UNIVERSITAT POLITÈCNICA DE CATALUNYA

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room

Josep R. Casas
UPC – Image Processing Group

Doctorado en Ingeniería Informática y de Telecomunicación
Escuela Politécnica Superior – UAM
May 19th 2006

CHIL

UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

UPC Smart Room Team

Service, architecture, integration
- **Joachim Neumann**
- Jordi Salvador (Daniel Almendro, Shadi El-Hajj)

Video Technologies
- Cristian Cantón (Body & Gesture)
- **Josep R. Casas**
- Christian Ferran (Object Detection)
- Xavi Giró (Object Detection)
- José Luis Landabaso (Det/Track)
- Miriam León (Text Detection & OCR)
- Ferran Marqués (Face Det + ID)
- Ramon Morros (Face ID +Det)
- Montse Pardás (Activity & Emotion)
- Javier Ruiz (software APIs)
- Verónica Vilaplana (Face Det)

Audio Technologies
- Alberto Abad
- Mireia Farrus
- Javier Hernando
- Jordi Luque
- **Dušan Macho**
- **Climent Nadeu**
- Carlos Segura
- Andrey Temko

NLP Technologies
- Pere Comas
- Maria Fuentes
- Edgar González
- Mihai Surdeanu
- **Jordi Turmo**

CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

Outline

- Framework: CHIL
  - Vision, target, services
- Functionalities → Technologies
  - Multimodal interface technologies
- Smart Room and Sensor Setup
  - Data collection (Evaluation campaigns)
- Software Architecture
  - Data flows and distributed processing (CHIL ICE cube)
- Vision Technologies at UPC
  - Person tracking, Person ID, Body Analysis, Object Detection, Text Detection, Activity Analysis, Emotion Detection
- Conclusion & Discussion

Technology Transfer? Integration?

- Research Institutes & Companies
  - Researchers and Engineers
  - Scientific Papers vs. Products (Market/innovation? Patents?)
- Institutional Initiatives
  - FP6/IPs, CENIT, Profit…
    → target: integration/technology transfer
- Attitudes

"We perform high-level, forward looking, long term research…"

"This is not good for my PhD…"

"This is long term research, and will never be useful for a product in the market" (company)

"I'm sure someone will find it useful for something…" (researcher)

→ Researchers (in Engineering) should envision actual applications…

Framework: CHIL project

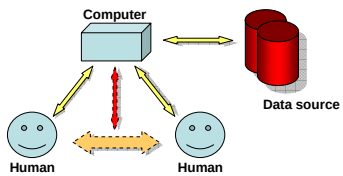
"Computers in Human Interaction Loop"

Instead of involving humans in the workflow and programmatic tasks defined and scheduled by machines (explicit operation, keying-in commands…)

The vision

→ ***put the Computers in the Loop of humans***

observing humans,
engaging and interacting with humans,
predicting and proactively providing services,
acting on perceived human need,
intruding as little as possible
*(hovering in the background as **electronic butlers**)*



The target

Computer services supporting Human-Human interaction

UPC Smart Room

Framework: CHIL services

- Provide computing services implicitly
  - … by putting Computers in the Interaction Loop of Humans
  - … by observing humans interacting with humans
  - … by predicting needs and proactively providing services
- CHIL services instantiated as demonstration prototypes
  - Connector
    Helps people to get in touch (avoids phone tag).
    It connects people at the **right time** through the **right device**.
  - Memory Jog
    Reminds you of things.
    It provides **pertinent information** at the **right time** (proactive/reactive, unobtrusive)
  - Socially Supportive Workspace
    Helps people to work together.
    It is a **Smart Table**, on which virtual paper is used to increase efficiency in group decisions
  - Relational Cockpit
    Analysis of group behavior to improve productivity

UPC Smart Room

Framework: CHIL contributions

- Expected societal outcome
  - Reduce preoccupation with technological artifact (techno-clutter)
  - Improve productivity by use of human context
  - Improve human experience
- Expected scientific outcome
  - **Perception:** Full description & understanding of all human communication signals across multiple modalities (audio, image, speech, language, signs…)
    - → **Functionalities: who, where, what (in/out), how, why…**
    - Robustness in perceptual user interfaces (always on)
  - **Synthesis:** from human-friendly to human-like interfaces
    - → **Functionalities: situation models, strategy, proactivity, politeness, privacy care…**
    - Progress in output interfaces and actuators
- European Project (6th FP / IST) http://chil.server.de
  - 2004 → 2006 → 2010 (2nd phase)
  - 25 M€ (1st phase)
  - Involves 15 partners from 9 countries
    Germany, France, Netherlands, Sweden, Italy, Check Rep, Greece, Spain, US

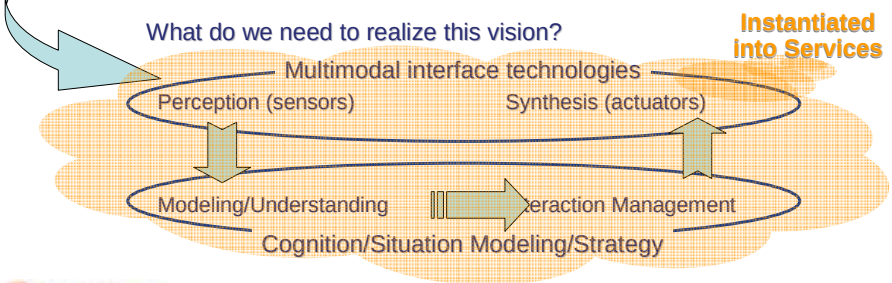
Perception to Action — … Realizing the CHIL Vision

CHIL vision

→ ***put the Computers in the Loop of humans***

observing humans,
engaging and interacting with humans,
predicting and proactively providing services,
acting on perceived human need,
intruding as little as possible

*(hovering in the background as **electronic butlers**)*



Technology framework

- Hypothesis
  "Multimodal interface technologies mature enough ***to get computers listening, watching, talking, helping…****"*
  → New generation of computer services
- Technology areas
  - Perception from sensors → who, where, what, how, why…
  - Modeling/Understanding → predict, interpret situation
  - Managing Interaction → proactive/reactive, natural, friendly, polite, privacy
  - Synthesis from actuators → audio, video, calls, signs, text
  - Software Architecture → integration, interoperation

  → Specific challenges in audio-visual technologies
- Scientific outcome: ***"Technology Push"***
  - Objective measures of progress & efficiency through open/well-defined technology evaluations

  → Technology catalogue
  - User studies and User evaluations

Multimodal Interface Technologies

Perception: from Sensors to Semantics

- Audio Signals – multiple microphones
  Enabling tech: Speech Activity Det (SAD)
  - Speaker Localization (SLOC)
  - Speaker Identification (Speaker ID)
    Combined: Speaker Tracking
  - Speech Recognition (ASR)
  - Acoustic Events (AEC)
    Enabling tech: Beamforming (e.g. for ASR)
- Video Signals – multiple cameras
  Enabling tech: Foreground Detection
  - Person Location & Tracking (PLT)
    Enabling tech: Face Detection
  - Face Identification (Face ID)
    Combined: ID tracking
  - Head-Pose Detection/Orientation
- Other (e.g. text) – multiple sources
  - Summarization, Question&Answering
- Multimodality (MM)
  - MM Location &Tracking (speaking/not)
  - MM Identification (Visible/speaking)
  - MM Head-Pose
  - MM Events
  - MM Activity
- Higher level analysis
  - Topic Detection
  - Attitude/Emotion Detection
  - Gesture Analysis
  - Group Activity Analysis
- Semantics
  - Situation Modeling

  … Ontology's for concepts and situations

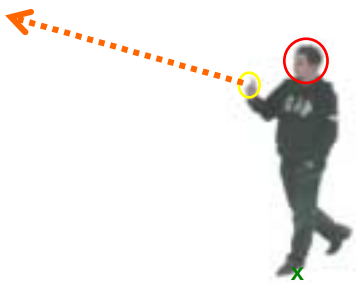
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Describing Human Activities

UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

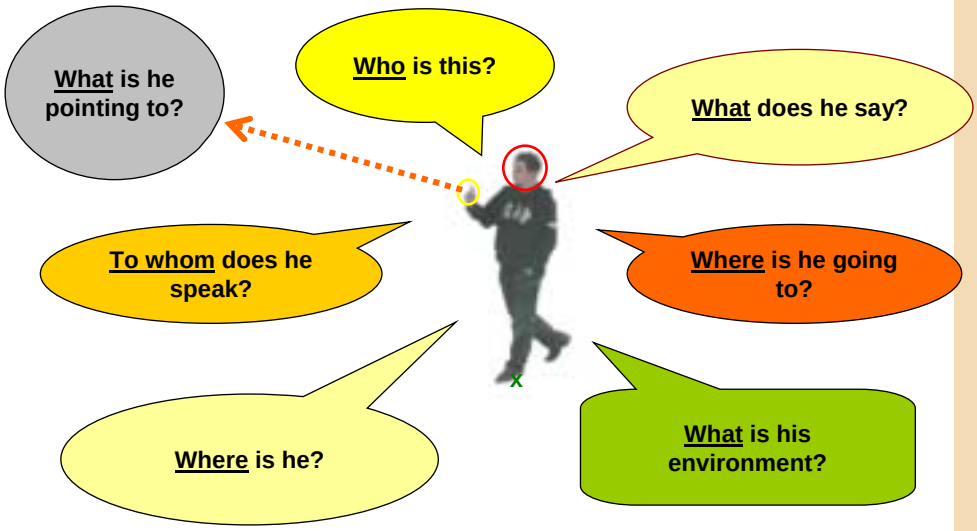
Describing Human Activities



CHIL

UPC Smart Room

Technologies/Functionalities



CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

Audio-Visual Perception

- Challenges for the "Who", "Where" (1st tier technologies)
  - Tracking people in natural, evolving, unconstrained scenarios
  - Persons behave without constraints, unaware of audio/video sensors
  - Location and tracking
    - Visual – background subtraction: error-prone (shadows, occlusion), feature based (e.g. color): difficult to initialize (color histogram)
    - Audio – high reverberation times (seminars & meeting rooms), impossible to rely on a direct path to microphones
  - Identification technologies
    - Audio – far field (noise, overlap)
    - Visual – wide angle (low-res), occlusions
    - A + V – unconstrained motion of the people, no assumptions on position/orientation to facilitate well-posed signals (frontal faces or speakers aiming at sensors)

CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

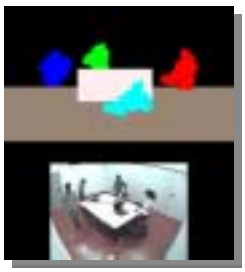
Facing challenges (I)

- Sensor fusion: Multi-view
  Probabilistic approach: product of single view likelihoods, generative model

5 targets, 1 cam, 3-10Hz

(ITC-irst) O. Lanz, "*Approximate Bayesian Multibody Tracking*," IEEE Trans. PAMI (accepted)

- Sensor fusion: 3D
  Background subtraction and shape from silhouette



(UPC) J.L. Landabaso, M. Pardas, *"Extraction of foreground regions towards real-time object tracking,"* MLMI 2005

CHIL

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Facing challenges (II)

- Feature fusion: Multi-modal
  AV speaker localization (fusion with particle filtering)

3D Head Position
Speaker Tracking
Frontal Face Detection
Side Face Detection
Upper Body Detection
Particles
Speaker Location

(UKA ISL) M. Wölfel, K. Nickel, J. McDonough, MLMI 2005

CHIL

Audio-Visual Perception

- Challenges for the "What" (2nd/3rd tier technologies)
  - Speech Recognition for continuous large vocabulary conversational speech, overlapped, competing acoustic events
  - Automatic Speech Recognition (ASR / AVASR)
    - Audio – far field, partly compensated with beamforming (subject to localization/tracking performance)
    - Audio – non-native English speakers
    - A + V – all the previous challenges for localization & ID
  - Summarization
    - (technology initially designed to work from written text input)
    - Unstructured textual input provided from transcriptions (ASR)

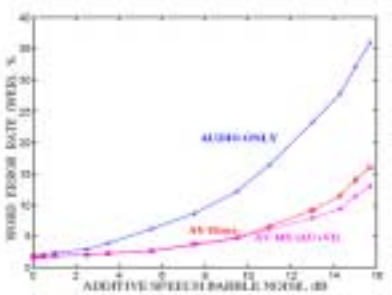
CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

Facing challenges (III)

- Multi-modal feature fusion
  Audio Visual Speech Recognition (speech +facial features – "lip reading")



(IBM, UKA) G. Potamianos et al, *"CHIL D5.2 Baseline System for Far-Field Audio-Visual Automatic Speech Recognition,"* 2005

CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

Facing challenges (IV)

- Multi-modal feature fusion +

  Microphone Array Driven Speech Recognition:

  Far field sensors (ubiquitous computers)

  → natural use of beamforming from a microphone array

Influence of Accurate Localization on the Word Error Rate

Tracking mode	WER
Close Talking Microphone	34.0%
Microphone Array	
single microphone	66.5%
estimated position (Audio only)	59.8%
estimated position (Video only)	59.1%
estimated position (Audio & Video)	58.4%
labeled position	55.8%

graph

(UKA ISL) M. Wölfel, K. Nickel, J. McDonough, MLMI 2005

Audio Visual Synthesis

- Synthesis (actuators)
  - Targeted Audio
  - Targeted Video

(DaimlerChrysler) D.Olszewski, K.Linhard, "*D5.6 Soundbox with steering capability*," 2005

(UKA ISL) Steerable beamer + camera

Animated talking agent

→ Managing interaction

Synthesis samples, expressive speech (KTH)

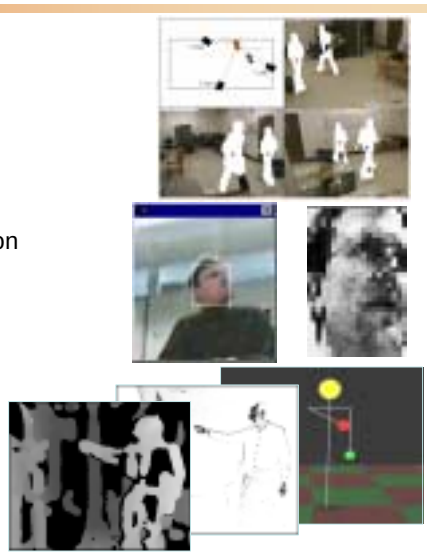
CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

CHIL Technologies at UPC

- Vision
  - Object/Body detection & tracking
  - Face Detection & ID
  - Body & Gesture Analysis
  - Object Detection & Analysis
  - Text Detection & Video OCR
  - Activity Analysis & Emotion Detection
- Speech
  - Speaker ID
  - Acoustic Source Localization
  - Acoustic Event Classification
  - Speech Activity Detection
- Natural Language Processing
  - Question Answering
  - Summarization



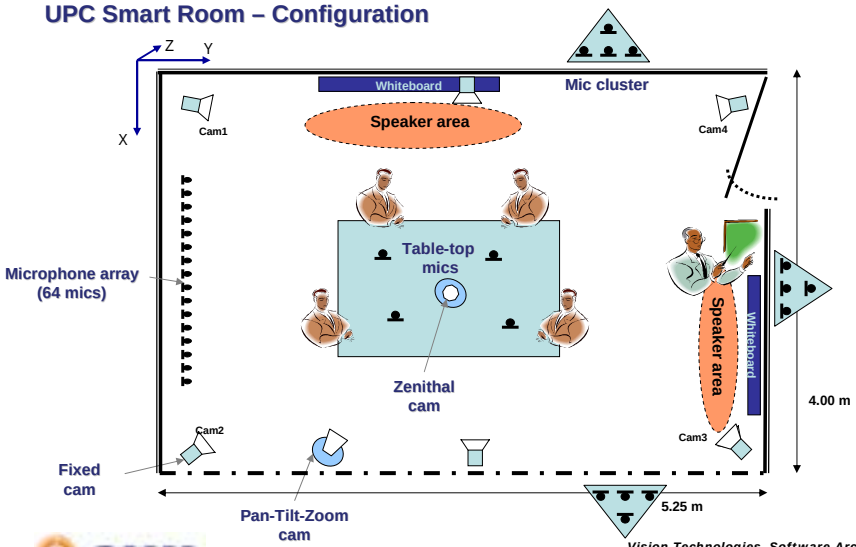
CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

Smart Room & Sensor Setup

UPC Smart Room – Configuration



UPC Smart Room

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room – Configuration

UPC Smart Room

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room – Video Equipment

Cameras: JVC TCK - 1481EG

- 25 fps, 768x576, interlaced, genlocked
- Frame Grabbers: Viewcast Osprey-210

Function & Lenses

- 4 Monitoring Cameras: 4 corners, wide angle lenses
  Computar HG2Z4516FCS-2: 1/2", 4.5-10mm (38º-81º)
- 1 Zenithal Camera: ceiling mounted, fish eye
  Fujinon lenses DV2.2x1.4,5SA2: 1/3", 1.4-3.1mm (84º-126º)
- 2 Person Cameras: mid walls, head & shoulders views
- 1 Active Camera PTZ:
  VideoTec PTH300, Pentax H6ZBME

Other

- Sync master: MOTU Digital Time Piece (genlock, Timecode labels for A/V)
- Video Selector MOXIE SVA-801: Real-time monitoring
- A/V distributor ELPRO (genlock signal, LTC)
- Ad-hoc Software for recording control

UPC Smart Room

UPC Smart Room – Audio Equipment

64-channel microphone array: NIST Mark III

- 64 ch. sample synchronized, 44.1 kHz
- Ethernet connection to acquisition computer

"T-shaped" microphone clusters

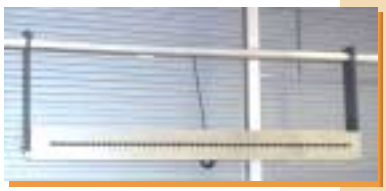
- 3x4 ch. sample synchronized, 44.1 kHz
- Hammerfall acquisition system

Close Talking and Table microphones

- "Invisible" close-talking mikes
  Countryman (wireless)
- Omni-directional table mikes
- Directional table mikes
- Hammerfall acquisition system

Other

- Hammerfall RME HDSP 9652 24 ch. sound-card
- OctaMic-D preamplifiers



UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

UPC Smart Room – Data Collection Scenario

Interactive seminar in small group

- Slightly scripted but natural
- Focus on interaction:
  - people in/out, latecomer
  - question interrupting talk
  - acoustic events (door, steps, coughs, keys, KB typing)
  - visual events (greetings, gestures, hand in face…)
  - coffee break (steps, laughs, phone ring, liquid pouring
  - question time

CHIL

UNIVERSITAT POLITÈCNICA DE CATALUNYA

UPC Smart Room – Camera images

CHIL

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Data collection – Presentation

Presentation starts
10” (cam1)

Latecomer enters
21” (Zcam5)

CHIL

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Data collection – Coffee break

Start coffee
25” (Zcam5)

Free chat, phone ring, laughs
30” (cam2)

CHIL

Data collection – CHIL Evaluation Campaigns

- Evaluations are Key to Assessing and Driving Progress
  - Benchmarks, Measures of Performance (MOPs)
  - User Studies, Measures of Effectiveness (MOEs)
  - Cooperation + Competition = ***"coopetition"***
- Functionalities & Technologies
  - Working Group in Each Area
  - Define Metrics, Databases and Benchmarks
  - Performance Benchmark Evaluations in Each Area
- Evaluation Campaigns
  - First "Dry-Run": completed June 2004
  - Year One: Completed January 2005 (Open to external sites)
  - Year Two: Completed March/April 2006 (Coordination with NIST)
    - **CLEAR 2006** http://www.clear-evaluation.org
    - **RT 2006** http://www.nist.gov/speech/tests/rt
  - Future:
    - **CLEAR 2007, RT07**
    - **CLEF** http://www.clef-campaign.org (Question/Answering)

UPC Smart Room – Software Infrastructure

- CHIL Distributed Processing Architecture
  - Provides
    - a programming framework,
    - programming tools and
    - programming environments
    - to build and evaluate CHIL services

    Rapid prototyping (to explore successful services)

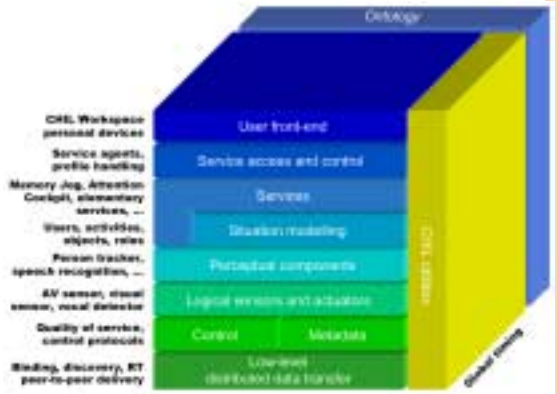
    Breadboard (agent-based): rapid insights & intermediary designs
  - Common reference for integrating multi-modal perceptual components to construct CHIL Services
    - Data flows exchange → NIST Smart Flow
    - Higher level modules → CHIL ICE cube

CHIL Architecture

- Quality Requirements
  - reliability
  - maintainability
  - portability
  - reusability
  - usability
  - efficiency

➔ Layered Architecture Model

- The ICE "Cube"
  Integrated Chil Exoskeleton

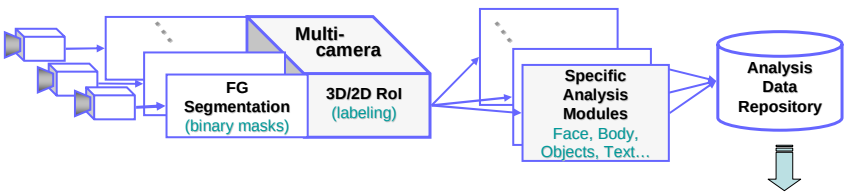


UPC Smart Room

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

Software Architecture for Perceptual Components

- Evolution: analysis DB → Flows
  - Initial proposal: low level architecture for CHIL analysis modules



All modules access the DB (XML vs SQL)
Flexible (queries)
DB access issues for real-time…

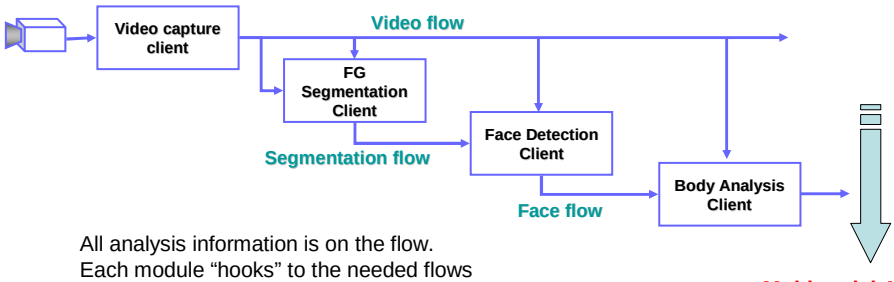
Multimodal Analysis Services

UPC Smart Room

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

Software Architecture for Perceptual Comps: Smartflow

- Evolution: analysis DB → Flows
  - Current proposal: completely SmartFlow based



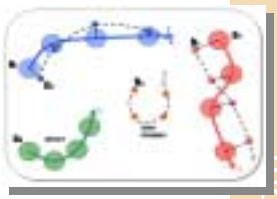
All analysis information is on the flow.
Each module "hooks" to the needed flows
Not flexible (flows must be defined at design time)
Real time
No common memory (each module stores any information needed)

UPC – Video Technologies

- General Object/Body detection & tracking
  José Luis Landabaso
- Face Detection & ID
  Verónica Vilaplana, Ramon Morros, Ferran Marqués
- Body & Gesture Analysis / Head Pose
  Cristian Canton
- Object Detection & Analysis
  Christian Ferran, Xavier Giró
- Text Detection & Video OCR
  Miriam León, Antoni Gasull
- Activity & Emotion Analysis
  José Luis Landabaso, Montse Pardàs

Localization & Tracking

- Motivation / Goal
  - Continuous monitoring of scene: ***"who-where"*** from all available sensors (A/V)
  - Support higher level tasks: ID, Head Pose, Activity Classification…
  - Fundamental for services: situation model, targeted audio/video… → elementary component for ***context awareness***
- Task definition
  - ***Locate*** people in scene
    - Single Person (speaker) / Multiple Person (everyone)
  - ***Track*** people positions in time (correspondence problem)
  - Input from 4 cameras (+zenithal) several microphones



- Metrics
  - MOTP: Multiple Object Tracking Precision
    → considers ***distance*** errors
  - MOTA: Multiple Object Tracking Accuracy
    → considers tracking ***correspondence errors*** over time (misses, false positives, mismatches)
  - Other metrics for reference/comparison (e.g. SLOC for Acoustic tracking)

UPC – Video Object/Body detection & tracking

L.-Q. Xu, J.L. Landabaso, M. Pardàs, *"Shadow Removal with Blob-based Morphological Reconstruction for Error Correction"*, ICASSP 2005, Philadelphia, USA

- Shape-from-silhouette (classic approach)

  Foreground camera points define rays in scene space intersecting object at some unknown depth. Union of visual rays for all points in silhouette defines a generalized cone within which the 3D object must lie

- Contribution: Cooperative Background Modeling

  Background models in each view are cooperatively learnt, using evidence from all cameras, in a Bayesian framework
  - Advantages
    - Better 2D foreground regions extracted
    - More accurate 3D foreground volumetric models
- 3D Location and tracking

  Spatially connected foreground voxels are grouped and tracking is done for 3D blobs

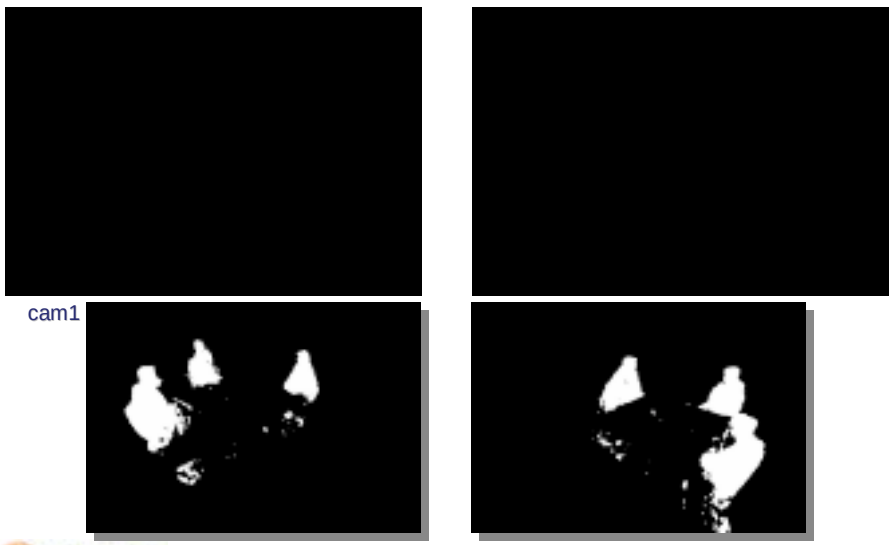
UNIVERSITAT POLITÈCNICA DE CATALUNYA

UPC – Video
Body detection & tracking Results
(showing probabilistic projections)

UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

UPC – Video
Body detection & tracking Results



UPC Smart Room

CHIL

UPC – Video Face Detection

F. Marqués, V. Vilaplana. *"Face segmentation and tracking based on connected operators and partition projection"*. Pattern Recognition, 35(3):601-614, 2002

Face Detection

- Low resolution images, small faces: use only color, size & shape descriptors, don't use texture
  - **Color**
    Constant color model in the (Cr,Cb) subspace, skin color modeled with a Gaussian distribution
  - **Shape**
    Aspect ratio of bounding box of region
    Haussdorf distance (between region contour and a face shape model)
- Exploiting temporal information:
  - For mask correction (to detect faces when the body tracking fails)
  - For face model adaptation (color and shape)

UPC Smart Room

UPC – Video Person ID

Face Recognition

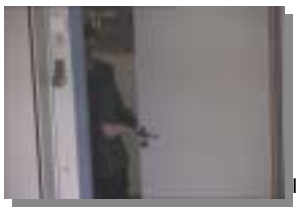
- Two different aspects:
  - Intra-session and Inter-session identification
  - Model updating
- Intra-session identification:
  - Lower variability: Principal Component Analysis
- Inter-session identification:
  - Higher variability: Bayesian Face Recognition
- Model updating:
  - A set of images is used to model every class

UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Person ID

- Motivation / Goal
  - ***"who-is-who"***: Identify people in multi-camera, multi-microphone far-field
- Task definition
  - Audio-only, video-only and audiovisual
  - Data:
    - Far field / low res: NIST Mark III 64 ch, 4 corner cameras (above eye level)
    - Varying conditions: different sites
      - Visual: Room size, lighting, BG clutter/occlusion, camera models
      - Audio: Different accents, distances from sensors
    - Data Base: ***26 individuals***
- Metrics
  - Percentage of wrong IDs
    - per training duration (15 or 30 sec)
    - per testing duration (1, 5, 10 and 20 sec)
- Last year status
  - Data from one site, 11/12 individuals/speakers
  - Different conditions (difficult comparison), not evaluated



ID

CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Person ID – Cont.

- Audio / Video / Audio-Visual systems evaluated (CLEAR)

	AIT	UKA	LIMSI	UPC
A – Speaker ID				
Model / other processing	MEL+D, 16G, deterministic EM / per Speaker PCA	128G (30s), 32G (15s) / Dereverb, feature warping	UBM, MEL+D+DD, 256G / Low energy filtering, feature warping	Frequency filtering +D +DD / None
V – Face ID				
Classifier / Face extraction Fusion	PCA-LDA / 200ms labels +interp+norm / Across time then classifier	Local Appearance-based using DCT / 200ms labs +norm / Cameras then time	–	PCA / 1s labels / Interpolation / Time
AV – Person ID				
Fusion / Modality trusted / Weights norm	Post-decision / slightly Audio / A/V same median for 1s tests	Post-decision / None / Sigmoid	–	Post-decision / Audio w/scores / z-score, histogram equalization

CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Person ID – Cont.

- Monomodal results for Person ID (CLEAR)
  (Percentage of wrong detections)

A–Speaker ID	15 sec training				30 sec training			
Test duration	1 s	5 s	10 s	20 s	1 s	5 s	10 s	20 s
AIT	26.92	9.73	7.96	4.49	15.17	2.68	1.73	0.56
CMU	**23.65**	**7.79**	7.27	3.93	**14.36**	**2.19**	**1.38**	**0.00**
LIMSI	51.71	10.95	**6.57**	**3.37**	38.83	5.84	2.08	**0.00**
UPC	24.96	10.71	10.73	11.80	15.99	2.92	3.81	2.81

V–Face ID	15 sec training				30 sec training			
Test duration	1 s	5 s	10 s	20 s	1 s	5 s	10 s	20 s
AIT	50.57	**29.68**	**23.18**	**20.22**	47.31	31.14	26.64	24.72
UKA	**46.82**	33.58	28.03	23.03	**40.13**	**23.11**	**20.42**	**16.29**
UPC	79.77	78.59	77.51	76.40	80.42	77.13	74.39	73.03

CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Person ID – Cont.

- Multimodal results for Person ID (CLEAR)

AV–Person ID	15 sec training				30 sec training			
Test duration	1 s	5 s	10 s	20 s	1 s	5 s	10 s	20 s
AIT	23.65	**6.81**	6.57	**2.81**	13.70	**2.19**	**1.73**	**0.56**
UKA / CMU	43.07	29.20	23.88	20.22	35.73	19.71	16.61	12.36
UPC	**23.16**	8.03	**5.88**	3.93	**13.38**	2.92	2.08	1.12
AIT	23.65	**6.81**	6.57	**2.81**	13.70	**2.19**	**1.73**	**0.56**

CHIL

Vision Technologies, Software Architecture & Processing Strategy in the UPC Smart Room
EPS-UAM, May 19th 2006

UPC Smart Room

Person ID – Cont.

- Progress
  - Hard to assess: changed evaluation conditions
    - 1 → 5 sites, 11 → 26 individuals
    - Face ID decoupled from face detection/tracking,
    - 15 sec training sequences instead of 5 training images
  - ***Audio***
    - For 30 sec training / 5 sec test, error rate dropped ***from 6.86% to 2.19%*** (CMU)
  - ***Video***
    - For 10 sec test, ***30%*** (AIT/UKA) improved to ***23%*** (AIT)
  - ***Multimodal***
    - Audio helps video when speaker is present
- Remarks
  - Far-field, unconstrained poses affect ***video performance*** greatly
  - ***Audio*** can be trusted, especially for long duration
  - When audio is present, video seems complementary
  - When audio is not present?
- Prospects
  - Check evaluation conditions to bring them closer to its use for services
  - Explore further fusion possibilities (multisensor/multimodality/integration)

UPC – Video Body & Gesture Analysis

C. Canton-Ferrer, J.R. Casas, M. Tekalp, M. Pardàs, *"Projective Kalman Filter: Multiocular Tracking of 3D Locations Towards Scene Understanding"*, MLMI2005, Edinburgh, July 2005

- Multiview video analysis to extract body pose and limbs position for gesture and scene understanding
  - Hierarchical human body model: geometry for analysis

Stick body model

↓

Gesture analysis, 3D tracking over multiple cameras,..).

Head Pose Estimation

- Motivation / Goal
  - "***who**'s looking **where***": estimate Pan and Tilt rotation of head position
- Task definition
  - ***Studio*** data: synthetic, high resolution captures of head rotations
  - ***Seminar*** data (NEW): CHIL seminar recordings with low resolution
- Several metrics
  - Pan / Tilt Mean Error [°]
  - Pan / Tilt Correct Classification [%]
    - Pan: -90°, -75°, -60°, …, +75°, +90°
    - Tilt: -90°, -60°, …, +60°, +90°
  - Pan Correct Classification within neighbour range [%]

New task: Acoustic (!) IRST demo on Speaker Loc + head orientation

Head Pose Estimation – Cont

- Status and progress on ***Studio data***
  - Mean error (pan/tilt): 12° / 15° → **12° / 10°**
  - Pan Correct classification: 45% → 51%
  - Tilt Correct classification: 43% → 50%
- Main progress this year: CLEAR Results on ***CHIL seminar data***
  (estimates with respect to the room coordinates)
  - Mean error (pan): **49º** (**34º** best system)
  - Pan Correct classification: 35% (45% best system)
- Prospects
  - ***Low resolution*** captures opens a complete new field for head pose estimation
  - Information from ***multiple views*** helpful AND necessary for stabilizing / confirming hypotheses
  - Classifiers and feature spaces used for high-resolution pose estimation are not feasible as standalone systems anymore → multimodal fusion approaches: body posture, tracking, speech detection, …

New Tasks: acoustic & multimodal head orientation (currently pilot experiments)

D4.7 *"3D tracking of several persons from multiple camera views. Head Orientation tracking"*

UPC – Video Object Detection

F. Marqués, M. Pardàs, O. Salerno, V. Vilaplana, *"Object recognition based on binary partition trees"*, ICIP2004, Singapur, October 2004

- Objects such as
  - Electronic devices: Laptop, PDA, mobile phones, etc.
  - Smart room objects: chairs, cups, bottles, etc.
- Features such as
  - Position, orientation, on/off, open/close, etc.
  - Owner, connected to, User, etc.
- Algorithms
  - Syntactic segmentation: Geometric & structural criteria for BPT creation
  - Description Graphs Detection: object modelled as a set of simpler semantic classes that satisfy certain structural relations

UPC – Video Activity detection

J.L. Landabaso, M. Pardàs, L.-Q. Xu, *"Hierarchical Representation of Scenes using Activity Information"*, ICASSP 2005, Philadelphia, USA

- Room activity analyzed using Stochastic Context Free Grammars
  - A set of rules are manually defined. Parsing is performed over series of events to effectively detect specific activities (in particular, static objects, moved chairs, etc.
  - High-Level information is not only seen as the aiming target, but also as a way to reinforce the basic Low-Level Tracking
- Background Modeling Using Video Understanding:

  Adaptive background modeling techniques usually fail under certain conditions.

  Suppose a person hovering in the background, which then stops, sits, or lays. During a period of time, the corresponding blob will still be active, but, little by little, the pixels of the blob will become part of the background.
  - The process of merging into the background, could be prevented once we positively know that the object has stopped. The instants when the objects stop could be determined by video understanding techniques.

Interesting outcome…

- Change in researchers' attitude

Previously: *"This recording/situation/scenario is not good because…"*

… the presenter gets out of the camera view
… the speaker does not talk to the microphone
… participants don't look at the cameras
… bad lighting/shadows, strong reverberation…

Now: *"Er… Well, we'll have to adapt… This is challenging"*

… cameras should cover the whole area
… ID profile views (challenging)
… cancel noise (classify acoustic events)
… far field, reverberation, wide views, noise, low res, shadows, lights,

→ *what if we combine? (signal data, features, scores, decision…)*
– exciting anticipation of the challenge –

→ Promising for Robust Perceptual Interfaces

Conclusion

- Framework: the CHIL project
  - The CHIL Vision → Computers **should** help, in a "naturally human" way
  - Proof of concept: instantiating services (demo)
- Multimodal Interface Technologies aim to fulfill the CHIL vision
  "Putting the **Computers in the Loop of Humans**" → instantiated in services
  - Robust technologies to understand human communication signals across multiple modalities, in natural, varying, unconstrained human interaction scenarios
- Facing challenges
  - Attitude change… Progress
  - Fusion: Multi-sensor, Multi-modal

→ Outcome: ***"Technology Push"***

- Smart Room and Sensor Setup
  - Equipment and Data collection
  - CHIL evaluation campaigns
- Software Architecture
  - Data flows and distributed processing (CHIL ICE cube)
- Vision Technologies at UPC
  - Person tracking, Person ID, Body Analysis, Object Detection, Text Detection, Activity Analysis, Emotion Detection
  - Most techniques published or to be published in 2004/2006 (http://chil.server.de → bibliography)

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Thanks for your attention!

Questions?

CHIL

UPC Smart Room