IM2 Phase III – IP1
Integrated Multimodal Processing

IP1 Report Summer Institute Sept. 2010

Aude Billard & Stephane Marchand-Maillet

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

IM
INTERACTIVE
MULTIMODAL
INFORMATION
MANAGEMENT

IP1: Integrated Multimodal Processing

Goals: IP1 covers primarily multimodal research

- 9 research teams pursue some of the fundamental research directions initiated in IM2-II.

→ Improving unimodal (audio and visual) processing.

→ Automatic speech recognition systems,

→ speaker diarization system (who spoke when),

→ visual scene analysis (body tracking, gesture recognition, gaze tracking, object recognition, etc).

→ Developing multimodal data processing and multi-modal applications.

→ multimodal object description,

→ multimodal emotion analysis in meeting data,

→ proactive navigation of multimodal data

IP1: Integrated Multimodal Processing

Goals: IP1 seeks the integration of research components into applications

- Improved development of two of the prototypes developed in IM2-II, namely, the *Automatic Content Linking Device* and the *Mobile Meeting Assistant.*
- Unimodal research applied in other projects within IP1
  - keyword spotting is used within the multimodal object description project,
  - automatic speech recognition is used in the Multimodal meeting assistants application).

IP1: Integrated Multimodal Processing

8 Reporting Subprojects

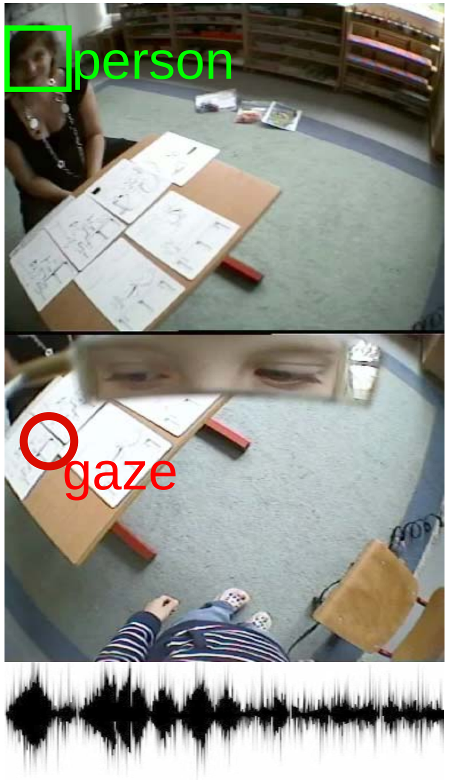
- EPFL-LASA & IDIAP
  - A/V gaze detection, object detection, kwd spotting for OoI identification
- EPFL-LTS5
  - Dynamic facial expression analysis for emotion recognition (meeting data)
- Idiap-Gatica-Perez
  - Multimodal analysis of human behaviour
- Idiap-Garner
  - Automatic Speech Recognition
- UniGE-Viper
  - Multimedia mining and navigation
- EPFL- MMSPG
  - Multimodal quality metrics for multimedia content, Multimodal Content Annotation
- Idiap-Popescu-Belis
  - Multimodal meeting assistants and content linking
- UniFR-DIVA
  - Economic hand & gesture recognition (capture & tracking) → Will be reported upon in IP2-report

AUDIO-VISUAL FUSION FOR OBJECT OF INTEREST DETECTION (*EPFL – LASA & IDIAP*)

- Combine video analysis (Gaze + Object detection) with speech analysis to extract information from a dyadic interaction between an adult experimenter and a child

  - Keyword spotting - Extract utterances of key words in the interaction (e.g. name of the child, name of objects in the environment)

  - Object detection - Extract the position of relevant people/objects in the environment and compare to gaze detection

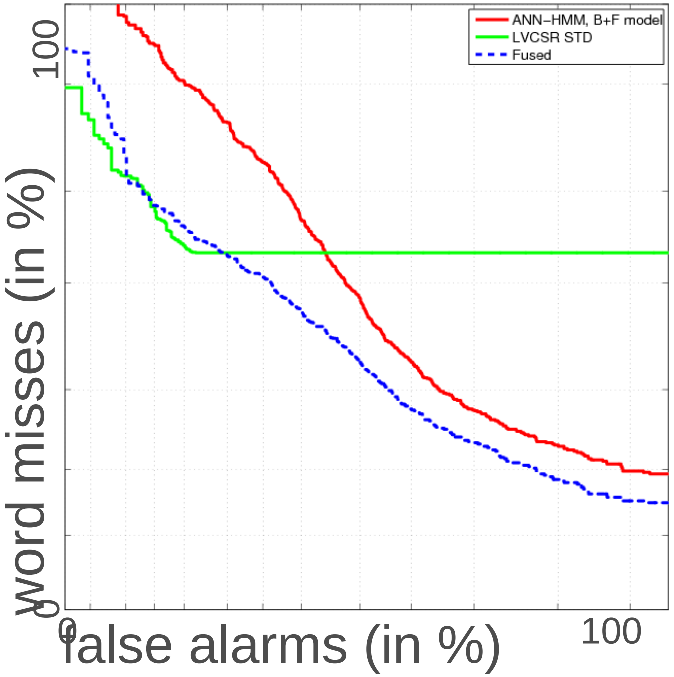


keyword

FNSNF
Fonds National Suisse
Schweizerischer Nationalfonds
Fondo Nazionale Svizzero
Swiss National Science Foundation

KEYWORD SPOTTING *(EPFL – LASA & IDIAP)*

- Acoustic Keyword Spotting (ANN-HMM)
  - Keyword vs. Background models
  - Very fast to compute
- Large Vocabulary Speech Recognition (LVCSR)
  - Use a large vocabulary to recognize speech
  - Look for keywords inside the extracted text
  - Very slow (multiple passes)
- ANN-HMM + LVCSR fusion improves accuracy



Testing Data

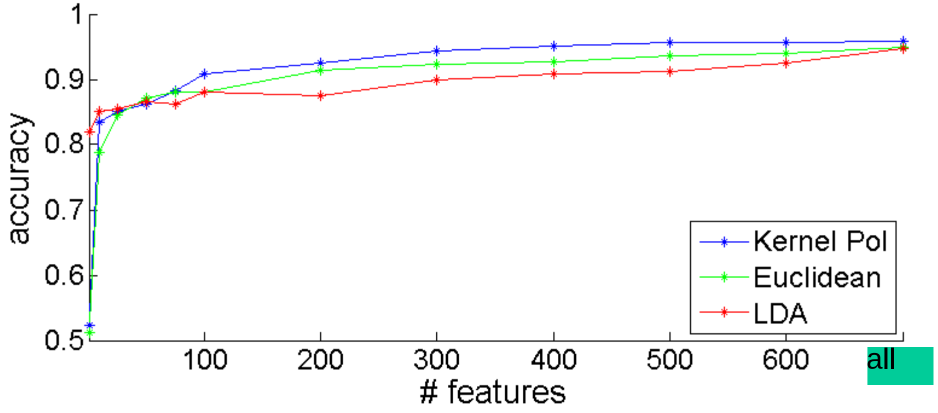
- 67 minutes (IM2 recordings from LASA)
- microphone recordings (16k), SNR~20dB
- 3 non-native, 3 native speakers
- 8 keywords (740 occurrences of keyword)

method	EER [%]	FOM [%]
LVCSR	12*	37.4
ANN-HMM	18	24.2
Fusion	13	42

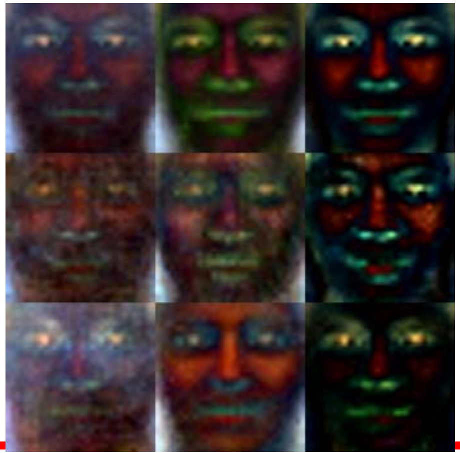
FEATURE SELECTION *(EPFL – LASA & IDIAP)*

- Find a subset of the training data that contains relevant information for finding ‘good features’
  - Support Vectors (SV) selected by a “naive” SVM
- Extract good discriminating features
  - Kernel polarization, Euclidean Distance, LDA
- Perform the feature selection procedure on a large clean dataset, and adapt to a new smaller dataset

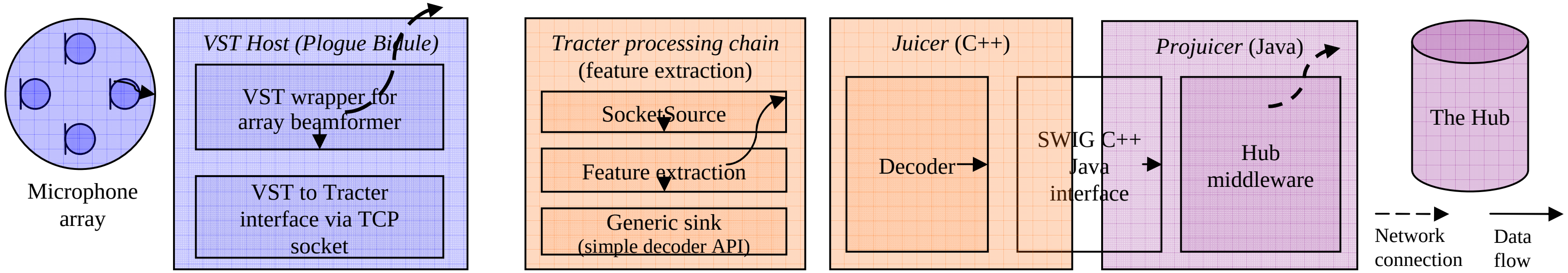
Feature Selection (from all samples)



Feature Weights (Color Feret Database)



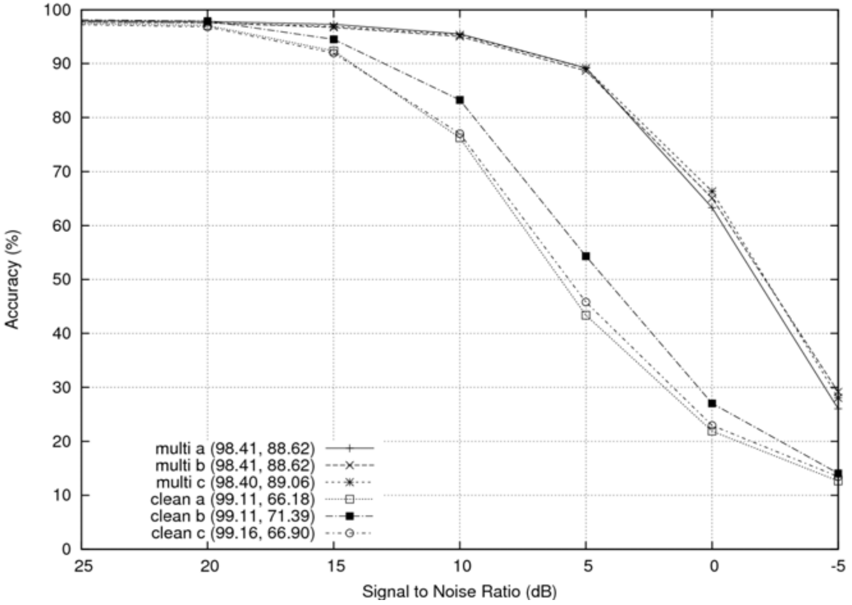
ASR *(IDIAP ASR group)*

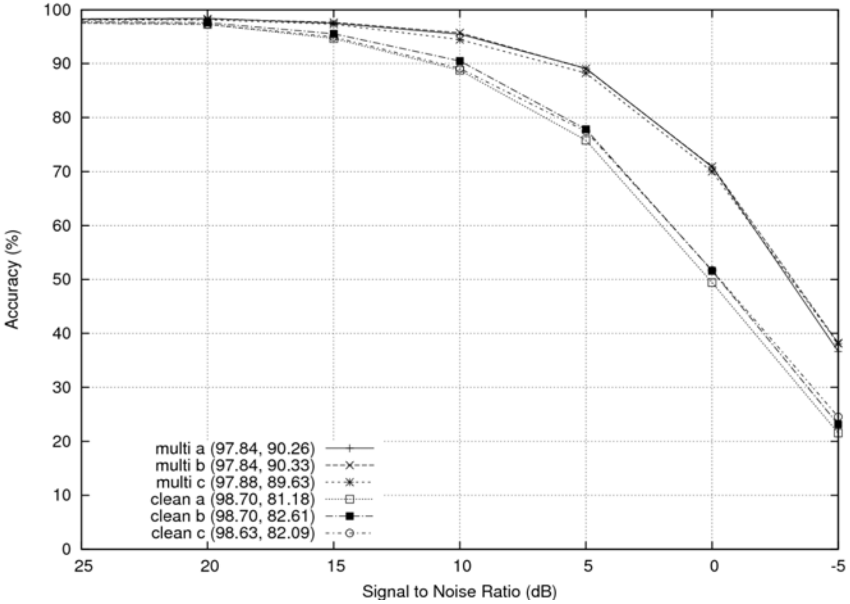


- An ASR system was developed in phase 2
  - Also under AMI/DA
  - Good performance in RT06/07/09 evaluation
  - Real-time capable
- Demonstrator exists on Mac
- Efforts put into commercialization:
  - joint licencing agreement between Idiap and theUniversities of Edinburgh and Sheffield
  - startup company called Koemei

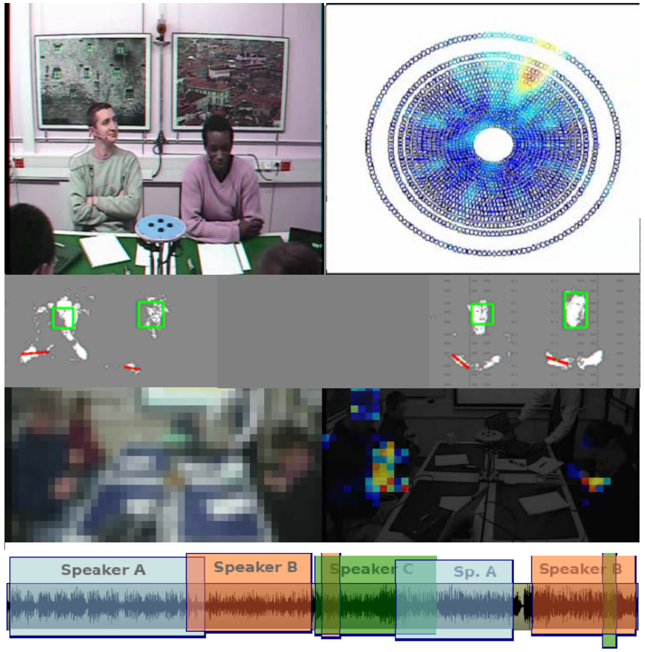
SNR features *(IDIAP ASR group)*

- Essentially an investigation in to the effect of cepstral normalisation
- Enhances the effect of cepstral variance





Audio-visual speech synchrony for multimodal speaker diarisation *(IDIAP ASR group)*



- Speaker diarisation motivation: correlation between facial movements and speech acoustics due to speech production

Estimation of Canonical Correlation Analysis and Mutual Information between:

- Acoustic features:
  - Speech energy
  - Mel Frequency Cepstral Coefficients
- Visual features:
  - Pixel-by-pixel motion intensity
  - Kanade Lucas Tomasi (KLT) motion tracking

Audio-visual synchrony features were compared with other visual cues: motion and head pose features:

Listeners are likely to look at the person who is talking

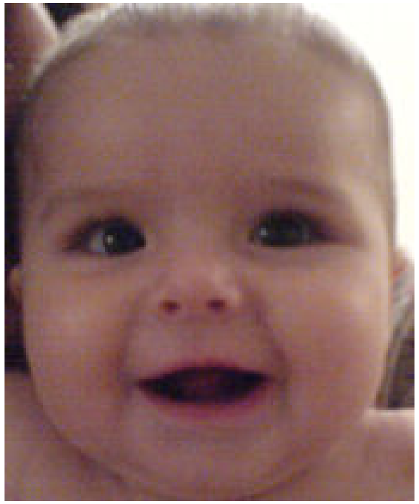
ASR *(IDIAP ASR group)*

Investigated the use of phoneme posterior probabilities estimated by multilayer perceptron for:

- Genre specific acoustic modelling (in the context of Mandarin ASR)
- Template-based ASR
- Language identification
- Grapheme-based ASR

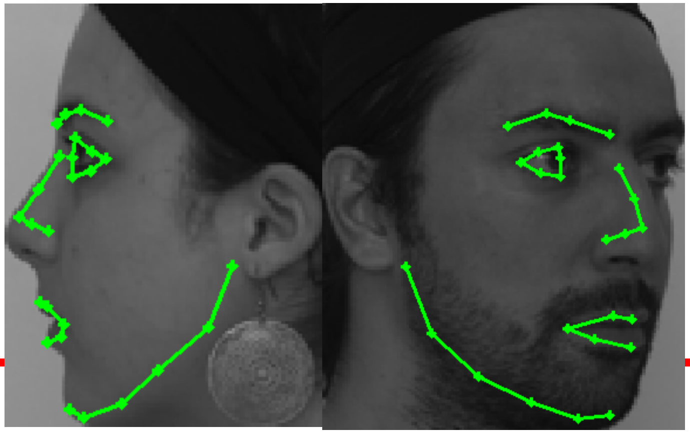
Multimodal Emotion Analysis in Meeting Data *(EPFL – LTS5)*

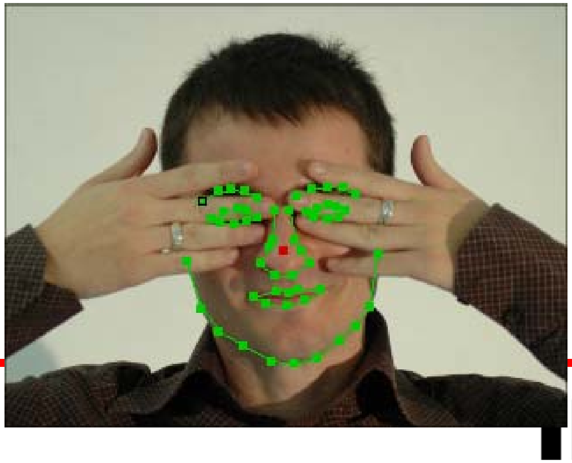
- Main task of the project is to perform dynamic facial expression recognition combined with audio/visual video analysis to infer emotions in meeting data



Multimodal Emotion Analysis in Meeting Data *(EPFL – LTS5)*

- Achievements so far:
  - Improving the face tracker to handle more general cases in order to use in meeting data
    - Combining the frontal model with side models to include various cases of head pose
    - Integrating with a model switching algorithm that utilises a fast robust PCA reconstruction method as a preprocessing step to facilitate tracking in cases of occlusions



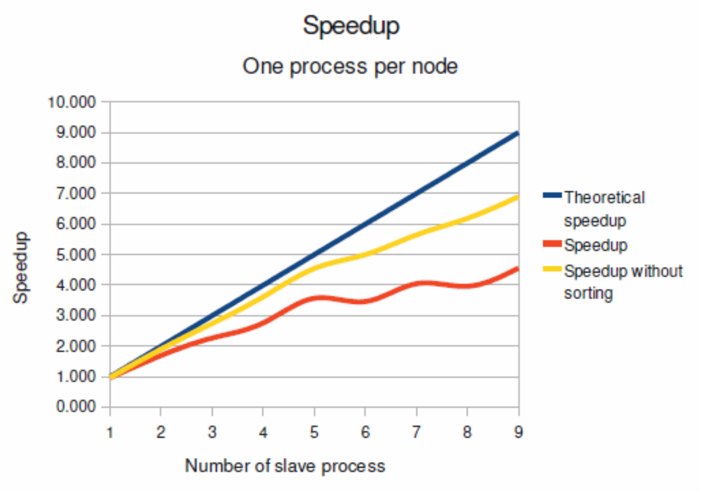


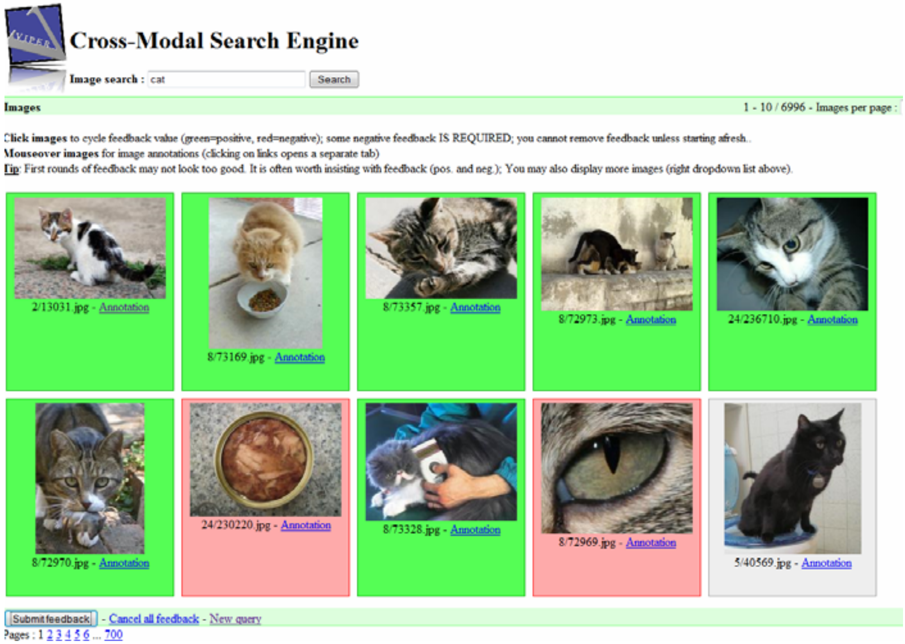
IP1: Integrated Multimodal Processing

- Improved development of two of the prototypes developed in IM2-II, namely, the *Automatic Content Linking Device* and the *Mobile Meeting Assistant.*

Mulimedia Information Retrieval (UniGE-Viper)

- Development of a **distributed version** of our Cross modal Search Engine (CMSE) → Public Demonstrator
  - Using MPICH over a cluster of PCs
  - Indexing of image retrieval benchmarks (ImageCLEF) to help to understand the content of the collection





Cross-Modal Search Engine

Image search : cat

Images 1 - 10 / 6996 - Images per page :

Click images to cycle feedback value (green=positive, red=negative); some negative feedback IS REQUIRED; you cannot remove feedback unless starting afresh..
Mouseover images for image annotations (clicking on links opens a separate tab)
Tip: First rounds of feedback may not look too good. It is often worth insisting with feedback (pos. and neg.); You may also display more images (right dropdown list above).

2/13031.jpg - Annotation
8/73169.jpg - Annotation
8/73357.jpg - Annotation
8/72973.jpg - Annotation
24/236710.jpg - Annotation
8/72970.jpg - Annotation
24/230220.jpg - Annotation
8/73328.jpg - Annotation
8/72969.jpg - Annotation
5/40569.jpg - Annotation

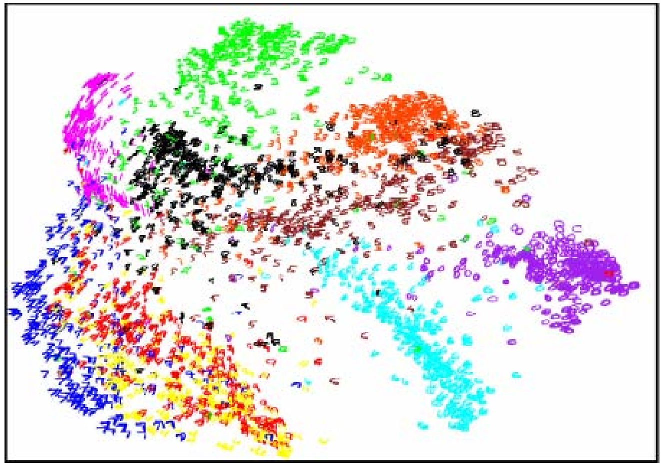
- Cancel all feedback - New query

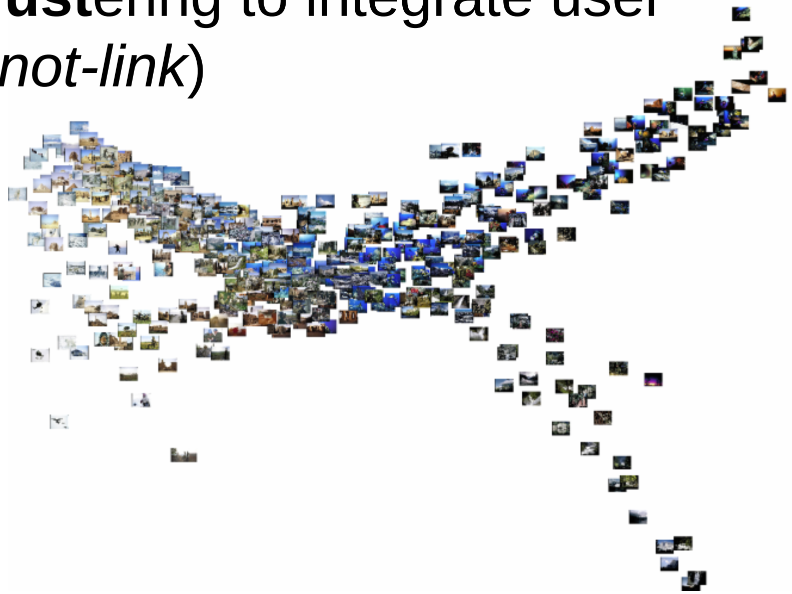
Pages : 1 2 3 4 5 6 ... 700

/lultimodal data mining and embedding (UniGE-Viper)

- Multimodal data embedding
  - Integration of our **structure preserving embedding** techniques for the construction of a multimedia browser

- Constrained/semi-supervised clustering
  - Extension of the **multiview clust**ering to integrate user constraints (eg *must-link, cannot-link*)





IM2 Phase III – IP1
Integration and HCI: Idiap contributions

Reported by: Andrei Popescu-Belis

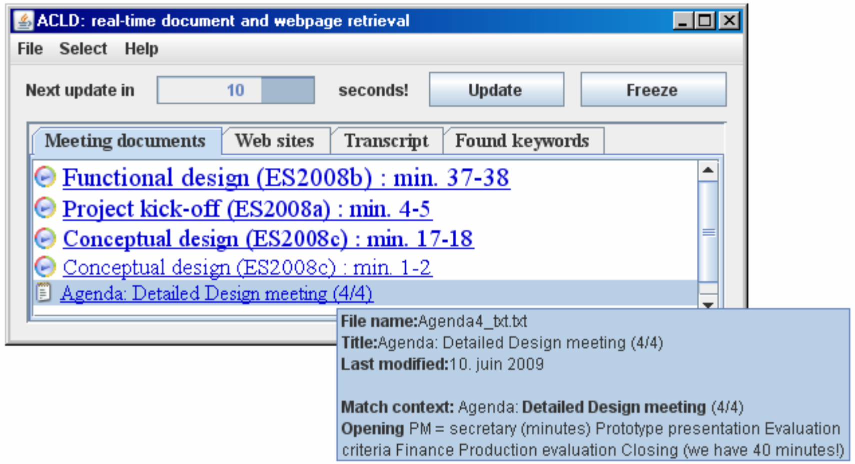
with: Alex Nanchen and Majid Yazdani

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

IM2
INTERACTIVE
MULTIMODAL
INFORMATION
MANAGEMENT

IP1 / Interfaces: 2010 achievements (IDIAP)

- ACLD: enrich in real-time a discussion with relevant documents, excerpts of multimedia recordings, and websites
  - “just-in-time retrieval” or “query-free search”



ACLD: real-time document and webpage retrieval

Functional design (ES2008b) : min. 37-38
Project kick-off (ES2008a) : min. 4-5
Conceptual design (ES2008c) : min. 17-18
Conceptual design (ES2008c) : min. 1-2
Agenda: Detailed Design meeting (4/4)

File name: Agenda4_txt.txt
Title: Agenda: Detailed Design meeting (4/4)
Last modified:10. juin 2009

Match context: Agenda: **Detailed Design meeting** (4/4) **Opening** PM = secretary (minutes) Prototype presentation Evaluation criteria Finance Production evaluation Closing (we have 40 minutes!)

- Software integration
  - installation of ACLD+rtASR on a laptop (dual core Mac)
    - → 1st portable ACLD demo
  - duplicated installation in the Idiap Show Room
  - frequent demos
  - packaging/documentation for new installations in IP2
    - specifically for the **Augmented Teams** application at CRAFT/EPFL
    - including a user-centric evaluation in IP2/UniFr

IP1 / Interfaces: 2010 achievements (IDIAP)

- Semantic disambiguation for semantic search
  - built very large network of documents from EN Wikipedia
  - computed relatedness in a random walk framework
    - several distances: hyperlinks, lexical similarity, or a mix (multi-stage)
  - stationary state was computed using either:
    - hitting time, visiting probability, personalized page rank
  - fast mapping of input text to network-based representation
- Used for three text similarity tasks
  - word similarity, paraphrase detection, document similarity
  - state-of-the-art results, but using *a unique resource*

Future work for end 2010 (IDIAP)

- Transfer ACLD to IP2 Augmented Teams
  - installation, documentation, change system for new requirements, including multimodal feedback
  - start user-oriented evaluation
- Help with transfer of other IP1 technology to AT
- Semantic distance for disambiguation
  - application to information retrieval
  - application to content linking with spoken input
    - integration into the Augmented Teams application

EPFL MMSPG

➢ Multimodal quality metrics for multimedia content

- **3rd year PhD student:** Francesca De Simone
- **Thesis objectives:** to provide a significant contribution to both subjective and objective aspects of the quality assessment challenge, focusing on two study cases:
  - quality assessment of high resolution data in the context of performance evaluation and comparison of compression algorithms
  - quality assessment of audio-visual sequences, compressed and transmitted over error-prone networks, in multimedia applications

➢ Multimodal content annotation

- **2rd year PhD student:** Ivan Ivanov
- **Thesis objectives:** the goal of this research is to address the challenge of efficient management and organization of image collections by enriching images with a semantic context; this thesis aims at developing a system to enrich images with automatic annotation based on multimedia content analysis and social network tagging

Multimodal quality metrics for multimedia content

(EPFL MMSPG)

- ➢ In the context of quality assessment for codec performance evaluation and comparison:
  - ✓ We performed the widest ever subjective test campaign for subjective quality assessment and performance comparison of video codecs in the framework of ISO and MPEG joint standardization effort for the development of the High Efficiency Video Coding (HEVC) standard.

- ➢ In the context of understanding human perception of multimedia quality:
  - ✓ We performed many psychovisual tests to study:
  - the impact of video transcoding artifacts on the perceived quality of video sequences;
  - the influence of different combinations of the scalability parameters in scalable video coding schemes on the overall perceived quality.

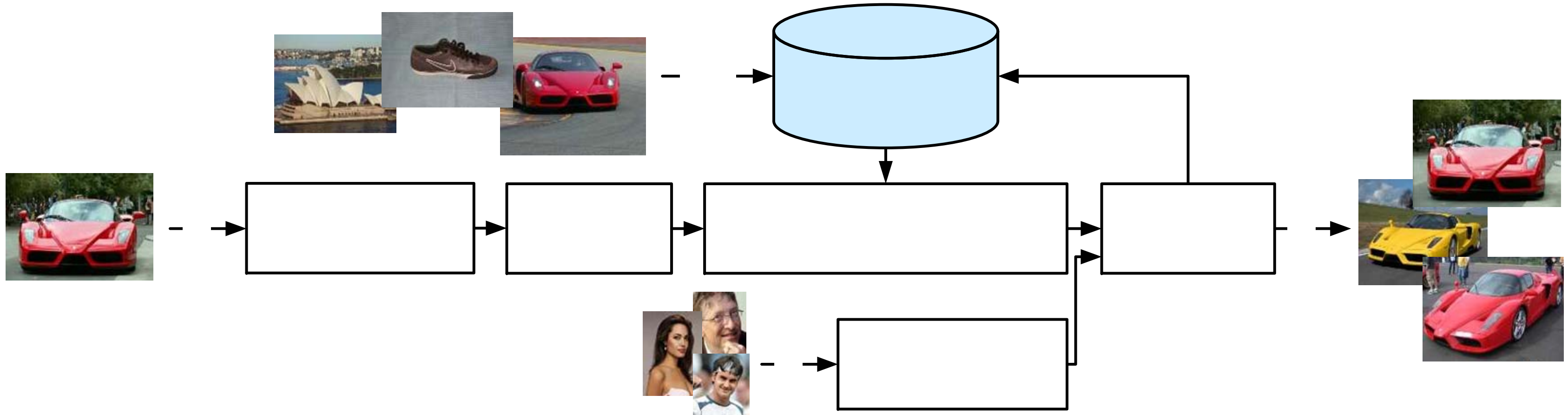
- ➢ In the context of reproducible research in the field of quality assessment:
  - ✓ We extended our publicly available database of impaired sequences and related subjective quality scores, by including 4CIF sequences.

- ➢ In the context of quality of experience in mobile scenario:
  - ✓ We we started our activities in the HAPQNET network of excellence, focused on the development of a quality of experience-based optimization strategy for bandwidth and energy consumption when considering video transmission for mobile devices.

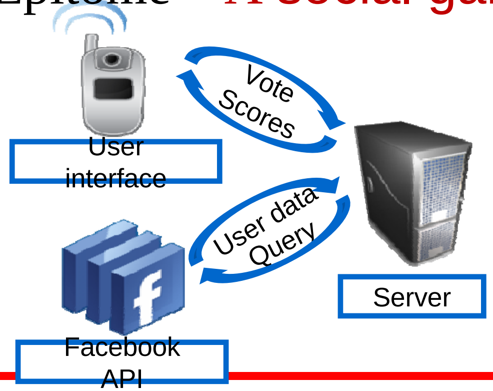
Multimodal Content Annotation

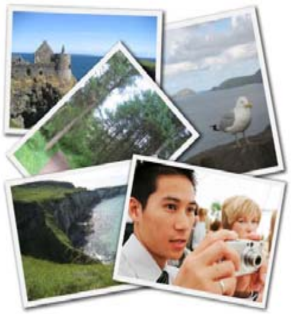
(EPFL MMSPG)

- Interactive online platform for semi-automatic object tagging in images
- Extension to geotagging considering user trust information



- Epitome – A social game for photo album summarization





IP1: Integrated Multimodal Processing

Collaborations within IM2:

- IP1 essentially driven by applications defined in IP2, both in terms of targets, evaluation and integration.
- Some of the speech and visual features detected by algorithms developed in IP1 will be exploited within IP2
- To ensure such a close collaboration across IP1 and IP2, two key representatives of IP2 attended the IP1 research meeting held at EPFL in April and June 2010.
- Other informal meetings across partners within IP1-IP2 have been organized in this first year to favor collaborations.
- Several groups declare activities in IP1 and IP2, demonstrating the fundamental and integrative aspects of their activities.