Methods and Techniques for MultiModal Information Fusion

Ling Guan

Ryerson Multimedia Laboratory &
Centre for Interactive Multimedia Information Mining
Ryerson University,
Toronto, Ontario Canada
lguan@ee.ryerson.ca
http://www.rml.ryerson.ca/

Acknowledgement

- The presenter would like to thank his former and current students *P. Muneesawang, Y. Wang, R. Zhang, Y. Tie, A. Bulzacki and M. T. Ibrahim* for their contributions to this research.
- This research is supported by
  - The Canada Research Chair (CRC) Program,
  - Canada Foundation for Innovations (CFI),
  - The Ontario Innovation Trust (OIT), and
  - Ryerson University

Major Publications

- Y. Tie and L. Guan, "Automatic face detection in video sequences using local normalization and optimal adaptive correlation," *Pattern Recognition*, vol. 42, no. 5, pp. 1859-1868, May 2009.
- P. Muneesawang, T. Amin and L. Guan, "A new learning algorithms for the fusion of adaptive audio-visual features for the retrieval and classification of movie clips," *Journal of Signal Processing Systems for Signal, Image and Video Technology*, DOI: 10.1007/s11265-008-0290-7 (12 pages), October 2008.
- Y. Wang and L. Guan, "Combining speech and facial expression for recognition of human emotional state," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936 - 946, August 2008.
- N. Joshi and L. Guan, "ASR with the combination of recognizers under non-stationary noise conditions", to appear in *Journal of Signal Processing Systems.*
- L. Guan, P. Muneesawang, Y. Wang, R. Zhang, Y. Tie, A. Bulzacki and M.T. Ibriham, "Multimedia Multimodal Technologies," *Proc. IEEE Workshop on Multimedia Signal Processing and Novel Parallel Computing* (In conjunction with ICME 2009), pp. 1600-1603, NYC, USA, Jul 2009 (Overview Paper).
- R. Zhang and L. Guan, "Multimodal image retrieval via Bayesian information fusion," *Proc. IEEE Int. Conf. on Multimedia and Expo*, pp. 830-833, NYC, USA, Jun/Jul 2009.

Why Multimedia Multimodal Methodology?

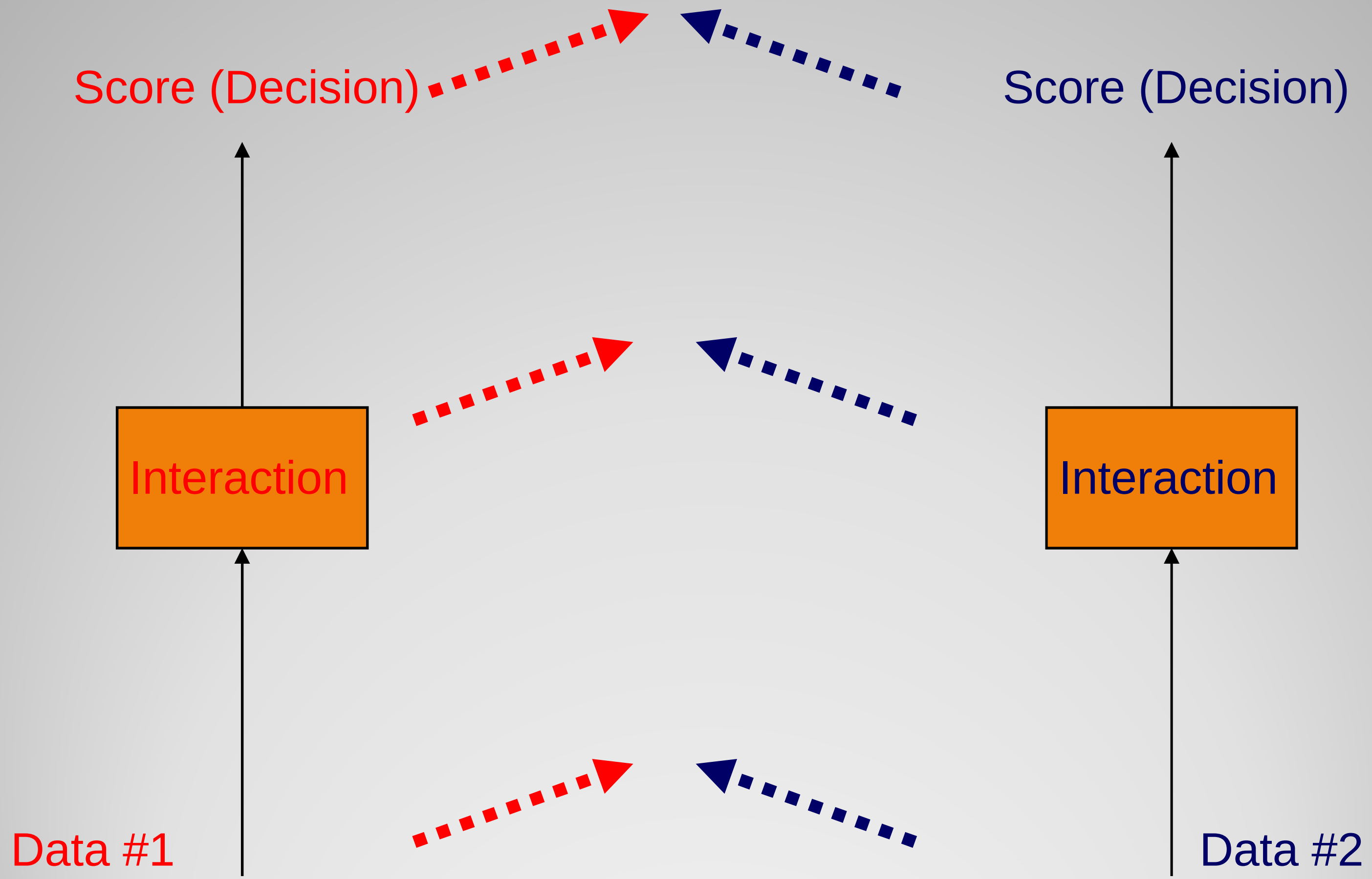
- Multimedia is a domain of multi-facets, e.g., audio, visual, text, graphics, etc.
- Easy to define each facet individually, but difficult to consider them as a combined identity
- A central aspect of multimedia processing is the coherent integration of media from different sources or multimodalities.
- Humans are natural and generic multimedia processing machines

Can we teach computers/machines to do the same (via Fusion methods and techniques)?

Potential Applications

- Human–Computer Interaction
- Learning Environments
- Consumer Relations
- Entertainment
- Digital Home, Domestic Helper
- Security/Surveillance
- Educational Software
- Computer Animation
- Call Centers

Source Fusion for Classification



Direct Data Fusion

Furthermore, let $\mathbf{S}_i \in R^q, i = 1, \cdots, N$ denote vectors comprising all the individual scores:

$$\mathbf{V}_i = \begin{bmatrix} v_i^{(1)} \\ v_i^{(2)} \\ \vdots \\ v_i^{(q)} \end{bmatrix} i = 1, \cdots, N \qquad (0.2)$$

Now, training data can be formed as the following input/teacher pairs

$$[\mathcal{V}, \mathcal{T}] = \{ \ [\mathbf{V}_1, \mathbf{t}_1], [\mathbf{V}_2, \mathbf{t}_2], \ldots, [\mathbf{V}_N, \mathbf{t}_N] \ \}$$

Prior knowledge can be incorporated into the fusion models by modifying

$$\mathbf{V}_i = \begin{bmatrix} \nu^{(1)} \mathbf{v}_i^{(1)} \\ \nu^{(2)} \mathbf{v}_i^{(2)} \\ \vdots \\ \nu^{(q)} \mathbf{v}_i^{(q)} \end{bmatrix}$$

Score Fusion

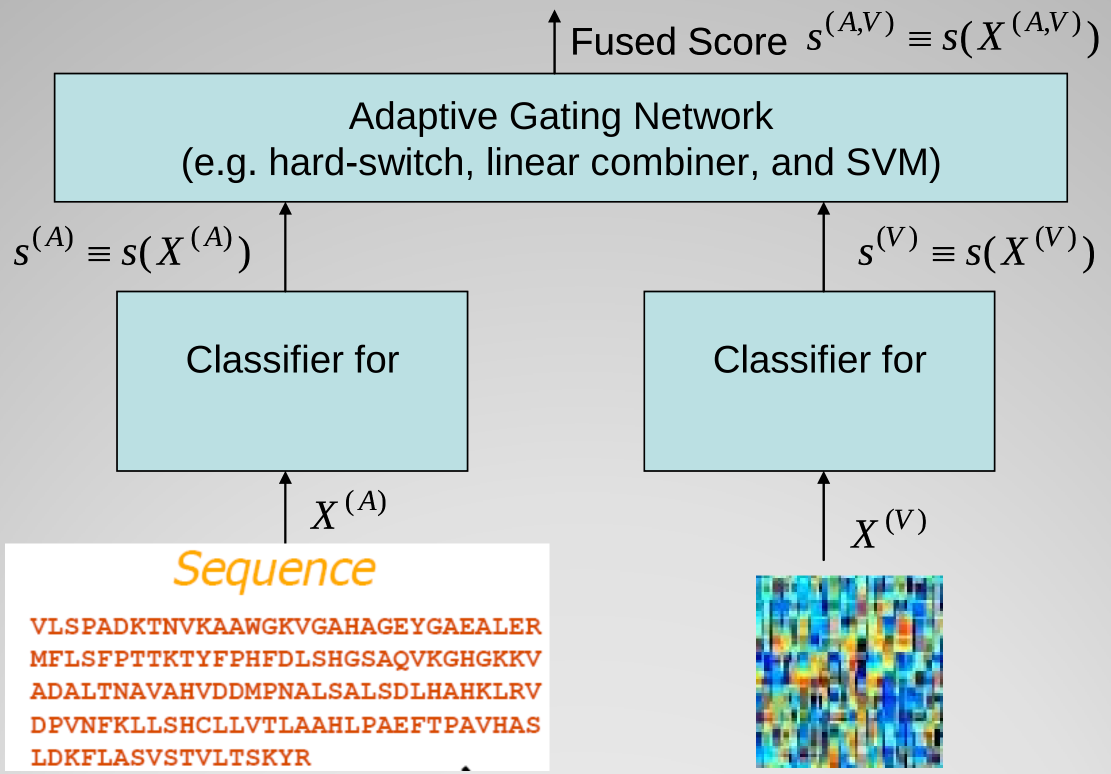
1a Score Fusion (w/o supervision)

- Linear Score Fusion (confidence/prior knowledge)
- Nonlinear Score Fusion (ROC-based)

1b Score Fusion (via supervision)

- Linear Score Fusion (adaptive supervision)
- Nonlinear Score Fusion (adaptive supervision)

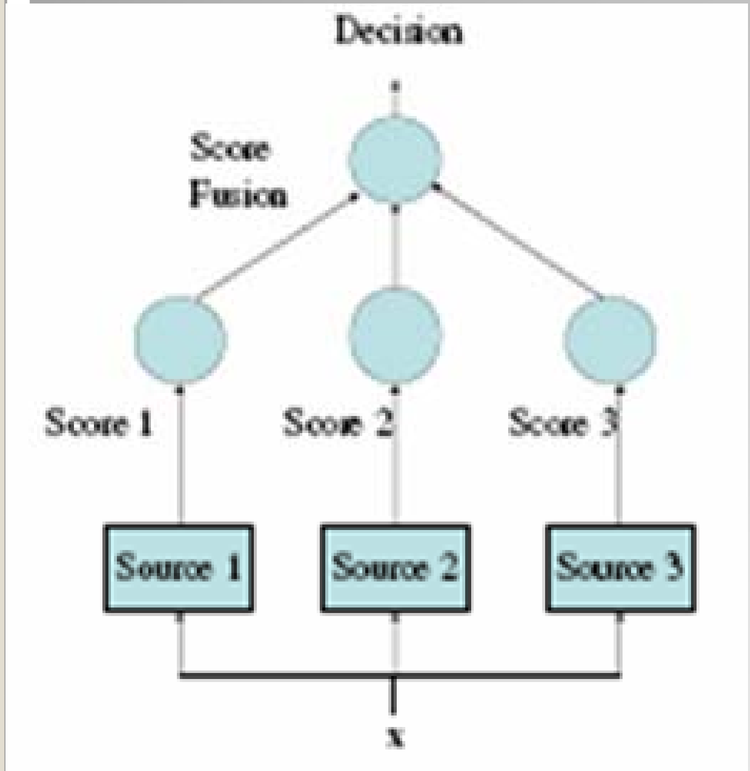
Score Fusion Architecture



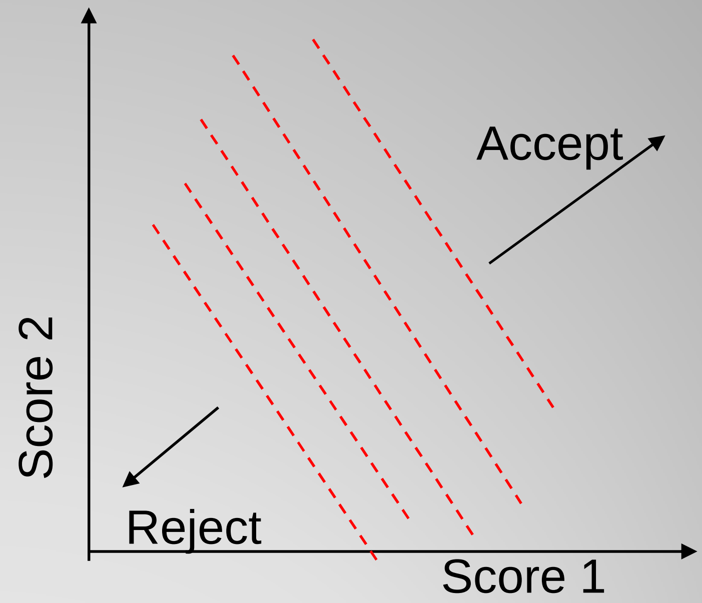
The scores are independently obtained, which are then combined.

- The lower layer contains local experts, each produces a local score based on a single modality
- The upper layer combines the score.

Linear Fusion



Non-uniformly weighted



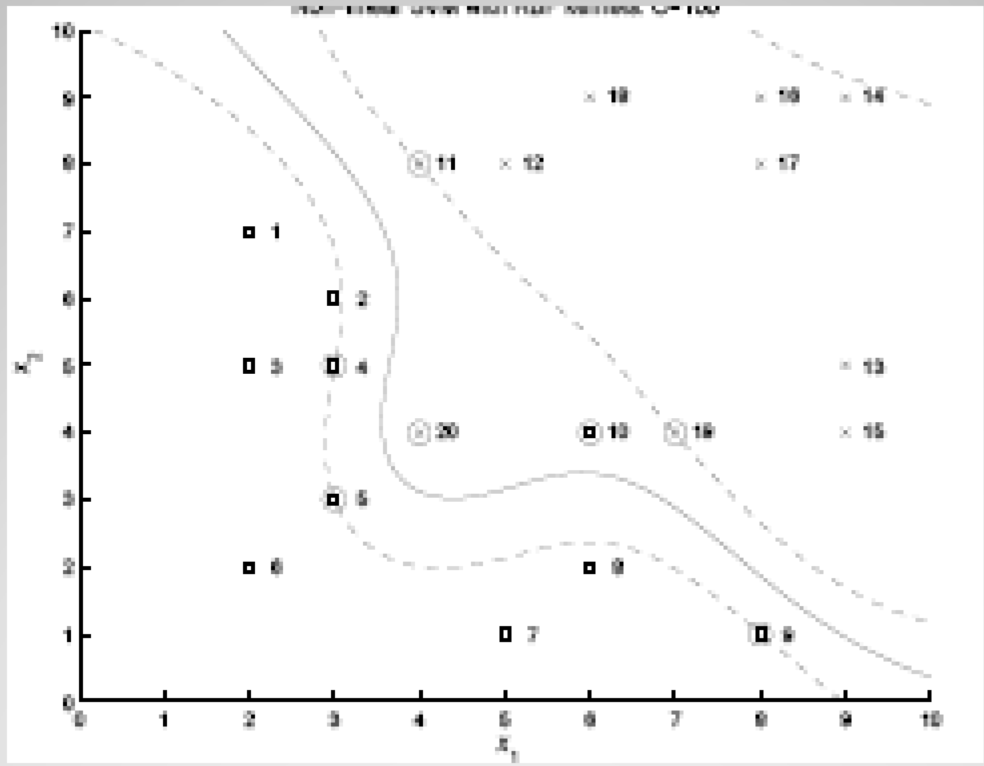
The most prevailing unsupervised approaches estimate the confidence based on prior knowledge or training data.

Linear SVM (supervised) Fusion is another appealing alternative.

Nonlinear Adaptive Fusion (via supervision)

(SVM)

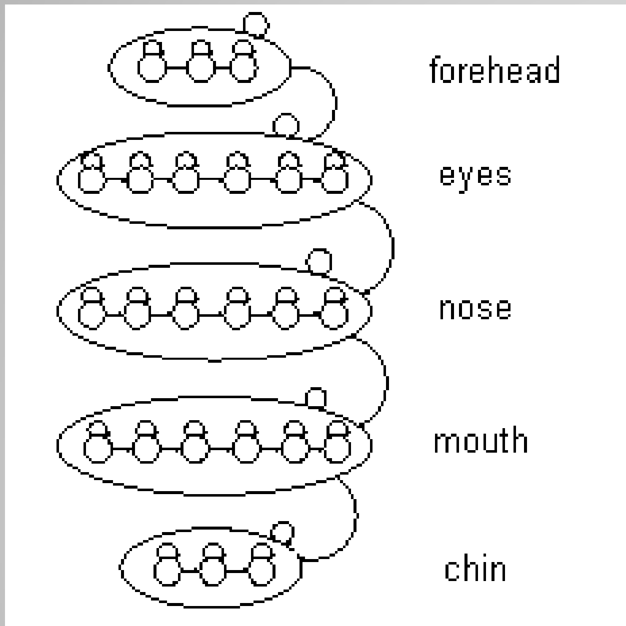
Score 2



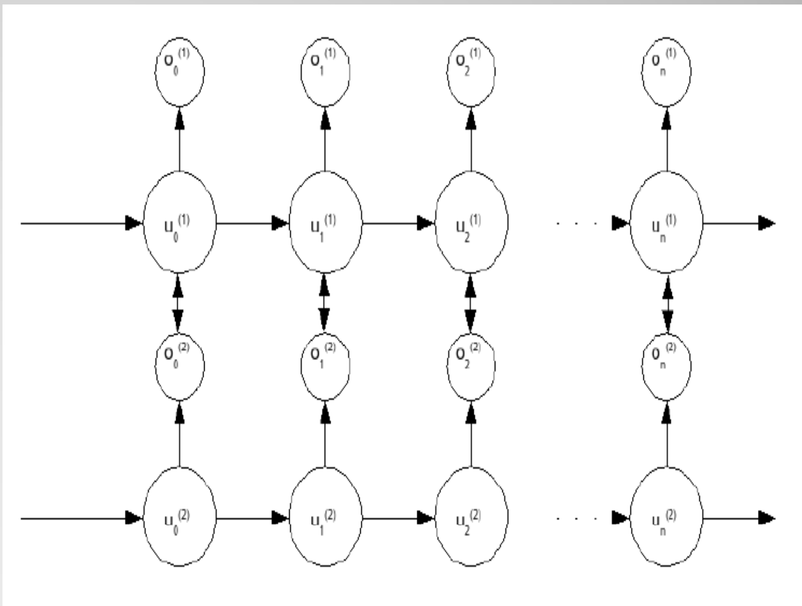
Score 1

Interaction Fusion

HHM



Fused HHM



Data (Feature) Fusion

- Simple and straightforward (Good)
- Curse of Dimensionality (Bad)
- Normalization issue

- Case study: Bimodal Human emotion recognition

Indicators of emotion

- Speech
- Facial expression

Major indicators of emotion

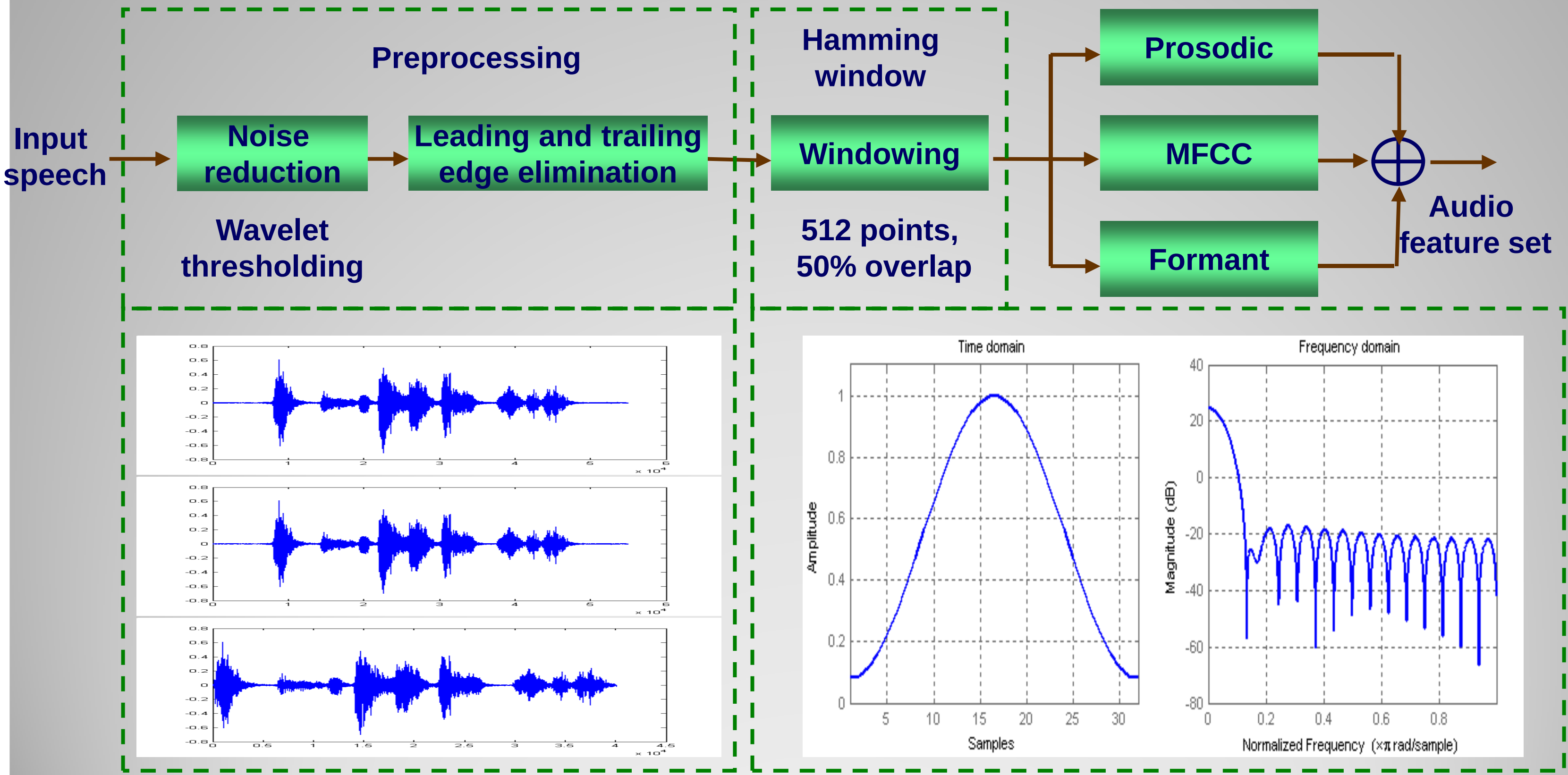
- Body language: highly dependent on personality, gender, age, etc
- Semantic meaning: two sentences could have the same lexical meaning but different emotional information

……

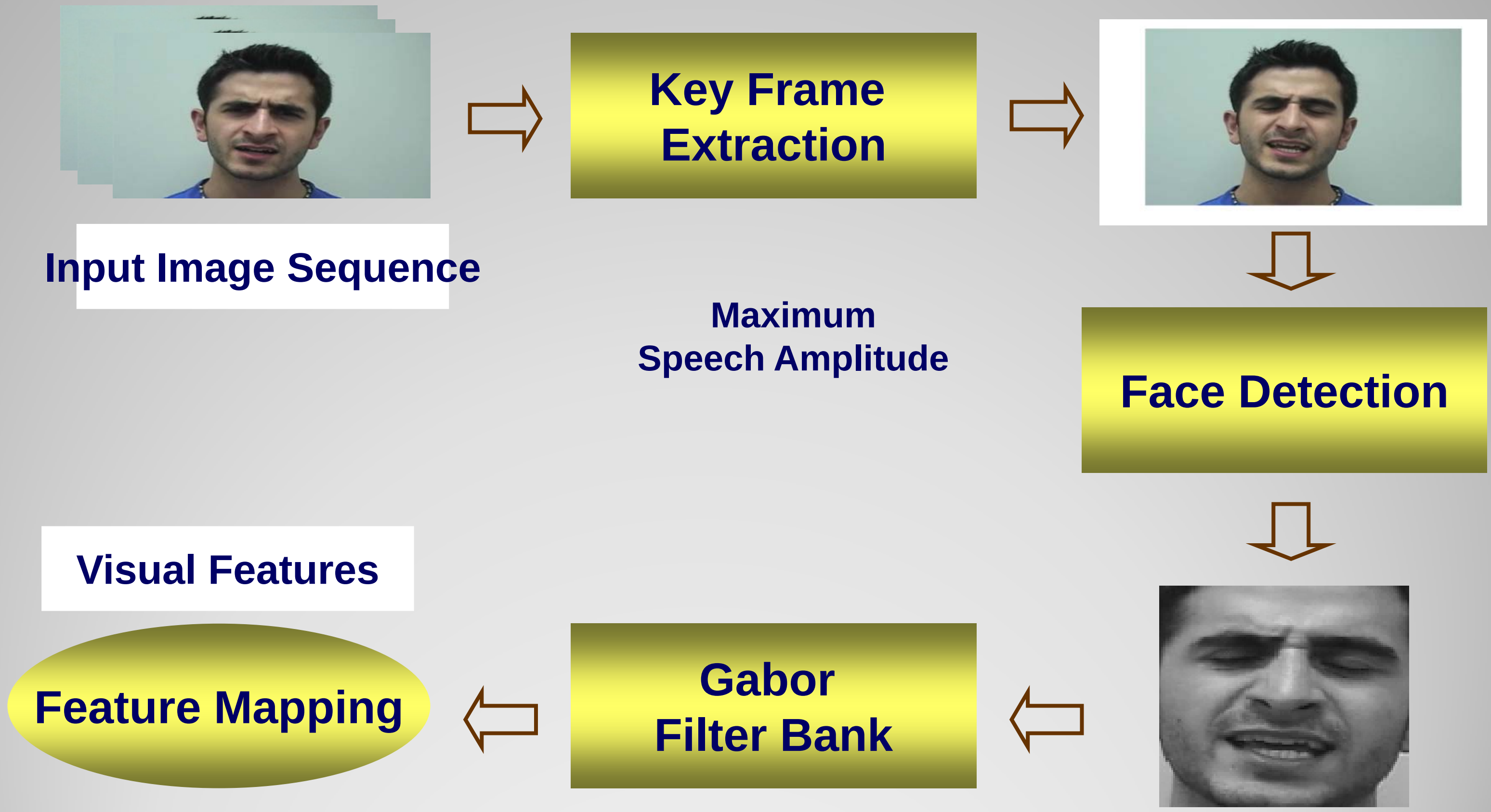
Objective

- **To develop a generic language and cultural background independent system for recognition of human emotional state from audiovisual signals**

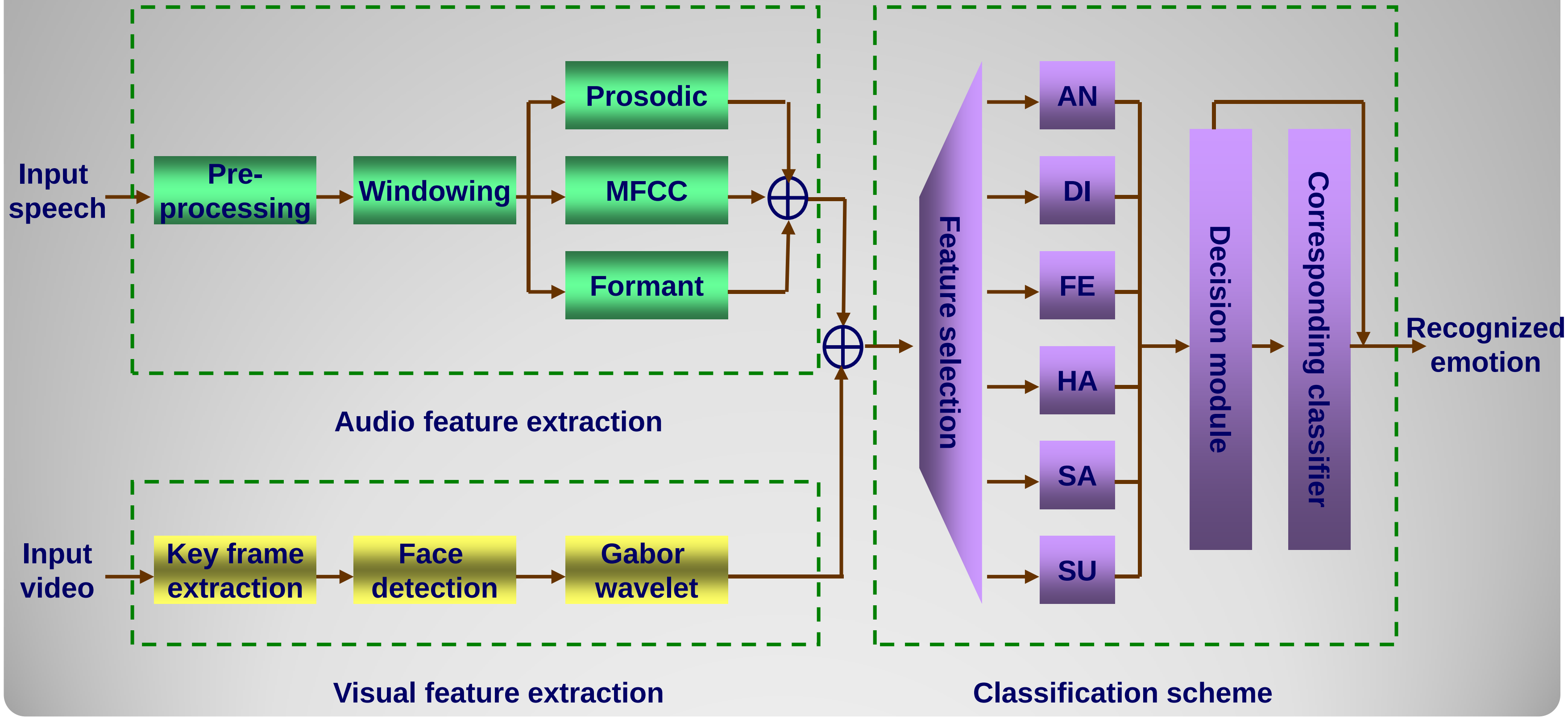
Audio feature extraction



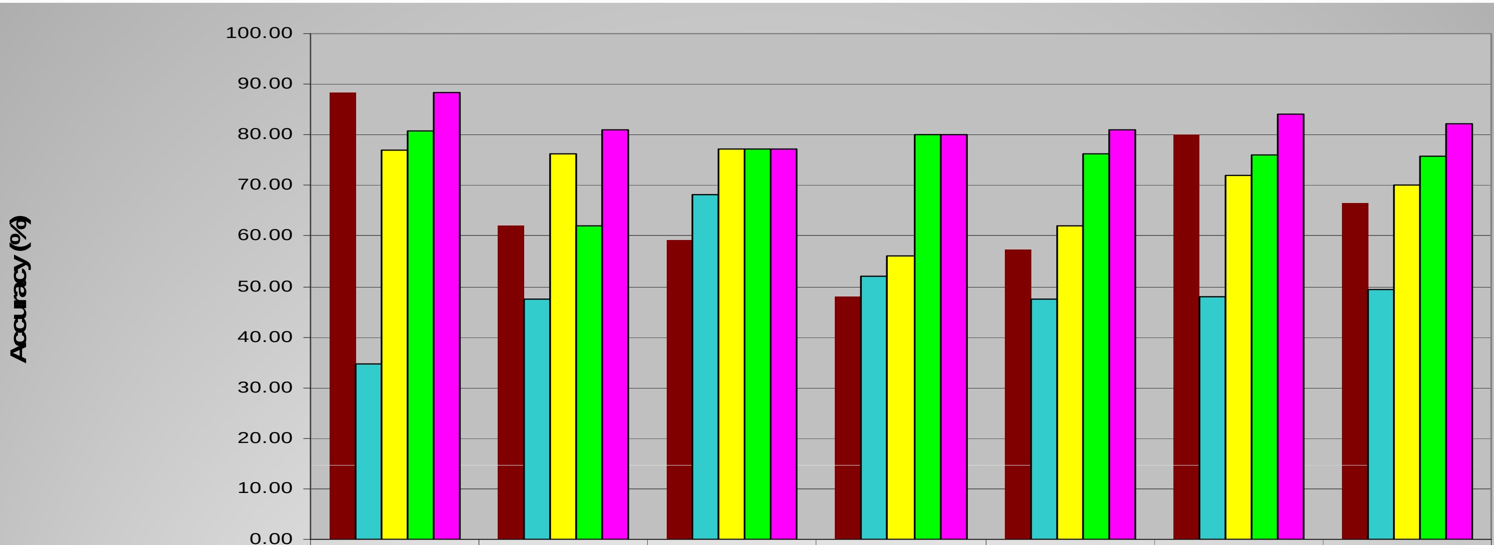
Visual feature extraction



The recognition system



Experimental results



	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Overall
Audio	88.46	61.90	59.09	48.00	57.14	80.00	66.43
Visual	34.62	47.62	68.18	52.00	47.62	48.00	49.29
Audiovisual	76.92	76.19	77.27	56.00	61.90	72.00	70.00
Stepwise	80.77	61.90	77.27	80.00	76.19	76.00	75.71
Multi-classifier	88.46	80.95	77.27	80.00	80.95	84.00	82.14

- **Experiments were performed on 500 video samples from 8 subjects, speaking 6 languages**
- **Six emotion labels: Anger, Disgust, Fear, Happiness, Sadness, and Surprise**
- **360 samples (from six subjects) were used for training, and the rest 140 (from the remaining two subjects) for testing, there is no overlap between training and testing subjects**

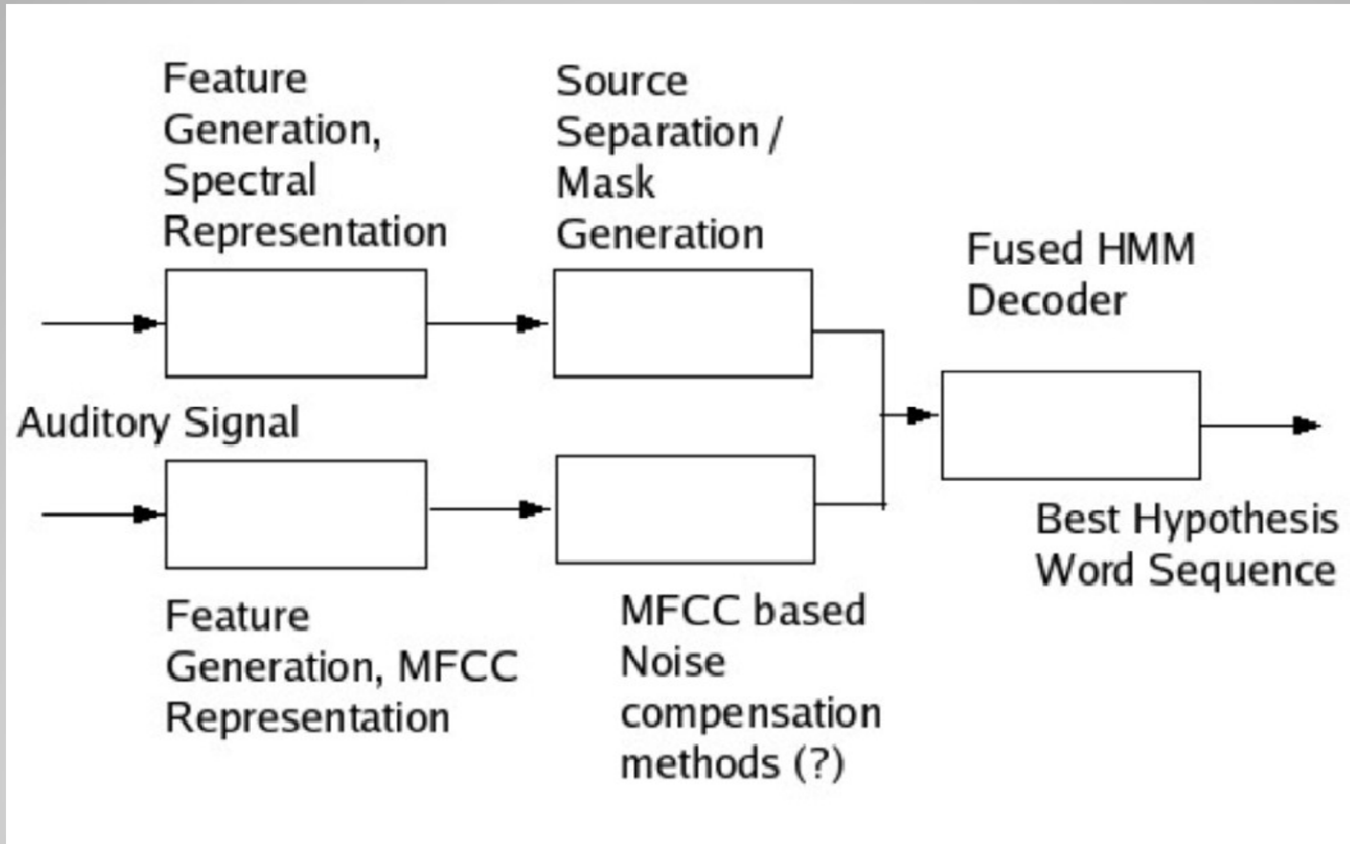
Interaction Fusion

- In general, not straightforward
- Scores obtained from different classifiers
- The scores may need to be normalized

- Case study:
  1. Speech Recognition
  2. Image Retrieval with Audio Information

Speech Recognition

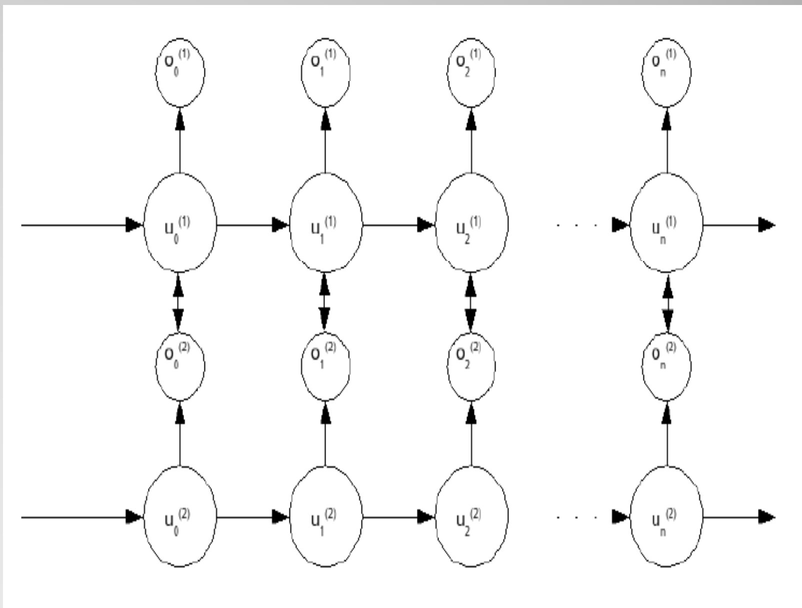
The System Diagram



Interaction Level Fusion

Two separate HMM based models:

- spectral features, missing data (MD),
- MFCC features.
- The Fused HMM model is used for the interaction level fusion.



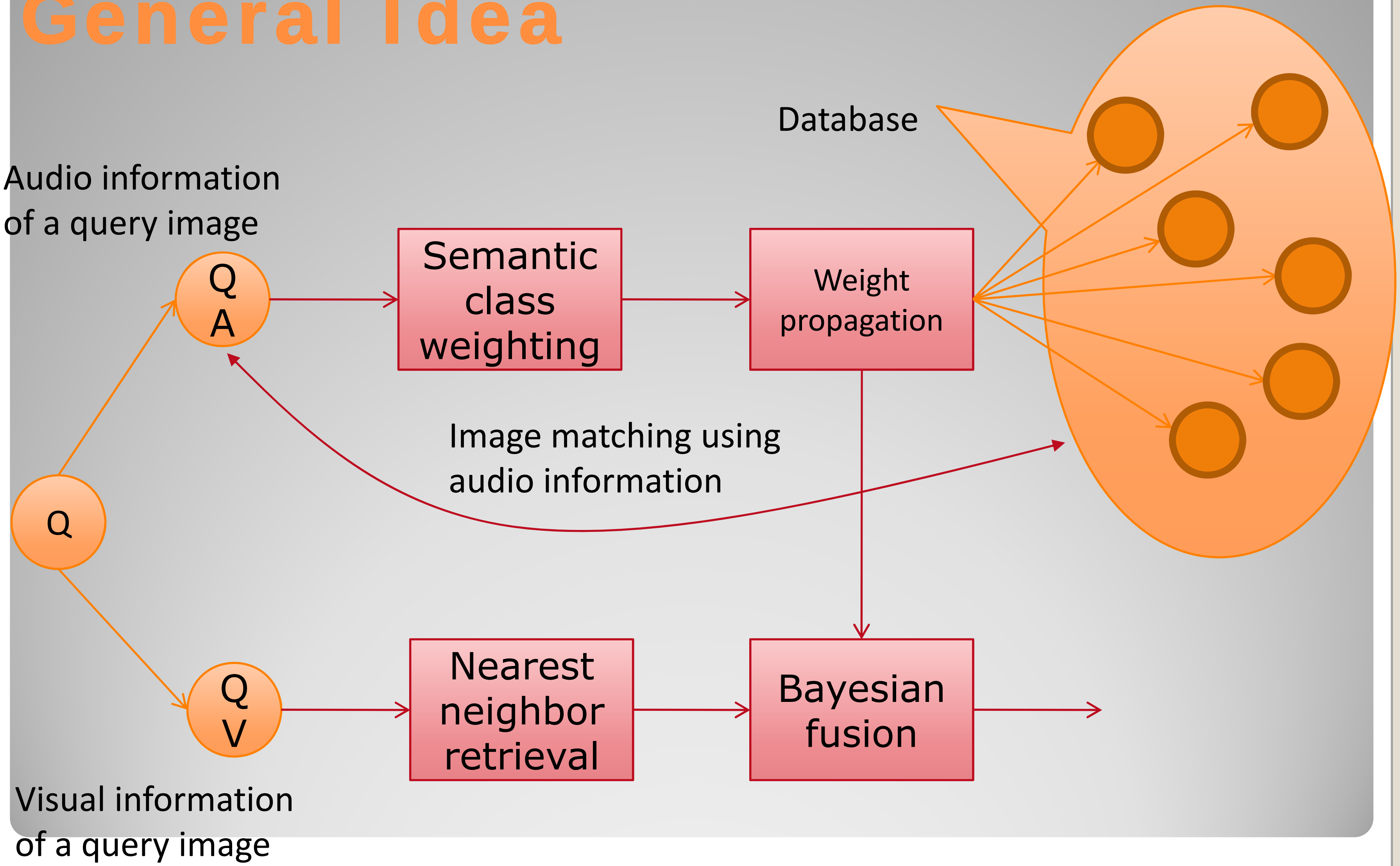
Speech Recognition

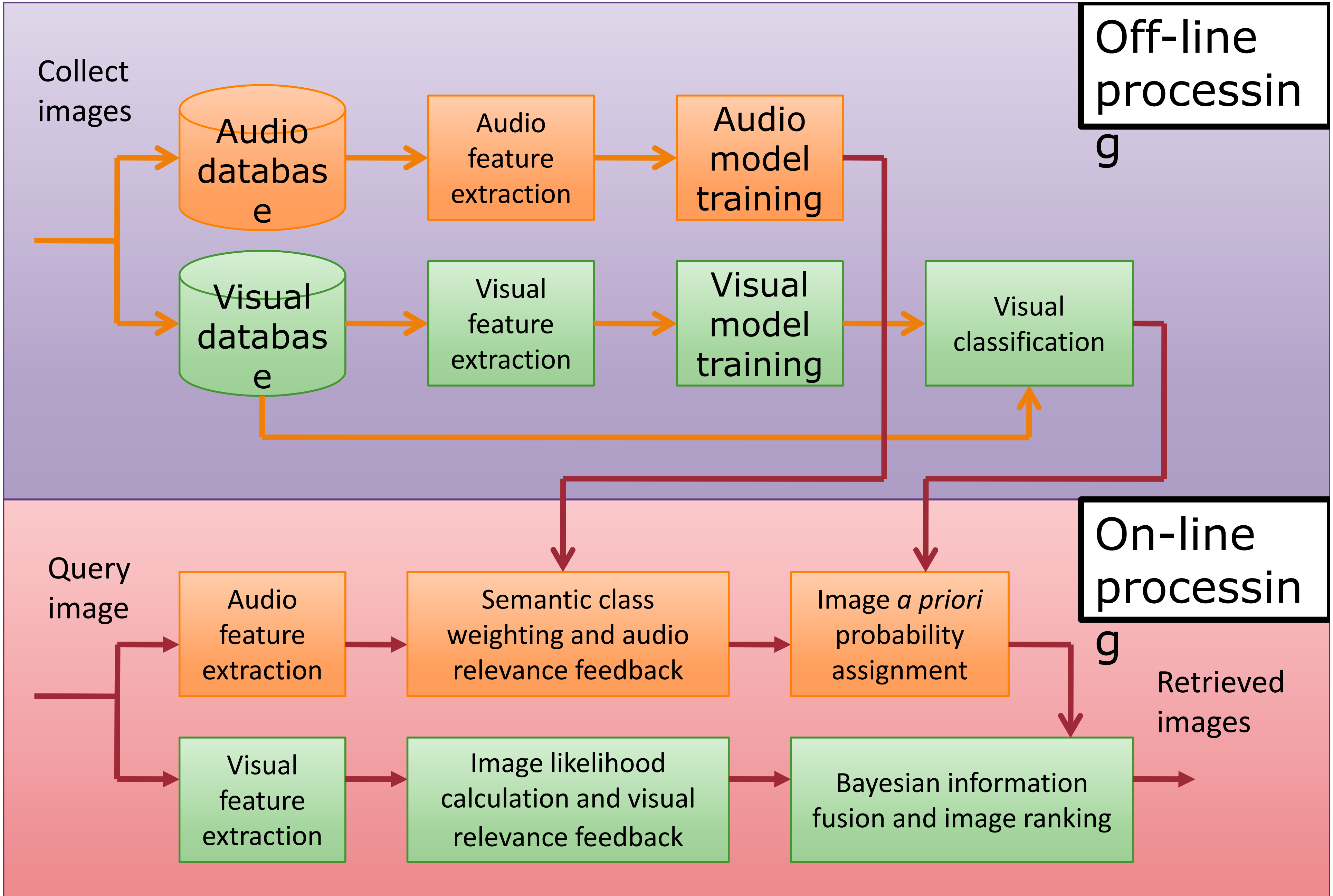
		SNR 18dB	SNR 6dB
Conventional	MFCC	**83.7**	64.7
Conventional	MFCC CMN	66.0	60.3
Conventional	Spectral Features, MD	76.5	**67.4**
COR	MFCC+MD	84.5	67.5
COR	MFCC, CMN+MD	**88.6**	**73.5**

TABLE I. RECOGNITION RESULTS WITH TEST CORPUS + FACTORY NOISE

Image Retrieval with Audio Cues

General Idea



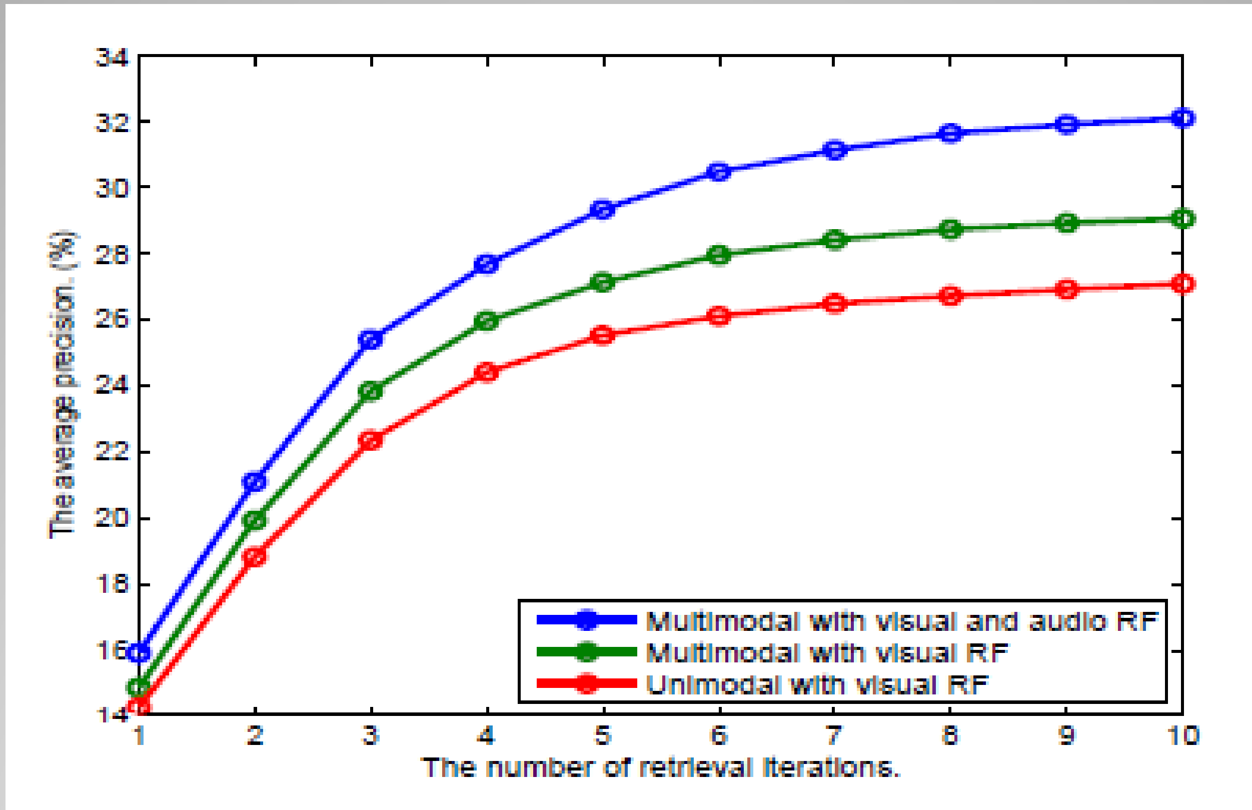

Off-line processing
Collect images
Audio database
Audio feature extraction
Audio model training
Visual database
Visual feature extraction
Visual model training
Visual classification
On-line processing
Query image
Audio feature extraction
Semantic class weighting and audio relevance feedback
Image *a priori* probability assignment
Retrieved images
Visual feature extraction
Image likelihood calculation and visual relevance feedback
Bayesian information fusion and image ranking

Experimental Setup

- Database
  - 4400 images collected from Flickr
  - Featuring 44 kinds of animals
- Visual feature selection
- Audio feature selection
  - MFCC with frame length equal to 256.

Color Feature	
Color Histogram	8, 4, 2 bins in H, S, V channels
Color Layout	An image is partitioned into 8×8 blocks, 6, 3, 3 coefficients in Y, Cb, Cr channels
Texture Feature	
Gabor Wavelet	4 scales and 6 orientations
Shape Feature	
Fourier Descriptors	10 coefficients

Experimental Results

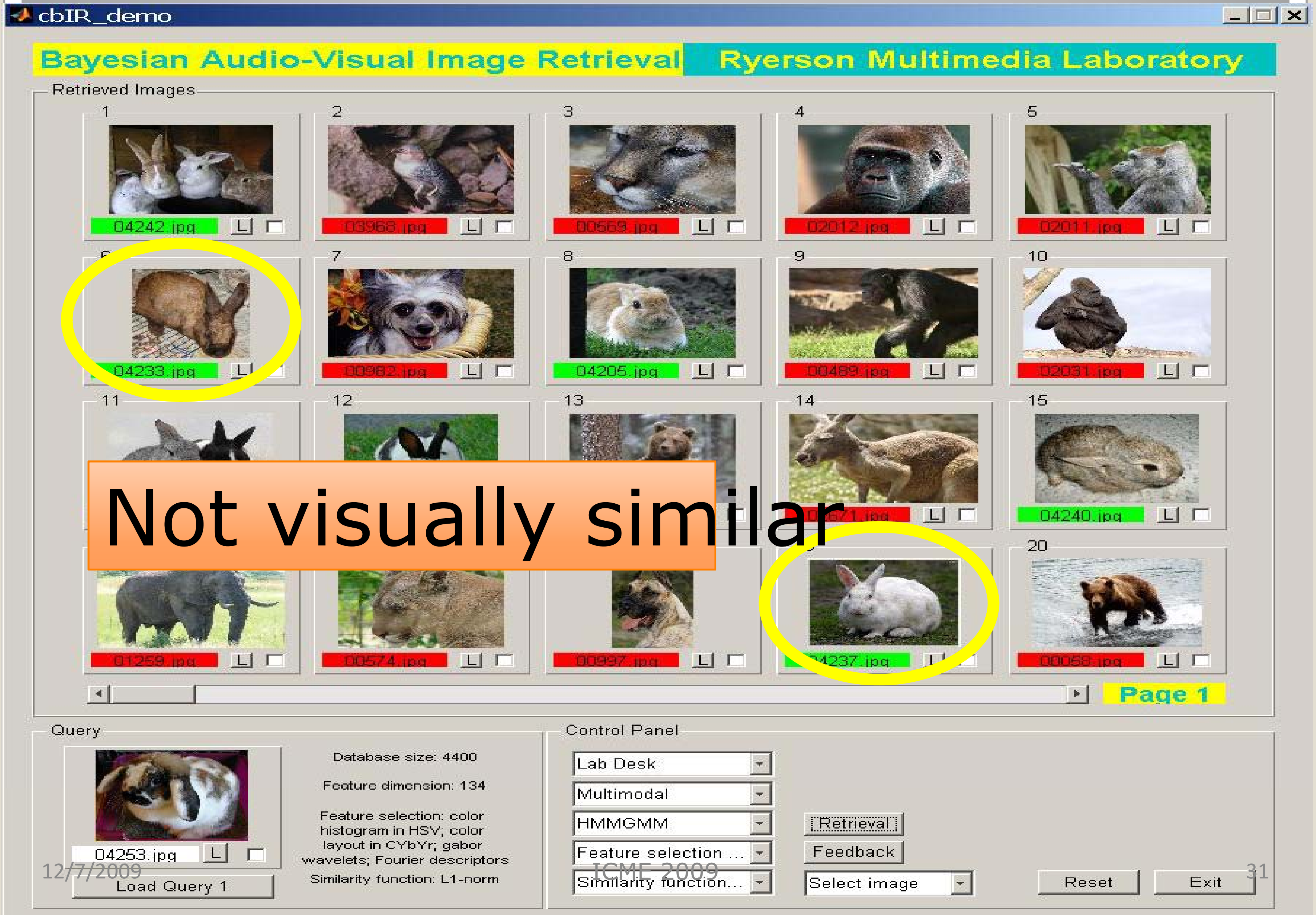


Precision as a function of the number of retrieval iterations.

Observations:

1. **Major improvement is obtained within the first iterations.**
2. **Information fusion improve the performance and the audio relevance feedback further improves it.**


cbIR_demo
Bayesian Audio-Visual Image Retrieval
Ryerson Multimedia Laboratory
Retrieved Images
03918.jpg
03142.jpg
01830.jpg
04255.jpg
03594.jpg
00963.jpg
03920.jpg
01852.jpg
01871.jpg
01487.jpg
01874.jpg
02183.jpg
03503.jpg
03225.jpg
02156.jpg
00519.jpg
00843.jpg
04322.jpg
01464.jpg
04319.jpg
Page 1
Query
04253.jpg
Load Query 1
Database size: 4400
Feature dimension: 134
Feature selection: color histogram in HSV; color layout in CYbYr; gabor wavelets; Fourier descriptors
Similarity function: L1-norm
Control Panel
Lab Desk
Unimodal
HMMGMM
Feature selection ...
Similarity function...
Retrieval
Feedback
Select image
Reset
Exit



Not visually similar

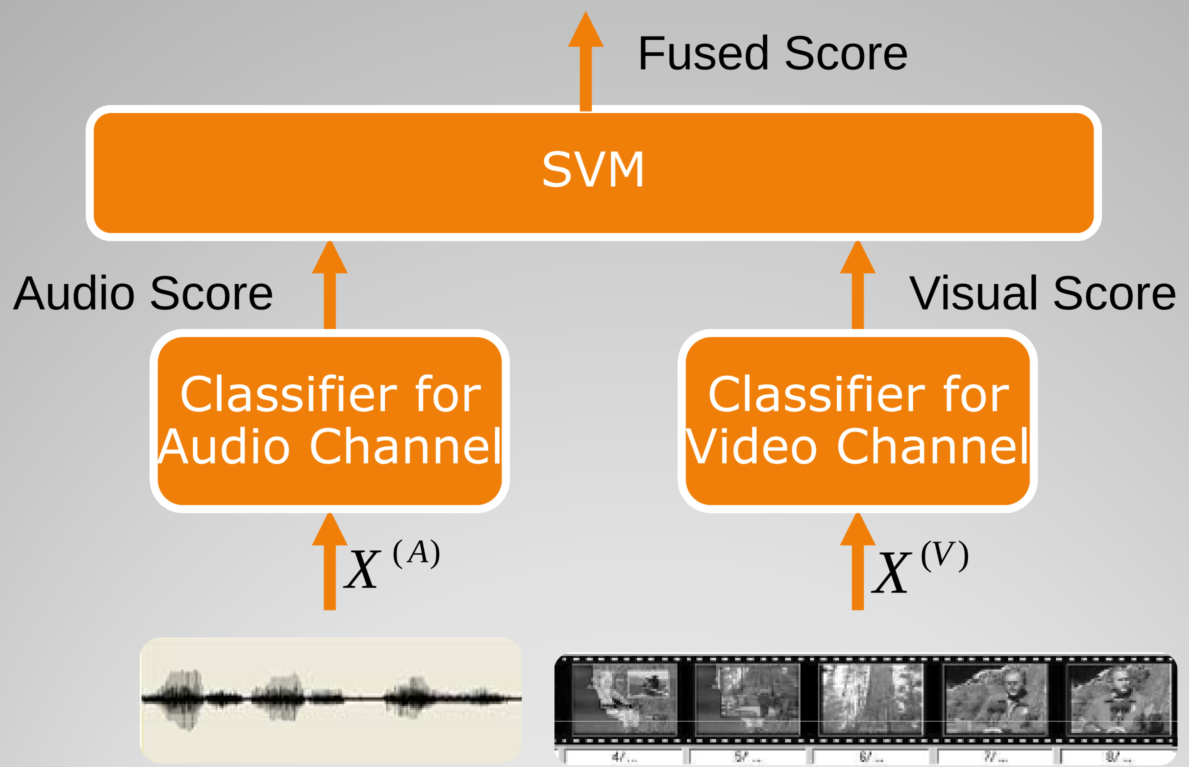
Score (Decision) Fusion

- Could be straightforward or involving more analysis.
- Rigid due to limit on information left

- Case study:
  1. Video Retrieval based on Audiovisual Cues

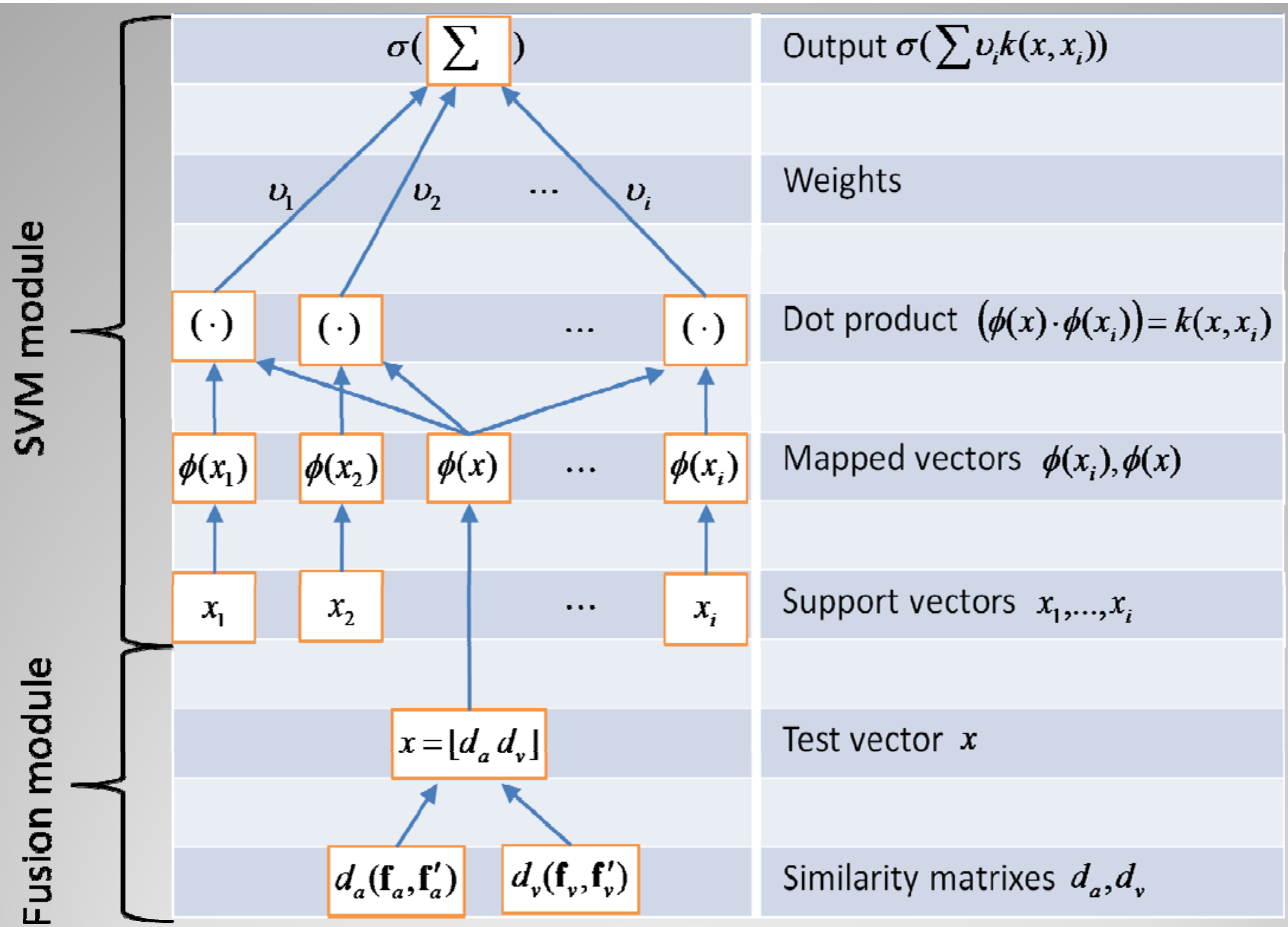
Video Retrieval by Audiovisual Cues

(Interaction or Decision?)



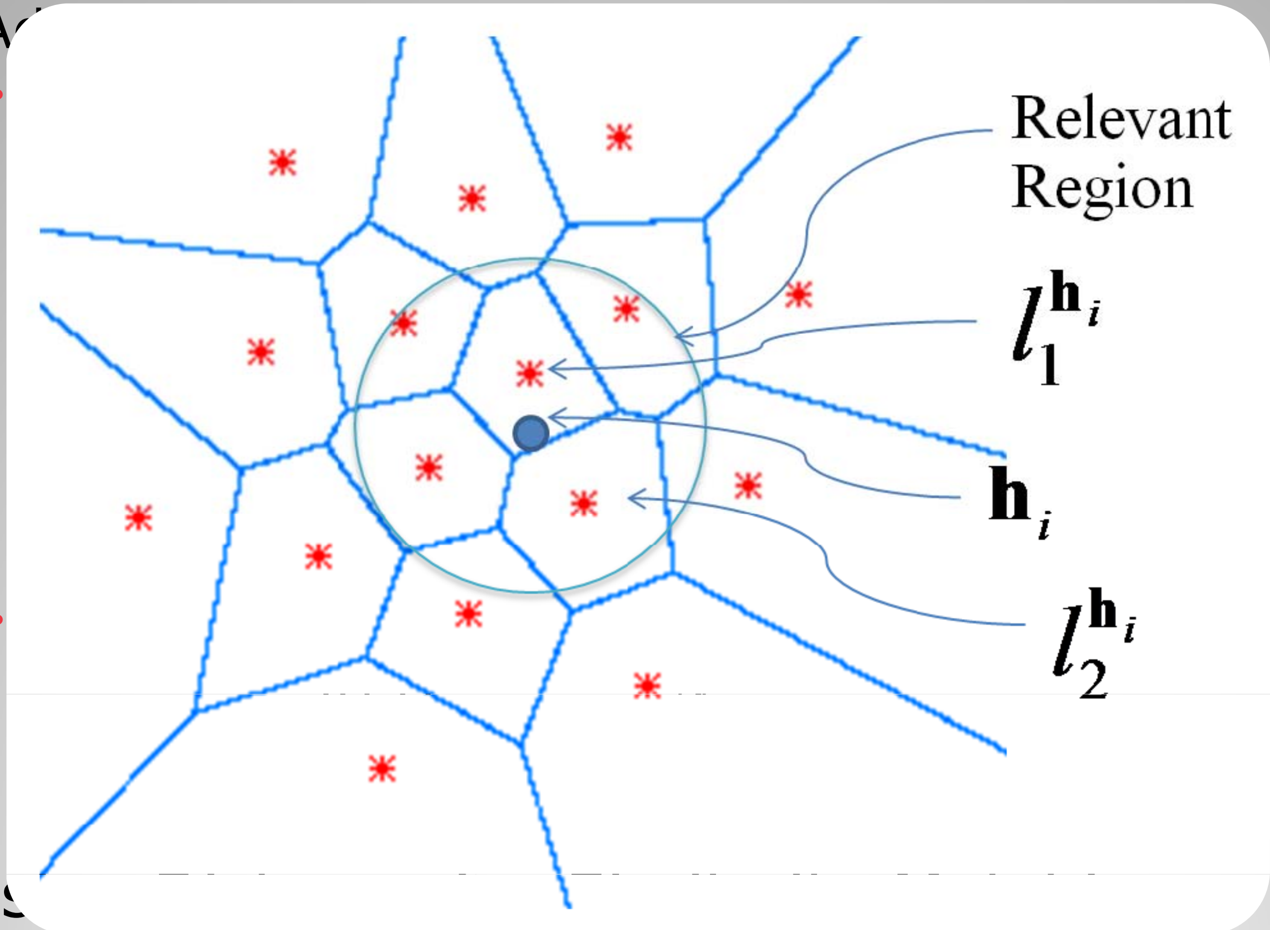
The scores are independently obtained, which are then combined

2. Video Retrieval


SVM module
Fusion module
$\sigma(\sum)$
υ_1 υ_2 ... υ_i
$(\cdot)$ $(\cdot)$ ... $(\cdot)$
$\phi(x_1)$ $\phi(x_2)$ $\phi(x)$ ... $\phi(x_i)$
x_1 x_2 ... x_i
$x = [d_a\ d_v]$
$d_a(\mathbf{f}_a, \mathbf{f}'_a)$ $d_v(\mathbf{f}_v, \mathbf{f}'_v)$
Output $\sigma(\sum \upsilon_i k(x, x_i))$
Weights
Dot product $(\phi(x) \cdot \phi(x_i)) = k(x, x_i)$
Mapped vectors $\phi(x_i), \phi(x)$
Support vectors $x_1, \ldots, x_i$
Test vector x
Similarity matrixes d_a, d_v

SVM Fusion

- Visual
  - A
- Co



Visual Feature Representation

- Laplacian Mixture Model (LMM) of wavelet coefficients of audio signal

$$p(\mathrm{w}_i) = \alpha_1 p_1(\mathrm{w}_i \mid b_1) + \alpha_2 p_2(\mathrm{w}_i \mid b_2)$$
$$\alpha_1 + \alpha_2 = 1$$

- Audio feature vector with model parameter (using EM estimator)

$$\mathbf{f}_a = \left[\{m_0, \sigma_0\}, \{\alpha_{1,i}, b_{1,i}, b_{2,i}\} \right], \quad i = 1,2,\ldots, L-1$$

- $\{\alpha_{1,i}, b_{1,i}, b_{2,i}\}$ are the model parameters obtained from the *i*-th high-frequency subband.

Audio Feature Representation

- Recognition rate obtained by the SVM based fusion model, using video database of 6,000 clips
- Five semantic concepts

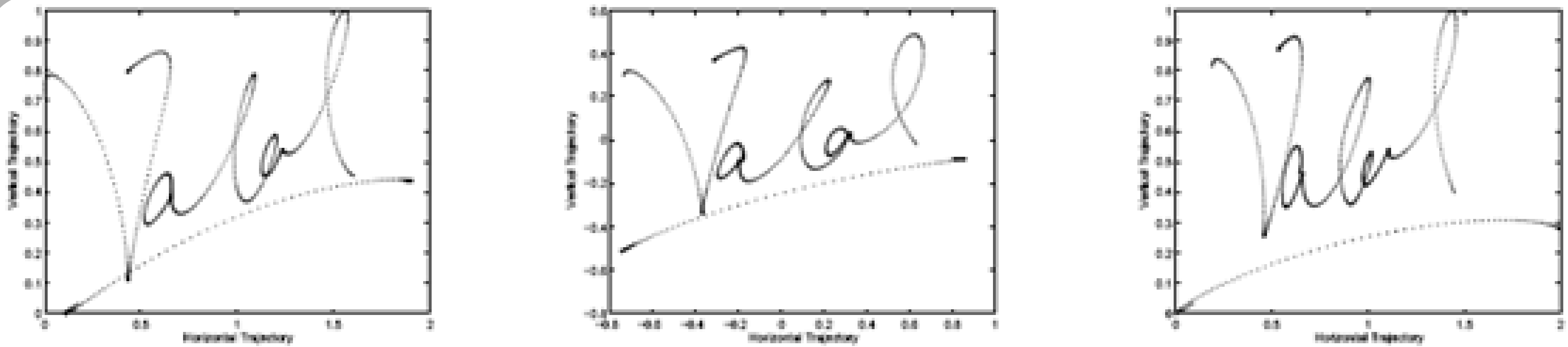
Type of concept	Accuracy (%)	False positive rate (%)	False negative rate (%)
Love scene	90.97	8.91	19.70
Music video	91.03	9.03	0
Fighting	84.68	25.65	14.55
Ship crashing	91.81	7.54	26.87
Dancing party	99.68	0.30	2.08
Average	91.63	10.29	12.64

Video Classification Result

Multimodal Human Authentication

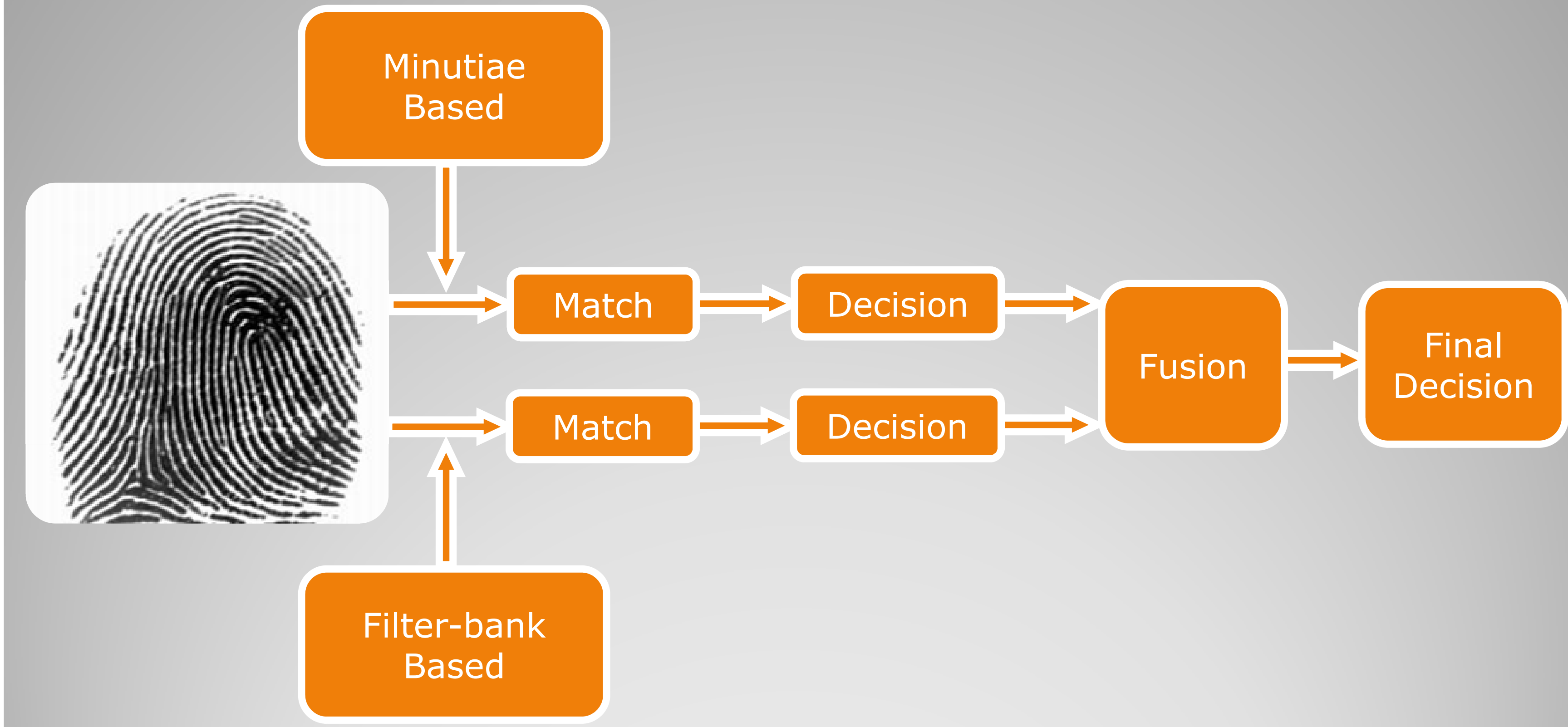
with Signature, Iris and Fingerprint

Ryerson University – Multimedia Research Laboratory
Ling Guan, Paisarn Muneesawang, Yongjin Wang, Rui Zhang, Yun Tie, Adrian Bulzacki, Muhammad Talal Ibrahim

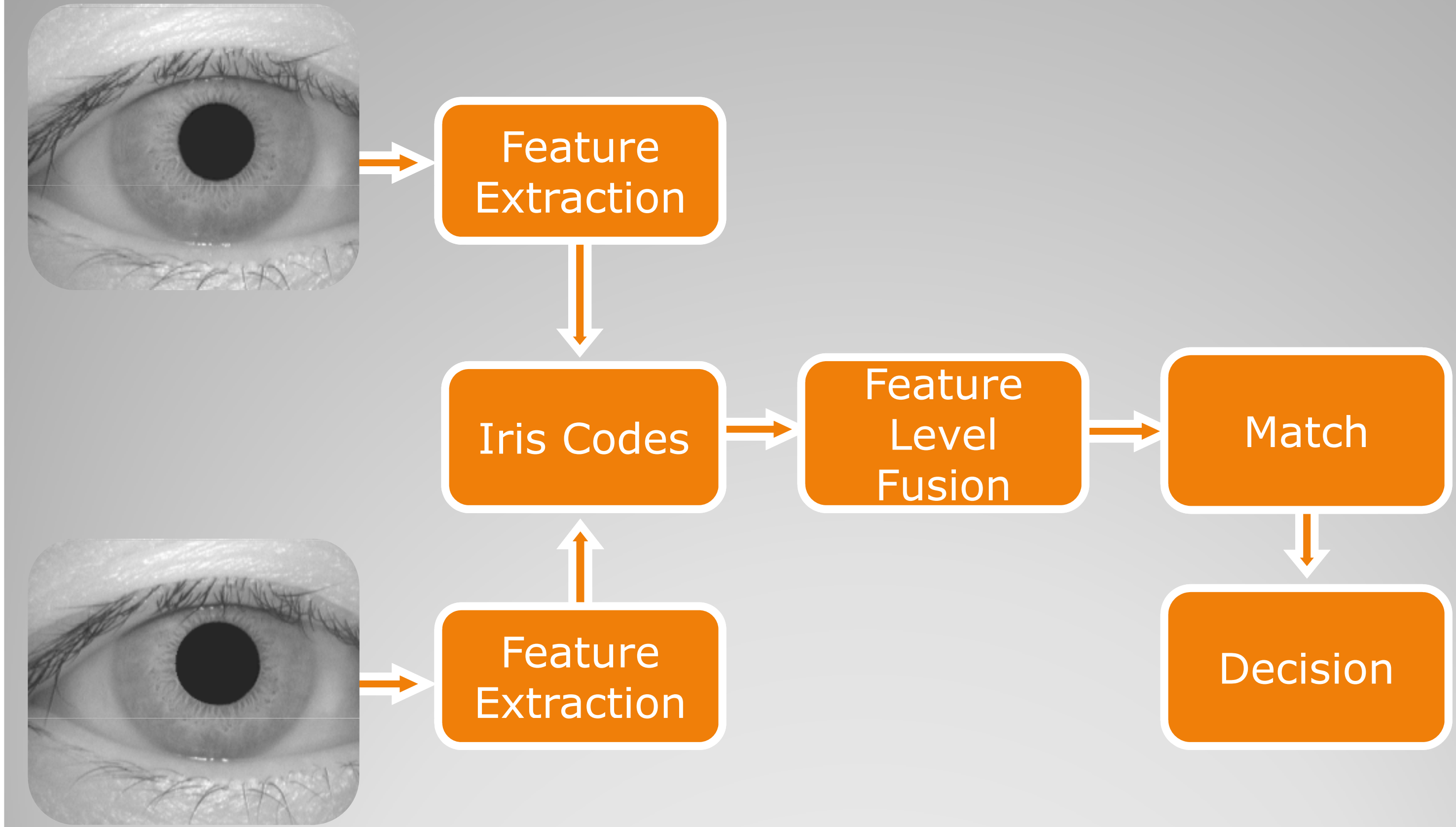


(a) (b) (c)

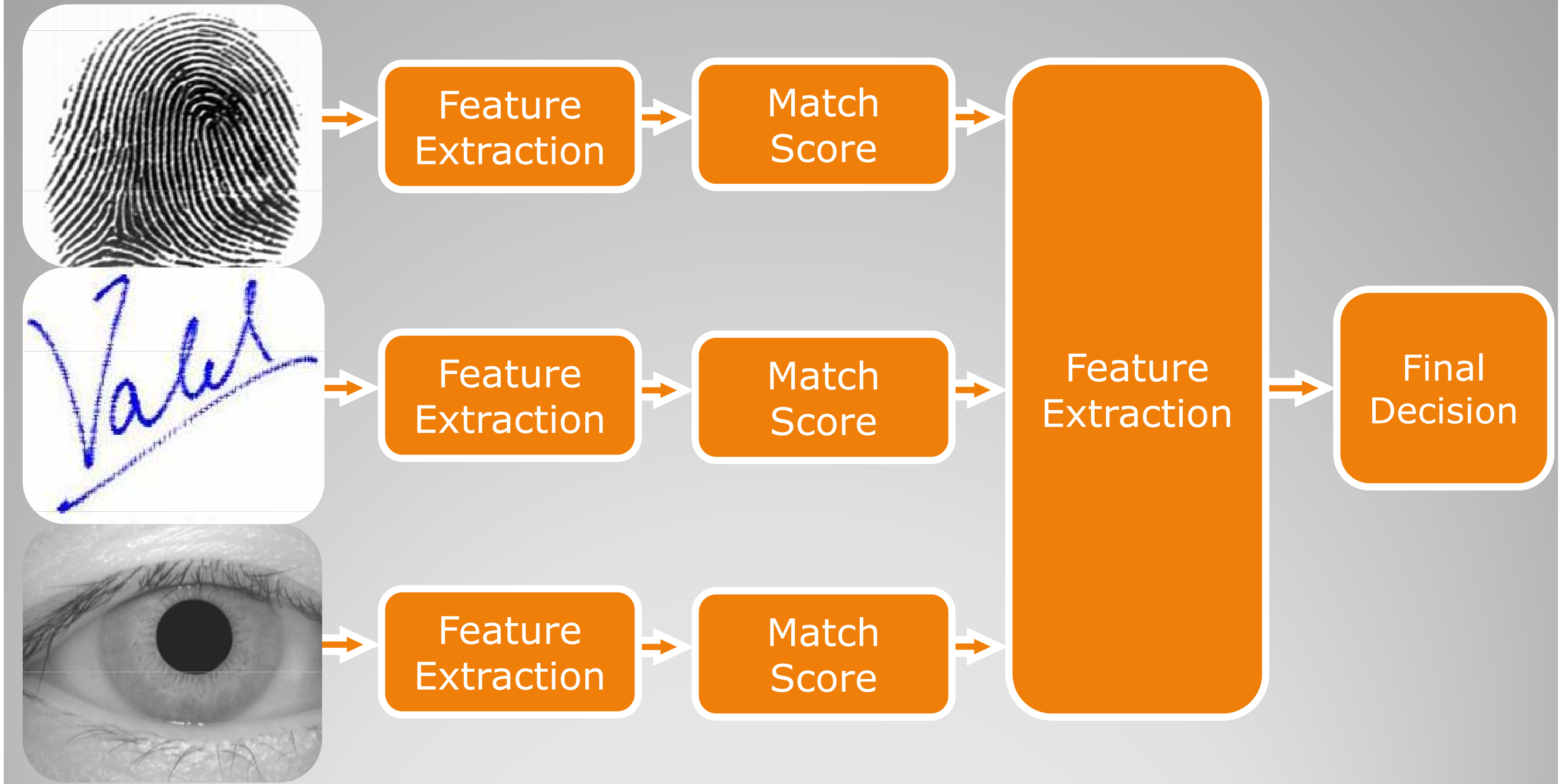
Signature Recognition



Fingerprint Image Enhancement System Overview

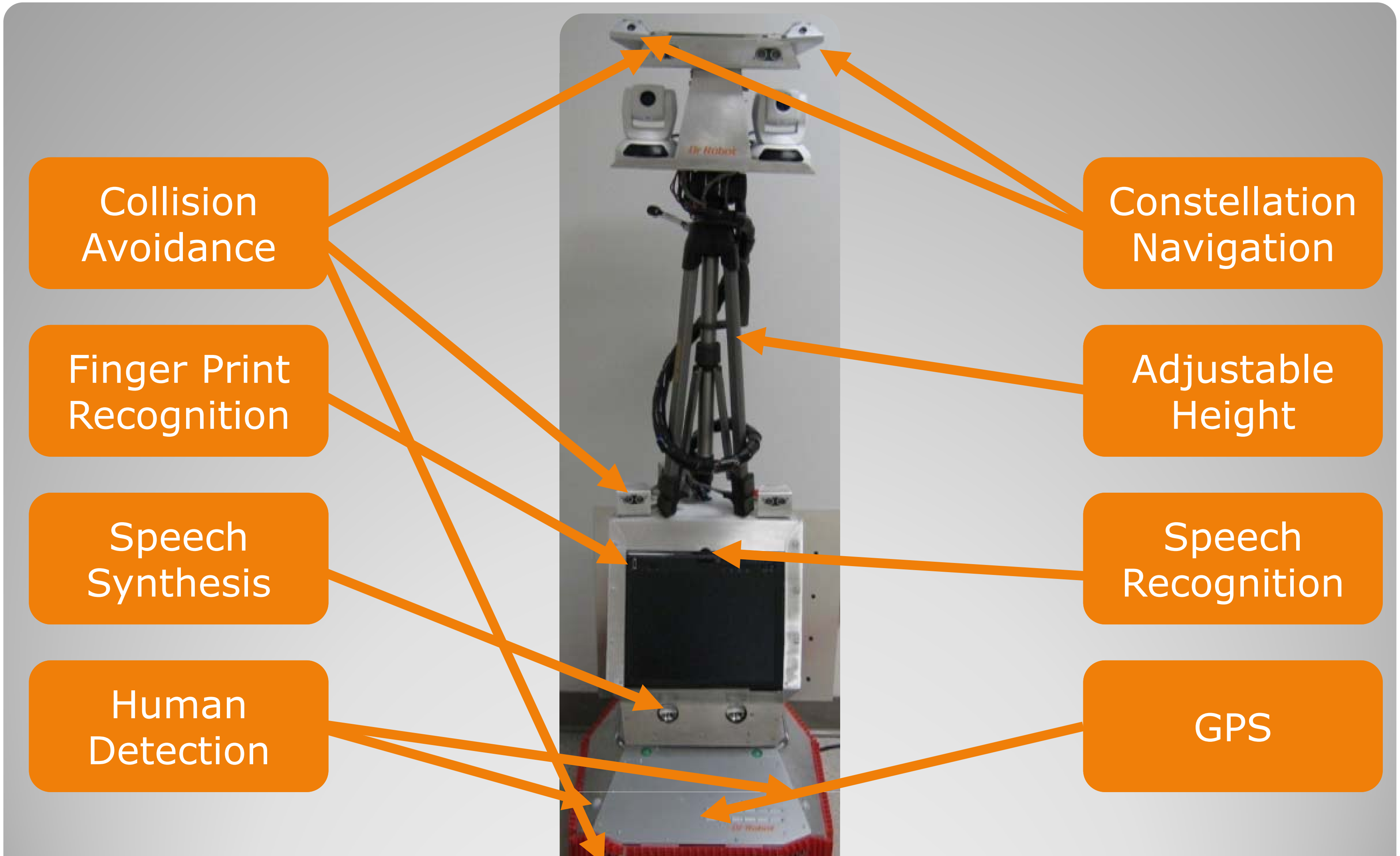


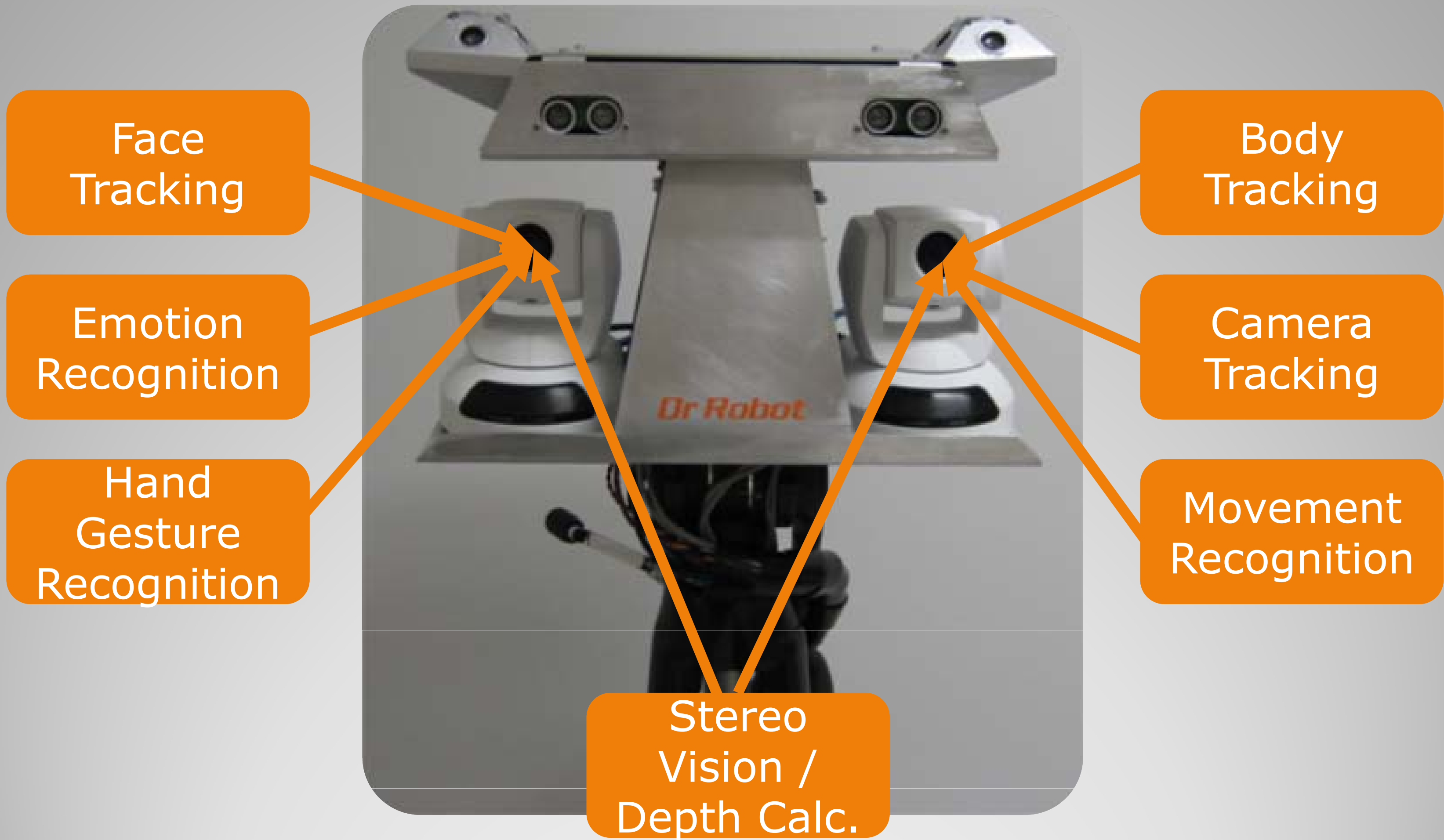
Iris Segmentation System Overview



Fingerprint / Signature / Iris Fusion

Robot Applications

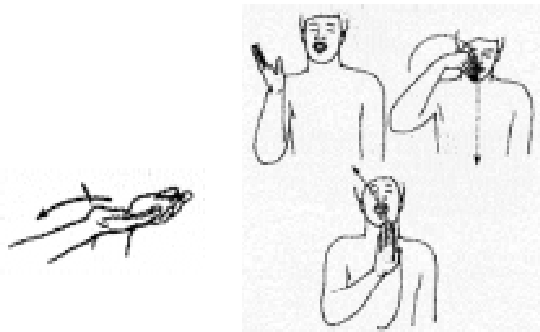




Robot Applications

Robot Application: Domestic Helper

- via emotion/intention recognition



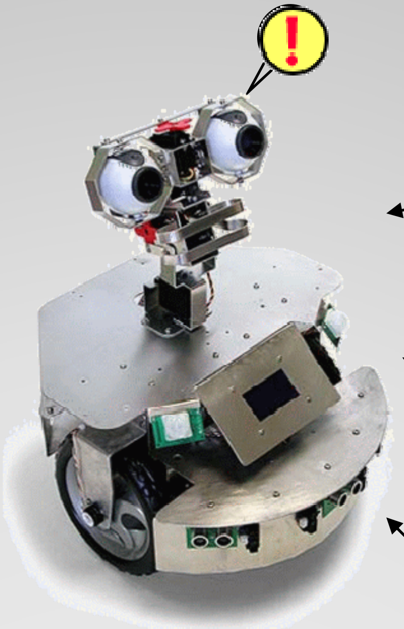
(a) " bring "   (b) " roller coaster "

1. Target people group: Elderly and disabled people at homes or community houses
2. Capable of simple gestures and body language
3. Capable of simple, and, sometime, incomplete verbal communications

Robot Application: Domestic Helper

- via emotion/intention recognition

1. Help the elderly and the disabled with their daily life.
2. Entertain the people they look after.
3. Call the nurse or emergency when in need.



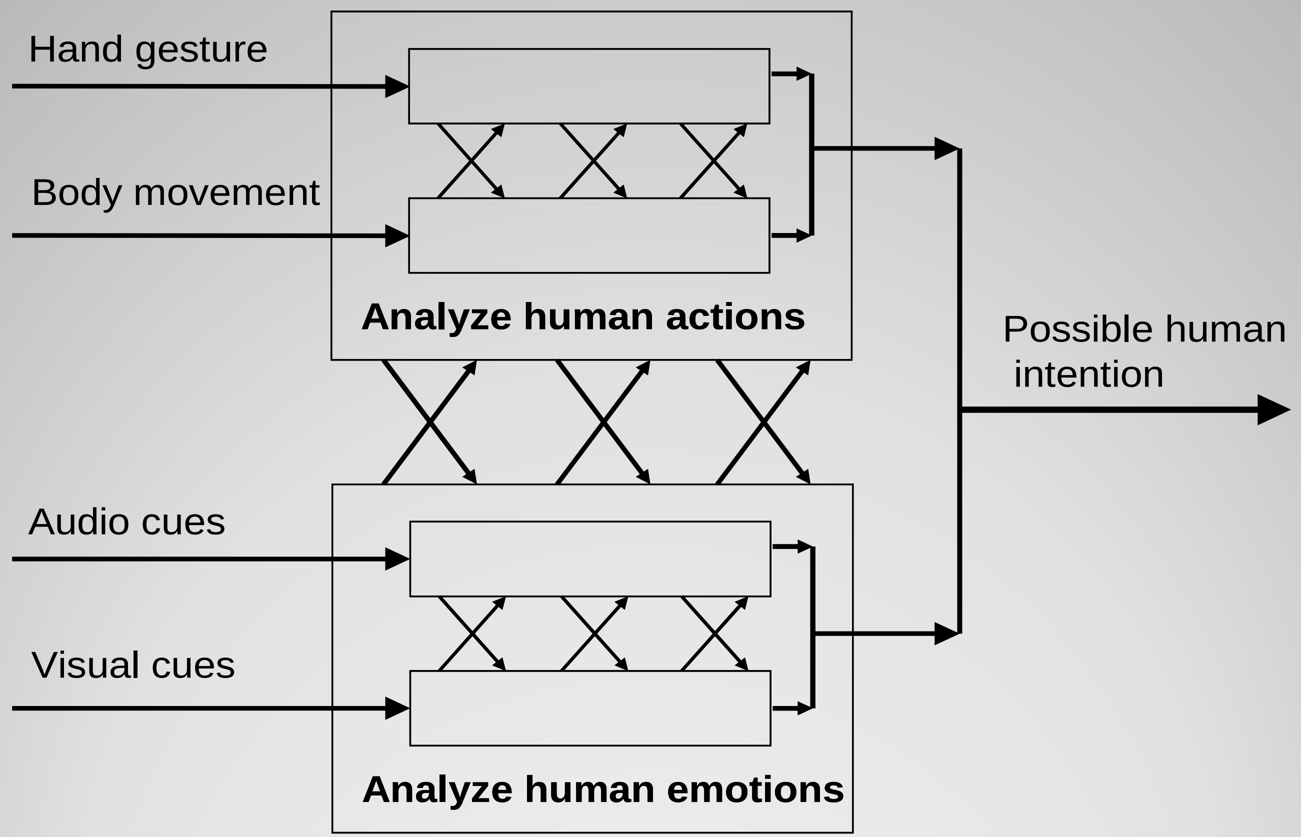
Head tracking

Emotion

Behavior

Such robots may also be deployed at airports, banks, subway stations, etc, to identify potential threats

Multimodal Fusion for Human Intention Recognition



Challenges

- Prof. Ming-ting Sun gave an indepth discussion about the major challenges in his workshop keynote speech yesterday.
- Stress on one issue: What is the best fusion model for a problem on hand?
  - Data level,
  - Interaction level,
  - Decision level,
  - Or multilevel.
- New data analysis and mining tools need to be developed to address the issue. Or the existing tools may be revisited.

Summary

- Fusion – coherent integration of multimedia multimodal information
- It is a natural process by human beings, but not straightforward for machines.
- It may be carried out at different information levels, but how to choose the right model?
- Several case studies are used to demonstrate the power of information fusion
- Multiple challenges are waiting to be addressed

Thank You

Any questions?