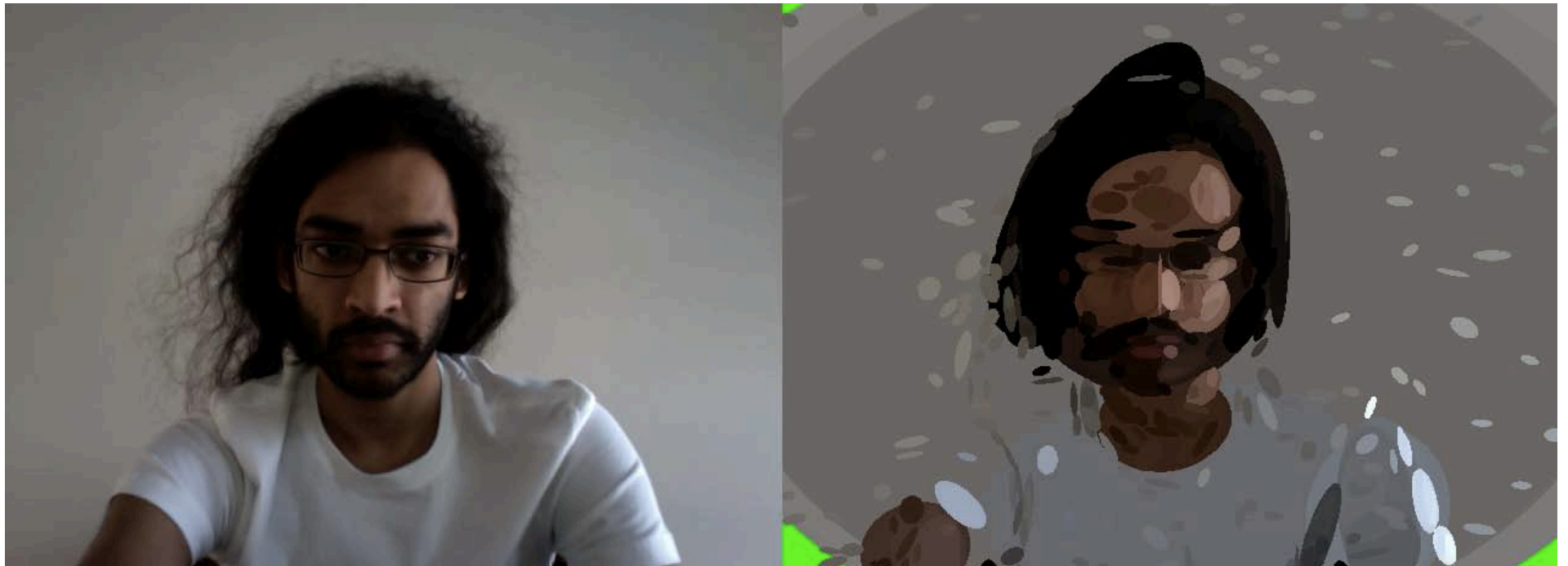
Model-based reconstruction (PCA, pLSA, LDA)



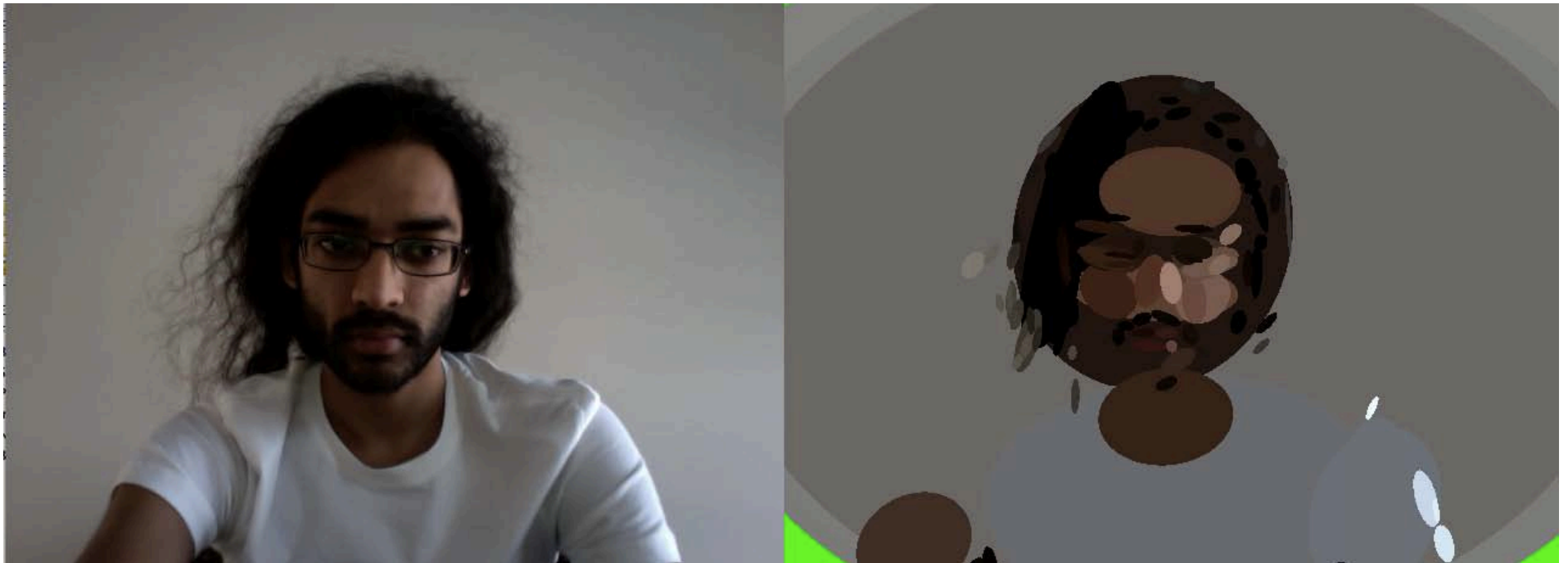
J. Matas, O. Chum, M. Urba, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions." Proc. Of British Machine Vision Conference, pp. 384-396, 2002.  
Stanislav Basovnik, Lukas Mach, Andrej Mikulik, and David Obdrzalek. "Detecting Scene Elements Using Maximally Stable Colour Regions" IEEE Computer Vision and Pattern Recognition, 2007.



J. Matas, O. Chum, M. Urba, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions." Proc. Of British Machine Vision Conference, pp. 384-396, 2002.  
Stanislav Basovnik, Lukas Mach, Andrej Mikulik, and David Obdrzalek. "Detecting Scene Elements Using Maximally Stable Colour Regions" IEEE Computer Vision and Pattern Recognition, 2007.



J. Matas, O. Chum, M. Urba, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions." Proc. Of British Machine Vision Conference, pp. 384-396, 2002.  
Stanislav Basovnik, Lukas Mach, Andrej Mikulik, and David Obdrzalek. "Detecting Scene Elements Using Maximally Stable Colour Regions" IEEE Computer Vision and Pattern Recognition, 2007.

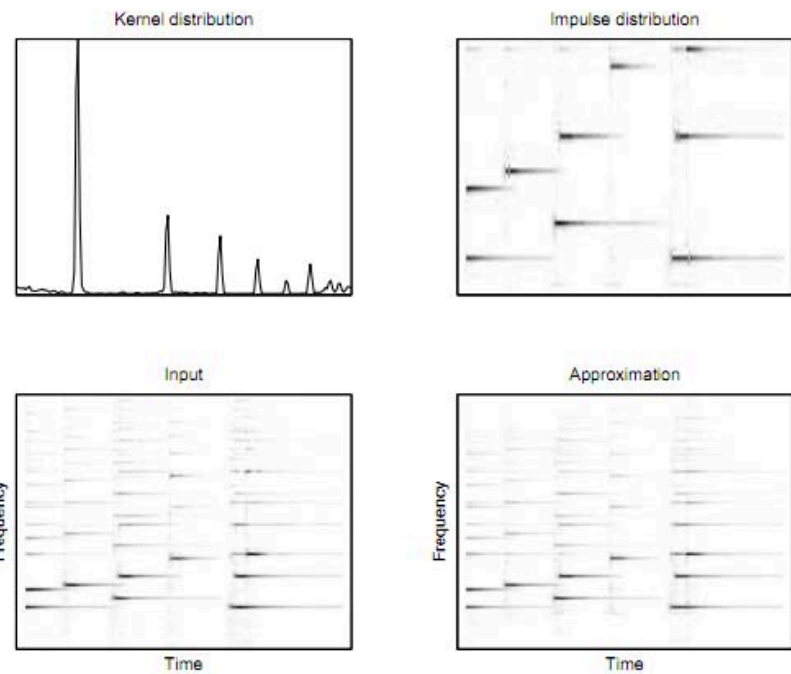


J. Matas, O. Chum, M. Urba, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions." Proc. Of British Machine Vision Conference, pp. 384-396, 2002.  
Stanislav Basovnik, Lukas Mach, Andrej Mikulik, and David Obdrzalek. "Detecting Scene Elements Using Maximally Stable Colour Regions" IEEE Computer Vision and Pattern Recognition, 2007.

# Source Separation

## Question

How can we describe a chunk of audio in terms of semantic factors?



## Modeling

Attention Prior

Spectral/Region Segmentation

*Temporal Event Segmentation*

## Synthesis

Retrieval/Indexing

Scene Reconstruction

## Modeling

Attention Prior

Spectral/Region Segmentation

~~Temporal Event Segmentation~~

## Synthesis

*Retrieval/Indexing*

Scene Reconstruction

# Interpreting the Model in Real-Time

## Question

How can technology employing  
cognitive models help us to  
better understand the model?

# Human-Computer Interaction

## Question

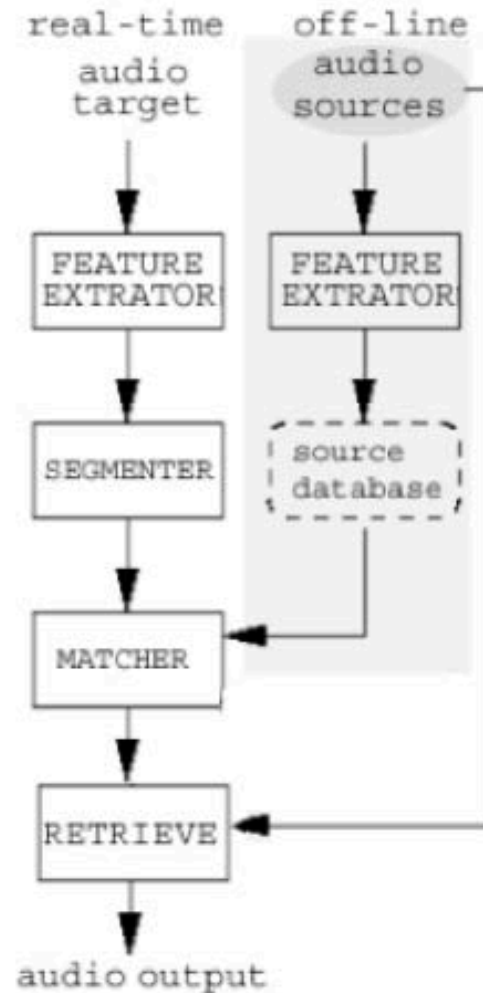
How can we build interfaces to our own perceptual processes?

- Augmented reality
- Interfaces for musical expression
- Robot perception

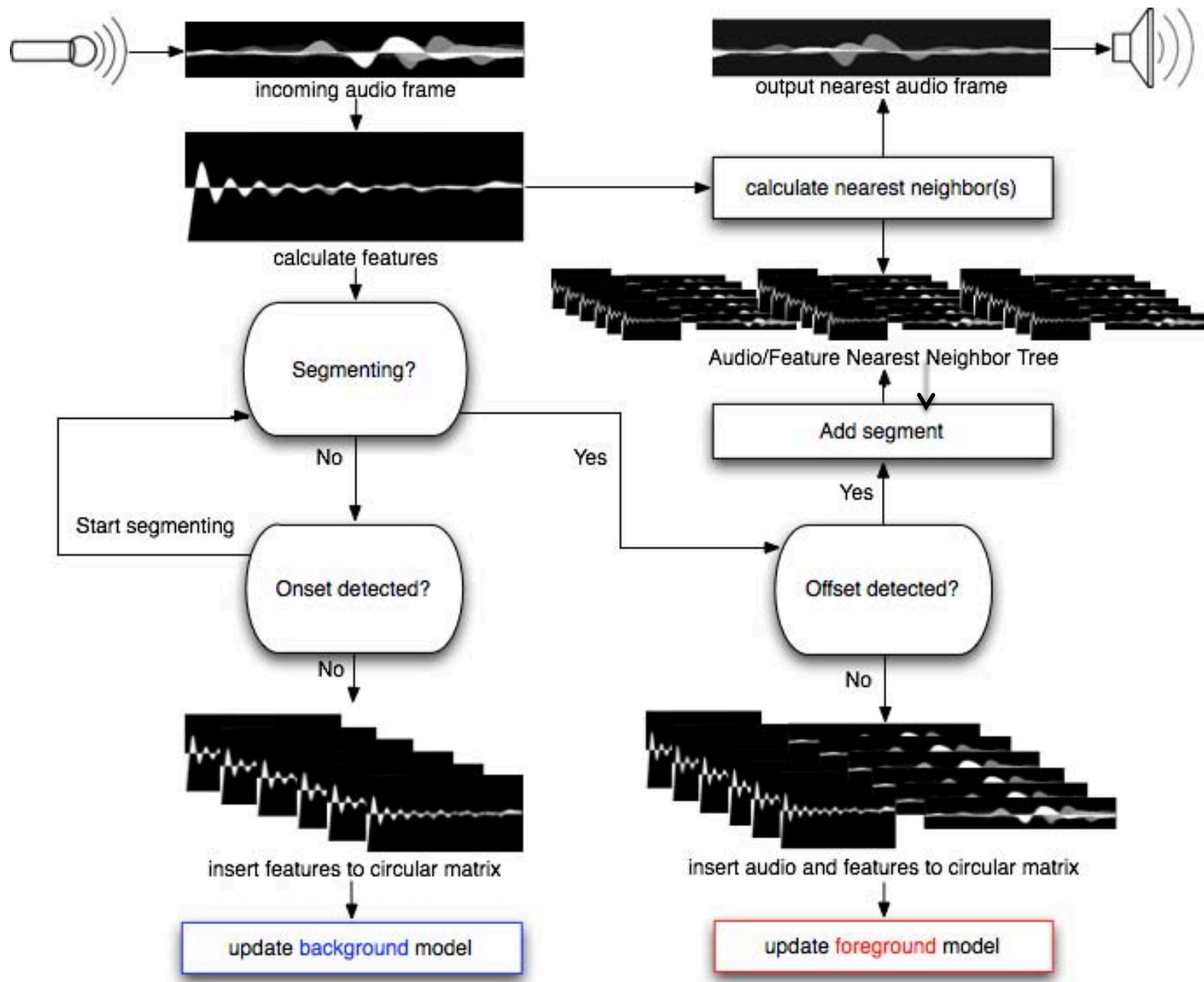
# Corpus based resynthesis

- Catart
- Soundspotter

“A new approach to creating musical streams by selecting and concatenating source segments from a large audio database using methods from music information retrieval” (Casey, 2009)



Casey, M. 2009. Soundspotting: a new kind of process?. In The Oxford Handbook of Computer Music, ed. R. Dean. 421–53. New York: Oxford University Press.



## Modeling

Attention Prior

Spectral/Region Segmentation

Temporal Event Segmentation

## Synthesis

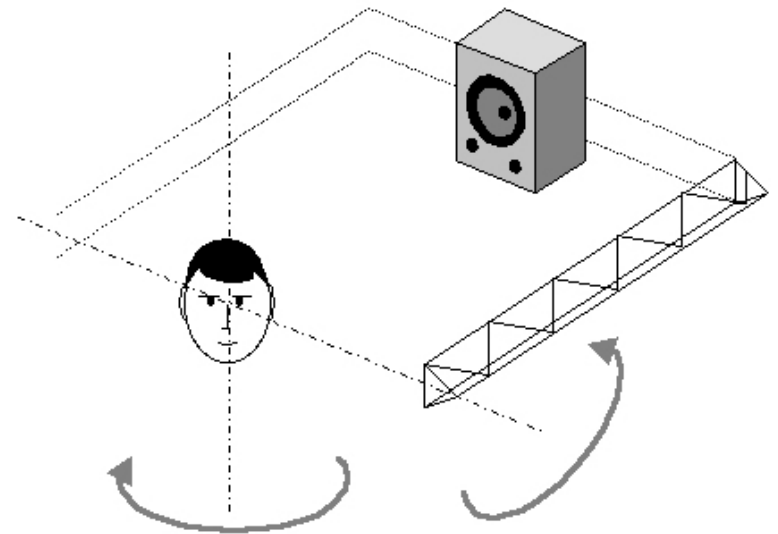
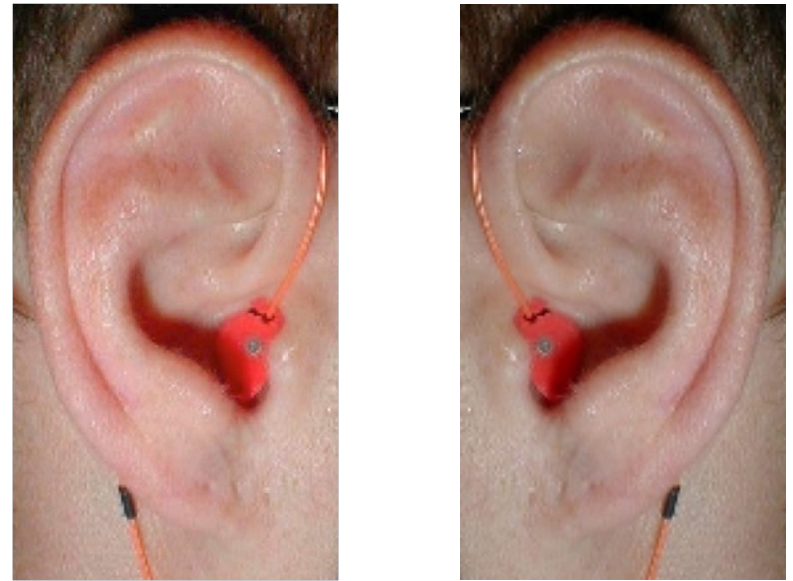
Retrieval/Indexing

*Scene Reconstruction*

# Sound Spatialization

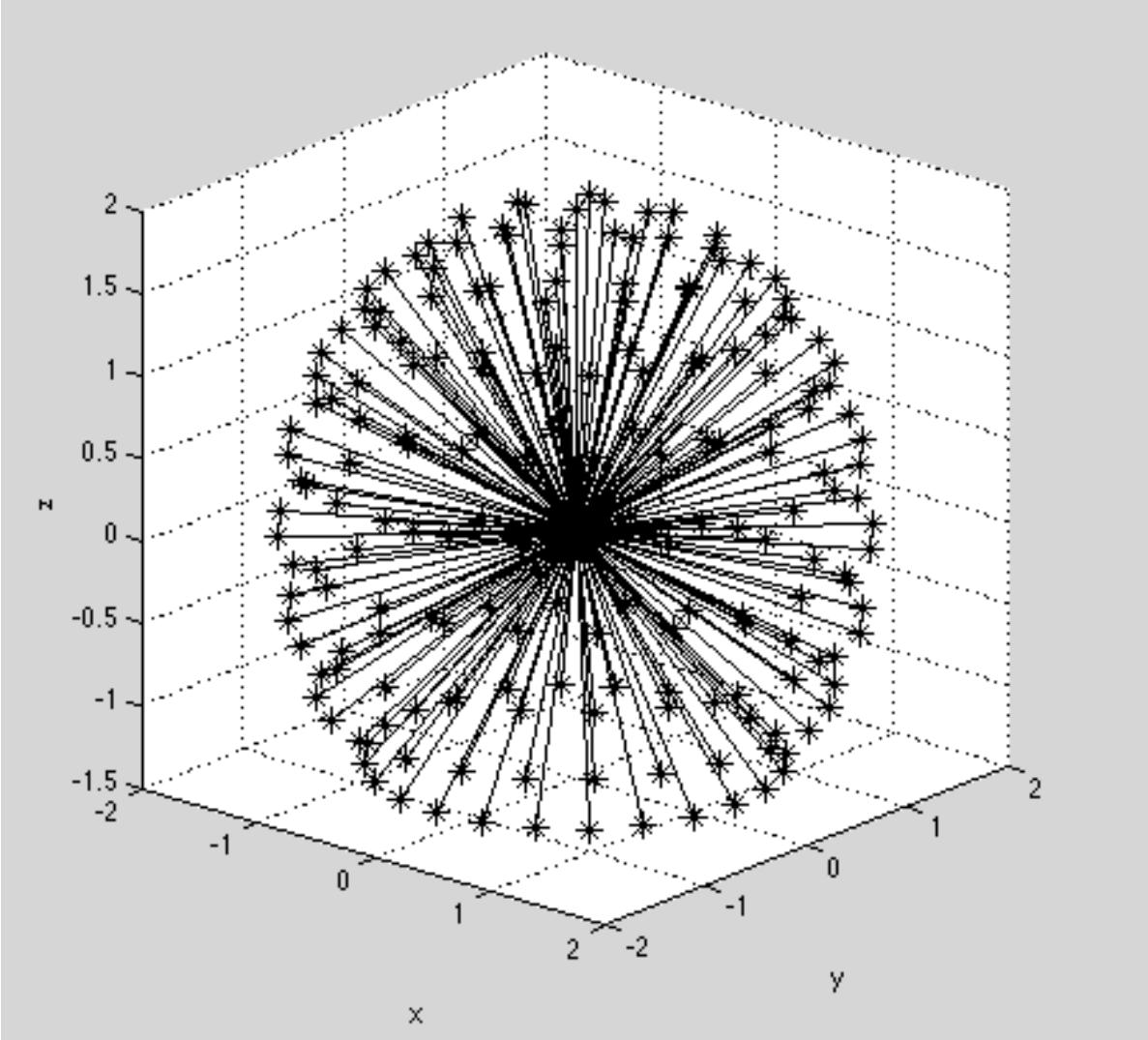
HRIR using both MIT and  
IRCAM LISTEN<sup>1</sup>

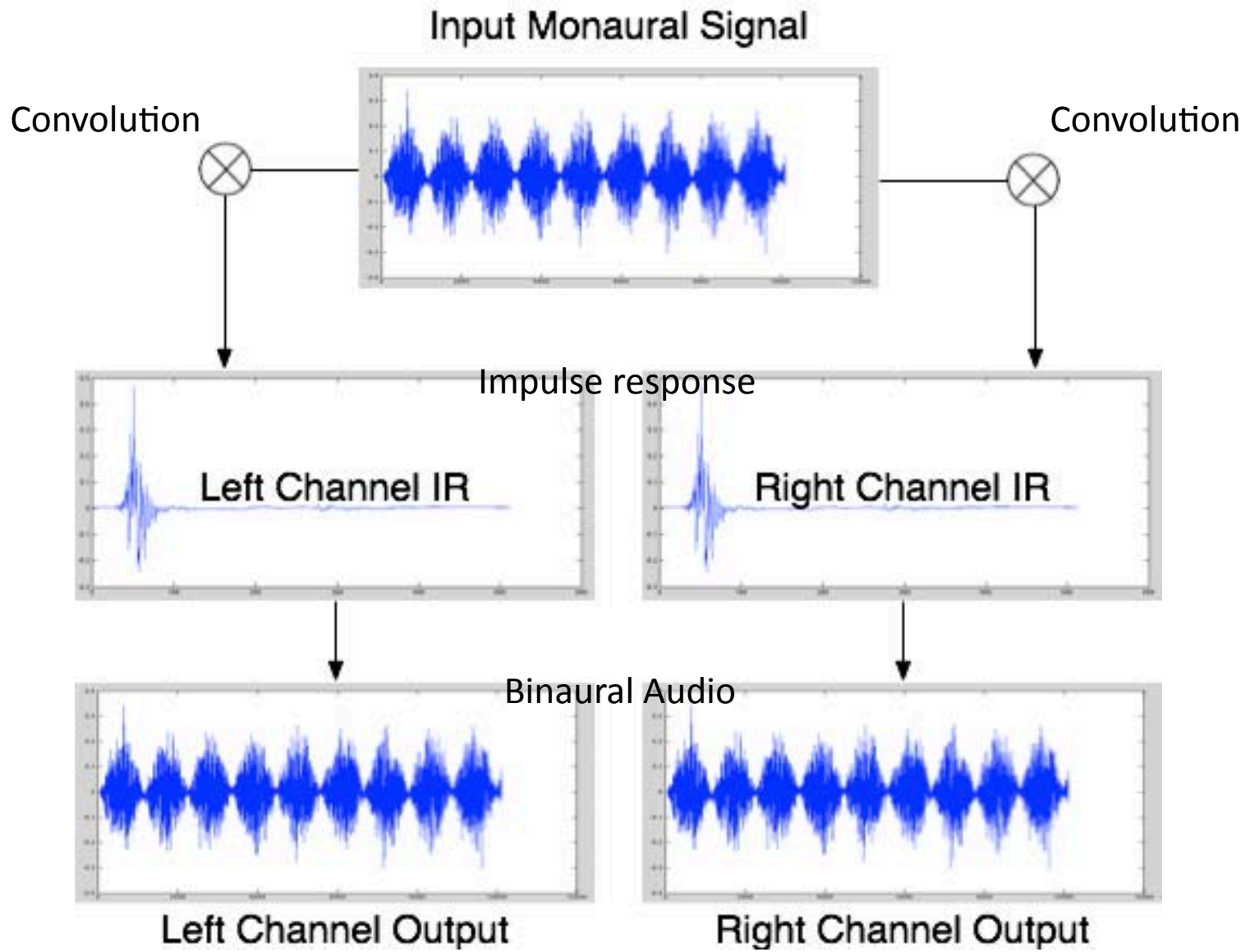
Perceptual filter encoding  
source of sound



[1]. <http://recherche.ircam.fr/equipes/salles/listen>

# Location of Impulse Responses







<http://pkmital.com>  
[parag@pkmital.com](mailto:parag@pkmital.com)