Language Technologies Institute

Carnegie Mellon University

Advanced Multimodal Machine Learning

Lecture 1.1: Introduction

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Your Instructor and TAs This Semester (11-777)

Louis-Philippe Morency
morency@cs.cmu.edu
Office: GHC-5411
Phone: 412-268-5508

Volkan Cirik
vcirik@andrew.cmu.edu
Office GHC-5713

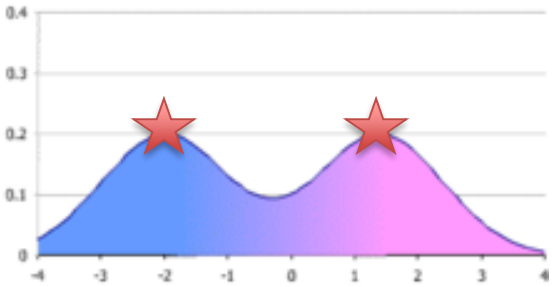
Soumya Wadhwa
soumyaw@andrew.cmu.edu

Lecture Objectives

- Introductions
- What is Multimodal?
  - Multimodal vs multimedia
- A historical view to multimodal research
- Core technical challenges
  - Representation learning, translation, alignment, fusion and co-learning
- Course syllabus and project assignments
  - Grades and course structure

What is Multimodal?

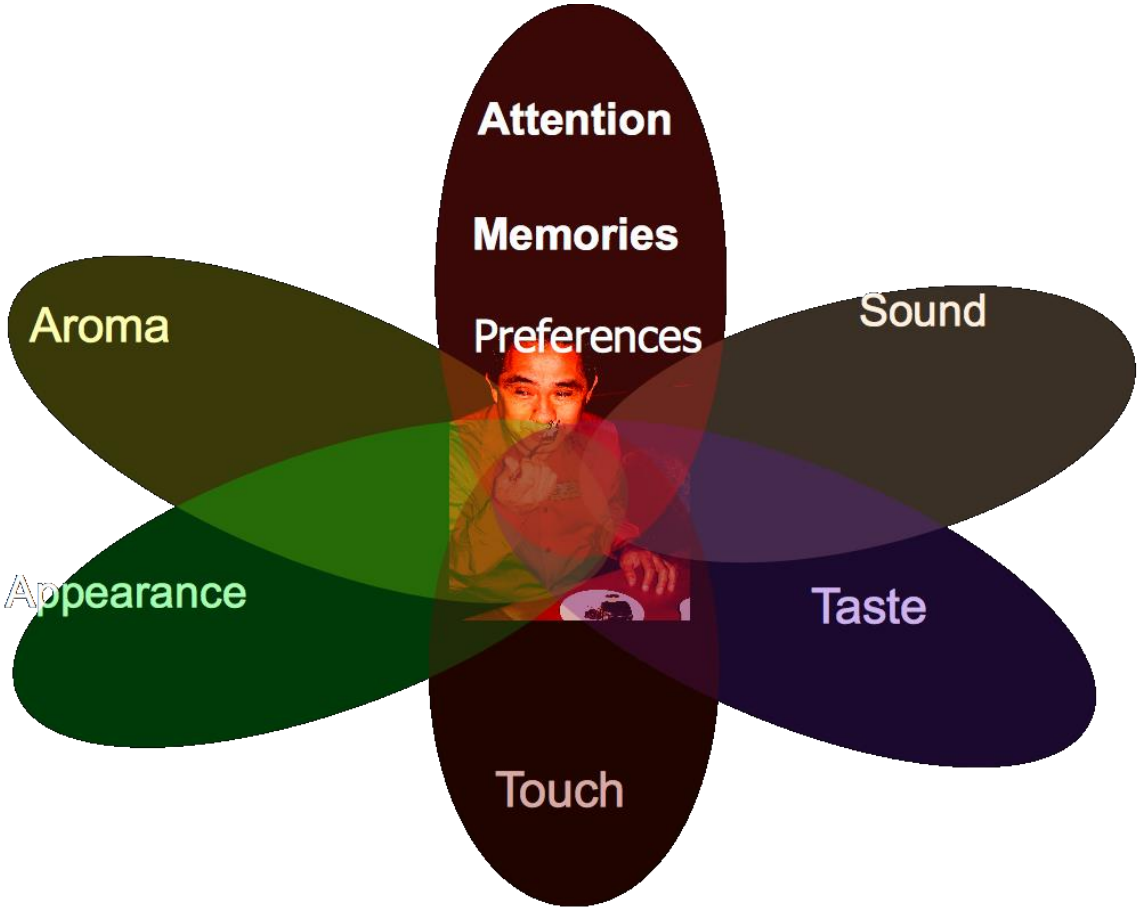
What is Multimodal?



Multimodal distribution

- Multiple modes, i.e., distinct “peaks” (local maxima) in the probability density function

What is Multimodal?



Sensory Modalities

What is Multimodal?

Modality

The way in which something happens or is experienced.

- Modality refers to a certain type of information and/or the representation format in which information is stored.
- Sensory modality: one of the primary forms of sensation, as vision or touch; channel of communication.

Medium (“middle”)

A means or instrumentality for storing or communicating information; system of communication/transmission.

- Medium is the means whereby this information is delivered to the senses of the interpreter.

Examples of Modalities

- Natural language (both spoken or written)
- Visual (from images or videos)
- Auditory (including voice, sounds and music)
- Haptics / touch
- Smell, taste and self-motion
- Physiological signals
  - Electrocardiogram (ECG), skin conductance
- Other modalities
  - Infrared images, depth images, fMRI

Multimodal Communicative Behaviors

Verbal

Lexicon
- Words

Syntax
- Part-of-speech
- Dependencies

Pragmatics
- Discourse acts

Vocal

Prosody
- Intonation
- Voice quality

Vocal expressions
- Laughter, moans

Visual

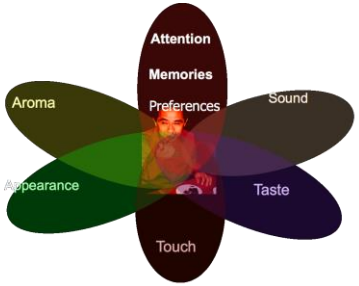
Gestures
- Head gestures
- Eye gestures
- Arm gestures

Body language
- Body posture
- Proxemics

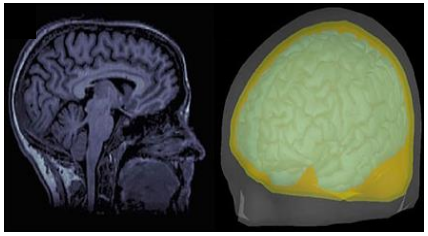
Eye contact
- Head gaze
- Eye gaze

Facial expressions
- FACS action units
- Smile, frowning

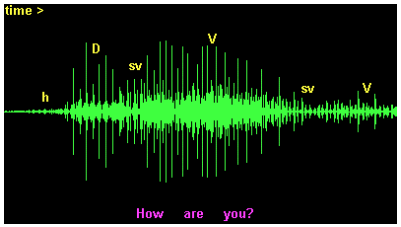
Multiple Communities and Modalities



Psychology



Medical

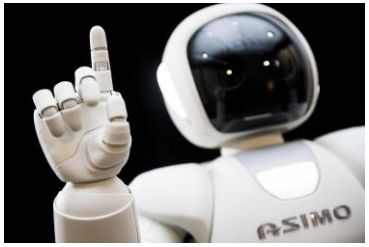


Speech

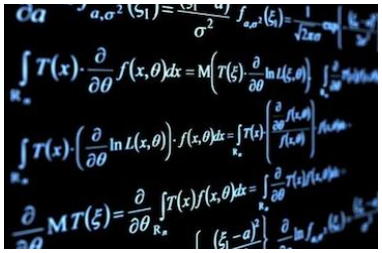
Vision

Language

Multimedia



Robotics



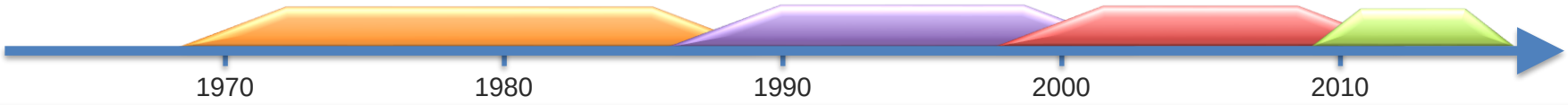
Learning

A Historical View

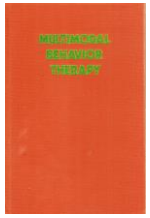
Prior Research on "Multimodal"

Four eras of multimodal research

- The "behavioral" era (1970s until late 1980s)
- The "computational" era (late 1980s until 2000)
- The "interaction" era (2000 - 2010)
- The "deep learning" era (2010s until …)
  - Main focus of this course



The "Behavioral" Era (1970s until late 1980s)



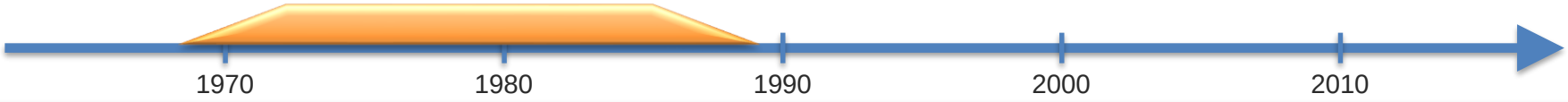
Multimodal Behavior Therapy by Arnold Lazarus [1973]

- 7 dimensions of personality (or modalities)

Multi-sensory integration (in psychology):

- Multimodal signal detection: Independent decisions vs. integration [1980]
- Infants' perception of substance and temporal synchrony in multimodal events [1983]
- A multimodal assessment of behavioral and cognitive deficits in abused and neglected preschoolers [1984]

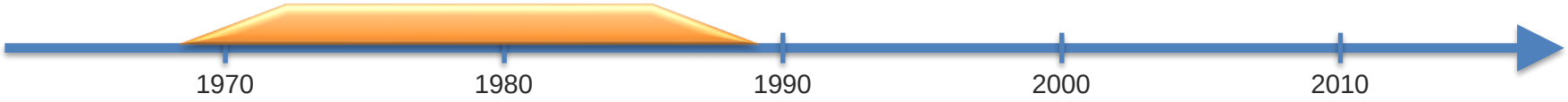
❑ TRIVIA: Geoffrey Hinton received his B.A. in Psychology ☺



Language and Gestures

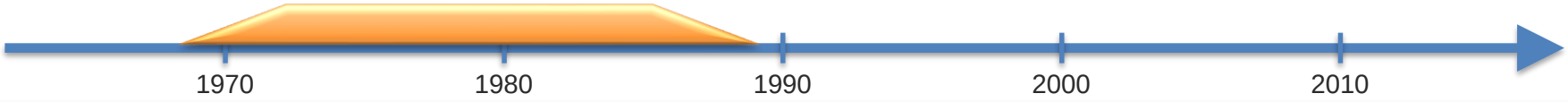
David McNeill
University of Chicago
Center for Gesture and Speech Research

"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."



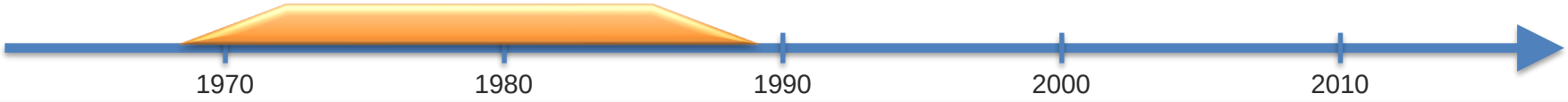
The McGurk Effect (1976)

Hearing lips and seeing voices – Nature



The McGurk Effect (1976)

Hearing lips and seeing voices – Nature



➢ The "Computational" Era(Late 1980s until 2000)

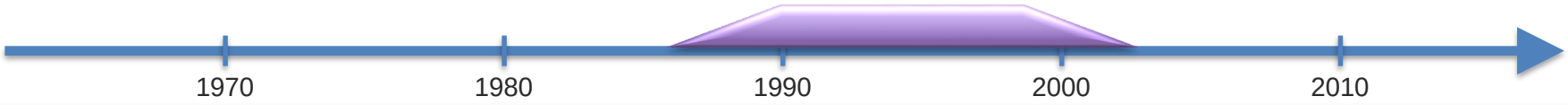
1) Audio-Visual Speech Recognition (AVSR)

- Motivated by the McGurk effect
- First AVSR System in 1986

  "Automatic lipreading to enhance speech recognition"
- Good survey paper [2002]

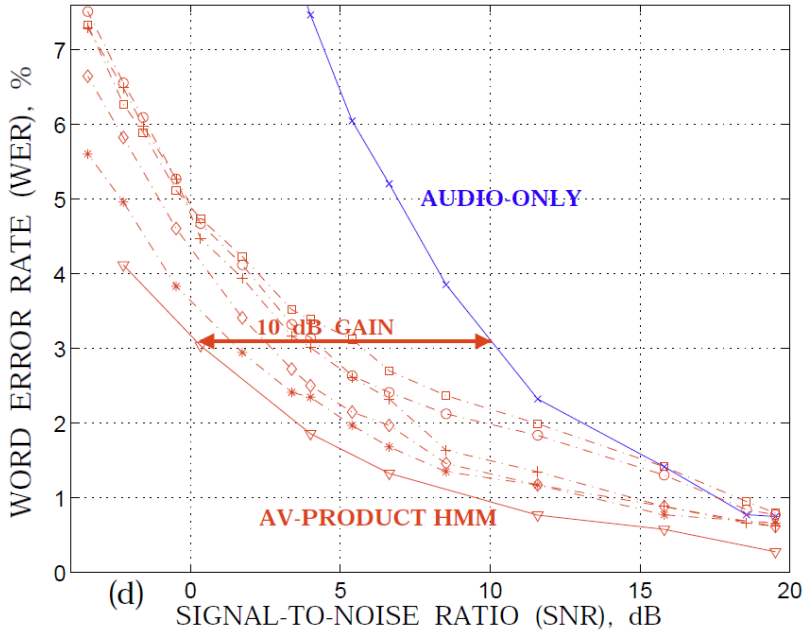
  "Recent Advances in the Automatic Recognition of Audio-Visual Speech"

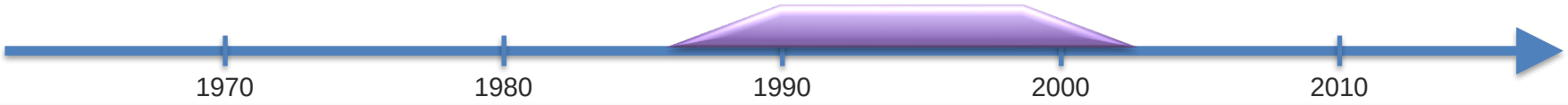
❑ TRIVIA: The first multimodal deep learning paper was about audio-visual speech recognition [ICML 2011]



➢ The "Computational" Era(Late 1980s until 2000)

1) Audio-Visual Speech Recognition (AVSR)



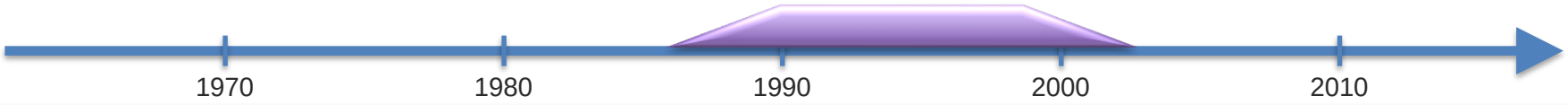


➢ The "Computational" Era (Late 1980s until 2000)

2) Multimodal/multisensory interfaces

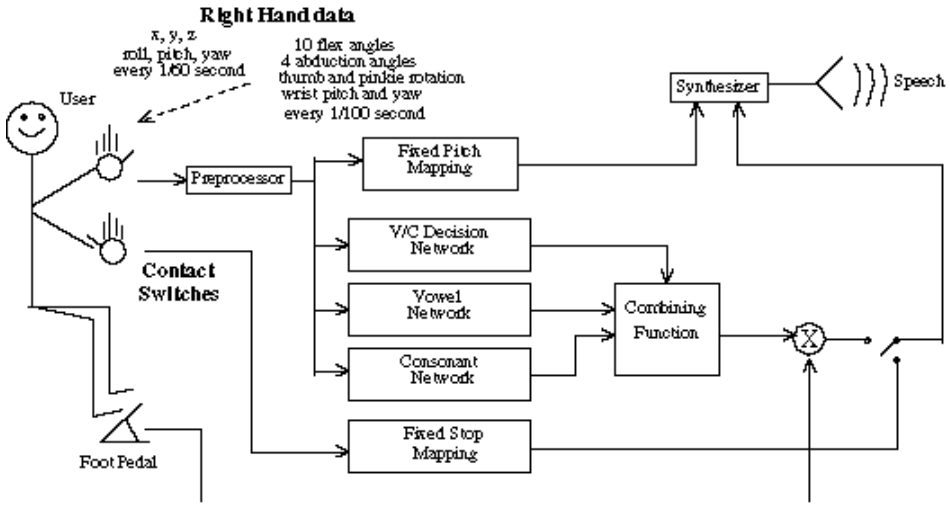
- Multimodal Human-Computer Interaction (HCI)

"Study of how to design and evaluate new computer systems where human interact through multiple *modalities, including both input and output modalities."*

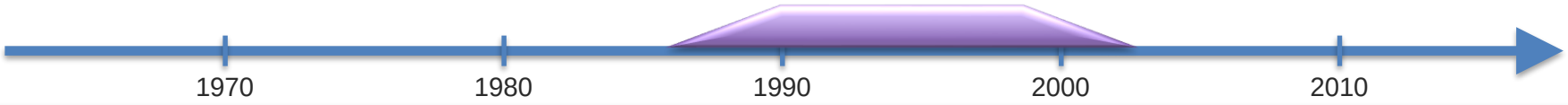


➢ The "Computational" Era (Late 1980s until 2000)

2) Multimodal/multisensory interfaces



Glove-talk: A neural network interface between a data-glove and a speech synthesizer By Sidney Fels & *Geoffrey Hinton [CHI'95]*

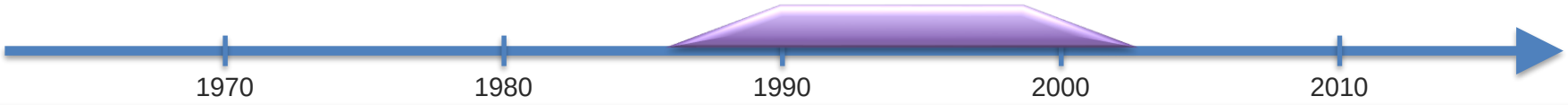


➢ The "Computational" Era (Late 1980s until 2000)

2) Multimodal/multisensory interfaces

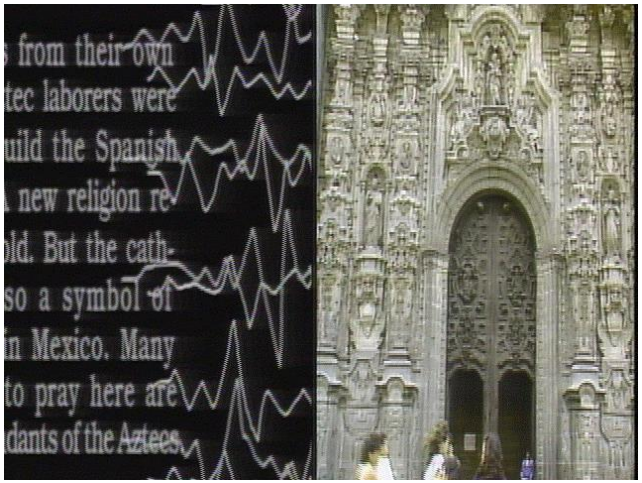
Rosalind Picard

Affective Computing *is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena.*



➢ The "Computational" Era (Late 1980s until 2000)

3) Multimedia Computing

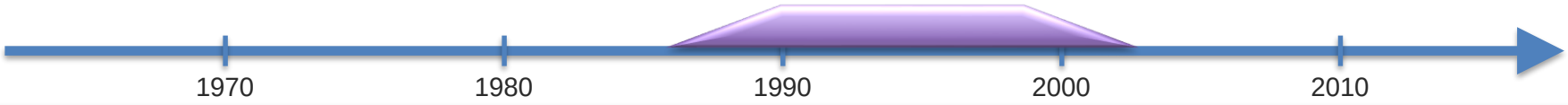


Carnegie Mellon University

informedia
digital video understanding

[1994-2010]

"The Informedia Digital Video Library Project automatically combines speech, image and natural language understanding to create a full-content searchable digital video library."

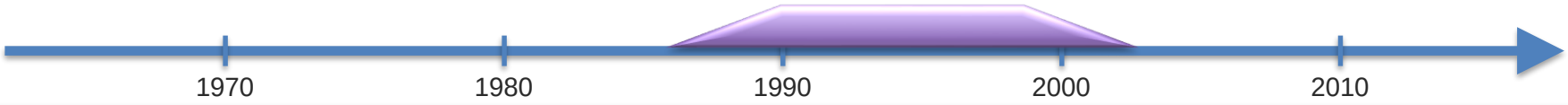


➢ The "Computational" Era (Late 1980s until 2000)

3) Multimedia Computing

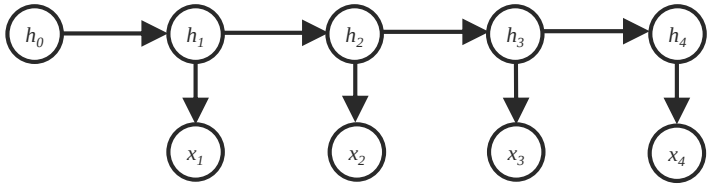
Multimedia content analysis

- **Shot-boundary detection (1991 -)**
  - Parsing a video into continuous camera shots
- **Still and dynamic video abstracts (1992 -)**
  - Making video browsable via representative frames (keyframes)
  - Generating short clips carrying the essence of the video content
- **High-level parsing (1997 -)**
  - Parsing a video into semantically meaningful segments
- **Automatic annotation (indexing) (1999 -)**
  - Detecting prespecified events/scenes/objects in video

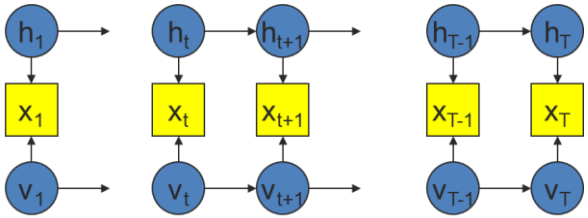


Multimodal Computation Models

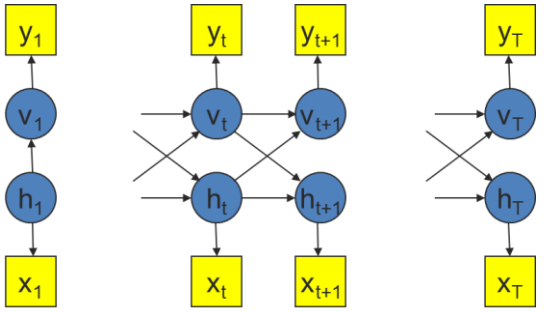
- Hidden Markov Models [1960s]

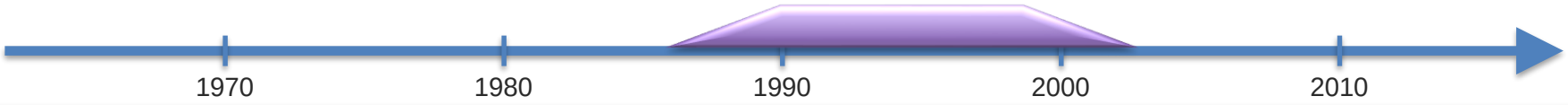


- ❑ Factorial Hidden Markov Models [1996]



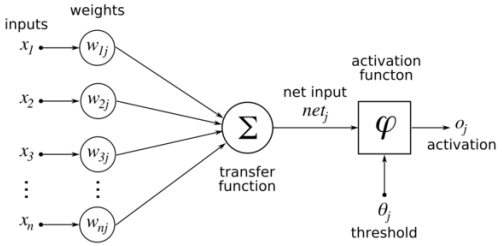
- ❑ Coupled Hidden Markov Models [1997]



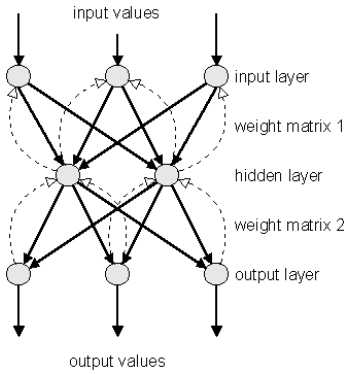


Multimodal Computation Models

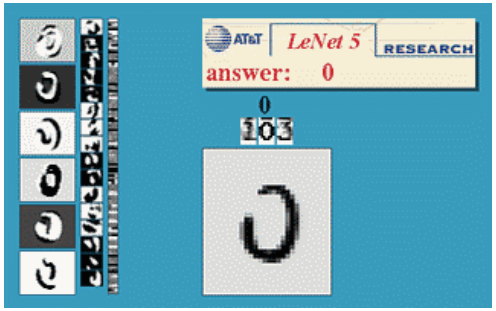
- Artificial Neural Networks [1940s]

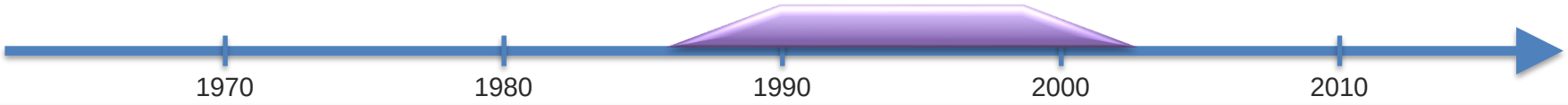


❑ Backpropagation [1975]



❑ Convolutional neural networks [1980s]



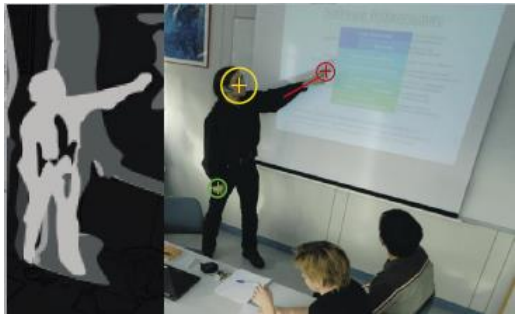


➢ The "Interaction" Era (2000s)

1) Modeling Human Multimodal Interaction

AMI Project [2001-2006, IDIAP]

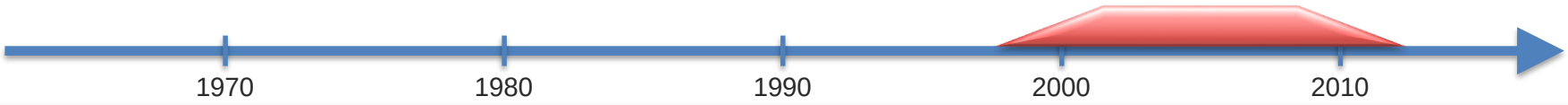
- 100+ hours of meeting recordings
- Fully synchronized audio-video
- Transcribed and annotated



CHIL Project [Alex Waibel]

- Computers in the Human Interaction Loop
- Multi-sensor multimodal processing
- Face-to-face interactions

❑ TRIVIA: Samy Bengio started at IDIAP working on AMI project



➢ The "Interaction" Era (2000s)

1) Modeling Human Multimodal Interaction

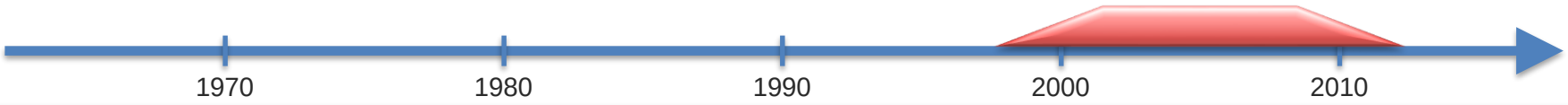
CALO Project [2003-2008, SRI]

- Cognitive Assistant that Learns and Organizes
- Personalized Assistant that Learns (PAL)
- Siri was a spinoff from this project

SSP Project [2008-2011, IDIAP]

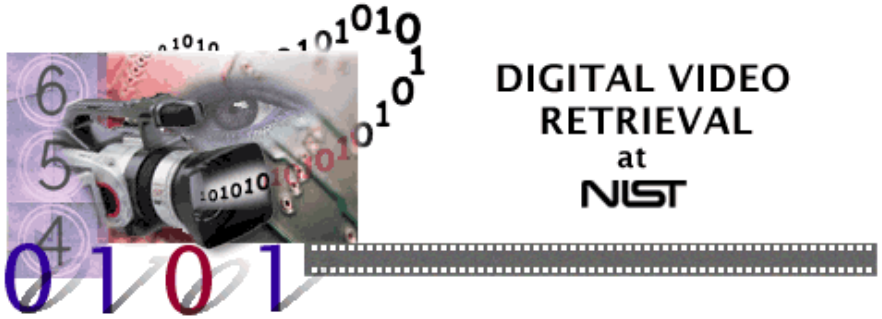
- Social Signal Processing
- First coined by Sandy Pentland in 2007
- Great dataset repository: http://sspnet.eu/

❑ TRIVIA: LP's PhD research was partially funded by CALO ☺



➢ The "Interaction" Era (2000s)

2) Multimedia Information Retrieval

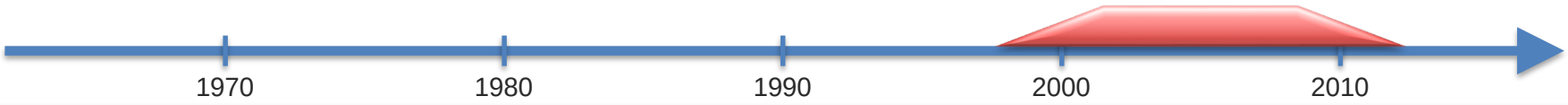


"Yearly competition to promote progress in content-based retrieval from digital video via open, metrics-*based evaluation"*

[Hosted by NIST, 2001-2016]

Research tasks and challenges:

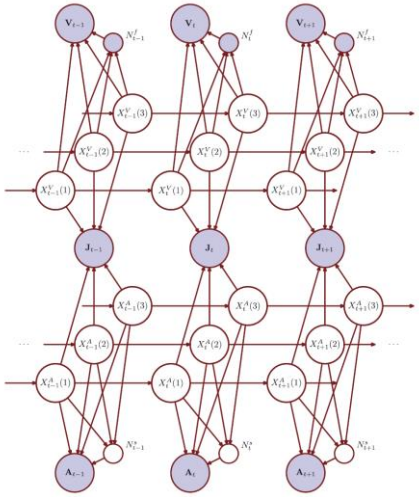
- Shot boundary, story segmentation, search
- "High-level feature extraction": semantic event detection
- Introduced in 2008: copy detection and surveillance events
- Introduced in 2010: Multimedia event detection (MED)

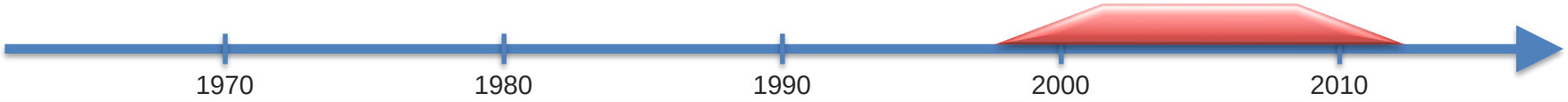


Multimodal Computational Models

- Dynamic Bayesian Networks
  - Kevin Murphy's PhD thesis and Matlab toolbox
  - Asynchronous HMM for multimodal [Samy Bengio, 2007]

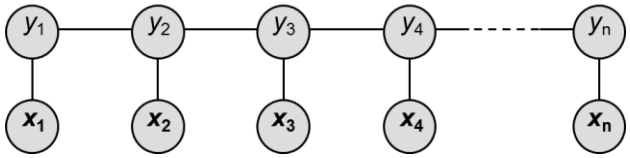
Audio-visual speech segmentation



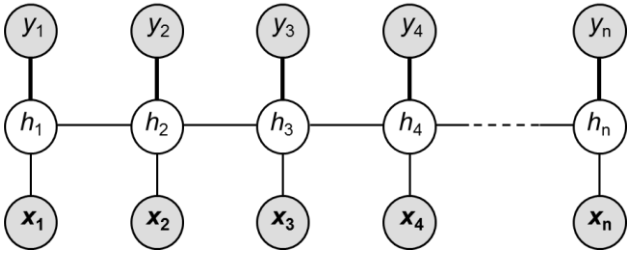


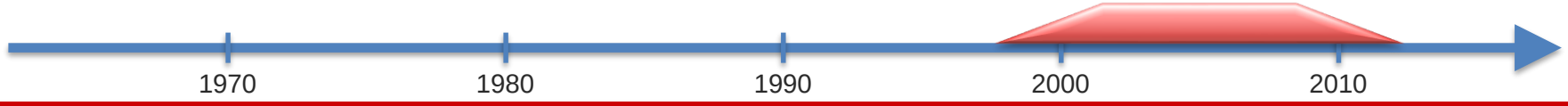
Multimodal Computational Models

- Discriminative sequential models
  - Conditional random fields [Lafferty et al., 2001]



  - Latent-dynamic CRF [Morency et al., 2007]





➢ The "deep learning" era (2010s until …)

Representation learning (a.k.a. deep learning)

- Multimodal deep learning [ICML 2011]
- Multimodal Learning with Deep Boltzmann Machines [NIPS 2012]
- Visual attention: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [ICML 2015]

Key enablers for multimodal research:

- New large-scale multimodal datasets
- Faster computer and GPUS
- High-level visual features
- "Dimensional" linguistic features

Our tutorial focuses on this era!

➢ The "deep learning" era (2010s until …)

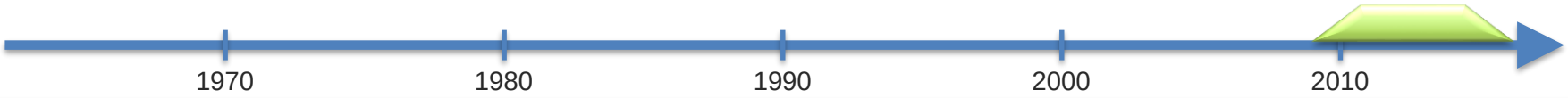
Many new challenges and multimodal corpora !!

Audio-Visual Emotion Challenge (AVEC, 2011-)

- Emotional dimension estimation
- Standardized training and test sets
- Based on the SEMAINE dataset

Emotion Recognition in the Wild Challenge (EmotiW 2013-)

- Emotional dimension estimation
- Standardized training and test sets
- Based on the SEMAINE dataset



➢ The "deep learning" era (2010s until …)

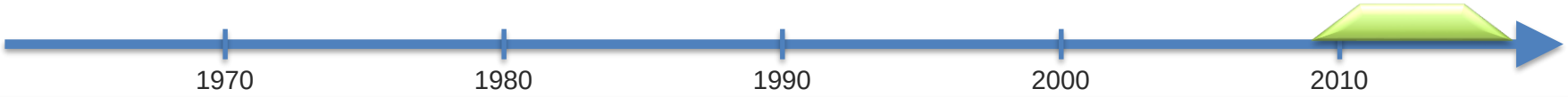
Renew of multimedia content analysis !

- Image captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

- Video description
- Visual Question-Answer



Real-World Tasks Tackled by Multimodal Research

- Affect recognition
  - Emotion
  - Persuasion
  - Personality traits
- Media description
  - Image captioning
  - Video captioning
  - Visual Question Answering
- Event recognition
  - Action recognition
  - Segmentation
- Multimedia information retrieval
  - Content based/Cross-media

Core Technical Challenges

Core Challenges in “Deep” Multimodal ML

Representation

Alignment

Fusion

Translation

Co-Learning

Multimodal Machine Learning: A Survey and Taxonomy

By Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

https://arxiv.org/abs/1705.09406

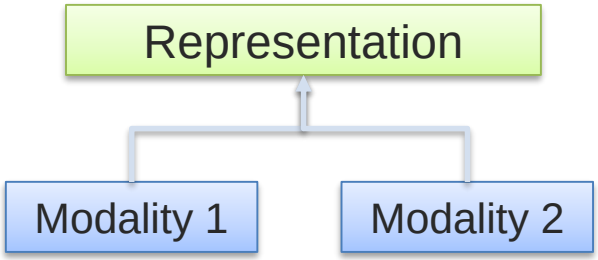
- ☑ **5 core challenges**
- ☑ **37 taxonomic classes**
- ☑ **253 referenced citations**

These challenges are non-exclusive.

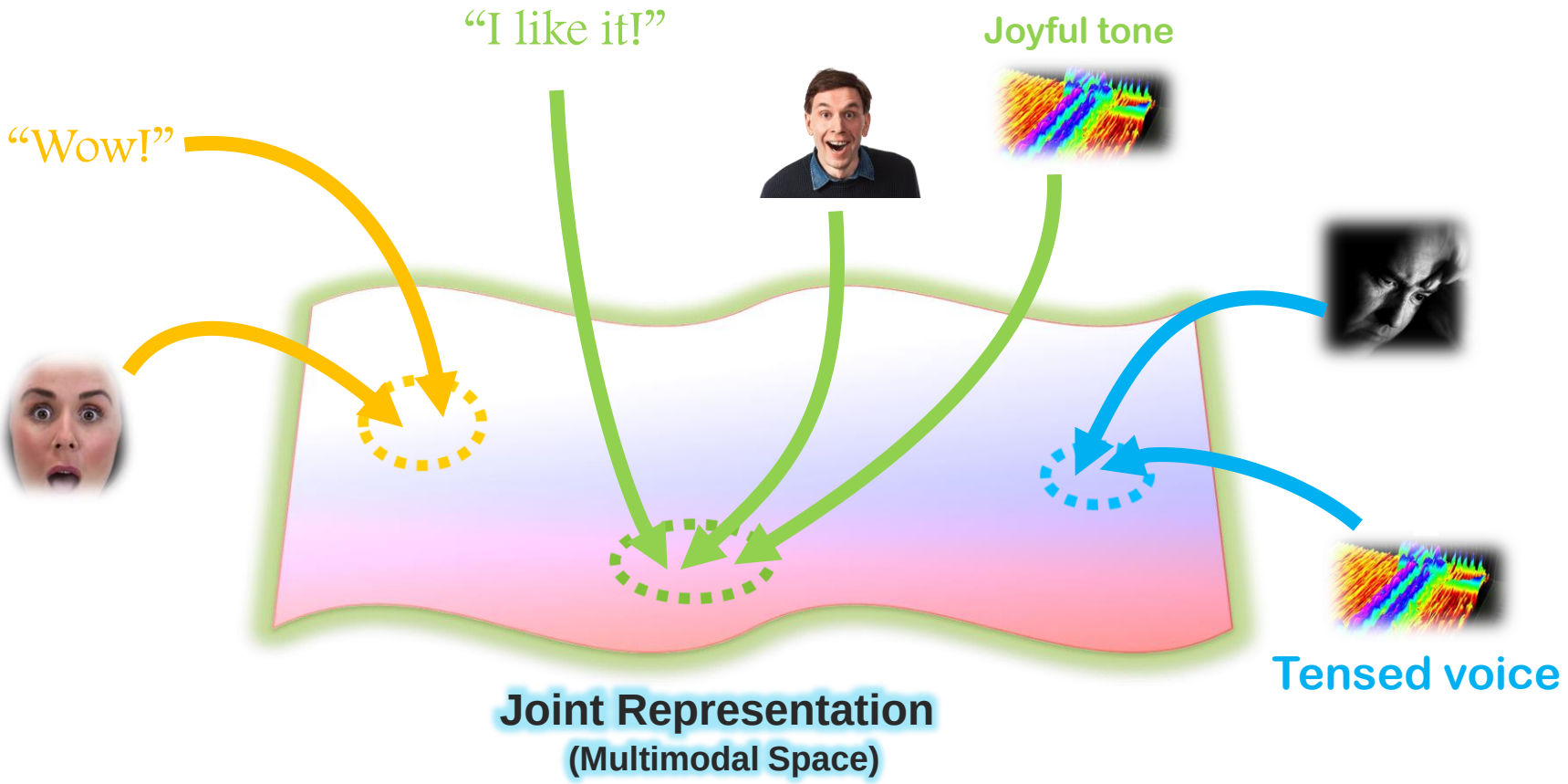
Core Challenge 1: Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

Ⓐ Joint representations:



Joint Multimodal Representation



Joint Multimodal Representations

Audio-visual speech recognition

[Ngiam et al., ICML 2011]

- Bimodal Deep Belief Network

Image captioning

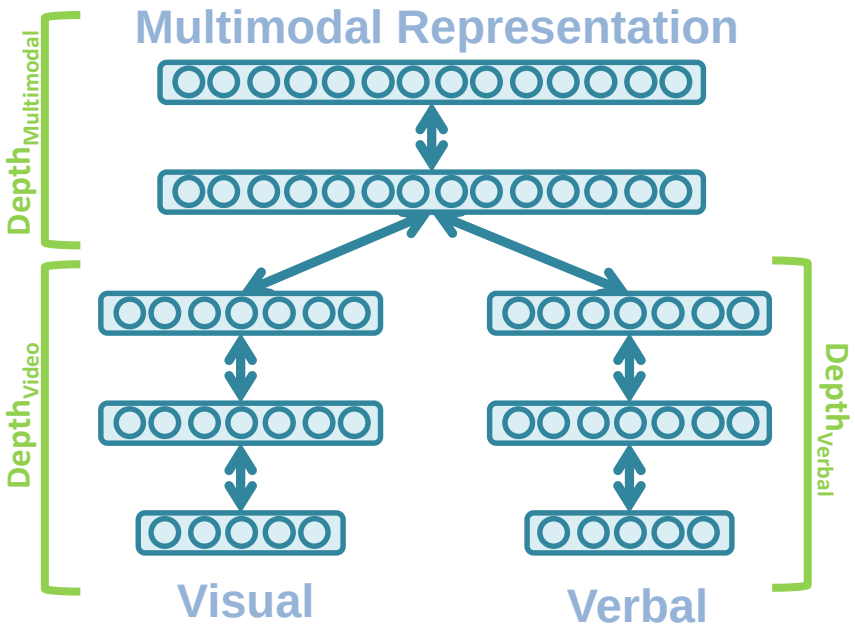
[Srivastava and Salahutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

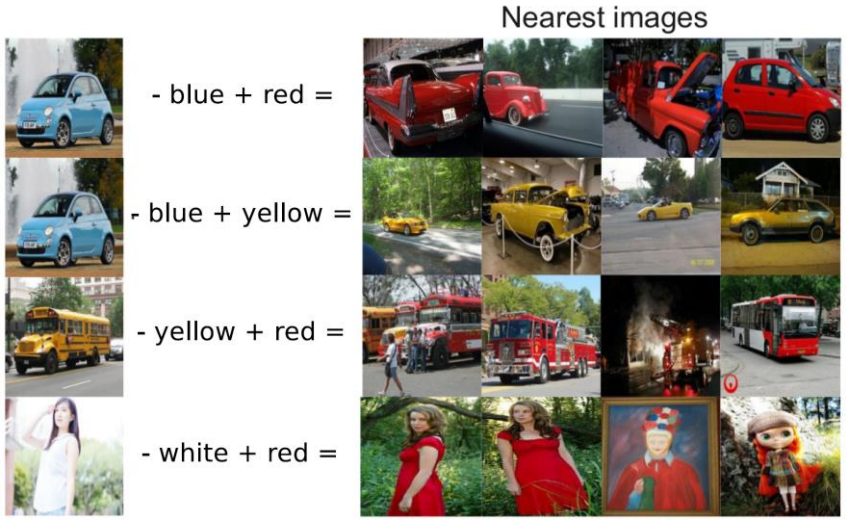
Audio-visual emotion recognition

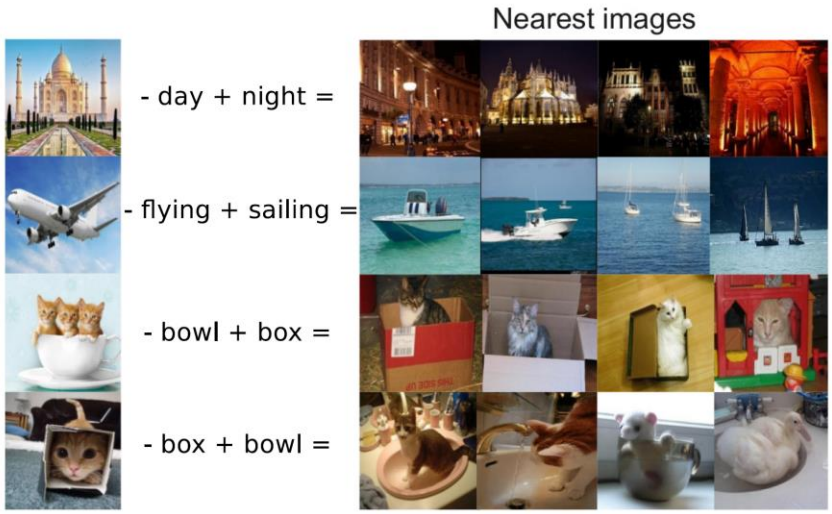
[Kim et al., ICASSP 2013]

- Deep Boltzmann Machine



Multimodal Vector Space Arithmetic



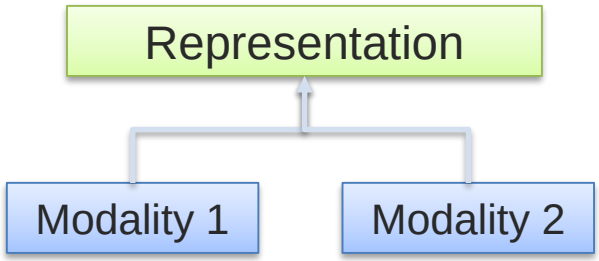


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

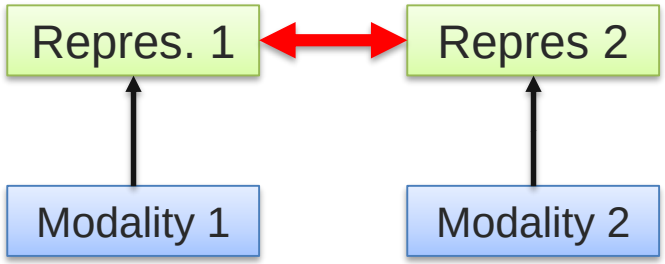
Core Challenge 1: Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

A Joint representations:



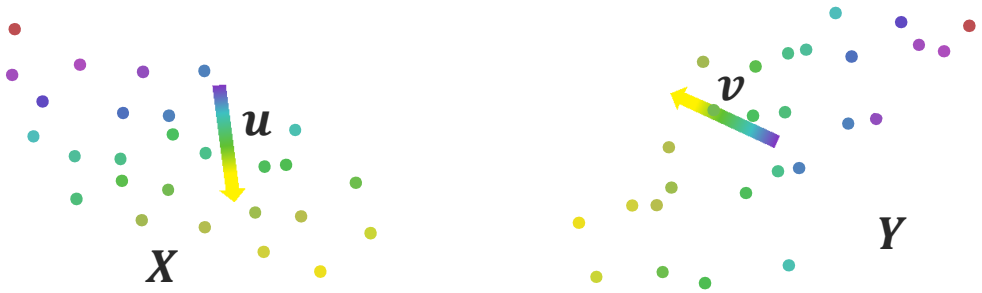
B Coordinated representations:

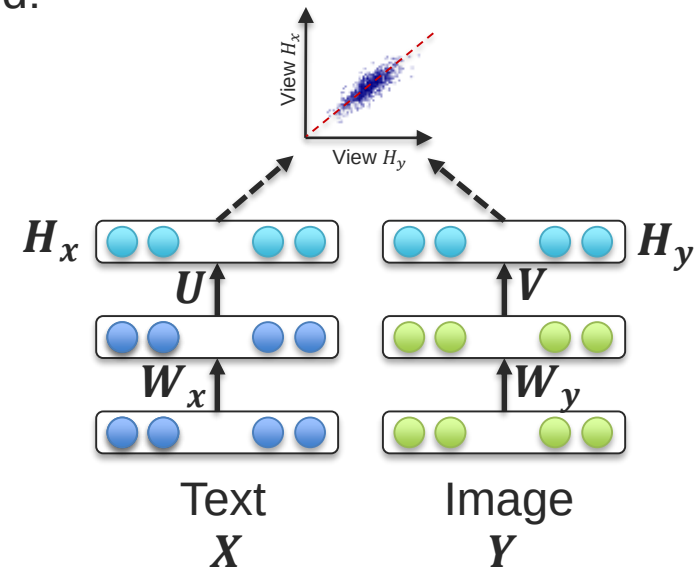


Coordinated Representation: Deep CCA

Learn linear projections that are maximally correlated:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr\left(\boldsymbol{u}^T\boldsymbol{X}, \boldsymbol{v}^T\boldsymbol{Y}\right)$$

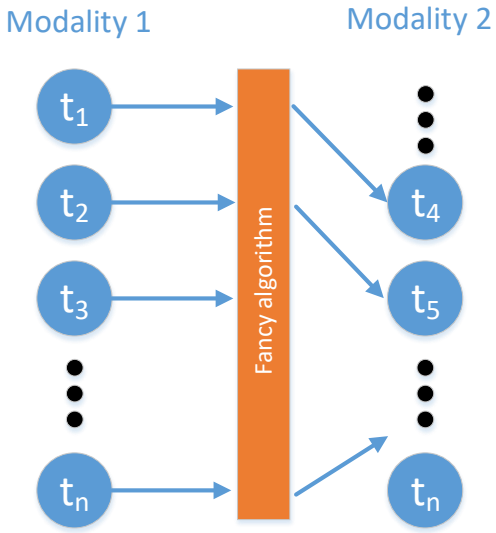




Andrew et al., ICML 2013

Core Challenge 2: Alignment

Definition: Identify the direct relations between (sub)elements from two or more different modalities.



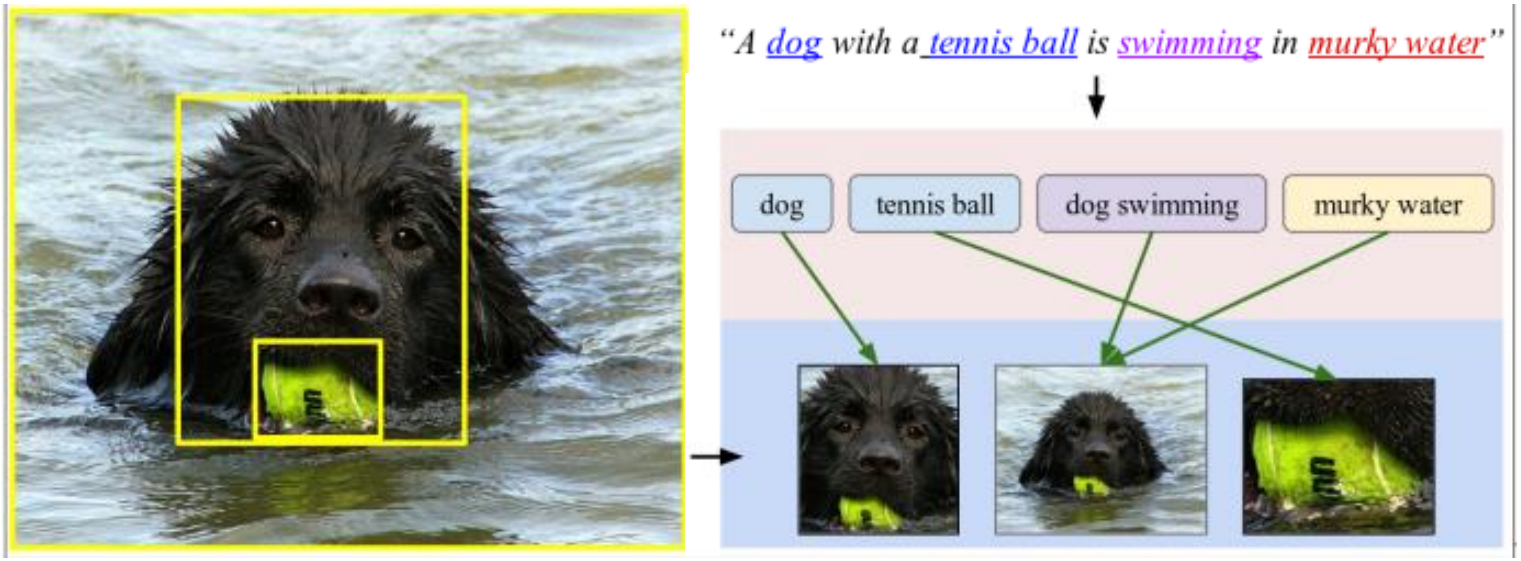
A Explicit Alignment

The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

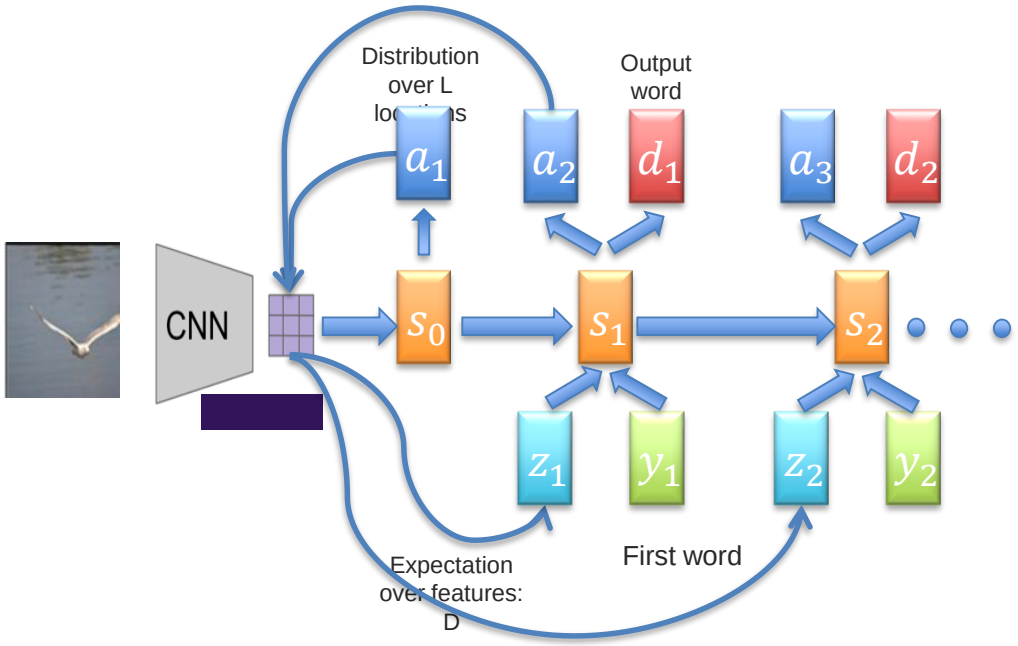
Uses internally latent alignment of modalities in order to better solve a different problem

Implicit Alignment



Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, https://arxiv.org/pdf/1406.5679.pdf

Attention Models for Image Captioning



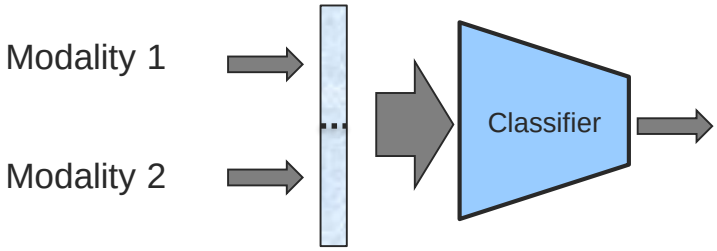
Xu et.al., ICML 2015

Core Challenge 3: Fusion

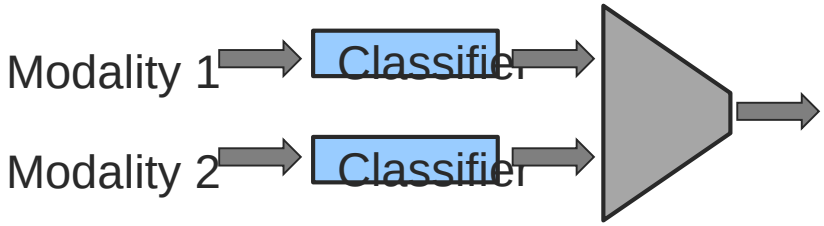
Definition: To join information from two or more modalities to perform a prediction task.

A Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion



Core Challenge 3: Fusion

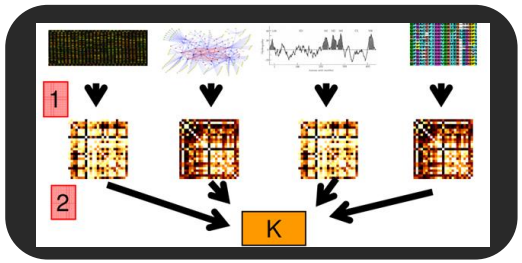
Definition: To join information from two or more modalities to perform a prediction task.

B **Model-Based (Intermediate) Approaches**

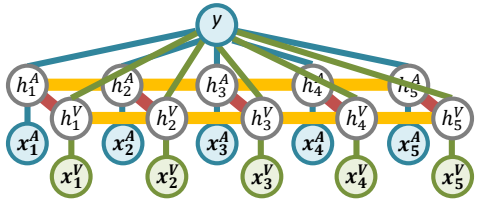
1) **Deep neural networks**

2) **Kernel-based methods**

3) **Graphical models**



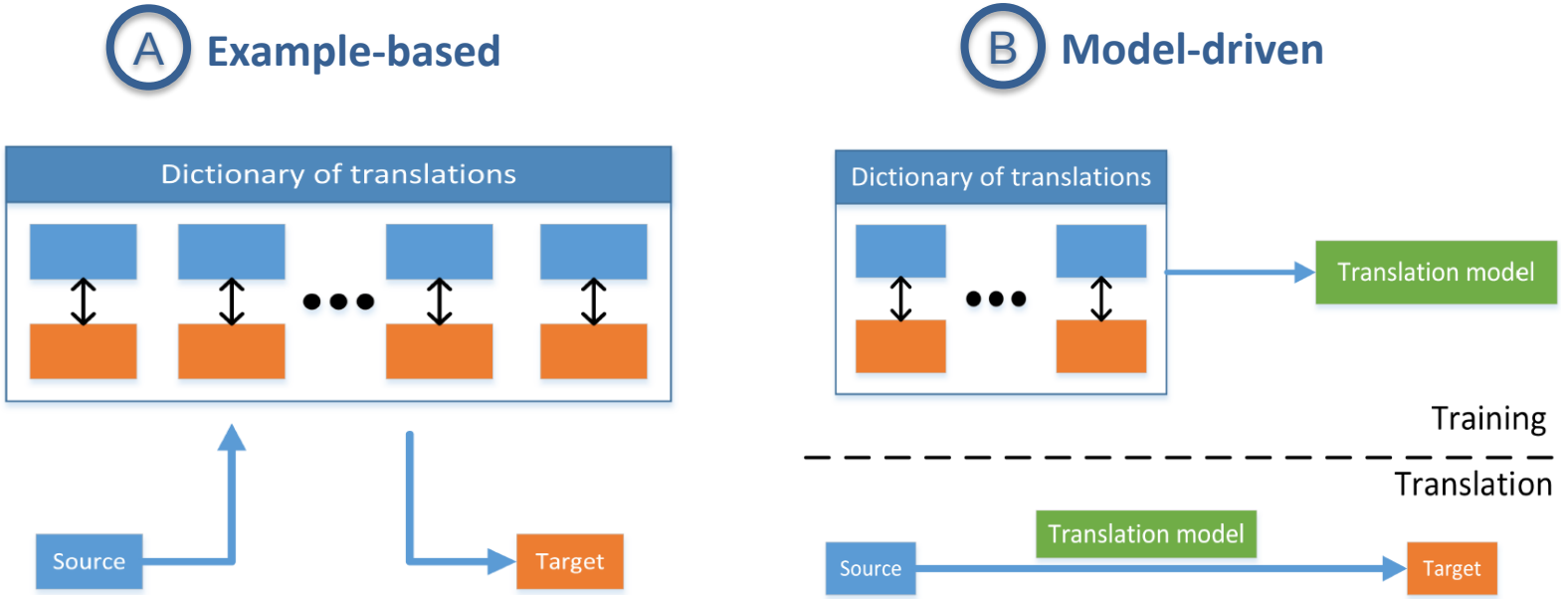
Multiple kernel learning



Multi-View Hidden CRF

Core Challenge 4: Translation

Definition: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.



Core Challenge 4 – Translation

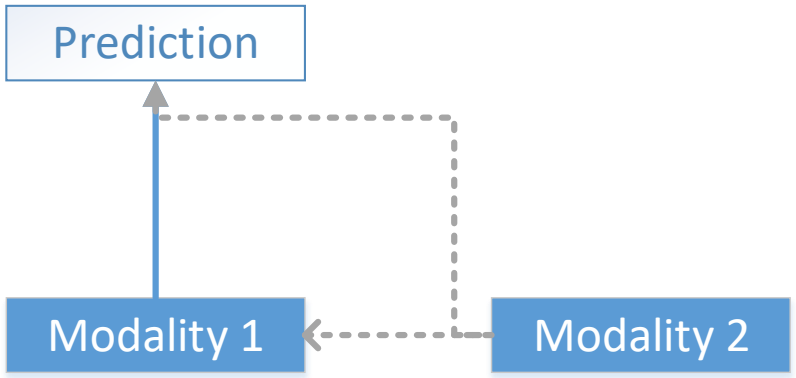
Visual gestures
(both speaker and listener gestures)

Transcriptions
+
Audio streams

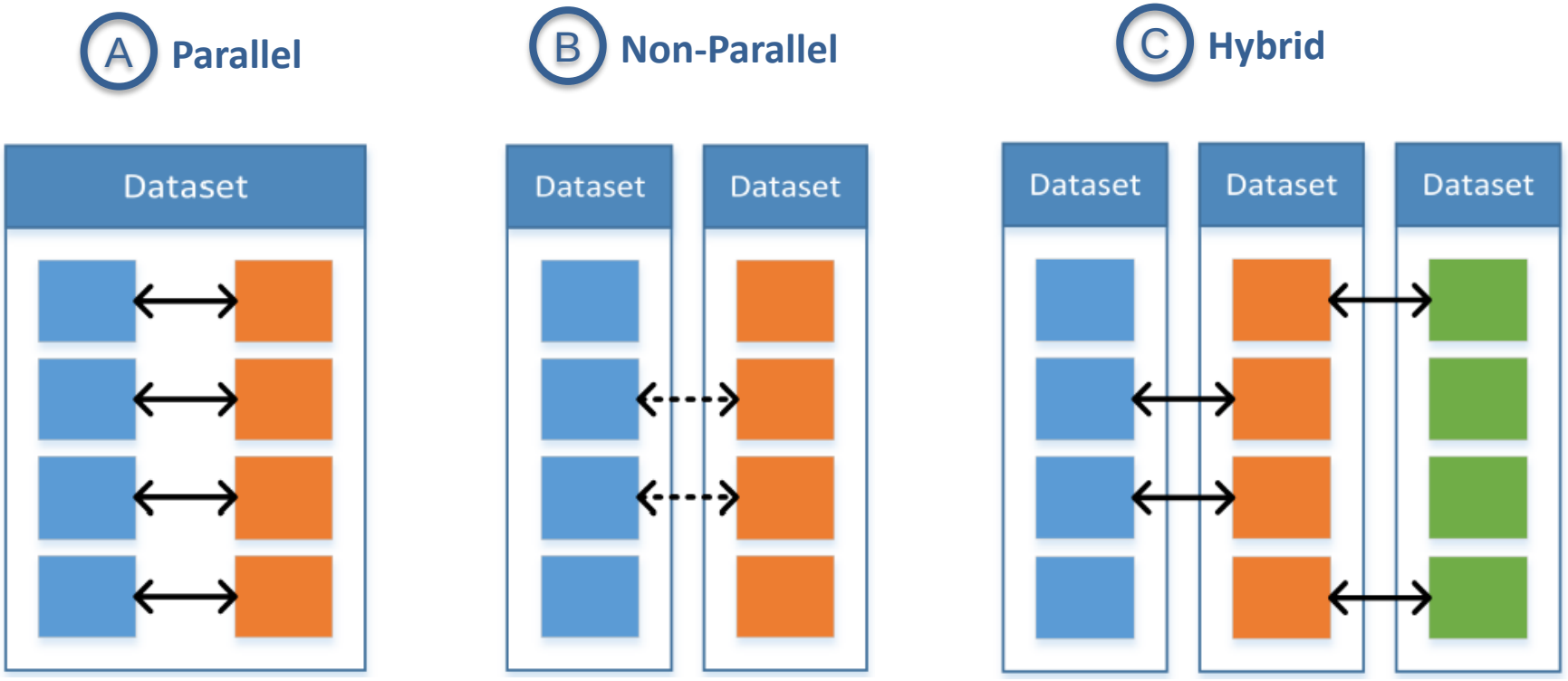
Marsella et al., Virtual character performance from speech, SIGGRAPH/Eurographics Symposium on Computer Animation, 2013

Core Challenge 5: Co-Learning

Definition: Transfer knowledge between modalities, including their representations and predictive models.



Core Challenge 5: Co-Learning



Taxonomy of Multimodal Research [https://arxiv.org/abs/1705.09406]

Representation

- Joint
  - Neural networks
  - Graphical models
  - Sequential
- Coordinated
  - Similarity
  - Structured

Translation

- Example-based
  - Retrieval
  - Combination
- Model-based
  - Grammar-based
  - Encoder-decoder
  - Online prediction

Alignment

- Explicit
  - Unsupervised
  - Supervised
- Implicit
  - Graphical models
  - Neural networks

Fusion

- Model agnostic
  - Early fusion
  - Late fusion
  - Hybrid fusion
- Model-based
  - Kernel-based
  - Graphical models
  - Neural networks

Co-learning

- Parallel data
  - Co-training
  - Transfer learning
- Non-parallel data
  - Zero-shot learning
  - Concept grounding
  - Transfer learning
- Hybrid data
  - Bridging

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

Multimodal Applications

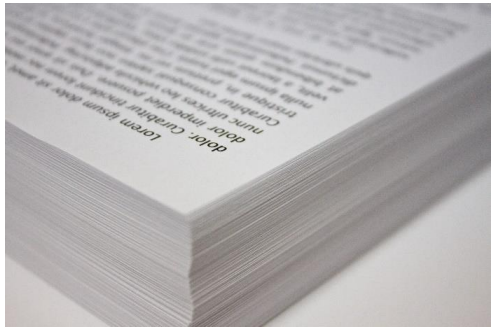
[https://arxiv.org/abs/1705.09406]

	Challenges				
Applications	Representation	Translation	Fusion	Alignment	Co-learning
Speech Recognition and Synthesis					
Audio-visual Speech Recognition	✓		✓	✓	✓
(Visual) Speech Synthesis	✓	✓			
Event Detection					
Action Classification	✓		✓		✓
Multimedia Event Detection	✓		✓		✓
Emotion and Affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media Description					
Image Description	✓	✓		✓	✓
Video Description	✓	✓	✓	✓	✓
Visual Question-Answering	✓		✓	✓	✓
Media Summarization	✓	✓	✓		
Multimedia Retrieval					
Cross Modal retrieval	✓	✓		✓	✓
Cross Modal hashing	✓				✓

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

Course Syllabus

Three Course Learning Paradigms



Research assignments and course participation
(40% of your grade)

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

Course project assignments
(60% of your grade)

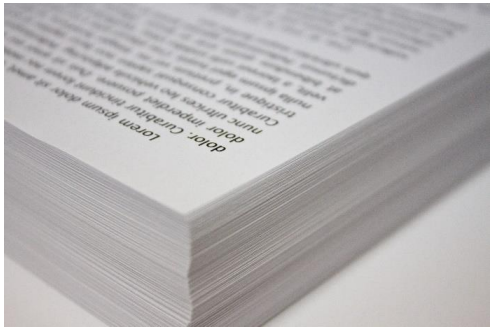
Course lectures
(including guest lectures)

Course Structure

Tuesdays

Course lectures

Thursdays



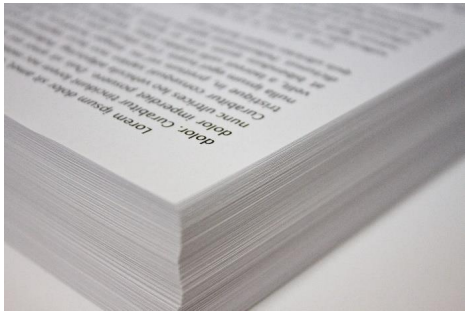
Group discussions

** with some exceptions

Course Recommendations and Requirements

1. Ready to read at least 9 papers this semester !
   - 9 research papers as part of the weekly reading assignments
   - Asked to answer research questions about papers
2. Already taken a machine learning course
   - Strongly recommended for students to have taken an introduction machine learning course
   - 10-401, 10-601, 10-701, 11-663, 11-441, 11-641 or 11-741
3. Motivated to produce a high-quality course project
   - Three course project assignments
   - Designed to enhance state-of-the-art algorithms

Course Grades



$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

- Discussion and lecture participation 20%
- Reading assignments 20%
- First project assignment
  - Report and presentation 15%
- Mid-term project assignment
  - Report and presentation 15%
- Final project assignment
  - Report and presentation 30%

Course Project

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

- Pre-proposal (in 2 weeks)
  - Define your dataset and research task
- First project assignment (in 5 weeks)
  - Experiment with unimodal representations
  - Explore/discuss simple baseline model(s)
- Midterm project assignment (in 10 weeks)
  - Implement and evaluate state-of-the-art model(s)
  - Discuss new multimodal model(s)
- Final project assignment (in 14 weeks)
  - Implement and evaluate new multimodal model(s)
  - Discuss future directions

Course Project Guidelines

- Dataset should have at least two modalities:
  - Natural language and visual/images
- Teams of 3 or 4 students are preferred
  - No individual projects
- The project should explore algorithmic novelty
- Possible venues for your final report:
  - NAACL 2018, ACL 2018, IJCAI 2017, ICML 2018
- We will discuss on Thursday about project ideas

Process for Selecting your Course Project

- Thursday 8/31: Lecture describing available multimodal datasets and research topics
- Monday 9/4: Submit a short paragraph listing your top 3 choices
- Tuesday 9/5: in later part of the lecture, we will do a “speed dating” session to meet teammates
- Sunday 9/17: pre-proposals are due. You should have selected your teammates and dataset.

Lecture Schedule

Classes	Lectures
Week 1 8/29 & 8/31	**Course introduction** • Research and technical challenges • Multimodal applications and datasets
Week 2 9/5 & 9/7	**Basic mathematical concepts** • Language, image and audio representat • Loss functions and basic neural networ
Week 3 9/12 & 9/14	**Convolutional neural networks and optimi** • Neural network optimization • Convolutional neural networks
Week 4 9/19 & 9/21	**Recurrent neural networks** • Backpropagation Through Time • Gated networks and LSTM

Thursday 8/31 in NSF 3305

Project preferences due on Monday night

Pre-proposal due on Sunday 9/17

Lecture Schedule

Classes	Lectures
Week 5 9/26 & 9/28	**Multimodal representation learning** • Multimodal auto-encoders • Multimodal joint representations
Week 6 10/3 & 10/5	***First project assignment - Presentat***
Week 7 10/10 & 10/12	**Multivariate statistics and coordinated representations** • Deep canonical correlation analysis • Non-negative matrix factorization
Week 8 10/17 & 10/19	**Multimodal alignment and attention models** • Explicit alignment and dynamic time warping • Implicit alignment and attention models

Thursday in NSH 1305, 5pm-6:20pm.
Proposal due: 10/8.

Lecture Schedule

Classes	Lectures
Week 9 10/24 – 10/26	**Multimodal optimization** • Practical deep model optimization • Variational approaches
Week 10 10/31 & 11/2	**Probabilistic graphical models** • Boltzmann distribution and CRFs • Continuous and fully-connected CRFs
Week 11 11/7 & 11/9	***Mid-term project assignment - Pre***
Week 12 11/14 & 11/16	**Multimodal fusion and new directions** • Multi-kernel learning and fusion • New directions in multimodal machine learning

Thursday in GHC-6115.
Midterm due on 11/12.

Lecture Schedule

Classes	Lectures
Week 13 11/21 & 11/23	***Thanksgiving week (+ Project preparation)***
Week 14 11/28 & 11/30	**Advanced multimodal representations** • Image and video description • Guest lecture
Week 15 **12/4** & 12/5 ** Final **	***Final project assignment - Present***

Monday in GHC-6115.
Final project due: 12/10.