Hes·so
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences Western Switzerland

MSE | MASTER OF SCIENCE IN ENGINEERING

Multimodal Processing, Recognition and Interaction

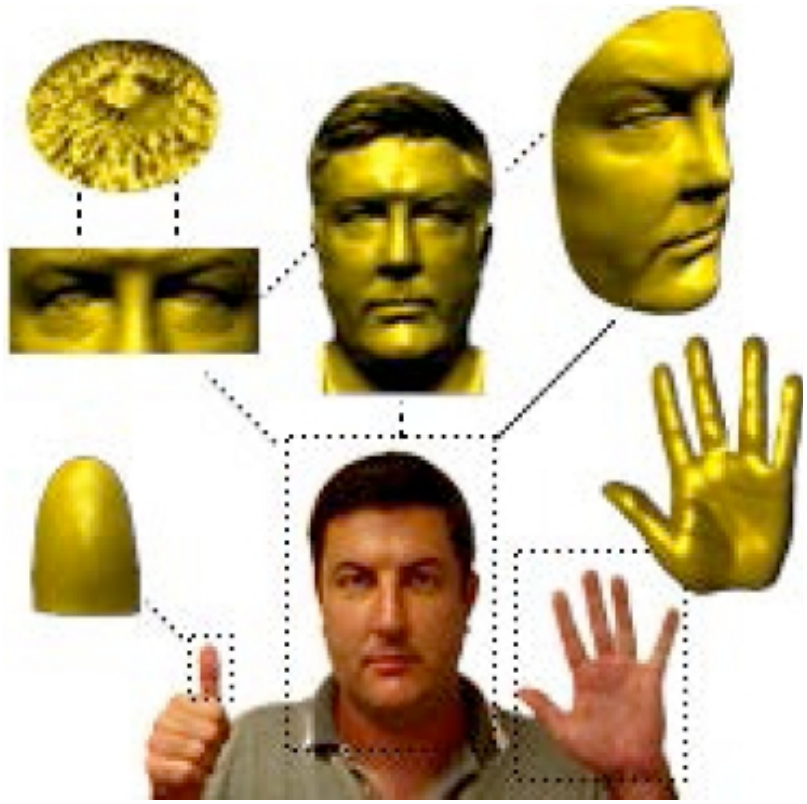
Multimodal fusion

Elena Mugellini, Stefano Carrino

Agenda

- Introduction to multimodal systems
  - Defining multimodality
  - Advantages and drawbacks of multimodal analysis
- CASE and CARE models
- Fusion models
  - pre and post classification fusion
  - fusion methods
- Conclusion

Introduction to multimodal systems



Defining Multimodality - I

- "*Multimodal interfaces process **two or more combined user input** modes (such as speech, pen, touch, manual gesture, gaze, and head and body movements) in a **coordinated** manner with multimedia system output. They are a new class of interfaces that aim to **recognise naturally occurring forms of human language** and behaviour, and which incorporate one or more recognition-based technologies (e.g. speech, pen, vision)*"

  S. Oviatt et al., 2002

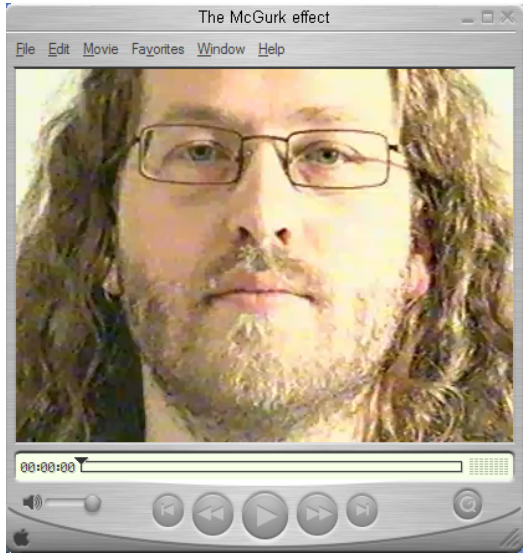
- Modality
  - Natural way of interaction
    - speech, vision, face expression, gesture, head and body movements, etc.
    - multimodal speech recognition (voice recognition + lip mouvement), multimodal biometric system (face recognition + voice recognition), multimodal emotion recognition, etc.

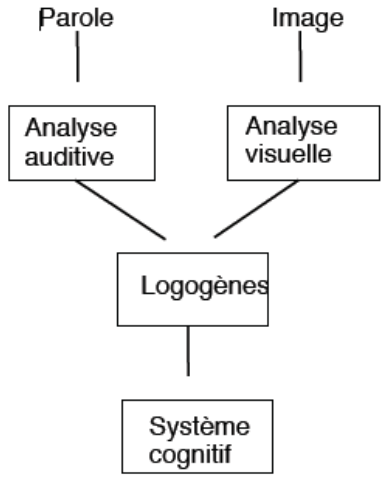
Defining Multimodality - II

- **Modality**
  - in HCI, a modality is a **natural way of interaction**
  - for psychologists, sensory modalities represent the **human senses** (sight, hearing, touch, etc.)
  - in signal processing, modalities are **signals** originating from the same physical sources but captured through different means
    - the number of modalities may not be same as the number of information streams
      - e.g. video modality offers information about the identity of the speaker (face analysis), his emotions (facial expression), what he is saying (lip-reading)
- Combine naturally complementary modalities in a manner that optimises the individual strengths of each, while simultaneously overcoming their weakness

Human fusion

- McGurk effect
  - image of “ga” + speech “ba”
  - perception of “da”
- Logogenes of Morton
- Action-perception loop and control





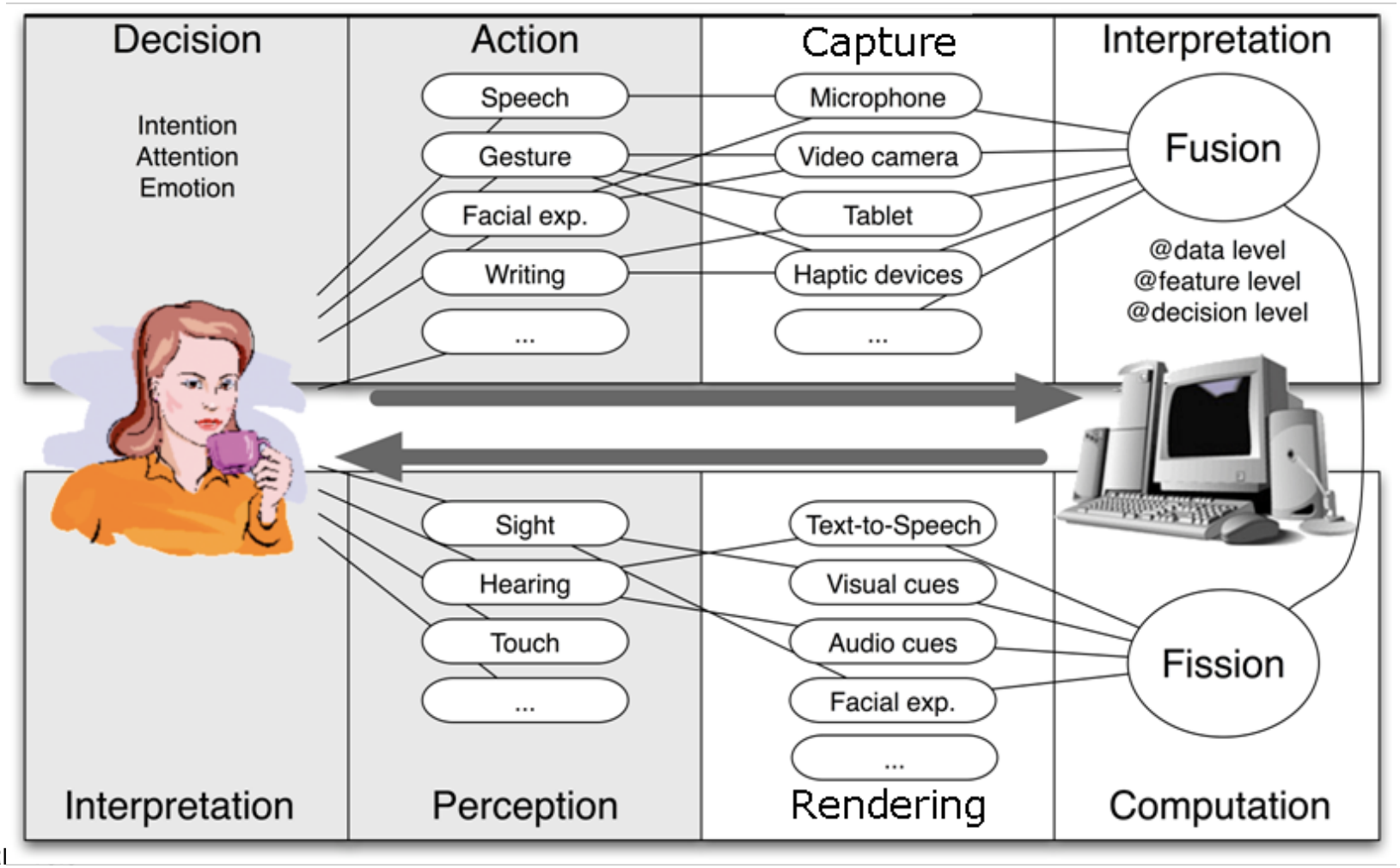
Multimodal systems - I

- Growth of interactive multimodal systems
  - combined natural human input modes (speech, pen, gaze, touch, gesture, etc.)
  - coordinated system multimedia output (audio, video, haptic, etc.)
- Put-that-there - Bolt
  - seminal demo system (1979)

Multimodal systems - II

- Advantages
  - **Accuracy**, complementarity of information
  - **Reliability**, mutual disambiguation of recognition errors [Oviatt, 1999a]
    - improved error handling & efficiency
      - 36% fewer errors, 10% faster task competition
  - **Usability**
    - greater expressive power
    - greater precision in visual-spatial tasks
    - support for users' preferred interaction style
    - accommodation of diverse users, tasks, and usage environments
      - e.g. accented speakers, mobile environments, etc.

Multimodal HCI

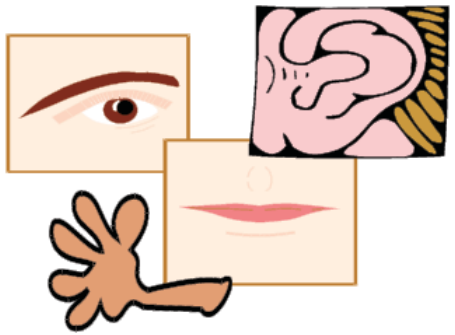


GUI vs MMI systems

- GUI systems
  - **single input** stream
  - atomic, **deterministic**
  - sequential processing
  - centralised architectures
- MMI systems
  - **multiple input** streams
  - continuous, **probabilistic**
  - parallel processing
  - distributed, time-sensitive architectures

Fundamental problems - I

- Adequation of tasks to modalities
- Adequation of usage to user profiles
- Fusion of inputs
  - Resolution of the co-reference: match the multimodal referents
- Fission of outputs
  - Resolution of the difference : activate the most suitable referent



Fundamental problems - II

- Challenges
  - extraction of **relevant features**
    - **relevant,** that contain information required to solve the problem
      - relevance is tied to the **context**!! (speech recognition vs speaker identification)
    - **compact,** feature vector with low dimensionality
      - avoid the **curse of dimensionality**
    - **balance!!**
  - **fusion** of the information
    - **integrating** the information coming from different modalities
      - different data rate, dimensionality, etc.
    - **dynamic** adaptation
    - **synchronisation**

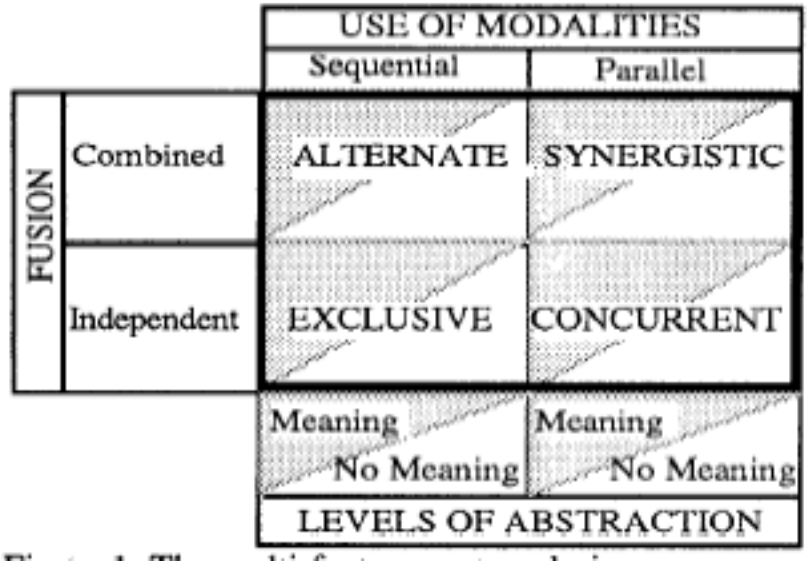
CASE and CARE models

CASE and CARE models

- Need to formalize human/machine multimodal interactions
- Conceptualize the different possible relationships between input and output modalities
- Two conceptual spaces:
  - Multimodal communication types
    - Machine-side (CASE)
  - Multimodal systems
    - Human-side (CARE) == Usability properties

CASE model: Multimodal System Communication Types

- 3 dimensions in the design space :
  - Levels of abstraction
  - Use of modalities
  - Fusion



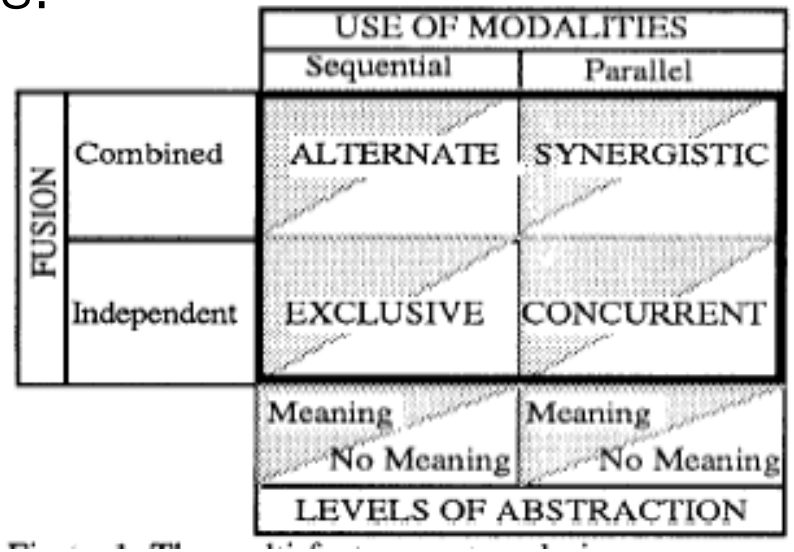
CASE model

- Use of modalities
  - Depends on temporal use
  - Parallel: multiple modalities employed simultaneously.
  - Or sequentially, one at a time
- Fusion
  - Possible combination of different types of data
  - Independent: absence of fusion, no coreference.
  - Combined: fusion necessary

		USE OF MODALITIES	
		Sequential	Parallel
FUSION OF MODALITIES	Combined	**A**LTERNATE	**S**YNERGISTIC
	Independant	**E**XCLUSIVE	**C**ONCURRENT

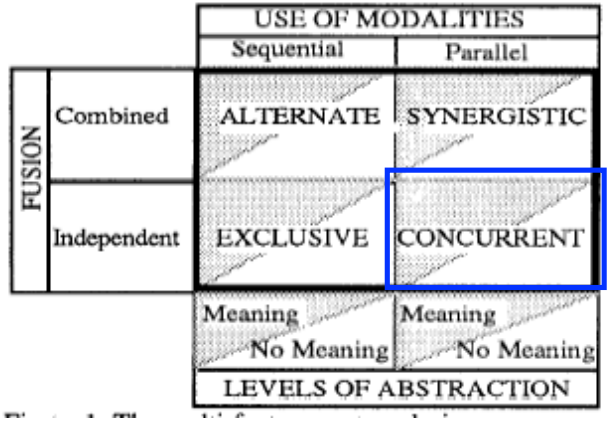
CASE model

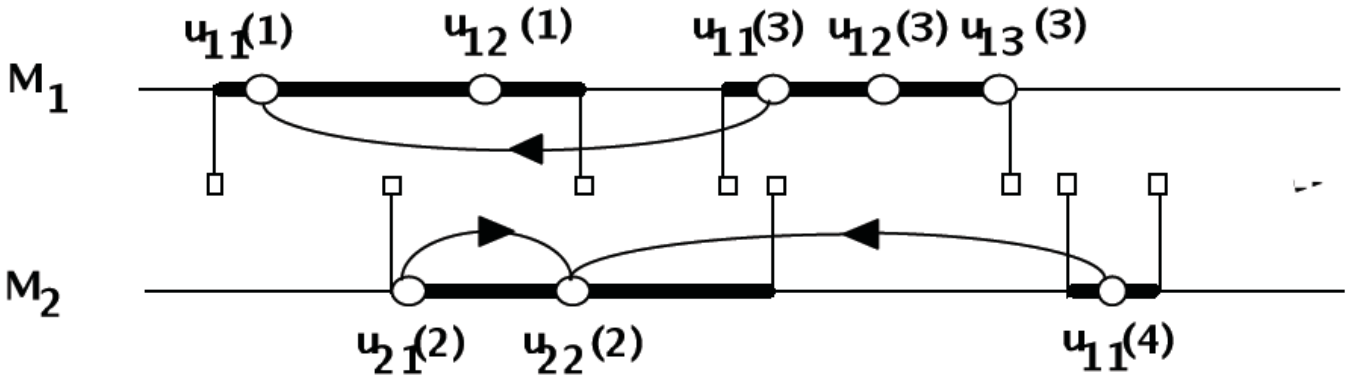
- Levels of abstraction
  - Data received from a device can be processed at multiple levels of abstraction;
- Exemple with speech analysis:
  - Signal level
  - Phonetic level
  - Semantic level



CASE - Concurrent

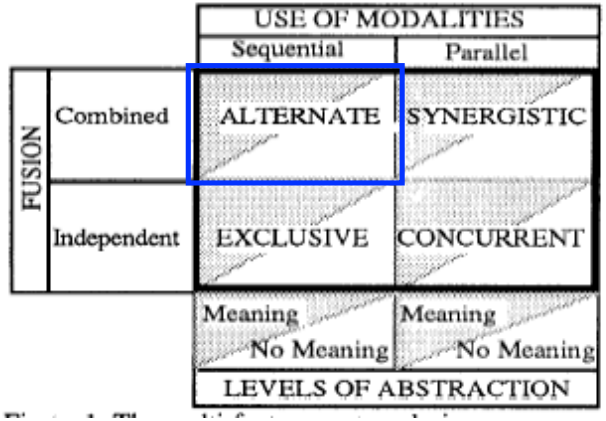
- C = Concurrent,
  - two distinct tasks in parallel,
  - No co-reference,
  - No temporal constraint

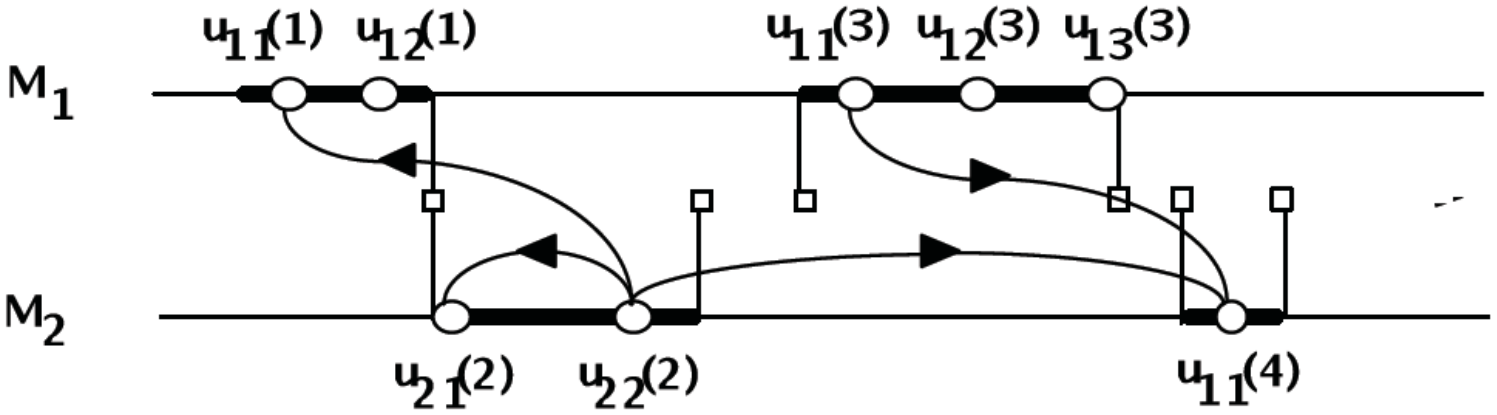




CASE - Alternate

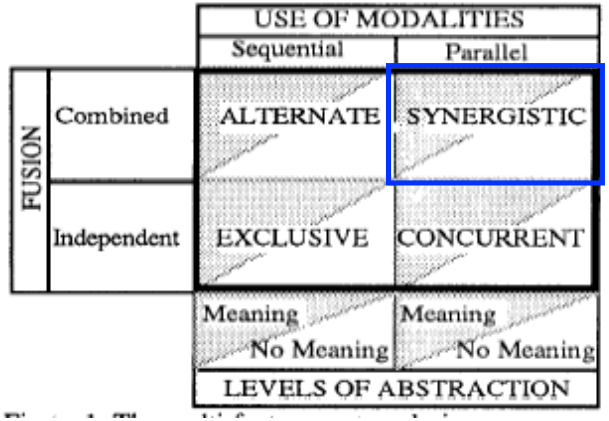
- A = Alternate,
  - A task with temporal alternation of modalities, using coreferences

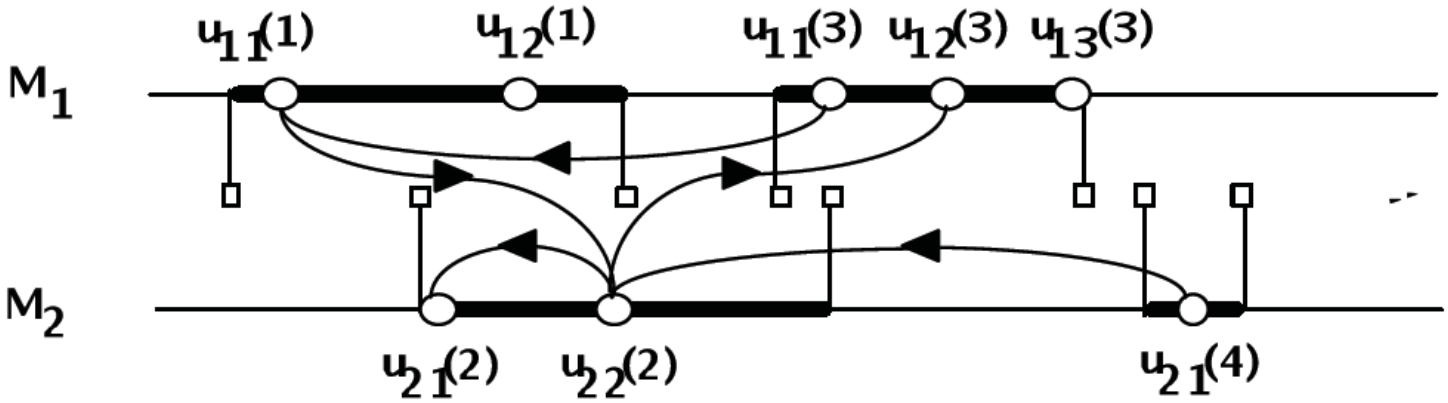




CASE - Synergistic

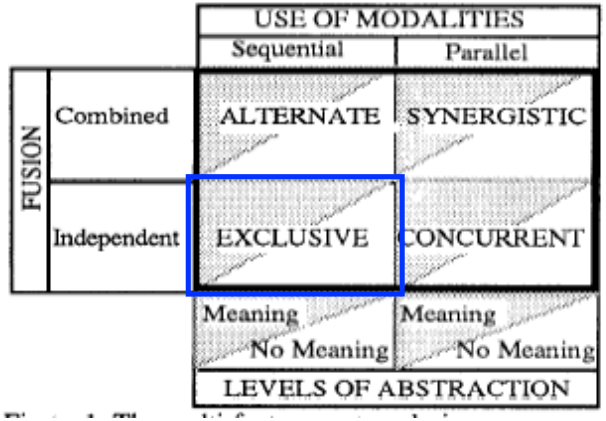
- S = Synergistic,
  - A task, in parallel, using several coreferent modalities





CASE - Exclusive

- E = Exclusive,
  - One task after the other using one modality at a time,
  - No coreference



- Find an example of application for each categories of CASE model

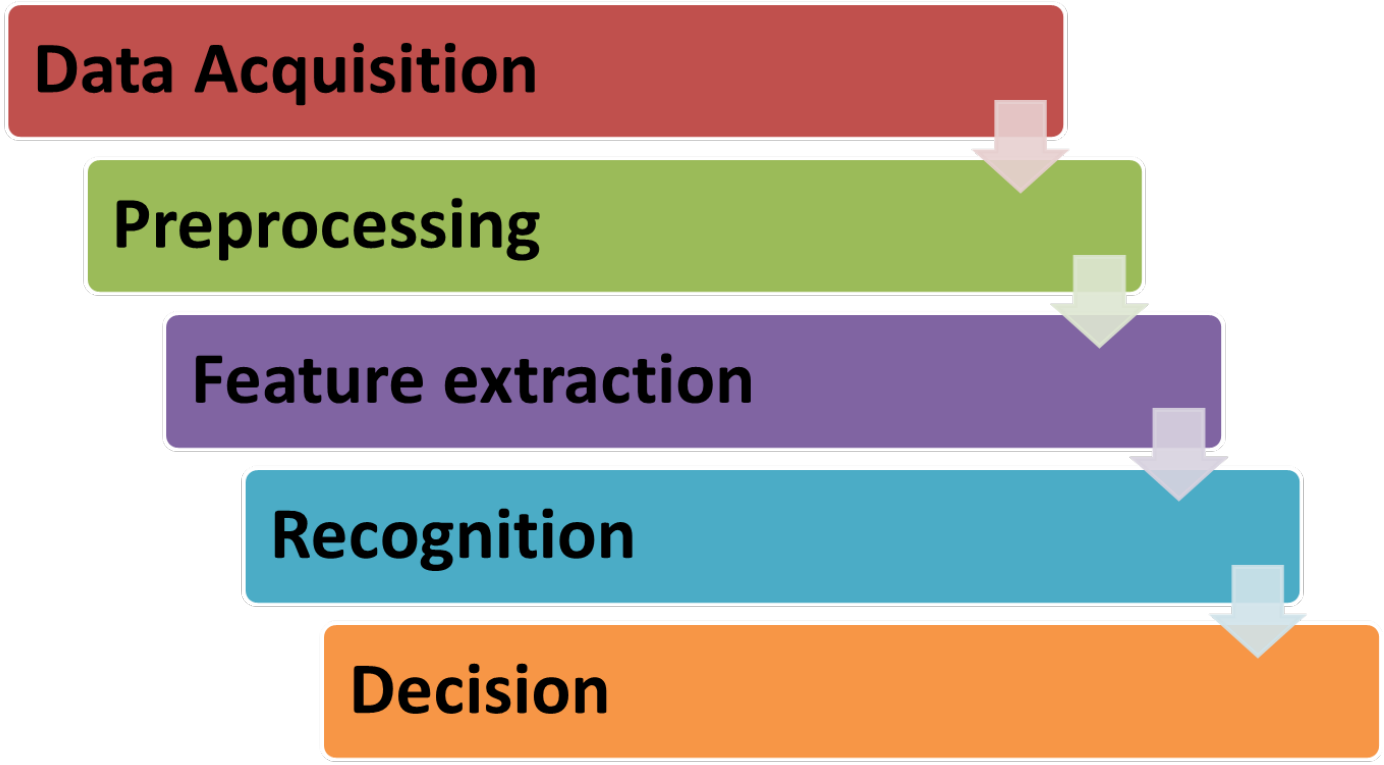
CARE model: Multimodal Systems Usability Properties

- 4 properties of multimodal interaction in order to characterise and assess usability in multimodal interaction:
  - **Complementarity**
    - If multiple modalities are to be used within a temporal window to reach a given state
    - e.g. « Please give me details about this list » and <point at a « list of flights » label
  - **Assignment**
    - Only one modality can be used to reach a given state
    - e.g. Movement of mouse to change the position of a window
  - **Redundancy**
    - If multiple modalities have the same expressive power (-> equivalent) and if they are all used within the same temporal window
    - e.g. « Could you show me the list of flights? » **And** <clic on « list of flights » button>
  - **Equivalence**
    - Necessary and sufficient to use any one of the available modalities
    - e.g.: « Could you show me the list of flights? » **Or** <clic on « list of flights » button>

Fusion models

- Sketch a general multimodal architecture (generic pipeline with multimodal fusion)
  - hint: think to “Information processing general schema”

General information processing schema



Fundamental problems - II

- Challenges
  - extraction of **relevant features**
    - **relevant,** that contain information required to solve the problem
      - relevance is tied to the **context**!! (speech recognition vs speaker identification)
    - **compact,** feature vector with low dimensionality
      - avoid the **curse of dimensionality**
    - **balance!!**
  - **fusion** of the information
    - **integrating** the information coming from different modalities
      - different data rate, dimensionality, etc.
    - **dynamic** adaptation
    - **synchronisation**

“Complexity” factors for Multimodal fusion

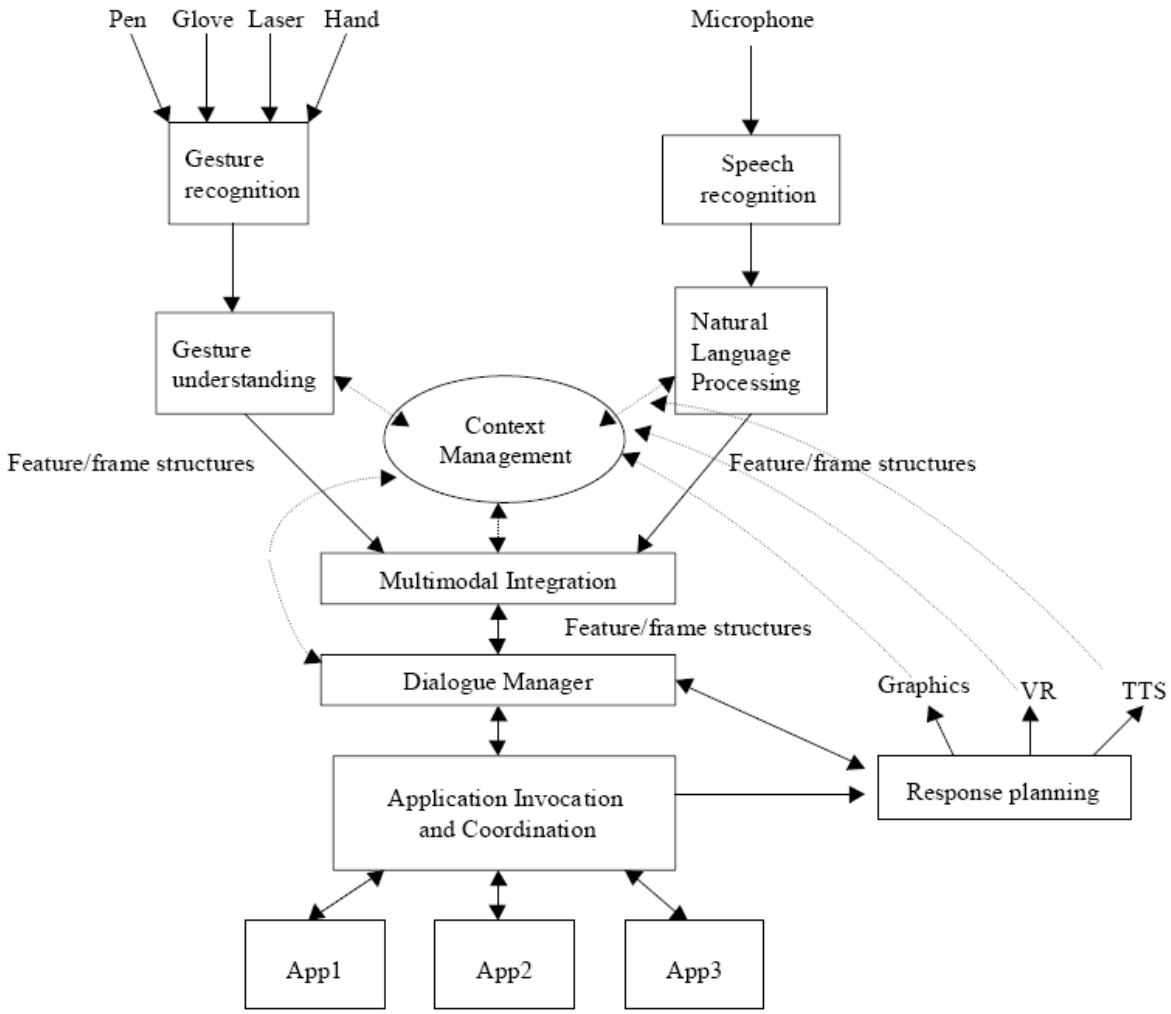
- Formats and rates
- Processing time
- Correlation or not of modalities
- Confidence levels
- Capturing and processing costs

Multimodal fusion challenges

- Levels of fusion
- How to fuse?
- When to fuse?
- What to fuse?

Multimodal architecture

- Smartkom project

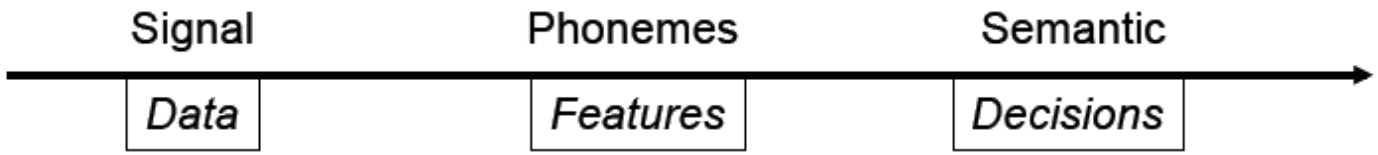


Fusion

- Definition : « Resolution of the co-reference: match the multimodal referents »
- Theoretically, three types (levels) of fusion :
  - Raw Data fusion
  - Features fusion
  - Decision fusion

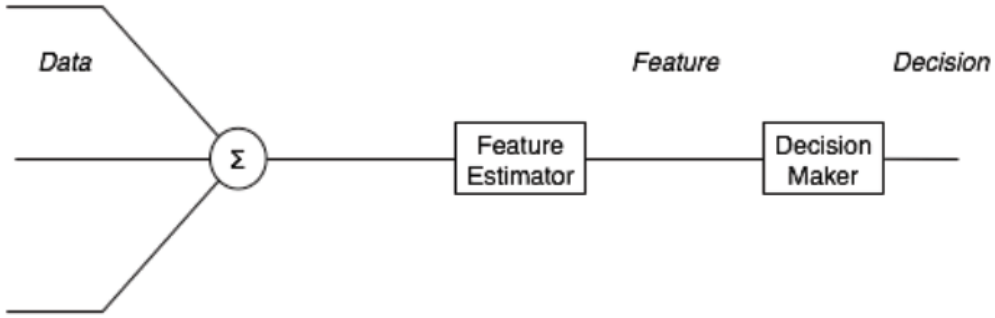
  Pre-classification fusion

  Post-classification fusion
- Example with speech :



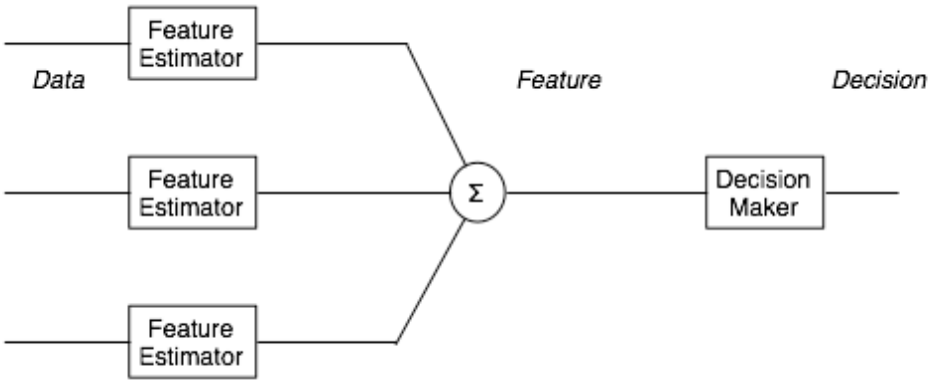
Data level fusion

- Lowest level of fusion
- Integration of raw observations
- Used to merge data from same type of sensors (2 cameras, e.g.)
- Adds
  - No loss of information
- Cons
  - Too much sensitive to specific nature of sensor to be of real generic use

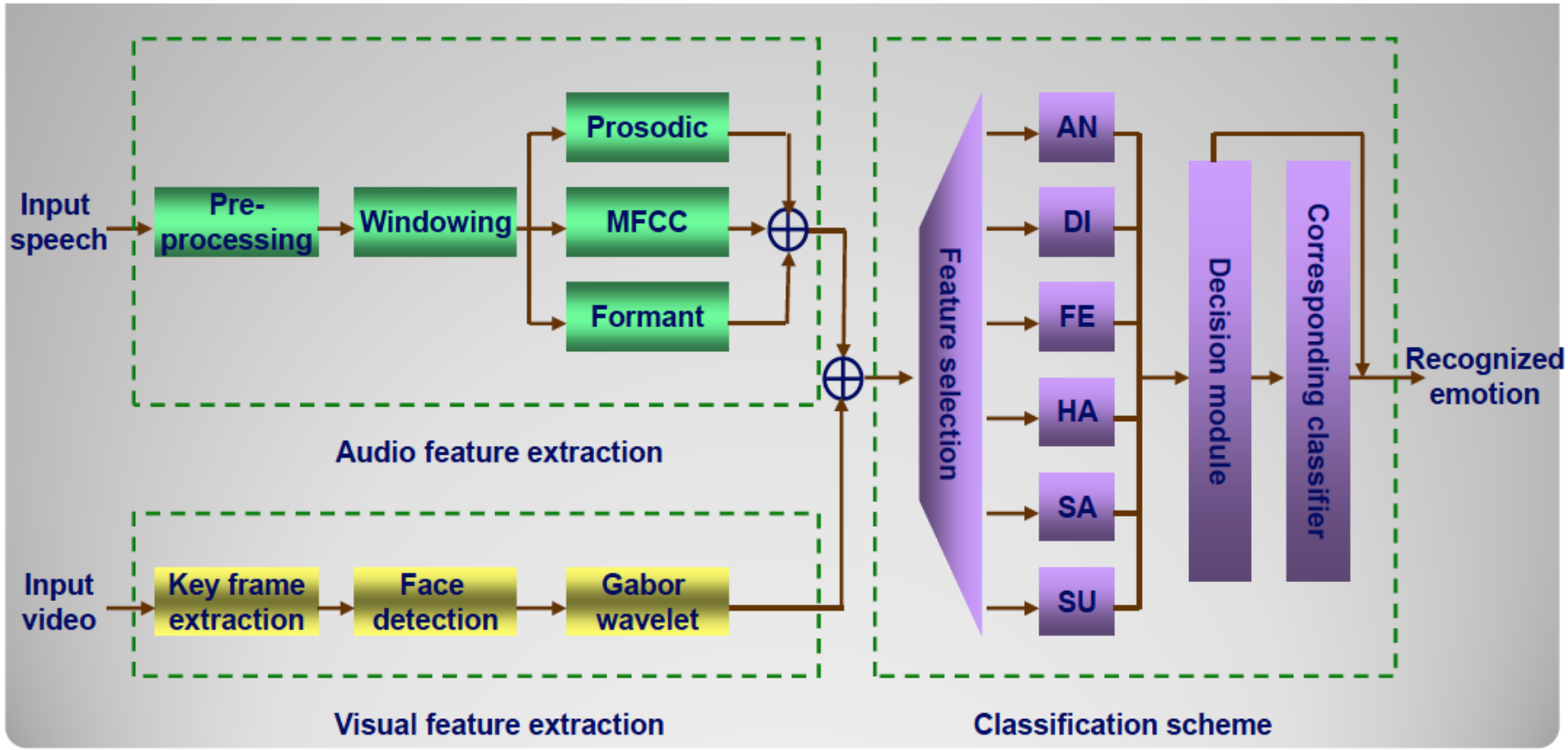


Feature-level fusion

- Also known as « **early fusion** »
- Each stream of sensory data is first analysed to extract features, then features are fused
- Appropriate for closely coupled and synchronised modalities (e.g. speech and lips)
- Most commonly used methods : ANNs, GMMs, HMMs, etc.

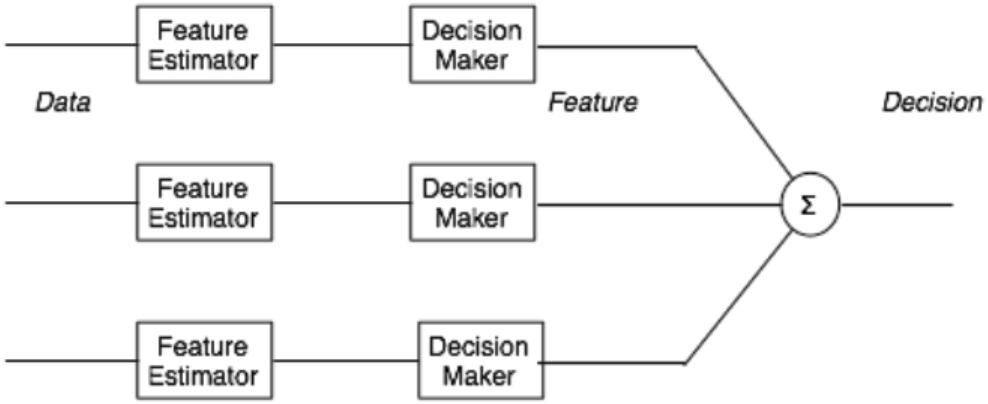


Ex: Feature Level Fusion for Emotion Recognition System



Decision-level fusion

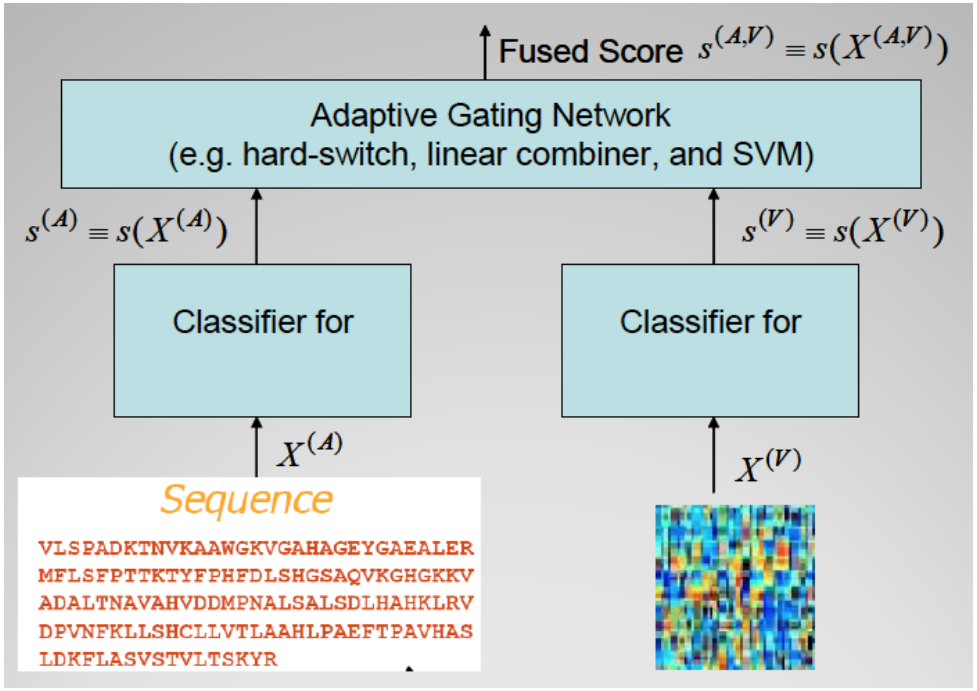
- Also known as « **late fusion** »
- Most commonly found type of fusion
- Fusion of individual decisions or interpretations (-> works at the semantic level)
  - e.g. speech and gesture
- Robust, but cannot recover loss of information that happened at lower levels



- Modify the schema of the Emotion Recognition System (slide 34) to implement a late fusion approach.

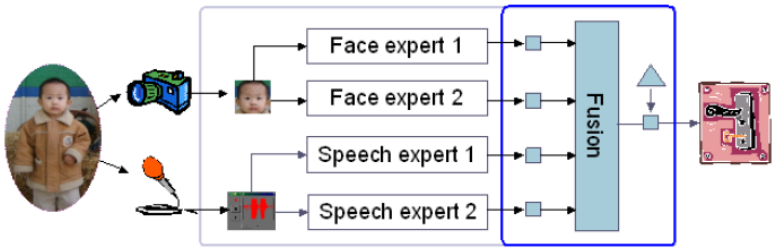
Fusion Architecture

- 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

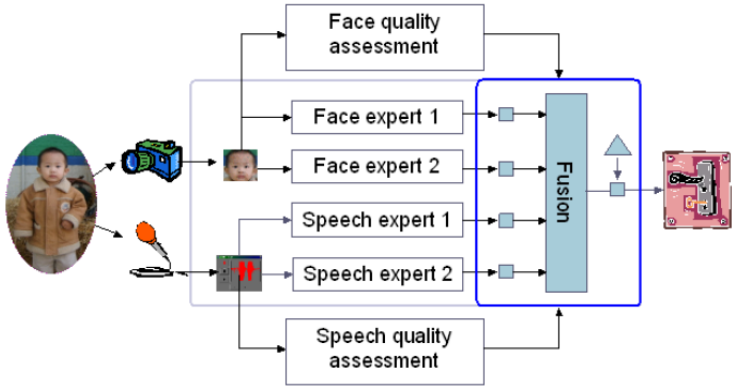


Adaptive vs non-adaptive fusion

- Adaptive, or quality-based fusion attempts to change the weight associated with a modality as a function of the signal quality measured on the modality
  - Ex. quality measures for face images are face detection reliability, presence of glasses, brightness, contrast. For speech, this would be signal-to-noise ratio and speech-likeness (versus noise).



(a) Non-adaptive fusion



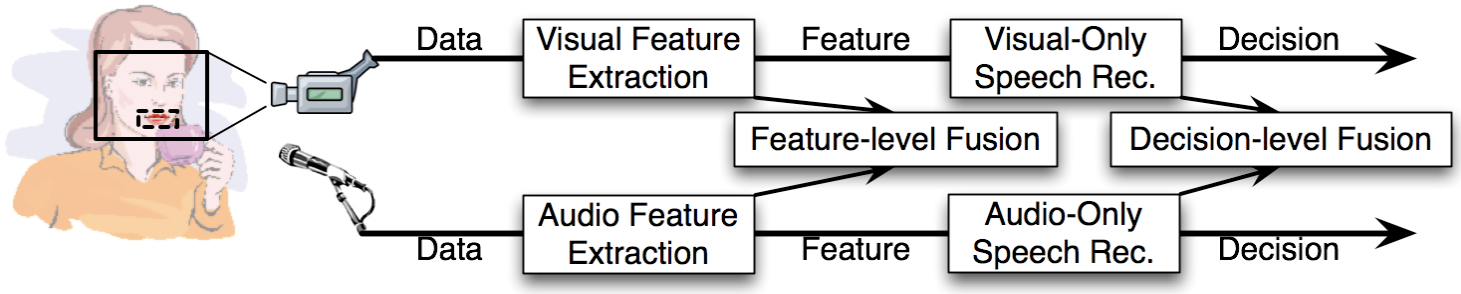
(b) adaptive fusion with quality

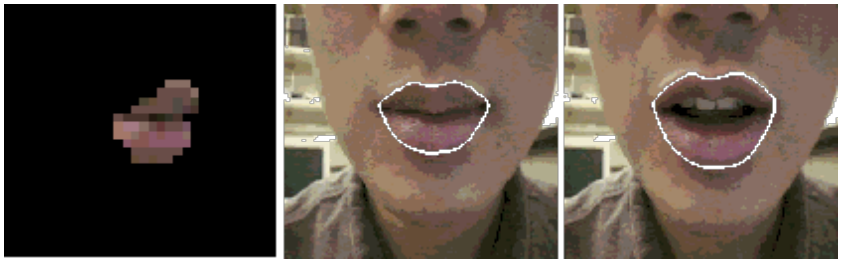
Exemple: Multimodal Fusion for Audio-Visual Speech Recognition - I

- Multimodal classification problem
  - Audio, sampling rate tens of kilohertz, 1 temporal dimension
  - Video, sampling rate tens of samples per second, 2 spatial and 1 temporal dimensions
  - HMMs classifiers
    - small vocabulary: each word modeled with an HMM
    - bigger vocabulary: each speech sounds modeled with a HMM
  - Fusion at feature or decision level

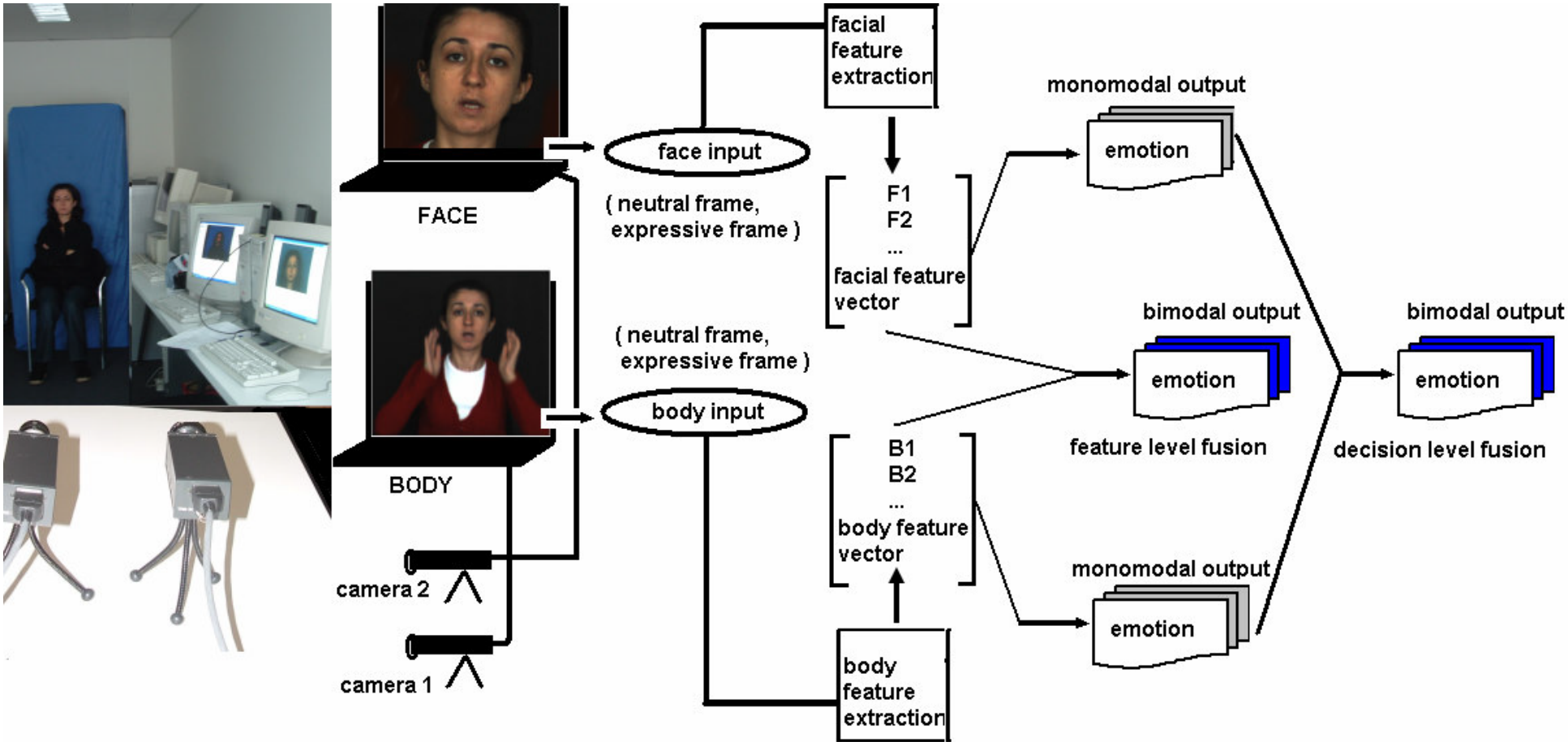
Exemple: Multimodal Fusion for Audio-Visual Speech Recognition - II

- Augmenting Speech
  - Speech Recognition degrades in noisy environments
  - Use of Image based modeling of the lips can improve accuracy





Affect Recognition from Face and Body: Early Fusion vs. Late Fusion, H. Gunes & M. Piccardi



Affect Recognition from Face and Body: Early Fusion vs. Late Fusion

Table 3. Emotion recognition results for 3 subjects using 156 training and 50 testing samples.

	Attributes	Number of Classes	Classifier	Correctly classified
Face*	67	8	C4.5	78 %
Body*	140	6	BayesNet	90 %

* These classification results were later used for late fusion.

Table 4. Emotion classification for the combined feature vector with C4.5 and BayesNet into eight emotion categories (anger, disgust, fear, happiness, sadness, surprise, happy_surprise and uncertainty).

	Training	Testing	Attributes	Classifier	Correctly classified (%)
Face& Body	156	50	206	BayesNet	88
Face& Body	156	50	206	C4.5	94
Face& Body	156	50	45	BayesNet	96
Face& Body	156	50	45	C4.5	82

Table 5.Emotion recognition results with BayesNet using the reduced feature vector (45 features) for 50 testing samples.

Emotion	Recognition results (%)
Overall	96
Anger	80
Disgust	100
Fear	100
Happiness	100
Sadness	100
Surprise	100
Happy_surprise	100
Uncertainty	100

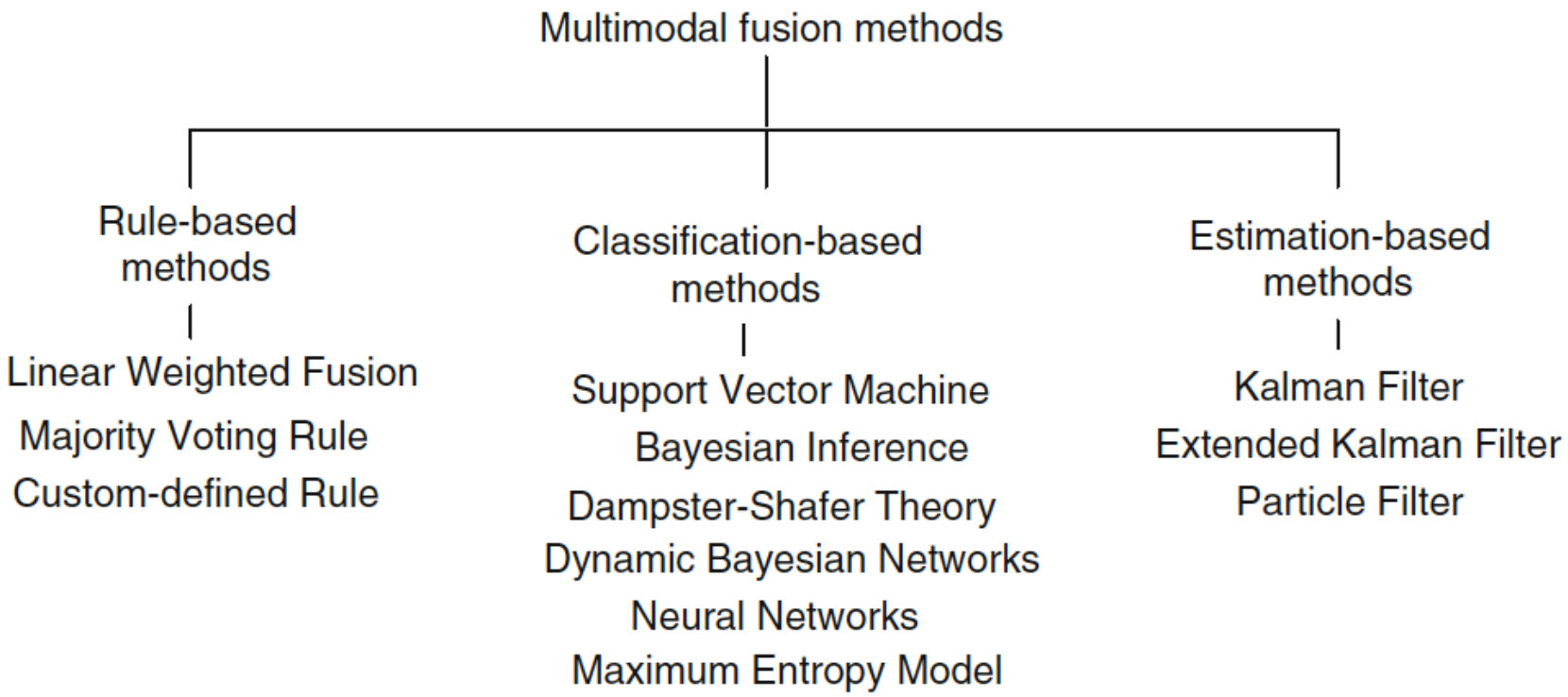
Table 6. Emotion recognition results for late fusion using the product rule, sum and weight criteria on the testing set of 50 samples.

	Recognition rates on the testing set (%)		
Emotion	**Product Rule**	**Sum Rule**	**Weight criterion ($\lambda_f = 0.70$, $\lambda_b = 0.30$)**
Overall	80	86	82
Anger	60	70	70
Disgust	87	100	100
Fear	100	100	100
Happiness	80	80	80
Sadness	40	60	40
Surprise	91	91	91
Happy_surprise	100	100	100
Uncertainty	100	100	86

Fusion types comparison

	Data-level fusion	Features-level fusion	Decision-level fusion
Used for…	Raw data of same type	Closely coupled modalities	Weakly coupled modalities
Level of information	Highest level of information detail	Moderate level of information detail	Cannot recover from previous loss of information
Noise/failures sensitivity	Highly susceptible to noise or failures	Less sensitive to noise or failures	Highly resistant to noise or failures
Usage	Not really used for MMI	Used for fusion of particular modes	Most widely used type of fusion

Fusion Methods



"Multimodal fusion for multimedia analysis: a survey", Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, Mohan S. Kankanhalli, Multimedia System, Springer, 2010.

Rule-based methods

- Includes a variety of basic rules of combining multimodal information
  - statistical rule-based methods such as **linear weighted fusion** (sum and product), MAX, MIN, AND, OR, majority voting
  - custom defined rules
- Work well when good quality of temporal alignement

Table 1 A list of the representative works in the rule-based fusion methods category

Fusion method	Level of fusion	The work	Modalities	Multimedia analysis task
Linear weighted fusion	Feature	Foresti and Snidaro [40]	Video (trajectory coordinates)	Human tracking
		Wang et al. [136]	Video (color, motion and texture	Human tracking
		Yang et al. [152]	Video (trajectory coordinates)	Human tracking
		Kankanhalli et al. [67]	Video (color, motion and texture)	Face detection, monologue detection and traffic monitoring
	Decision	Neti et al. [87]	Audio (phonemes) and visual (visemes)	Speaker recognition
		Lucey et al. [78]	Audio (MFCC), video (Eigenlip)	Spoken word recognition
		Iyenger et al. [57, 58]	Audio (MFCC), video (DCT of the face region) and the synchrony score	Monologue detection, semantic concept detection and annotation in video
		Hua and Zhang [55]	Image (six features: color histogram, color moment, wavelet, block wavelet, correlagram, blocked correlagram)	Image retrieval
		Yan et al. [151]	Text (closed caption, video OCR), audio, video (color, edge and texture histogram), motion	Video retrieval
		McDonald and Smeaton [83]	Text and video (color, edge and texture)	Video retrieval
		Jaffre and Pinquier [59]	Audio, video index	Person identification from audio-visual sources
Majority voting rule	Decision	Radova and Psutka [108]	Raw speech (set of patterns)	Speaker identification from audio sources
Custom-defined rules	Decision	Babaguchi et al. [12]	Visual (color), closed caption text (keywords)	Semantic sports video indexing
		Corradini et al. [32]	Speech, 2D gesture	Human computer interaction
		Holzapfel et al. [49]	Speech, 3D pointing gesture	Multimodal interaction with robot
		Pfleger [100]	Pen gesture, speech	Multimodal dialog system

"Multimodal fusion for multimedia analysis: a survey", Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, Mohan S. Kankanhalli, Multimedia System, Springer, 2010.

Classification-based methods

- Classification techniques used to classify the multimodal observation into one of the pre-defined classes.
  - SVM, HMM, NN, etc.

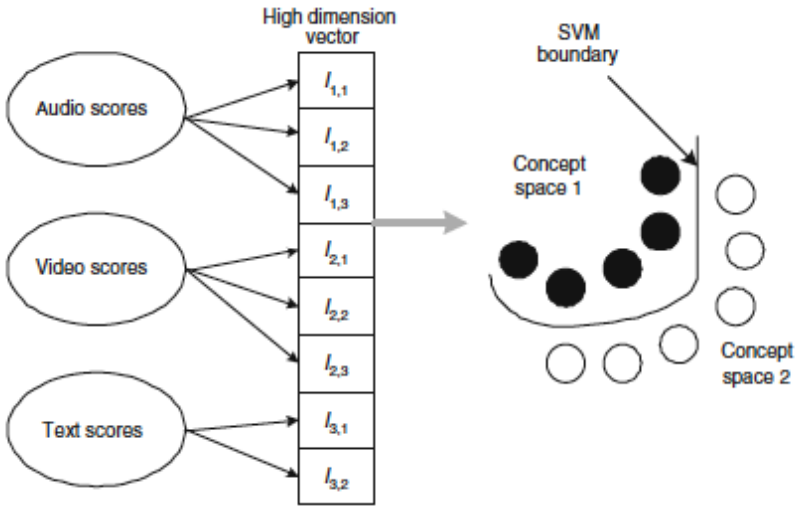


Table 2 A list of the representative works in the classification methods category used for multimodal fusion

Fusion method	Level of fusion	The work	Modalities	Multimedia analysis task
Support vector machine	Decision	Adams et al. [3]	Video (color, structure, and shape), audio (MFCC) and textual cues	Semantic concept detection
		Aguilar et al. [4]	Fingerprint, signature, face (MCYT Multimodal Database, XM2VTS face database)	Biometric verification
		Iyenger et al. [58]	Audio, video	Semantic concept detection
		Wu et al. [141]	Color histogram, edge orientation histogram, color correlogram, co-occurrence texture, motion vector histogram, visual perception texture, and speech	Semantic concept detection
		Bredin and Chollet [19]	Audio (MFCC), video (DCT of lip area), audio-visual speech synchrony	Biometric identification of talking face
	Hybrid	Wu et al. [143]	Video, audio	Multimedia data analysis
		Zhu et al. [156]	Image (low-level visual features, text color, size, location, edge density, brightness, contrast)	Image classification
		Ayache et al. [11]	Visual, text cue	Semantic indexing
Bayesian inference	Feature	Pitsikalis et al. [102]	Audio (MFCC), video (Shape and texture)	Speech recognition
	Decision	Meyer et al. [85]	Audio (MFCC) and video (lips contour)	Spoken digit recognition
	Hybrid	Xu and Chua [149]	Audio, video, text, web log	Sports video analysis
		Atrey et al. [8]	Audio (ZCR, LPC, LFCC) and video (blob location and area)	Event detection for surveillance
Dempster–Shafer theory	Feature	Mena and Malpica [84]	Video (trajectory coordinates)	Segmentation of satellite images
	Decision	Guironnet et al. [44]	Audio (phonemes) and visual (visemes)	Video classification
		Singh et al. [116]	Audio (MFCC), video (DCT of the face region) and the synchrony score	Finger print classification
		Reddy [110]	Audio (MFCC), video (Eigenlip)	Human computer interaction
	Hybrid	Bendjebbour et al. [16]	Video (trajectory coordinates)	Segmentation of satellite images
Dynamic Bayesian networks	Feature	Wang et al. [138]	Audio (cepstral vector), visual (gray-level histogram difference and motion features)	Video shot classification
		Nefian et al. [86]	Audio (MFCC) and visual (2D-DCT coefficients of the lips region)	Speech recognition
		Nock et al. [90, 91]	Audio (MFCC) and video (DCT coefficients of the lips region	Speaker localization
		Chaisorn et al. [25]	Audio (MFCCs and perceptual features), video (color, face, video-text, motion)	Story segmentation in news video
		Adams et al. [3]	Video (color, structure, and shape),audio (MFCC) and textual cues	Video shot classification
		Beal et al. [15]	Audio and video—the details of features not available	Object tracking
		Bengio et al. [17]	Speech (MFCC) and video (shape and intensity features)	Biometric identity verification
		Hershey et al. [46]	Audio (Spectral components), video (fine-scale appearance and location of the lips)	Speaker localization
		Zou and Bhanu [158], Noulas and Krose [92]	Audio (MFCC) and video (pixel value variation)	Human tracking
		Ding and Fan [38]	Video (spatial color distribution and the angle of yard lines)	Shot classification in a sports video

Table 2 continued

Fusion method	Level of fusion	The work	Modalities	Multimedia analysis task
	Decision	Wu et al. [142]	Image (color, texture and shape) and Camera parameters	Photo annotation
		Town [131]	Video (face and blob), ultrasonic sensors	Human tracking
	Hybrid	Xie et al. [145]	Text (closed caption), Audio (pitch, silence, significant pause), video (color histogram and motion intensity), Speech (ASR transcript)	Topic clustering in video
Neural networks	Feature	Cutler and Davis [34]	Audio (phoneme) and video (viseme)	Speaker localization
		Zou and Bhanu [158]	Audio (spectrogram) and video (blob)	Human tracking
	Decision	Gandetto et al. [41]	CPU load, Login process, Network load, Camera images	Human activity monitoring
	Hybrid	Ni et al. [88]	Image (features details not provided in the paper)	Image recognition
Maximum Entropy Model	Feature	Magalhães and Rüger [80]	Text and Image	Semantic image indexing

"Multimodal fusion for multimedia analysis: a survey", Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, Mohan S. Kankanhalli, Multimedia System, Springer, 2010.

Estimation-based methods

- Used to better estimate the state of a moving object based on multimodal data
  - e.g. object localisation and tracking using video and audio signals
  - Kalman filter, particle filter

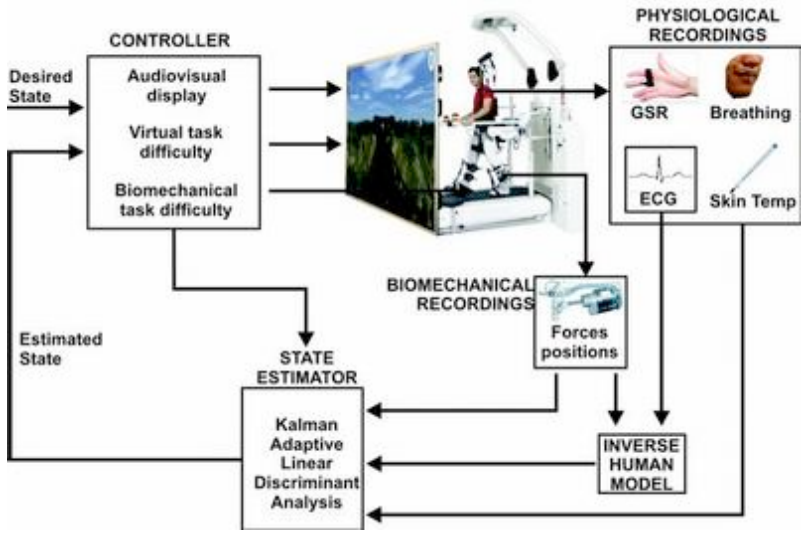


Table 3 A list of the representative works in the estimation methods category used for multimodal fusion

Fusion method	Level of fusion	The work	Modalities	Multimedia analysis task
Kalman filter and its variants	Feature	Potamitis et al. [107]	Audio (position, velocity)	Multiple speaker tracking
		Loh et al. [77]	Audio, video	Single speaker tracking
		Gehrig et al. [43]	Audio (TDOA), video (position of the speaker)	Single speaker tracking
		Zhou and Aggarwal [154]	Video [spatial position, shape, color (PCA), blob]	Person/vehicle tracking
	Decision	Strobel et al. [124]	Audio, video	Single object localization and tracking
		Talantzis et al. [125]	Audio (DOA), video (position, velocity, target size)	Person tracking
Particle filter	Feature	Vermaak et al. [132]	Audio (TDOA), visual (gradient)	Single speaker tracking
	Decision	Zotkin et al. [157]	Audio (TDOA), video (skin color, shape matching and color histograms)	Multiple speaker tracking
		Perez et al. [99]	Audio (TDOA), video (coordinates)	Single speaker tracking
		Nickel et al. [89]	Audio (TDOA), video (Haar-like features)	Single speaker tracking

"Multimodal fusion for multimedia analysis: a survey", Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, Mohan S. Kankanhalli, Multimedia System, Springer, 2010.

Conclusion

- Multimodality is natural but rise several challenges!
  - how to fuse data?, what to fuse?, when to fuse?, etc.
- CASE and CARE model
- Multimodal signal fusion levels and methods

References

- "Multimodal fusion for multimedia analysis: a survey", Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, Mohan S. Kankanhalli, Multimedia System, Springer, 2010.
- Cohen, P., Johnston, M., McGee, D., Oviatt, S., Pittman, J.,Smith, I., Chen, L. and Clow, J. Quickset: Multimodal interaction for distributed applications. Proceedings of the Fifth ACM International Multimedia Conference, New York, NY:ACM Press, 1997, 31-40.
- Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J. and Young, R.: Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE properties, Proceedings of the INTERACT'95 conference, S. A. Arnesen & D. Gilmore Eds., Chapman&Hall Publ., Lillehammer, Norway, June 1995, pp. 115-120.
- Dumas, B., Lalanne, D., Oviatt, S. Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In Denis Lalanne, Jürg Kohlas eds. (2009). Human Machine Interaction, LNCS 5440, Springer-Verlag, Berlin/Heidelberg, pp. 3-27.
- Lalanne, D., Nigay, L., Palanque, P., Robinson, P., Vanderdonckt, J., Ladry, J-F. Fusion Engines for Multimodal Interfaces: a survey. International Conference on Multimodal Interfaces and Workshop on Machine Learning for Multi-modal Interaction (ICMI-MLMI 2009), Cambridge, Massachusetts, USA, ACM, 2009.
- Nigay & Coutaz 1993 : Nigay, L., Coutaz, J. A design space for multimodal interfaces: concurrent processing and data fusion, in Proc. INTERCHI'93 Human Factors in Computing Systems (Amsterdam, April 24-29, 1993), ACM Press, pp. 172-178.
- Sharma, R., Pavlovic, V.I., & Huang, T.S. (1998). Toward multimodal human-computer interface. Proceedings IEEE, 86(5) [Special issue on Multimedia Signal Processing], 853-860.
- Vo, M. T. and C. Wood. 1996. Building an application framework for speech and pen input integration in multimodal learning interfaces. International Conference on Acoustics, Speech, and Signal Processing 1996, Atlanta, GA.
- Wu, Lizhong, Oviatt, S. L., Cohen, P. R., Multimodal Integration-A Statistical View, IEEE Transactions on Multimedia, Vol. 1, No. 4, December 1999, pp. 334-341.

What you should know

- Explain the concept of multimodality and discuss its advantages
- Illustrate the schema of multimodal communication?
- Discuss the fundamental problems to solve with multimodal systems?
- Explain the CARE and CASE models and provide examples
- Sketch a general architecture of a multimodal system
- Explain and compare the levels and methods for multimodal fusion?