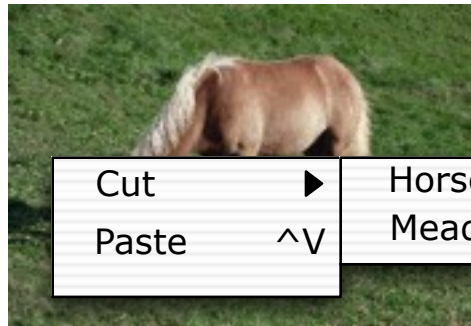


Introduction to Multimedia Semantics



Dr. Gerald Friedland
International Computer Science Institute
Berkeley, CA
fractor@icsi.berkeley.edu



Introduction to Semantic Multimedia Analysis

This presentation includes material from slides by:

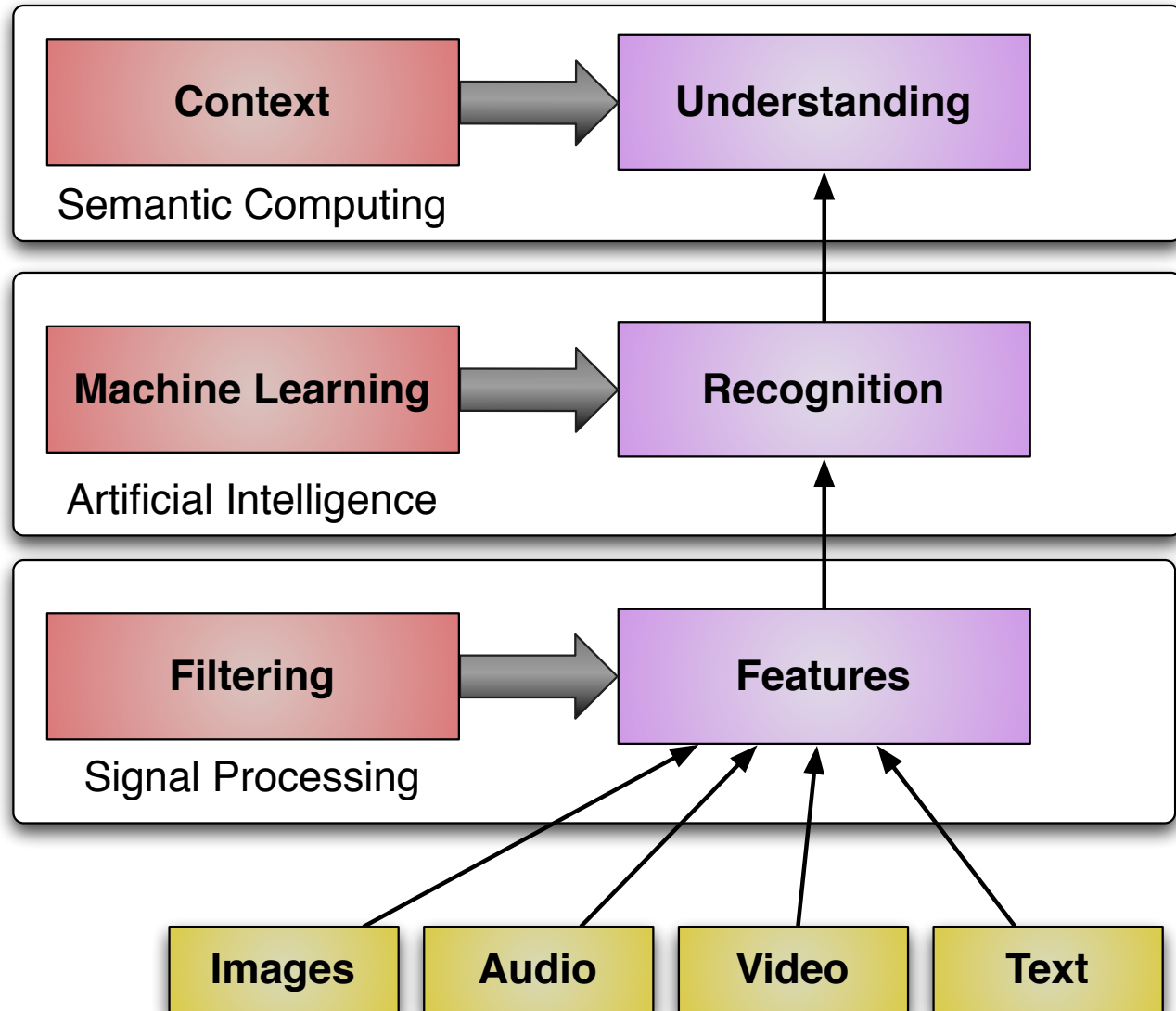
Dr. Shu-Ching Chen

Professor, School of Computing and
Information Sciences Florida International
University, Miami, Florida, USA

Overview

- Introduction and Examples
- Visual Representation
- Acoustic Content Representation
- Modeling Techniques
- Multimodal Integration
- Concluding Remarks

Multimedia Analytics





Part I

Definitions and Examples



Multimedia: Definition

Entry:

multimedia

Function:

noun plural but singular or plural in construction

Date:

1950

A technique (as the combining of sound, video, and text) for expressing ideas (as in communication, entertainment, or art) in which several media are employed; also: something (as software) using or facilitating such a technique.

(Merriam-Webster online dictionary)



Multimedia Content Extraction

Automatic analysis of the content (semantics) contained in data directly encoded for human perception (audio, images, video, touch) and its associated meta data (natural text, computer-encoded data).








Typical Problems

- Which videos contain person X
- Is this upload copyright infringement?
- Who does what when and where in this scene?
- What are the speakers, scenes, objects, narrative themes in this video collection?

Semantic Analysis of Multimedia Data

- enables automatic logical inference on perceptually encoded data
- enables more “natural” interaction with the computer: “do what the user means”
- assists in the creation of content: “do what the author means”

Image Retrieval

							
1) 0.00 29020.jpg	2) 8.16 29077.jpg	3) 12.23 29005.jpg	4) 12.64 29017.jpg	5) 13.82 20003.jpg	6) 14.52 53062.jpg	7) 14.70 29018.jpg	8) 14.78 29019.jpg

Y. Rubner, C. Tomasi, and L. J. Guibas: The Earth Mover's Distance as a Metric for Image Retrieval. *Int. Journal of Computer Vision*, 40(2):99–121, 2000.



Image Retrieval: Surfing Shoes

ShoeSurfer: Random

http://www.shoesurfer.com/beta/

Aktuelle Nachrichten... Gmail dba db24 Bank of America Google Recent Earthquakes ... ABC.com Full Episod... JFerret - Home -- M... #118 (mac os x Run...

ShoeSurfer
by Implicit Interfaces

Displaying: Random shoes (click on one!)

Search coming soon...

BRUNOMAGLI: Gabor
\$328.95



Image Retrieval: Surfing Shoes

ShoeSurfer: BRUNOMAGLI: Gabor

http://www.shoesurfer.com/beta/

Aktuelle Nachrichte... Gmail diba db24 Bank of America Google Recent Earthquakes ~... ABC.com Full Episod... JFerret - Home - M... #118 (mac os x: Run...

ShoeSurfer

by Implicit Interfaces

Displaying: Best matches for this shoe

Search coming soon...

BRUNOMAGLI: Gabor **\$328.95**
Sizes in stock: Unknown

Fertig

0.215s



Image Retrieval: Surfing Shoes

ShoeSurfer: Badgley Mischka: Shiloh

http://www.shoesurfer.com/beta/

Aktuelle Nachrichte... Gmail dba db24 Bank of America Google Recent Earthquakes ... ABC.com Full Episod... JFerret - Home -- M... #118 (mac os x; Run...

ShoeSurfer

by Implicit Interfaces

Displaying: Best matches for this shoe

Search coming soon...

Badgley Mischka: Shiloh **\$548.95**
Sizes in stock: Unknown

Quick-view this shoe at Zappos.com.

Fertig 0.807s



Image Retrieval: Surfing Shoes

ShoeSurfer: Badgley Mischka: Shiloh at Zappos

http://www.shoesurfer.com/beta/

Aktuelle Nachrichten... Gmail dba db24 Bank of America Google Recent Earthquakes ... ABC.com Full Episod... JFerret - Home -- M... #118 (mac os x: Run...

ShoeSurfer

by Implicit Interfaces

Displaying: Zappos.com

Search coming soon...

Return to ShoeSurfer Open Zappos.com in a new window

Zappos Sites: Zappos.com Couture KidsShop Running Outdoor Bags Shopping Cart | My Account | My Favorites | Help

Zappos

POWERED BY SERVICE™

Shoes Clothing Bags Accessories Brands On Sale Search

Women Men Kids Gift Ideas New Styles

Free Shipping & Free 365 Day Returns 24/7 Customer Service 1-800-927-7671 (Español 1-800-927-4104)

Search: Search Zappos Search by Shoe Size

Popular Searches: shoes, ugg, nike, nike shoes, womens shoes, ugg boots, ugg, wide shoes, heelys, dansko, keen shoes, crocs, clarks shoes, mtb shoes, frye boots, snow boots, cowboy boots, new balance, born shoes, stuart weitzman, dansko shoes, boots, donald pliner, donald pliner shoes, sandals, clothing, womens boots, leather shoes, mens shoes

Way to Pay! No Wallet. No Wait. No Worries.™ with **BillMeLater** Select to Credit Approval Decl.

[Back to Browsing](#) [Browse More Styles...](#)

Do you own this product? ** [Be the first one to review it!](#) **

BADGLEY MISCHKA FOOTWEAR

[Larger Image & Additional Views](#)



SKU #7339371

SIZE: WIDTH: M (standard)

COLOR: Mushroom

In Stock

FREE SHIPPING BOTH WAYS!
Receive your order within 4-5 business days with free standard shipping!
[Need it faster?](#)

- You won't go unnoticed in this dazzling rhinestone detailed sandal.
- Adjustable ankle strap.
- Combination leather and satin upper.
- Leather lining.

SKU #7339371

MULTI VIEW

FREE SHIPPING BOTH WAYS

CUSTOMER TESTIMONIALS SHOP FOR LATEST STYLES

2007 **MAJORITY** **REWARDS** **2007**

365-DAY RETURN POLICY

Canadian Shoppers
Visit [canada.zappos.com](#)

Zappos Rewards Visa Card
 Earn 1,500 points with first use.

Contact Us - open 24 hours!
800-927-7671
[help@zappos.com](#)
24/7, 365-Days a year!
(or, order by fax)

[Home](#)
[Brand List](#)
[Latest Styles](#)
[Shoes](#)
[Clothing](#)

Examples: Cut & Paste in Images

**SIOX
Button**



G. Friedland, K. Jantz, T. Lenz, F. Wiesel, R. Rojas: "Object Cut and Paste in Images and Videos", International Journal of Semantic Computing Vol 1, No 2, pp. 221-247, World Scientific, USA, June 2007.

Speaker Diarization: Who Spoke When?

Audio Track:



Segmentation:



Clustering:



G. Friedland, O. Vinyals, Y. Huang, C. Müller: “Prosodic and other Long-Term Features for Speaker Diarization”, IEEE Transactions on Audio, Speech, and Language Processing, Vol 17, No 5, pp 985--993, July 2009.

Analyzing Meetings

The screenshot displays the 'Ferret for ISSCO Meeting 024' interface. At the top, three video windows show different views of a meeting. Below the videos are 'Video' and 'Slides' tabs. The main interface is divided into several panels:

- Left Panel:** Features the 'Ferret!' logo (a stylized 'A' with a paw print and a tiger) and a timeline for the 'ISSCO Meeting' from 0:06-23. It includes 'Play', 'Pause', and 'Stop' buttons.
- Center Panel:** A transcript of the meeting with time-coded text. Examples include:
 - [03:43] o
 - [03:45] there's this furniture
 - [03:45] yeah
 - [03:46] yeah
 - [03:47] o
 - [03:47] that we're talking about
 - [03:51] yeah
 - [03:52] have that and and uh and i decorations and i get clients is or reduce batch is the decorations i think we need to get the french data away for since they get from guaz it into space and
 - [03:53] then it what are they going on the yeah in the space
 - [04:04] meeting in as many
 - [04:04] you know
- Right Panel:** A list of 'Personal Preferences' including:
 - Designer woods
 - Carvings or engravings on order
 - Solar cells
 - Microphone and Speaker on advanced chip
 Below this is a 'Real Reaction' section and a 'Findings' section with the text 'Materials for curved case'.

Wellner, Pierre, Flynn, Mike, Tucker, Simon and Whittaker, Steve (2005): “A meeting browser evaluation test”, Proceedings of ACM CHI 2005 Conference on Human Factors in Computing Systems 2005. pp. 2021-2024.



Copyright Detection

Experiment.swf (application/x-shockwave-flash-Objekt)

http://ec2-72-44-49-80.z-1.compute-1.amazonaws.com/Experiment.swf


Aktuelle Nachrichte... Gmail diba db24 Bank of America Google Recent Earthquakes ~... ABC.com Full Episod... JFerret - Home - M... #118 (mac os x: Run...)

ShoeSurfer: Camper: Minie-29... Experiment.swf (application/x-... Gmail - Inbox - fractor@gmail... SIOX: Simple Interactive Object ...

Cruxle Copyright Detector Demo

VIDEOS TO BE TAKEN DOWN

Perfect Match

 Everyone_likes_Apple


Very High Probability Match


High Probability Match

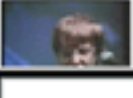
Medium Probability Match (0)

Low Probability Match (1)

Videos in the database

 Bill_Gates_and_Steve_Jobs


 David_Letterman_-_Earthqu


 Everyone_likes_Apple_Com

Upload

Provide Youtube URL in the above box


Copyrighted videos to be monitored

 Bill_Gates_Praising_Apple_C

 Letterman_s_tribute_to_Bill

Upload

Provide Youtube URL in the above box



Bill_Gates_Praising_Apple_Comp_20s.flv

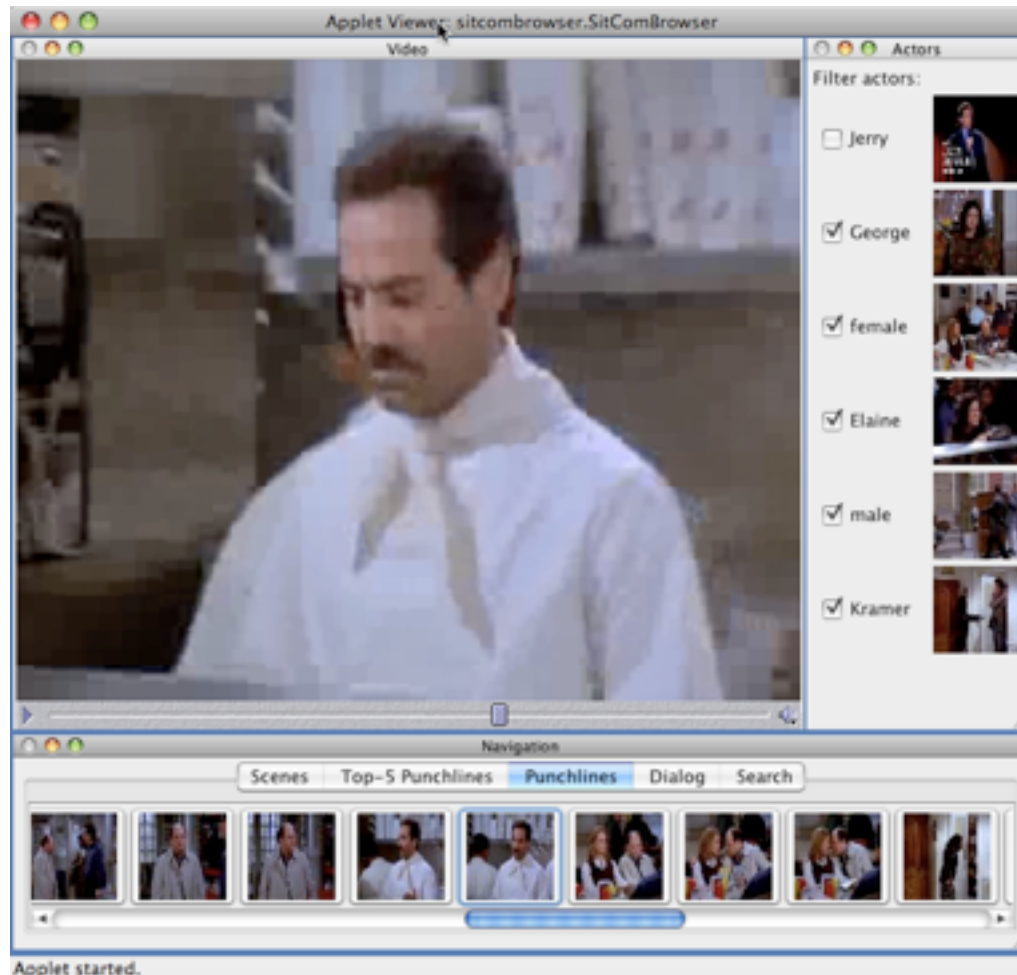
1 copyright violations identified. Please find those videos in the right panel. Do you want to take them down now ?

Yes, Delete All Delete Later

Delete Video Select the video to be deleted and click "Delete Video" button

ec2-72-44-49-80.z-1.compute-1.amazonaws.com/geisen 12.101s

Narrative Theme Navigation



G. Friedland, L. Gottlieb, A. Janin: "Joke-o-mat: Browsing Sitcoms Punchline by Punchline", Proceedings of ACM Multimedia, Beijing, China, pp.1115-1116, October 2009.

Location Estimation



G. Friedland, O. Vinyals, T. Darrell: *Multimodal Location Estimation*, accepted as full paper at ACM Multimedia, Florence, Italy, October 2010.



Part II

Visual Content Representation

An Image...

An Image:

8x8 pixel
block

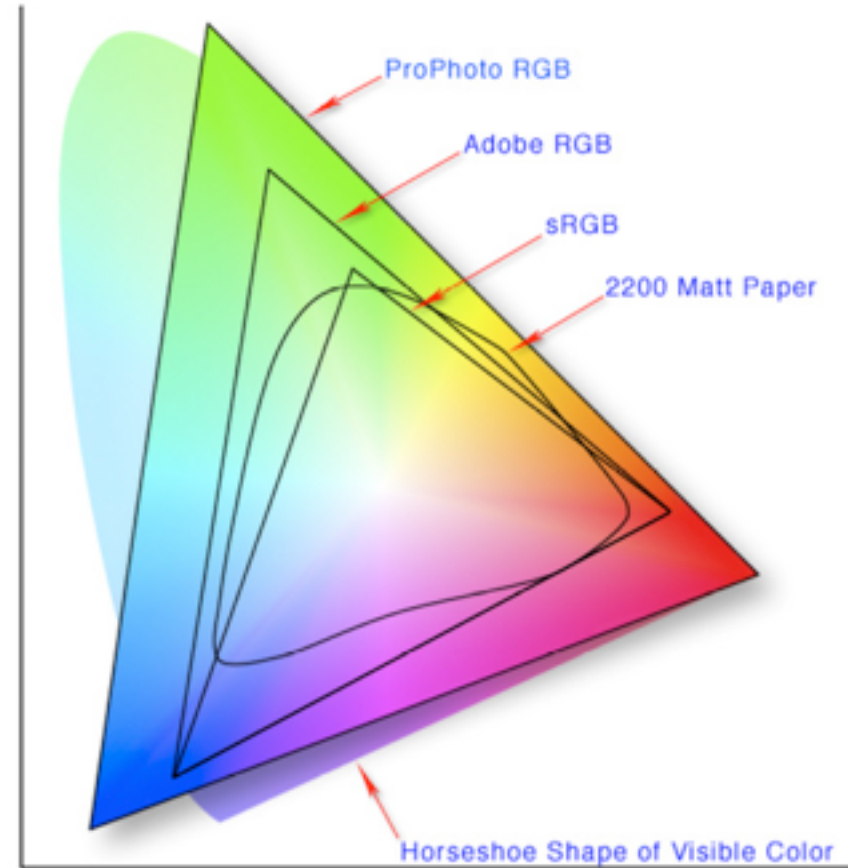
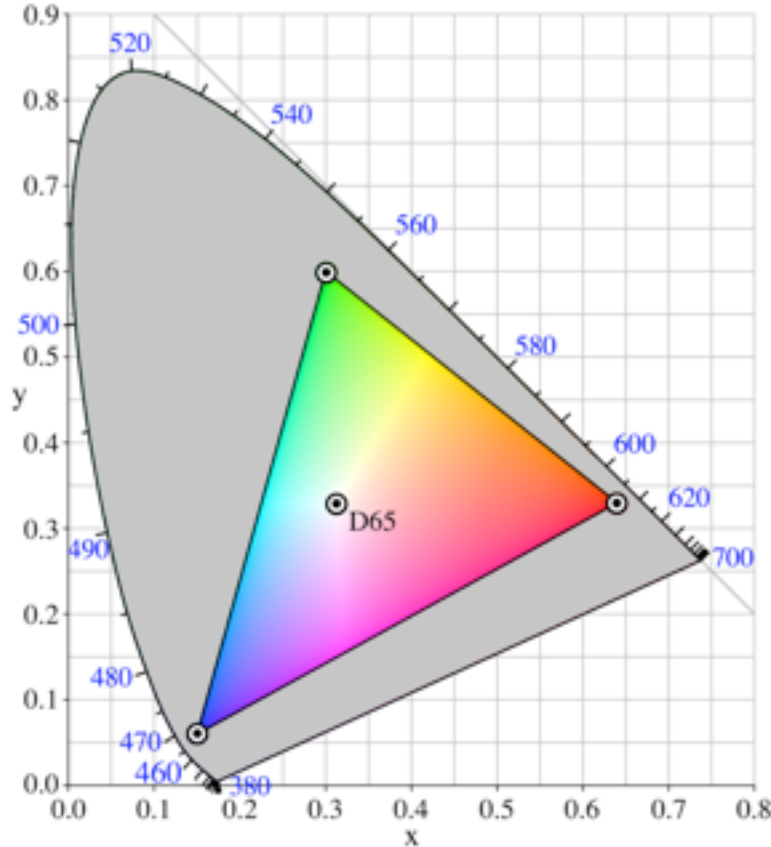


Source: bigfoto.com



RGB Color Space

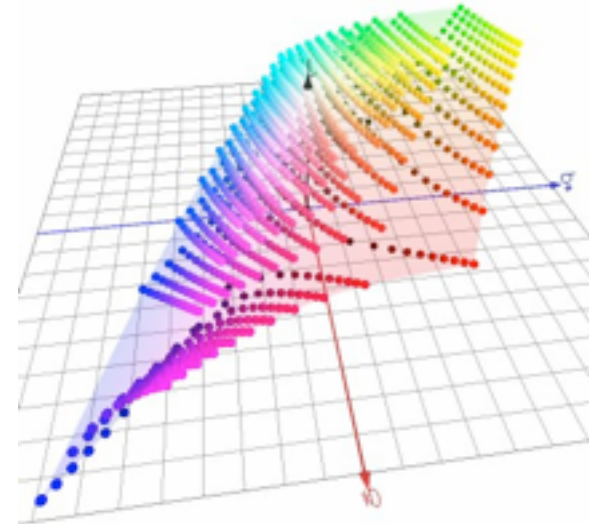
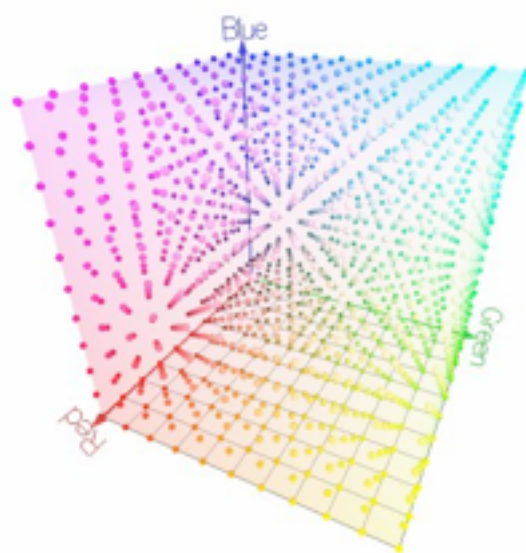
INTI
COMI
I N



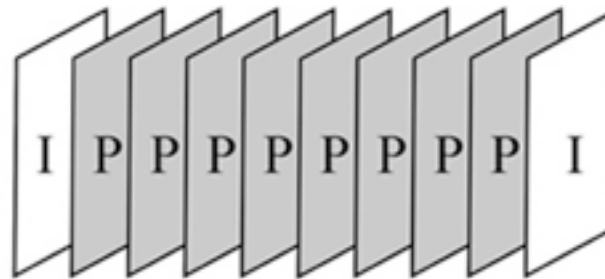
- RGB (Red-Green-Blue) encoding three bytes enabling $(2^8)^3=2^{24}$ colors.

Color Spaces

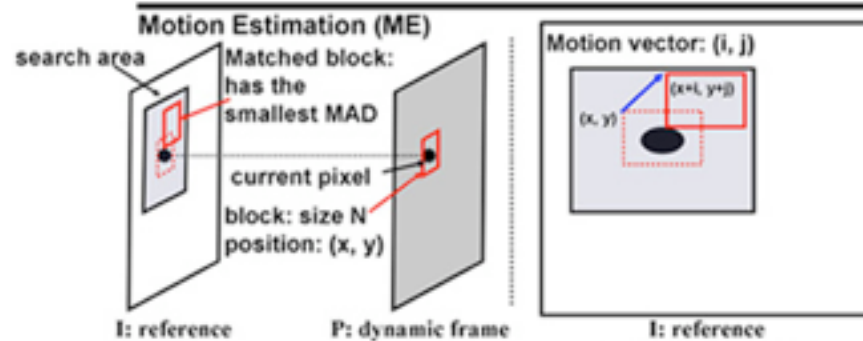
- CMY(K): cyan, magenta, yellow, (and black)
- YCbCr
- HSV (hue, saturation, and value)
- CIE L*a*b*, CIE L*u*v*



From Image to Video

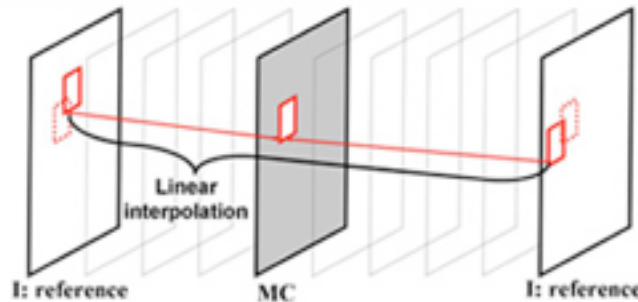


(a)



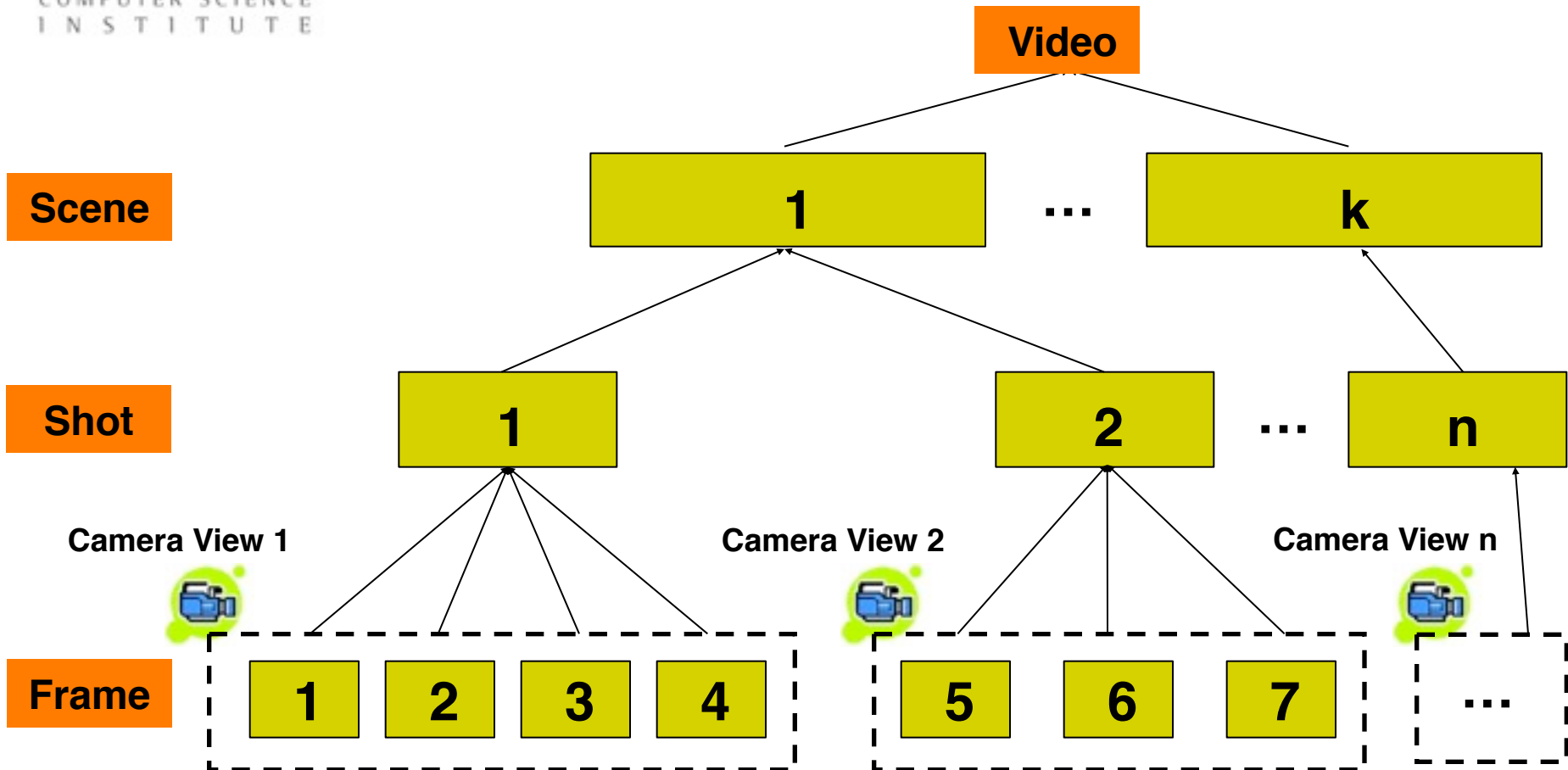
(b)

Motion Compensation (MC):
 when two reference frames are available



(c)

“Natural” Segmentation



Typical Video Features

Low-level features:

- Color features: color dominant, color histogram, color moment, etc.
- Texture features: structural features, statistical features
- Shape features: edge detectors, boundary-based, region-based, etc.

Typical Video Features

Color moments

- First order: mean $\mu_i = \frac{1}{N} \sum_{j=1}^N f_{ij}$
- Second order: variance $\sigma_i = \left(\frac{1}{N} \sum_{j=1}^N (f_{ij} - \mu_i)^2 \right)^{\frac{1}{2}}$
- Third order: skewness $s_i = \left(\frac{1}{N} \sum_{j=1}^N (f_{ij} - \mu_i)^3 \right)^{\frac{1}{3}}$
- Forth order: kurtosis $k_i = \left(\frac{1}{N} \sum_{j=1}^N (f_{ij} - \mu_i)^4 \right)^{\frac{1}{4}}$

Typical Video Features

Texture features

- Co-occurrence matrix $P_d[i, j] = n_{ij}$
- Energy $C(k, n) = \sum_i \sum_j P_d[i, j]^2$
- Contrast $C(k, n) = \sum_i \sum_j (i - j)^k P_d[i, j]^n$
- Entropy $C_e = - \sum_i \sum_j P_d[i, j] \ln P_d[i, j]$
- Homogeneity $C_h = \sum_i \sum_j \frac{P_d[i, j]}{1 + |i - j|}$
- Correlation $C_c = \frac{\sum_i \sum_j [ijP_d[i, j]] - \mu_i \mu_j}{\sigma_i \sigma_j}$

Typical Video Features

Edge Detectors

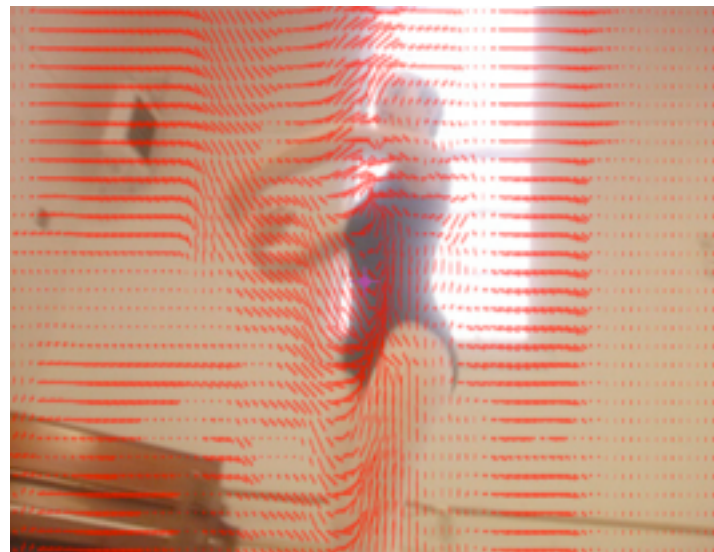
- Sobel,
- Prewitt,
- Canny,
- ... and others



Typical Video Features

Medium-level features

- Difference frames, optical flow
- Region detection: shapes patterns, skin-color, textures
- Scale-Invariant Feature Transform (SIFT)



Typical Video Features

High-level features:

- Face detection:
number of faces, location of face, etc.
- Categories:
indoor and outdoor, play and non-play, etc.
- Metadata: GPS coordinates, compression rate, time

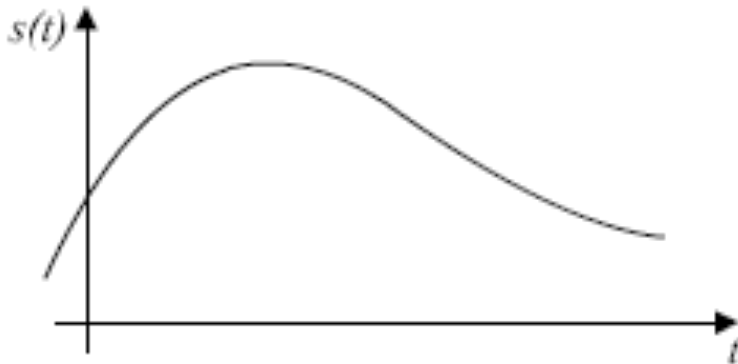


Part III

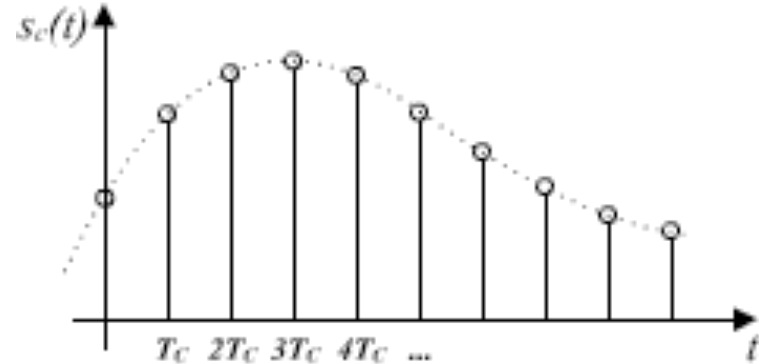
Audio Content Representation

A Sound Wave...

Analog wave:



Digitized wave:



Typical formats:

- 8000Hz, 8-bit log. companded (μ -law): telephone
- 16000Hz, 16-bit linear: speech (Skype)
- 44100Hz, 16-bit linear, stereo: Compact Disk
- 48000Hz, 32-bit linear, stereo: Digital Audio Tape

Typical Audio Features

Amplitude space:

- Energy (usually rms)
- Zero-Crossing Rate
- Tresholding
- “delta” and “delta-delta”s
- Entropy

$$x_{\text{rms}} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

Typical Audio Features

Frequency space:

- Pitch
- Voicedness/Unvoicedness (HNR)
- Long-Term Average Spectrum (LTAS)
- Formants F1...F5
- Speaking rate

Typical Audio Features

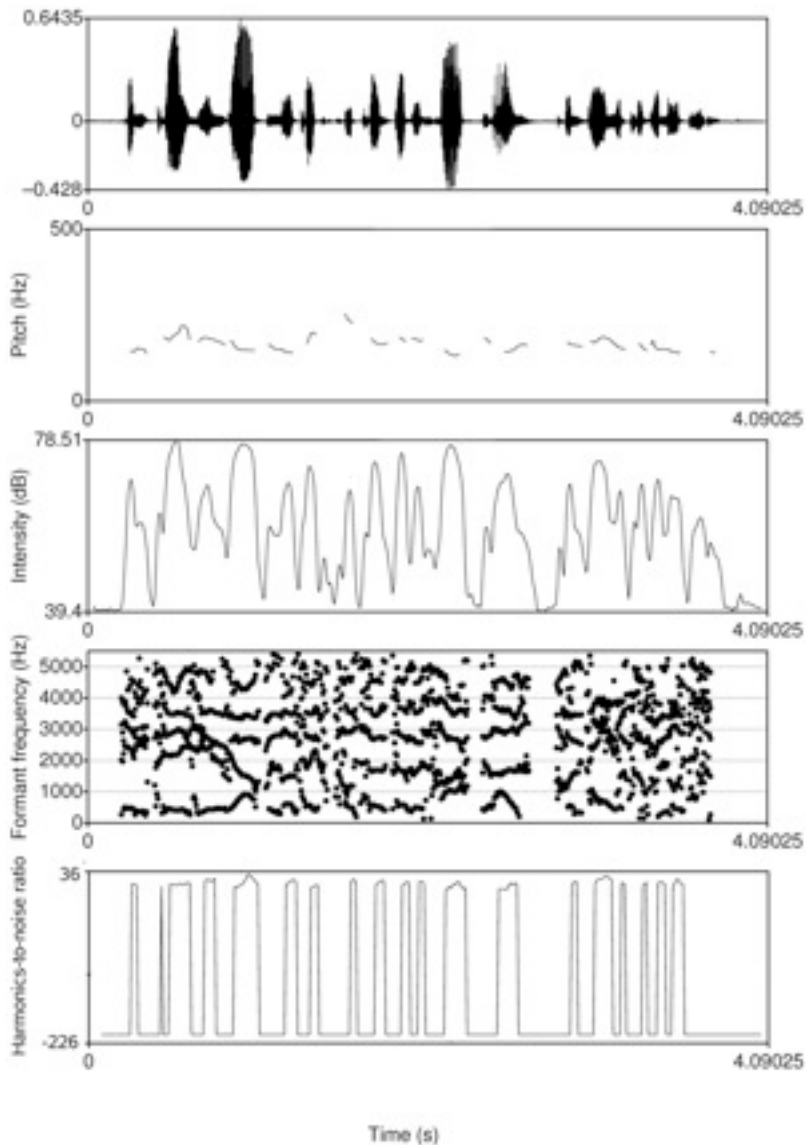
**A Speech
Signal:**

Pitch:

Intensity:

Formants:

HNR:

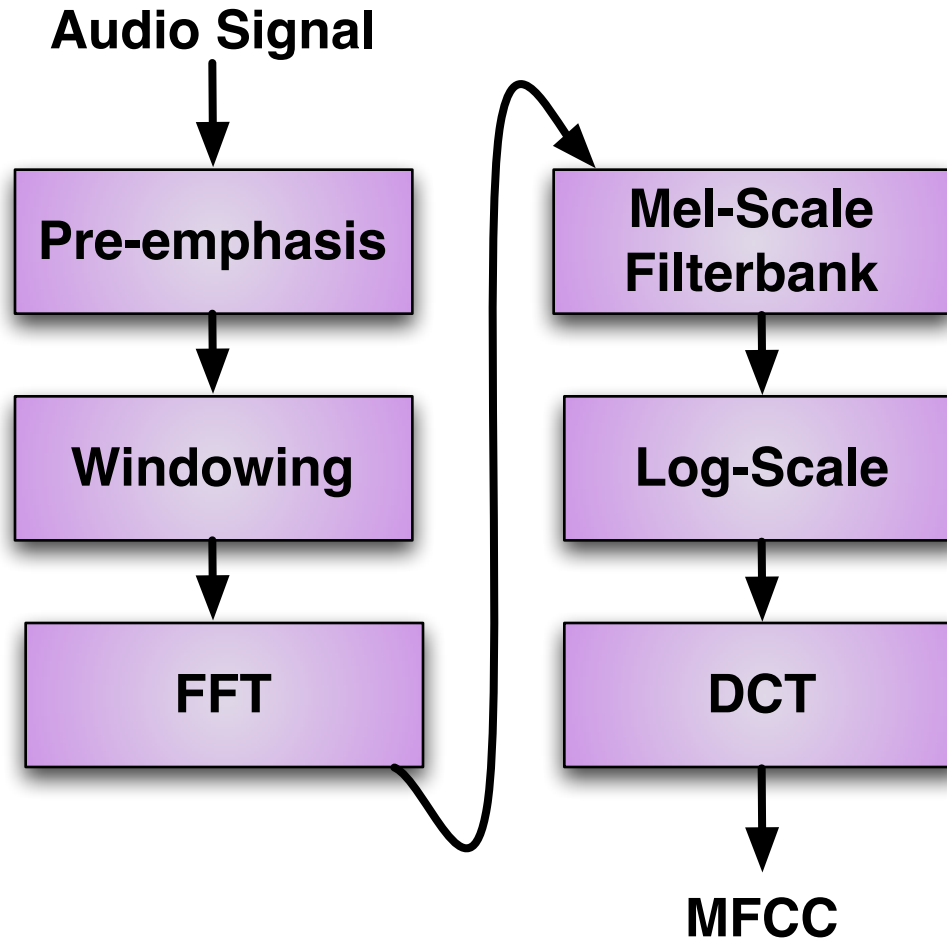


Typical Audio Features

Other spaces:

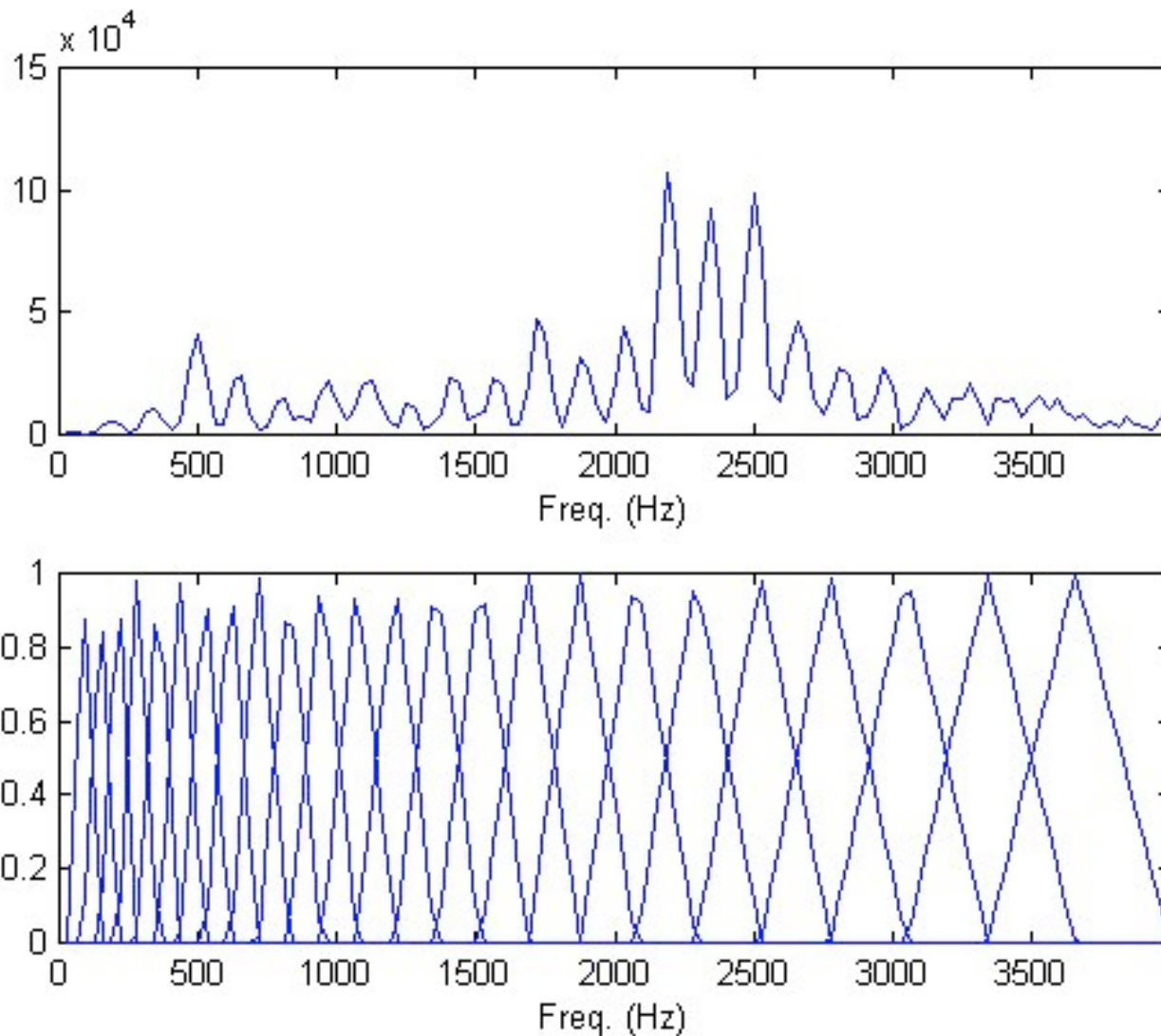
- Linear-Prediction Coefficients (LPC)
- Mel-Frequency-Scaled Coefficients (MFCC):
MFCC12, MFCC19, MFCCxx+delta+deltadelta
- PLP (Perceptual Linear Prediction)
- RASTA, RASTA-PLP, MSG

MFCC: Idea

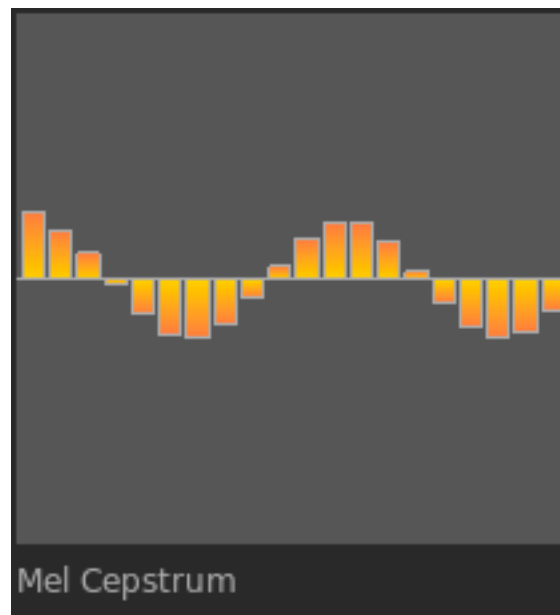
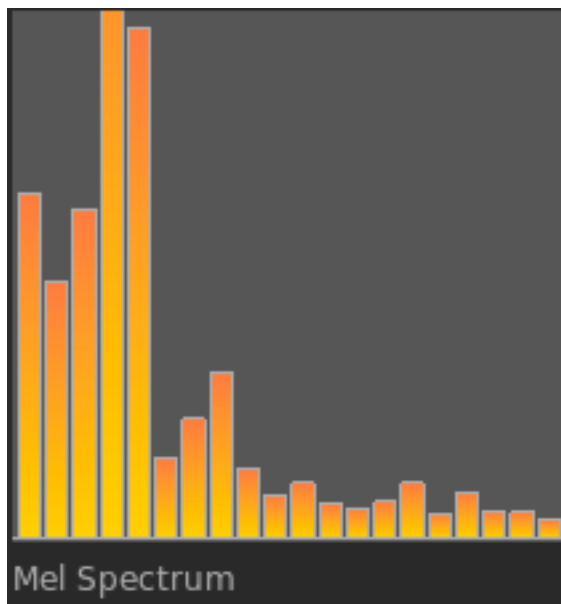
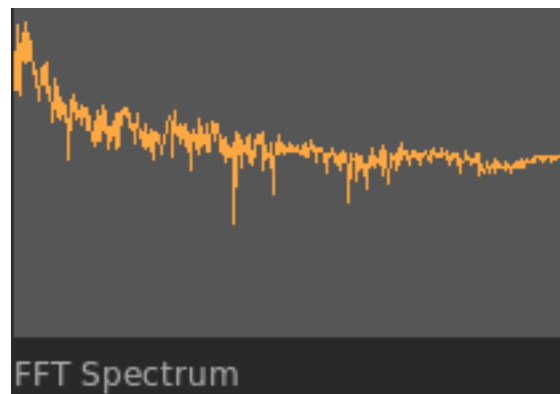
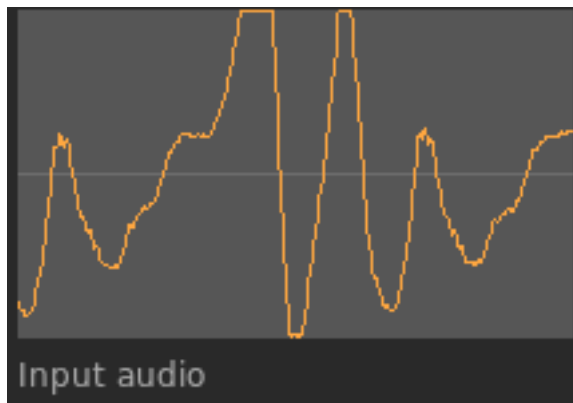


$$\text{power cepstrum of signal} = \left| \mathbf{F} \left\{ \log \left(\left| \mathbf{F} \left\{ \text{the signal} \right\} \right|^2 \right) \right\} \right|^2$$

MFCC: Mel Scale



MFCC: Result

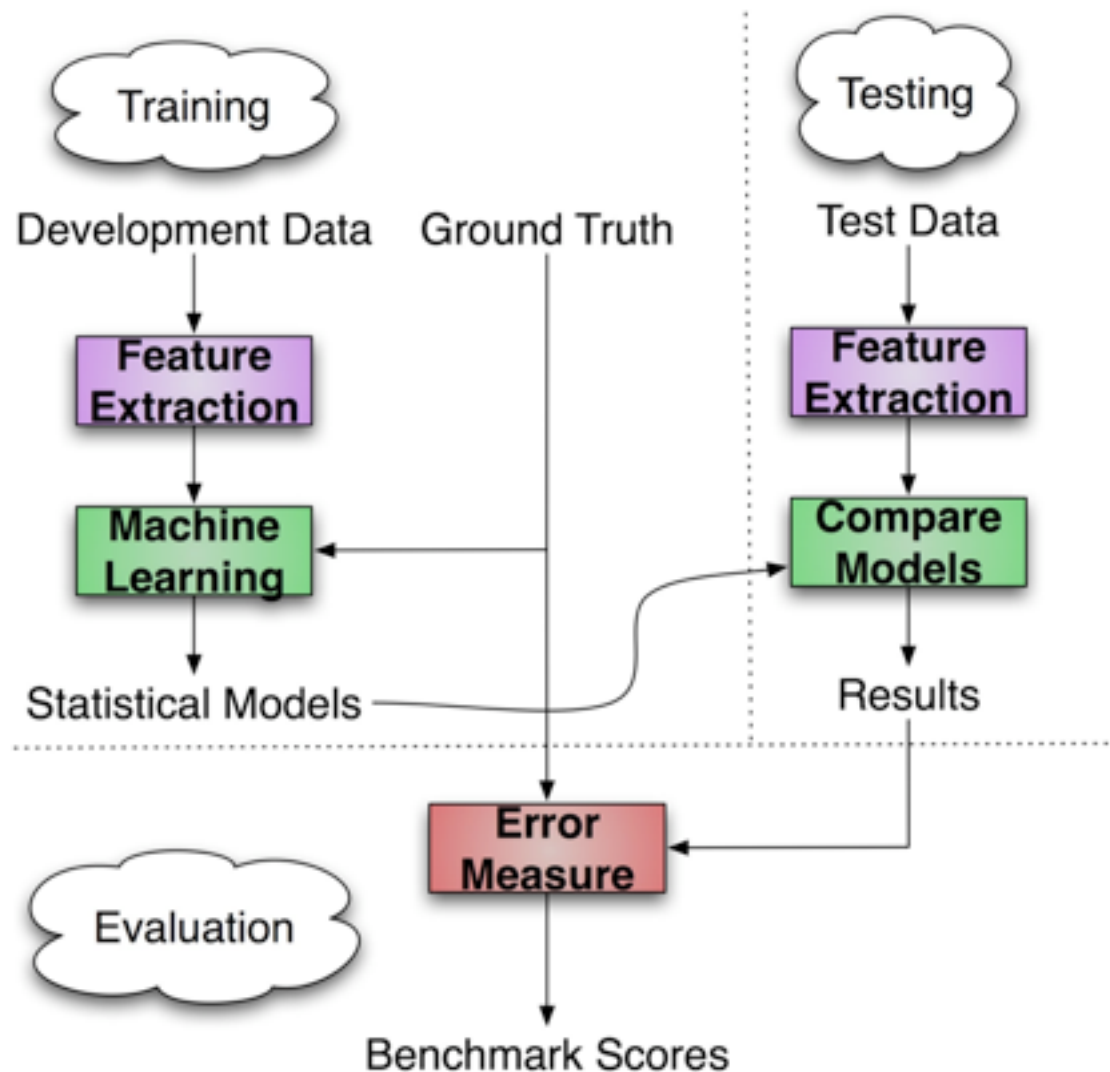




Part IV

Modeling Techniques

Development of a Content Analysis Algorithm



Modeling Techniques

Unsupervised Techniques:

- K-Means, X-Means
- PCA

Supervised Techniques:

- Gaussian Mixtures
- Neuronal Networks
- Support-Vector machines
- Hidden-Markov Models



K-Means

Algorithm Outline (Expectation Maximization)

Choose k initial means μ_i at random

loop

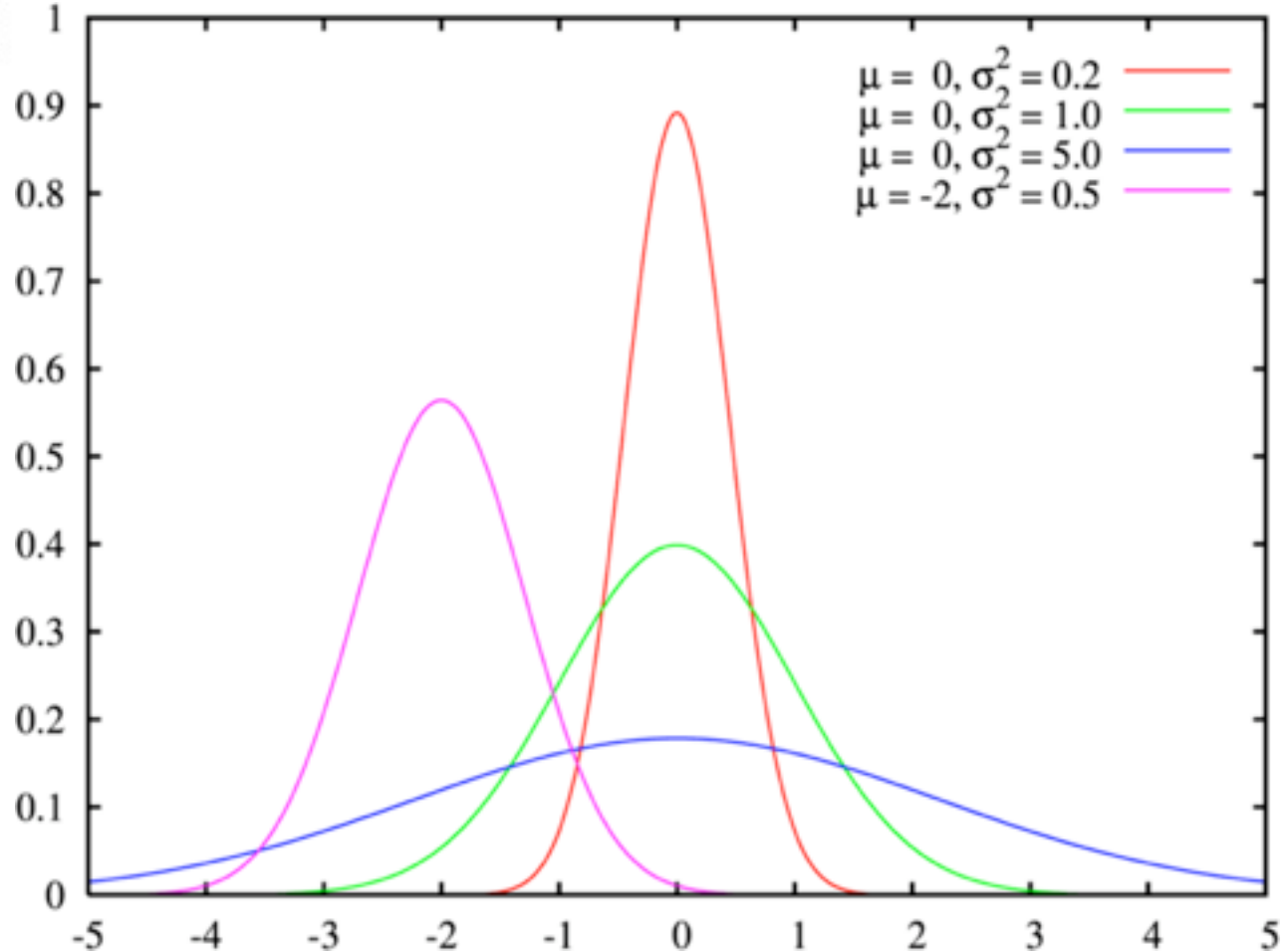
 for all samples x_j :

 assign membership of each element to a mean (closest mean)

 for all means μ_i

 calculate a new μ_i by averaging all values x_j that were assigned members
until means μ_i are not updated significantly anymore

Gaussian Mixtures



Training of Mixture Models

Goal: Find a_i for $f_X(x) = \sum_{i=1}^n a_i f_Y(x; \theta_i)$.

Expectation: $y_{i,j} = \frac{a_i f_Y(x_j; \theta_i)}{f_X(x_j)}$.

Maximization: $a_i = \frac{1}{N} \sum_{j=1}^N y_{i,j}$

$$\mu_i = \frac{\sum_j y_{i,j} x_j}{\sum_j y_{i,j}}$$



Part V

Multimodal Integration

Multimodal Integration

- ... is a field of cognitive psychology.
- Before 1960: Unimodal approach
- Initial results in the 1960's, recently hyped again (2003+)

Multimodal Integration

Human psychology suggests:

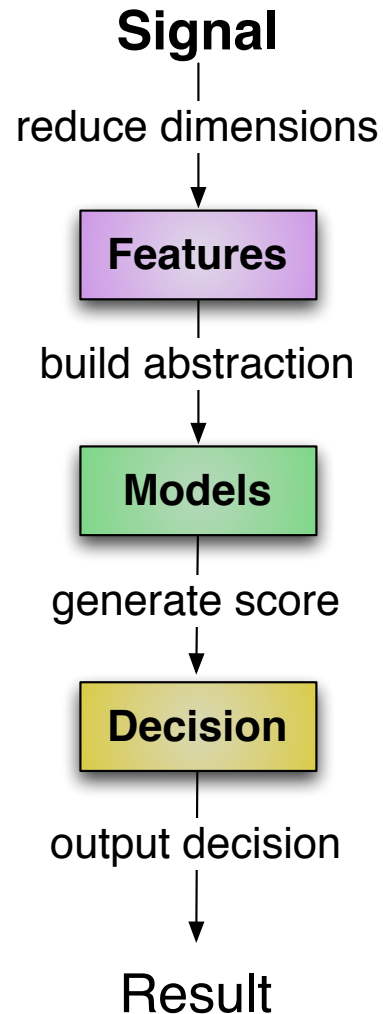
- Uncertainty in sensory domains results in increased dependency of multisensory integration (Alais & Burr 2004)
- Multiple sensory inputs increase the speed of the output (Hershenson 1962)

Multimodal Integration

In computer science:

- How to create systems that benefit from multimodal integration in similar ways the brain does, i.e. they are
 - more accurate, robust, and/or faster than unimodal state of the art and/or
 - offer qualitative improvements over unimodal approaches

Generic Scheme of a Classification Algorithm



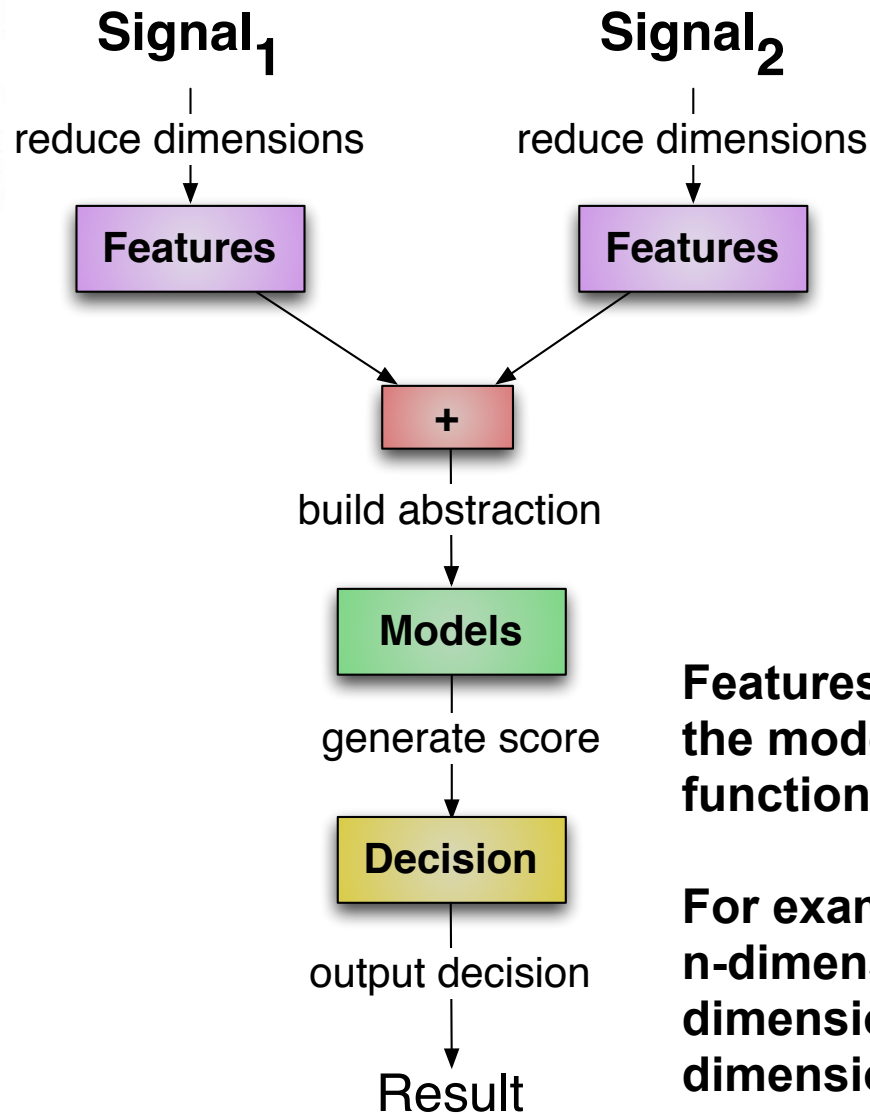
Some signal is observed and reduced...

...to the essentials relevant to the problem, ...

...statistical models are used to compute a score (e.g. probabilities) for the given observations, ...

... so that a decision function can decide on the classification.

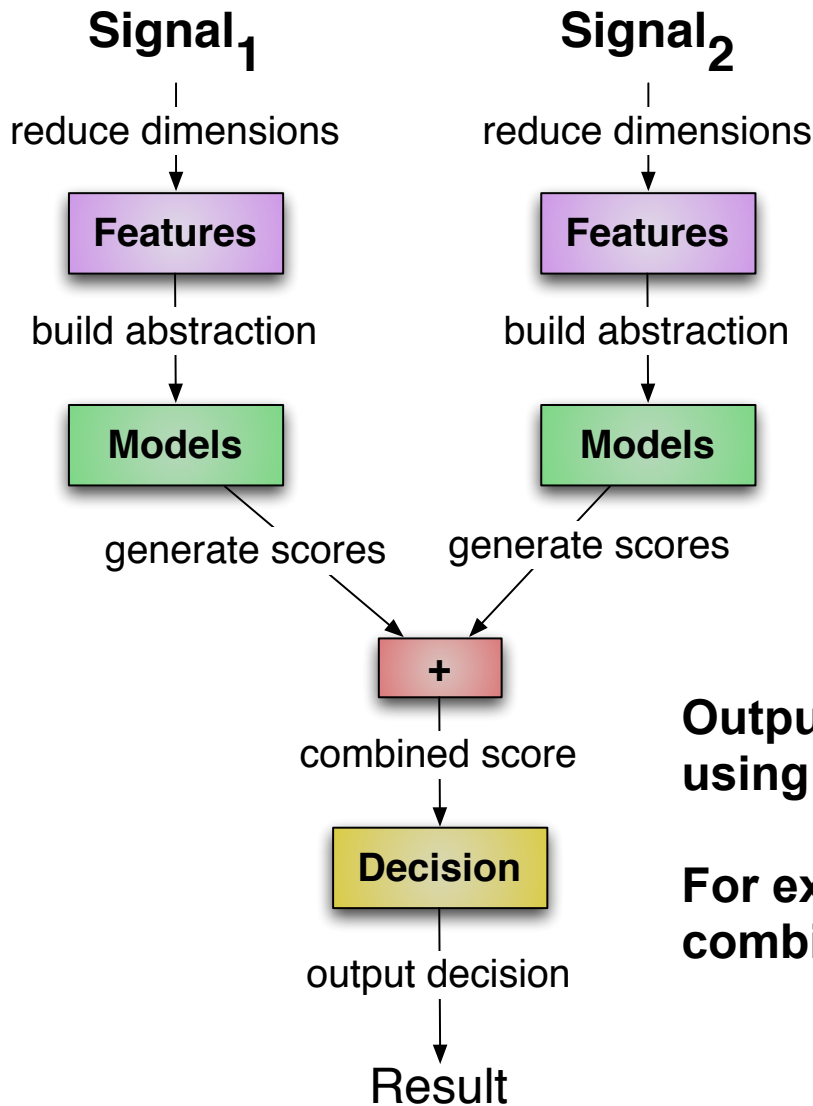
Feature-Level Integration



Features are integrated before the model layer using a function '+'.

**For example concatenation:
n-dimensional vector '+' m-
dimensional vector = n+m-
dimensional vector**

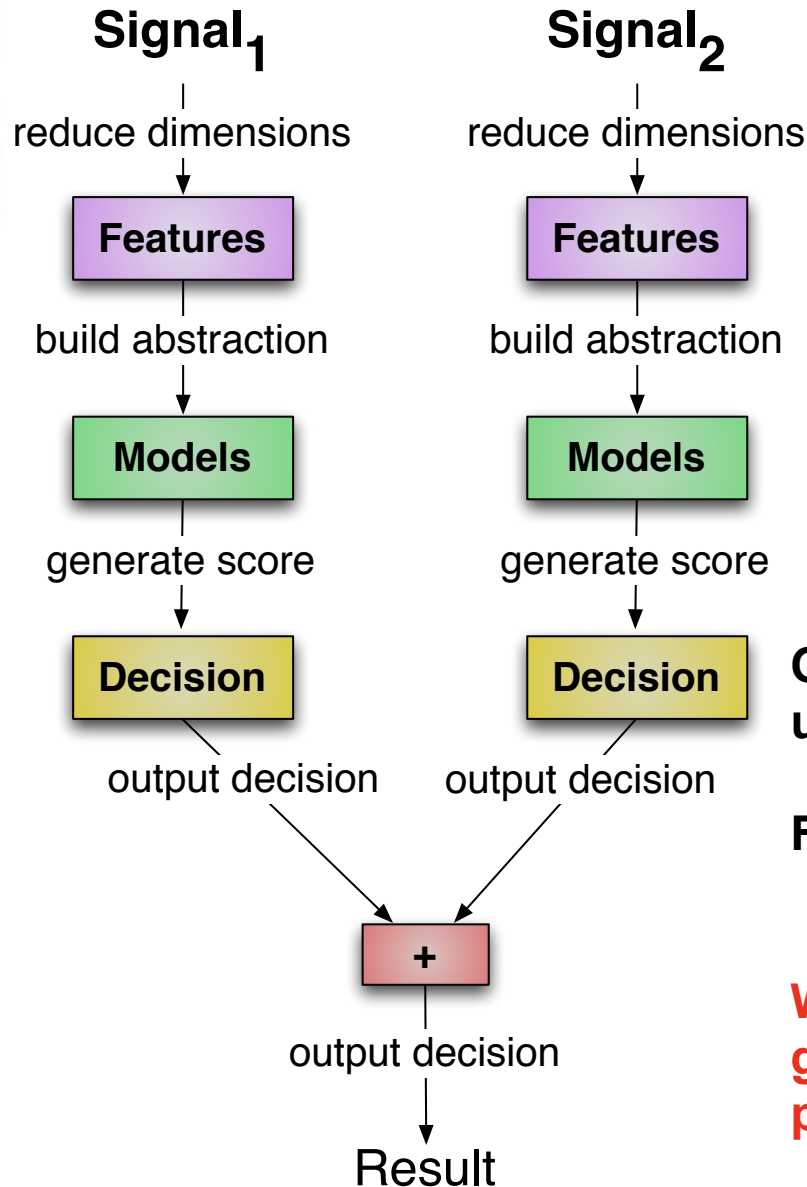
Model-Level Integration



Output scores are integrated using a function '+'.

For example weighted combined log-likelihoods.

Decision-Level Integration



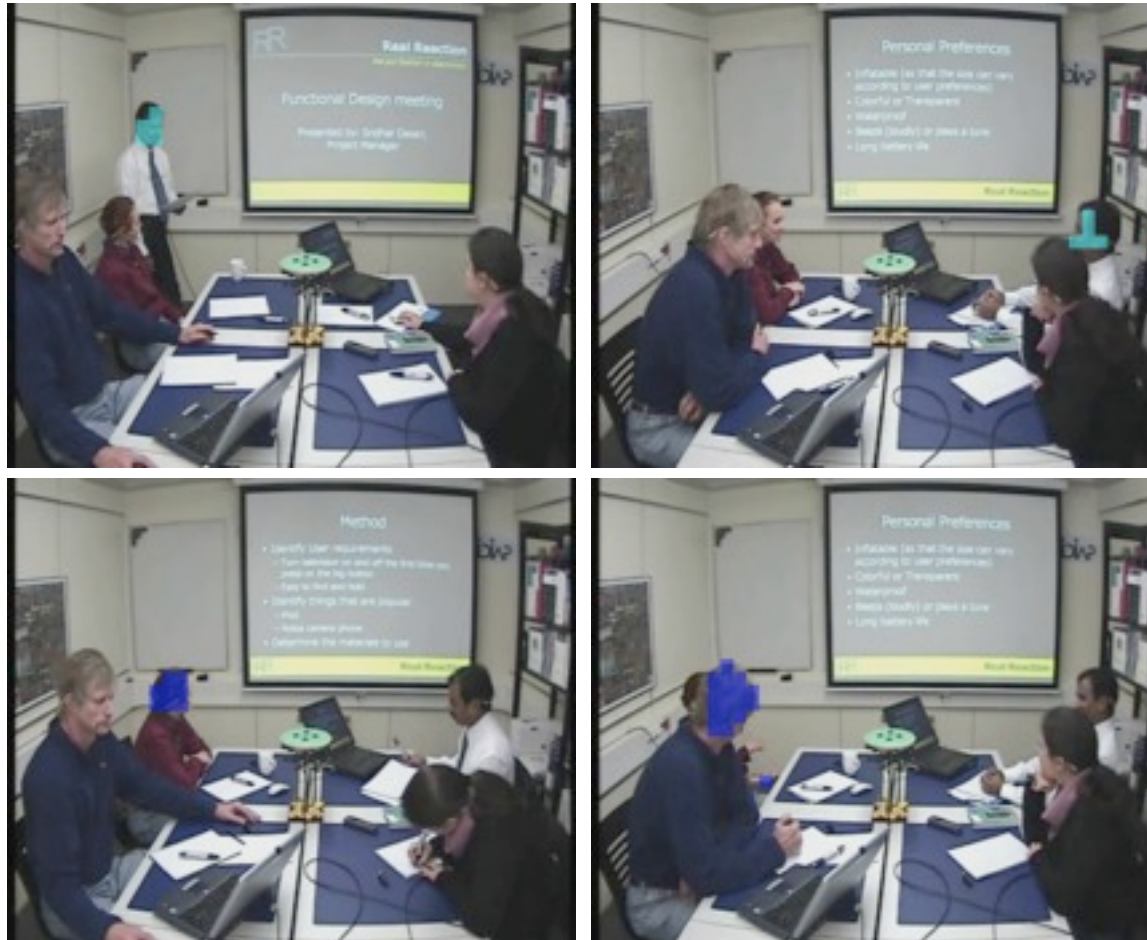
Output decision are fused using a function '+'.
For example majority voting.

WARNING: Meta-data fusion in general is a difficult research problem.

Remarks

- Signal-level integration is unlikely because of intractable data dimensionality.
- Multi-Level integration is also possible.
- In reality, a classification algorithm is more complicated than this scheme (eg. feedback loops)
- The integration function ‘+’ may also be learned automatically.

Example: "3D from Audio"



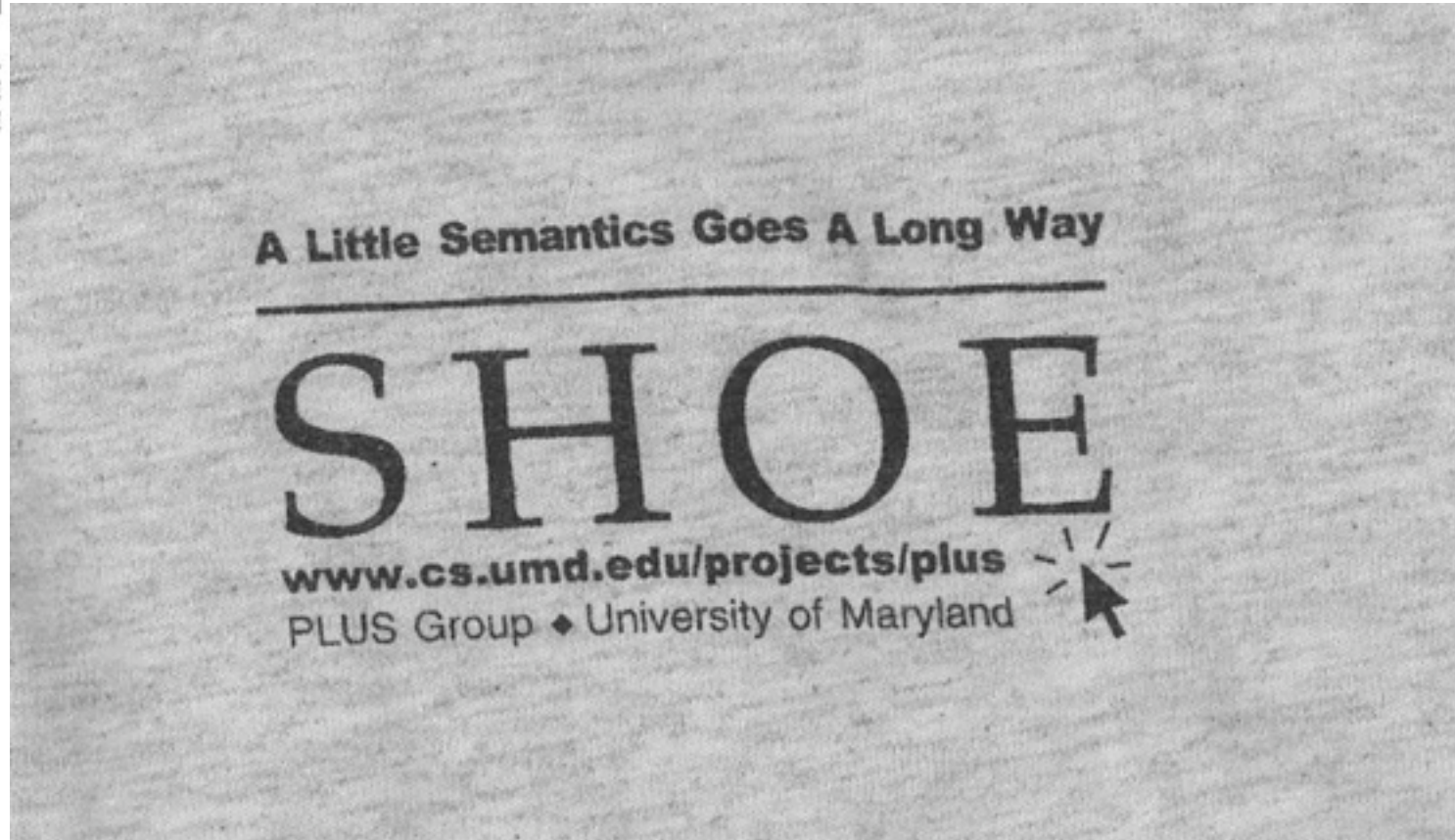
G. Friedland, C. Yeo, H. Hung: "Visual Speaker Localization Aided by Acoustic Models" (full paper), Proceedings of ACM Multimedia, Beijing, China, to appear October 2009.



Part VI

Concluding Remarks

A note...



James A. Hendler

Most Important Lesson

Multimedia content analysis is hard.

Therefore *every possible cue* should be considered for a solution, including context and user presumptions.

Semantic Computing

Computing methods become **semantic** computing methods when they are indistinguishable from **understanding** to the user.

In other words, when the computer does “**what the user means**” with the minimum communication possible.

Outlook

Coming up: Hands-on Session

