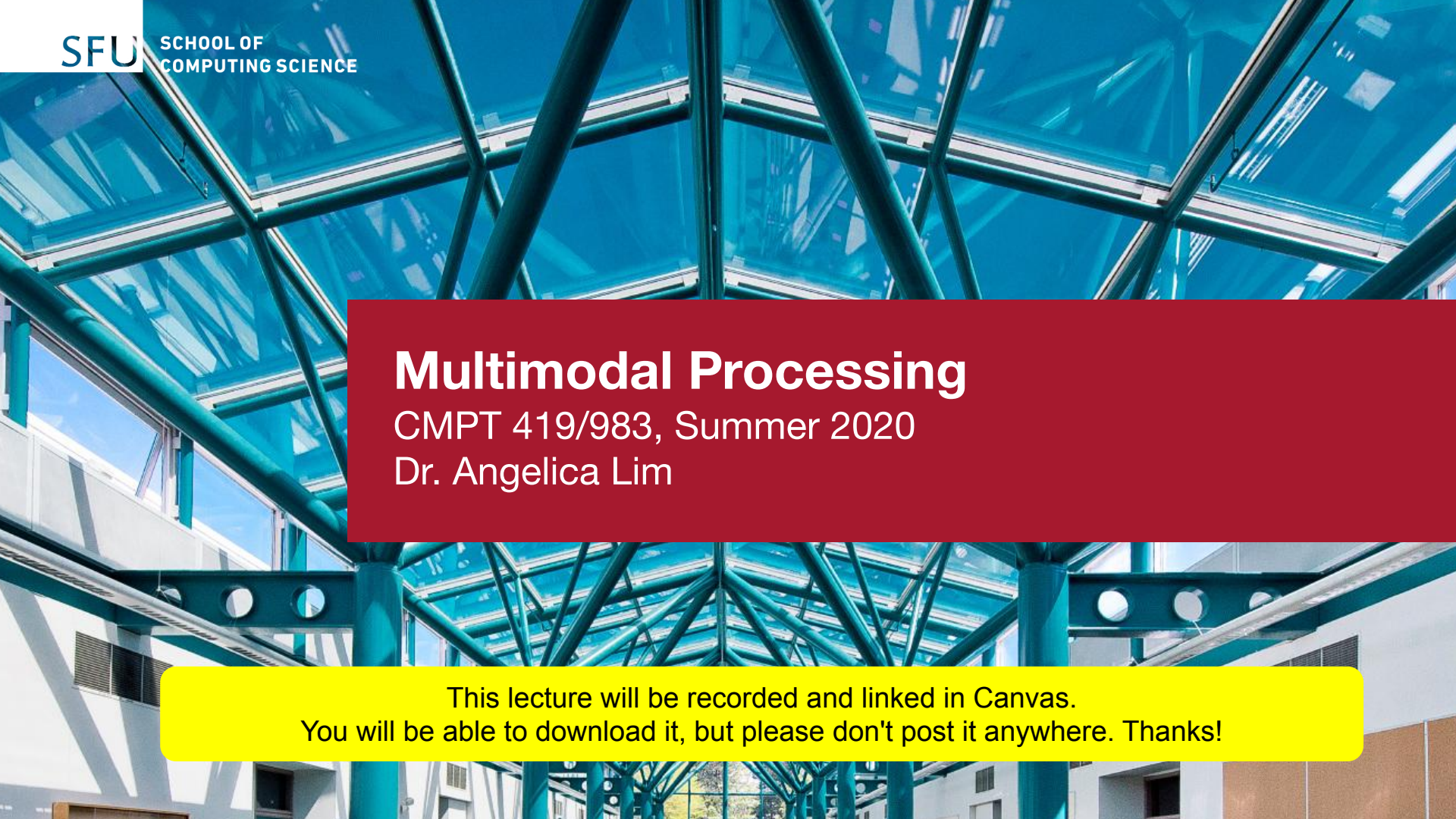


SFU SCHOOL OF COMPUTING SCIENCE

Multimodal Processing

CMPT 419/983, Summer 2020

Dr. Angelica Lim

This lecture will be recorded and linked in Canvas.
You will be able to download it, but please don't post it anywhere. Thanks!

Today's Topics

- K-Nearest Neighbours
- Multimodal Introduction
- Early fusion, late fusion, hybrid fusion, event-based fusion
- Related research by TA Payam Jome Yazdian

Coming up next week...

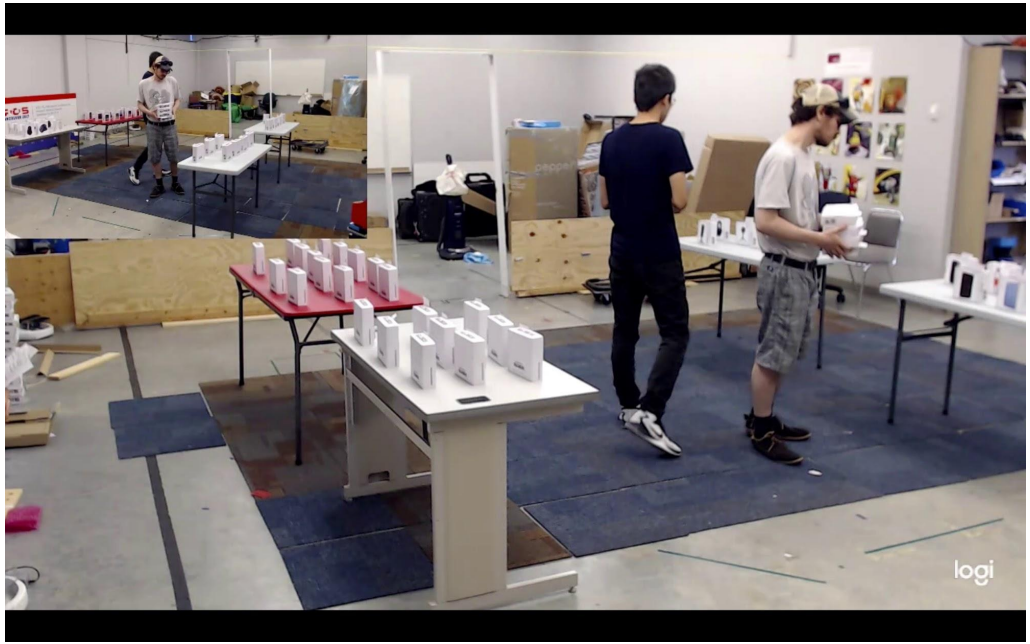
Some project ideas

- Confusion prediction*
- Object-related affect*
- Unity system interaction (more info on Tuesday)

*To apply for these projects, send us a Piazza DM by July 1.

Confusion prediction

Special TA: PhD student Zhitian Zhang



At which timestamp do you think the person needs help in a shopping scenario?

Annotation Task:

- Annotated each frame on a scale of 0~5.
- 0 represents the person does not need any help (not confused).
- And 5 represents the person needs immediate help (confused).

Human Confusion Prediction:

- Train a neural network model to predict the confusion level of a person.

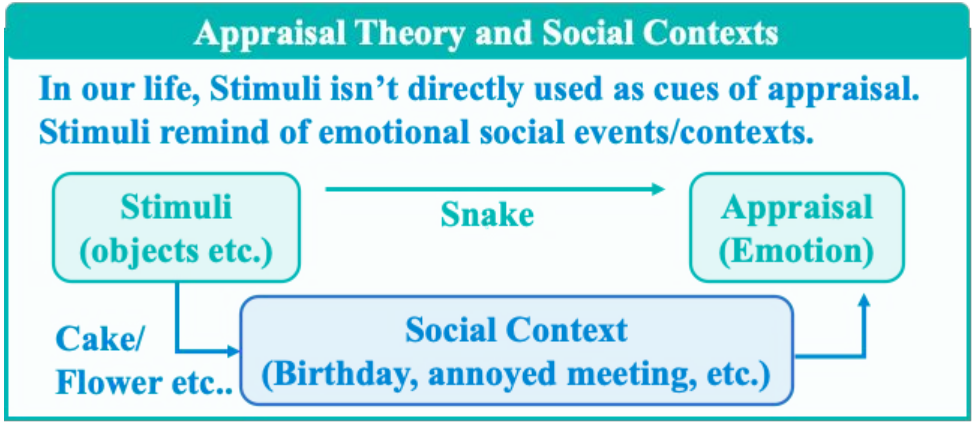
Object-related Affect

Annotation Task

Story making/Describing person's thinking

Please include below.

- **What emotion do you appraise as person's emotion?**
- **What objects do you think Stimuli (cues) of emotion?**
- **Are Human's gazed and handed information important ?**

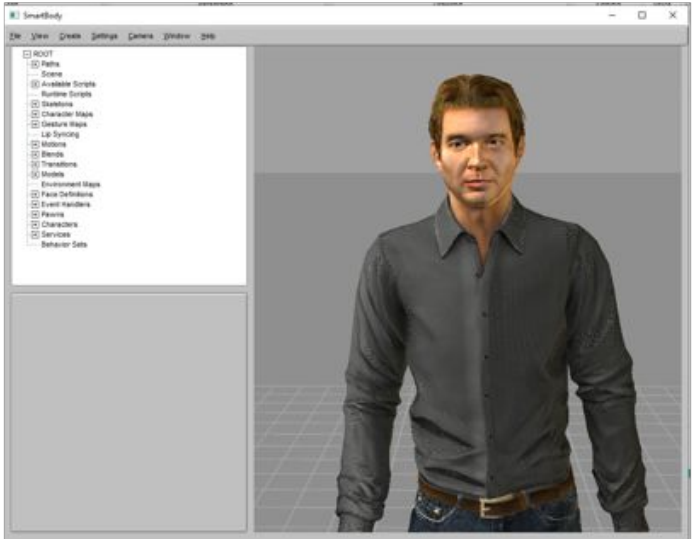


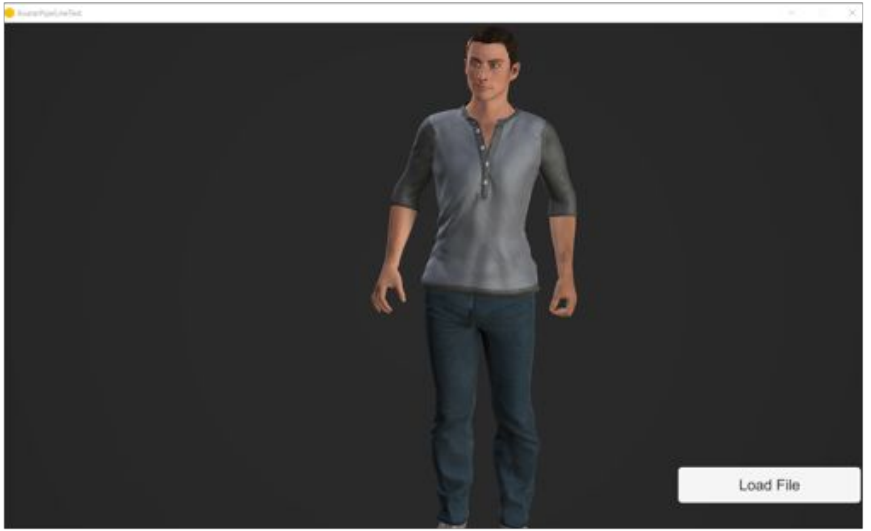
Ex:
The boy is gazing at a cake.
The situation seems his birthday and,
It seems he is celebrated by other people.
Thus, It seems the boy is happy.

Behavior Realizer

SMARTBODY : http://smartbody.ict.usc.edu/

iVizLab Unity Behavior Realizer: https://drive.google.com/file/d/1ec-D4QNn0VDoC_VCNxlaB-Rrd3UBKofZ/view?usp=sharing **more avatars available for this version





The Great Time Series Bake-Off

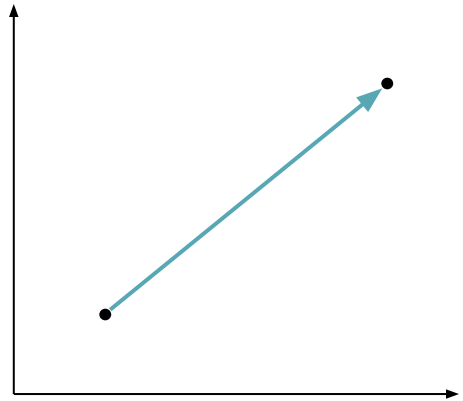
- Multiple rigorous studies show that for time series classification, **Nearest Neighbour DTW** is very hard to beat

- Where NN-DTW can be beaten, it's typically by a small margin at a large cost in code complexity, time and space overhead.

*This paper conducts 35 million experiments, on 85 datasets, with dozens of rival methods
The Great Time Series Classification Bake Off: An Experimental Evaluation of Recently Proposed Algorithms (Bagnall, 2016).

Nearest Neighbors

- Relies on the assumption that similar examples should be classified the same
- Examples represented as points in an n-dimensional space
- Similarity is dependent on the distance to other examples on the graph
- Simplest version: Give your new example the same classification of the “closest” training example



Nearest Neighbors Steps

1. Plot labelled training set examples using selected features

2. Given a new, unseen example, compute the distance from the new example to all our **n** training examples O(n)

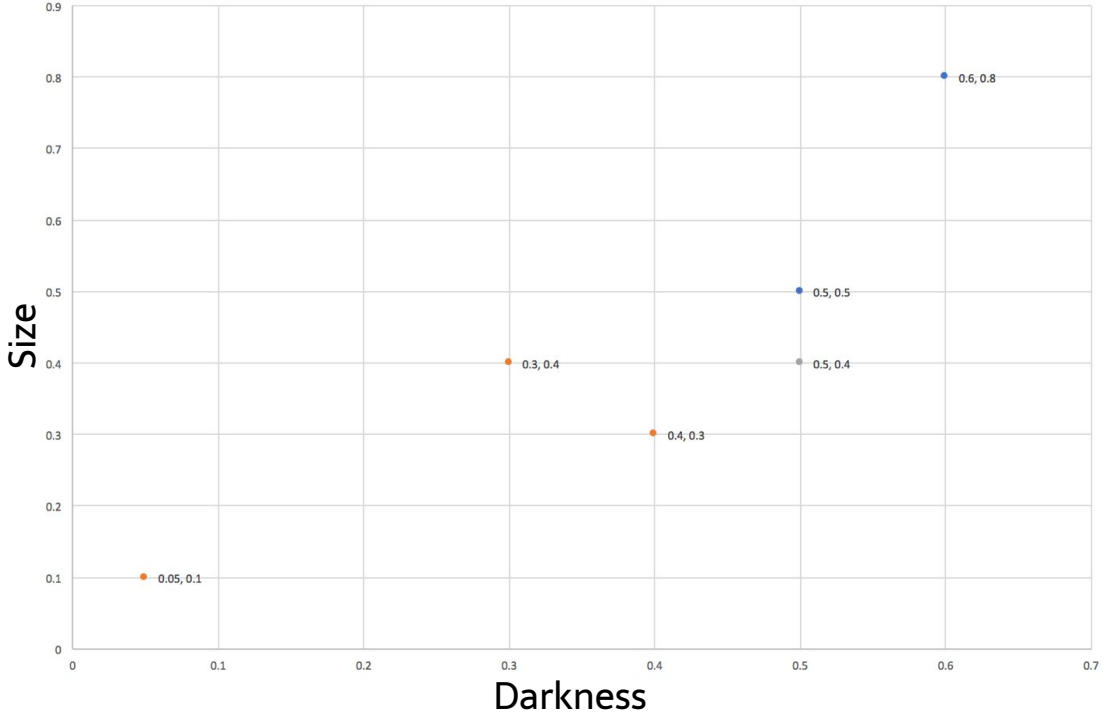
3. Assign the example a classification based on the classification of the closest example

Nearest Neighbors Example

New example

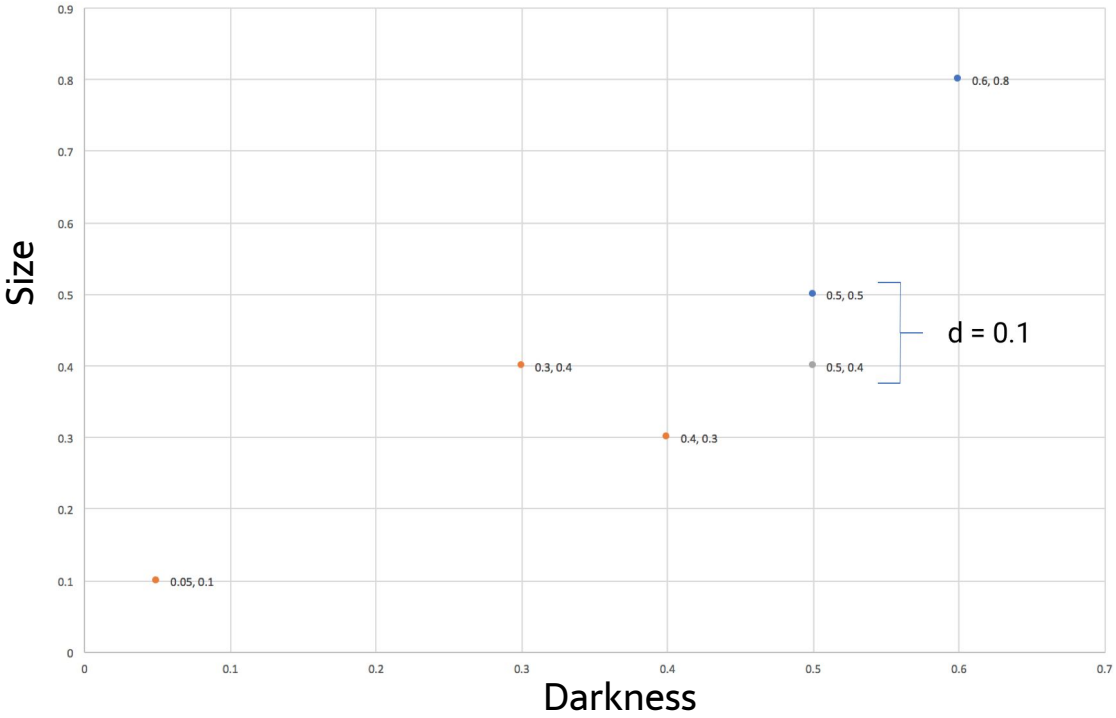
Non-cancerous

Cancerous



Nearest Neighbors Example

- New example
- Non-cancerous
- Cancerous



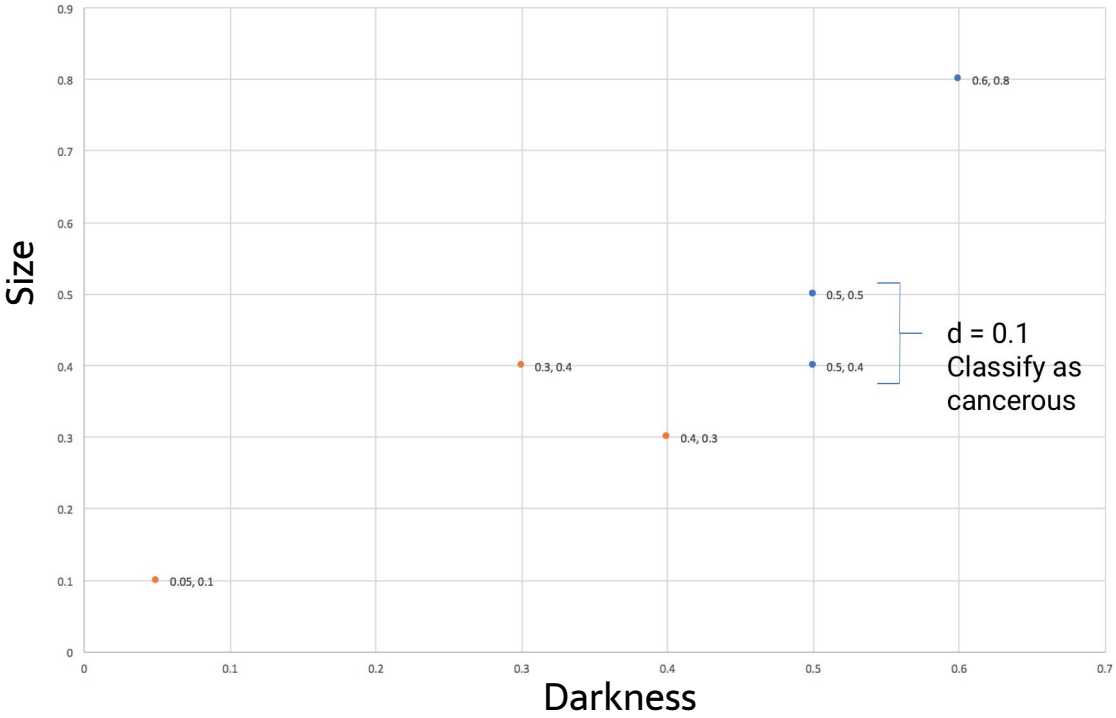
Distance metric (e.g. Euclidean distance) $\mathrm{d} = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$

Nearest Neighbors Example

New example

Non-cancerous

Cancerous



Distance metric (e.g. Euclidean distance) $\mathrm{d} = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$

K-Nearest Neighbors (K-NN)

- An extension of Nearest Neighbors
- **Classification**: Use most common label of the *k* nearest examples
- **Regression**: Use the mean of the *k* nearest examples
- If there is a tie, base the classification off the *k - 1* nearest examples

Nearest Neighbors Summary

- KNN is a simple algorithm that classifies based on distance and labels of neighbor
- 1-NN may suffice, and completes in O(1)
- Sensitive to local distribution and noise

Context

Disgust face on different bodies

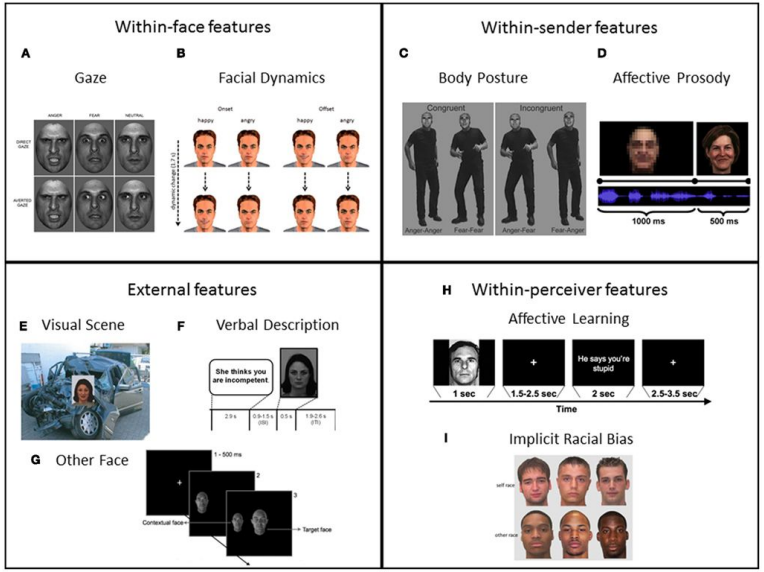
https://www.pnas.org/content/pnas/102/45/16518.full.pdf

Faces in context: a review and systematization of contextual influences on affective face processing

Matthias J. Wieser[1]* **and Tobias Brosch**[2]

[1]Department of Psychology, University of Würzburg, Würzburg, Germany
[2]Department of Psychology, University of Geneva, Geneva, Switzerland

Facial expressions are of eminent importance for social interaction as they convey information about other individuals' emotions and social intentions. According to the predominant "basic emotion" approach, the perception of emotion in faces is based on the rapid, automatic categorization of prototypical, universal expressions. Consequently, the perception of facial expressions has typically been investigated using isolated, de-contextualized, static pictures of facial expressions that maximize the distinction between categories. However, in everyday life, an individual's face is not perceived in isolation, but almost always appears within a situational context, which may arise from other people, the physical environment surrounding the face, as well as multichannel information from the sender. Furthermore, situational context may be provided by the perceiver, including already present social information gained from affective learning and implicit processing biases such as race bias. Thus, the perception of facial expressions is presumably always



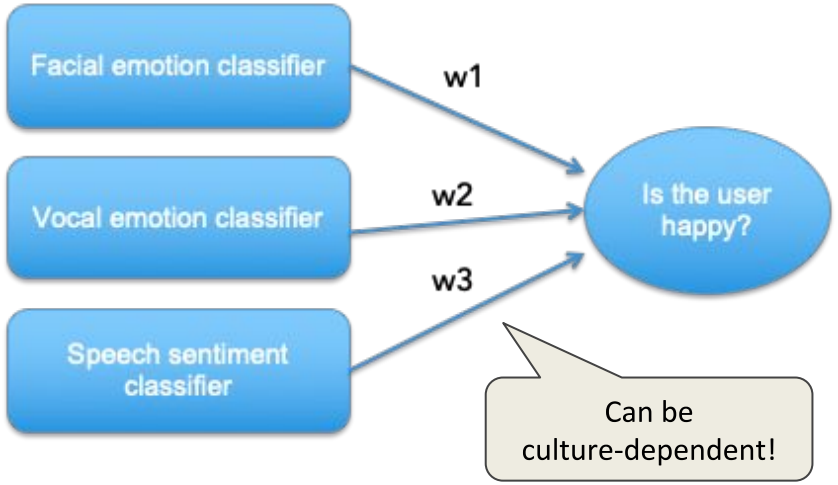
https://link.springer.com/article/10.1007/s10548-009-0099-0

McGurk Effect

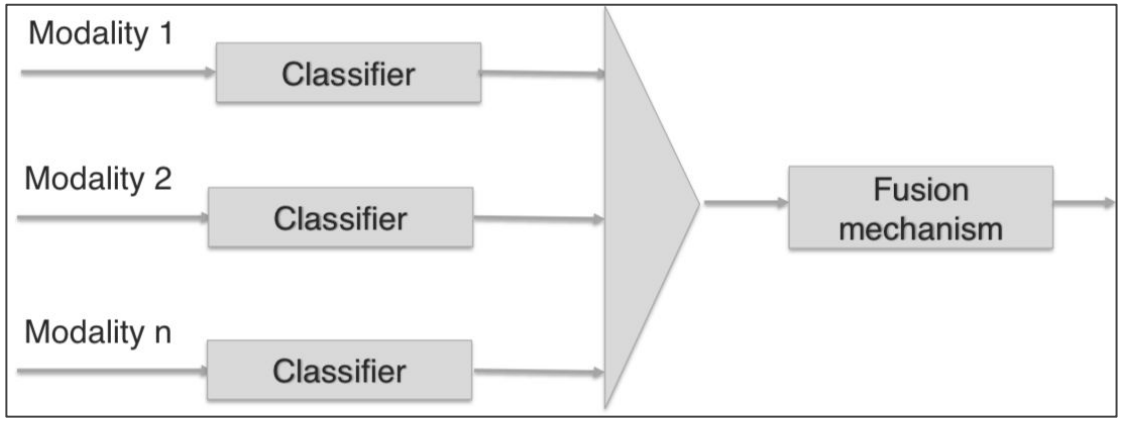
Multimodal Fusion

Multimodal Fusion

Simple weighted fusion:



Late Fusion



Optimizing Weights for Weighted Sum

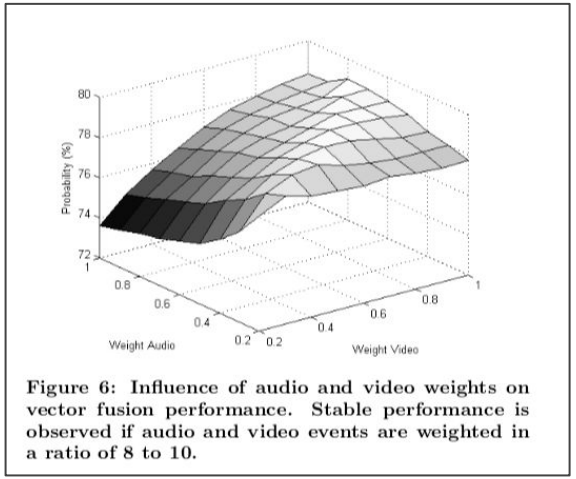
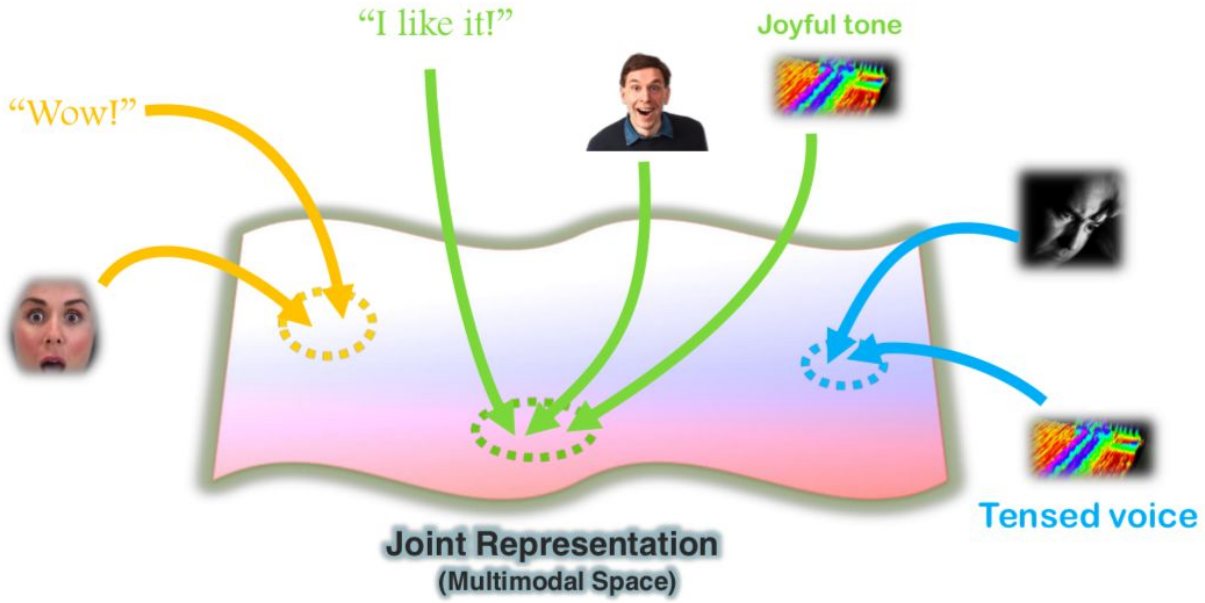


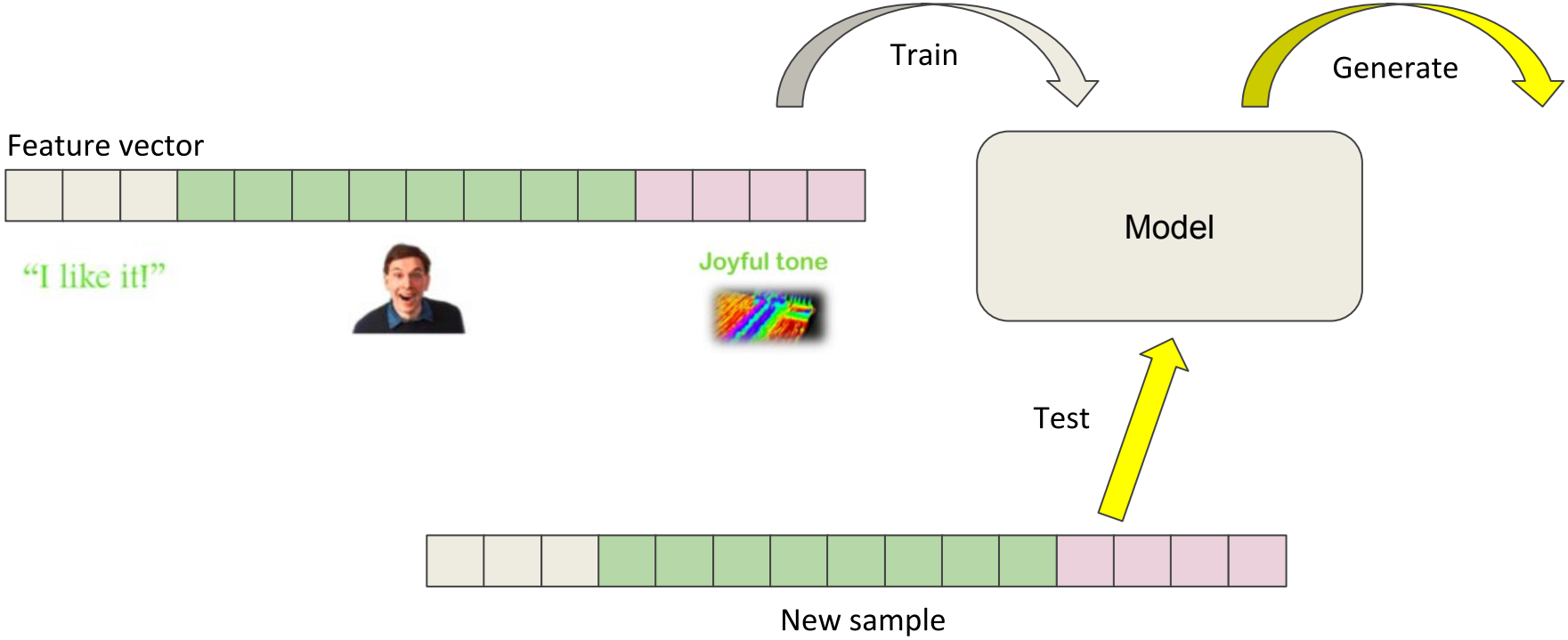
Figure 6: Influence of audio and video weights on vector fusion performance. Stable performance is observed if audio and video events are weighted in a ratio of 8 to 10.

- Train a unimodal predictor and a multimodal fusion one
- Requires multiple training stages
- Does not model low-level interactions between modalities
- Fusion mechanism can be voting, weighted sum or an ML approach

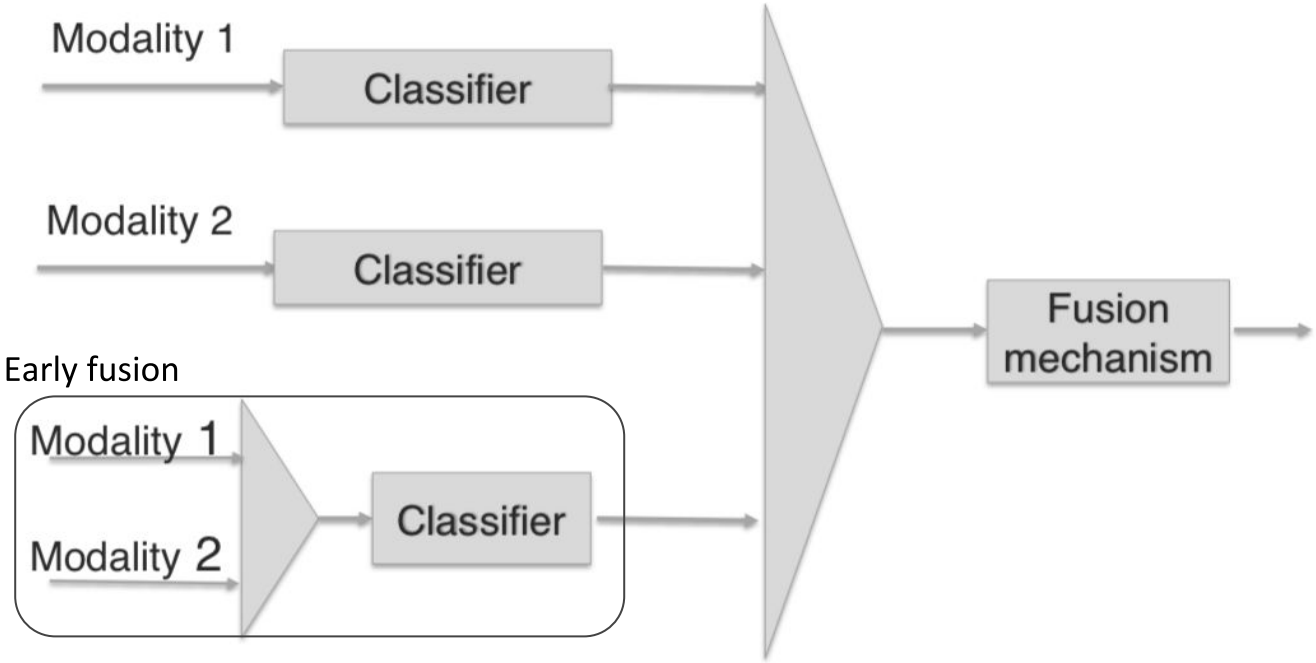
Early Fusion (Feature-level Fusion)



Early Fusion (Feature-level Fusion)



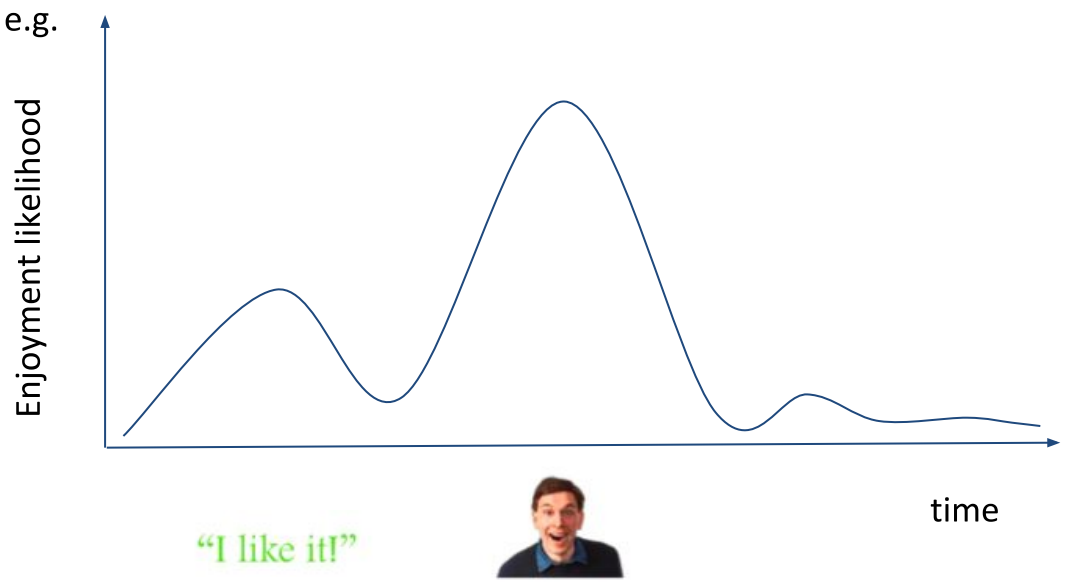
Hybrid Fusion



Event-based Fusion

Useful for **continuous** and/or **real-time** data, e.g. the videos you annotated at the beginning of this course, or a human-agent interaction.

This could be valence, arousal, the affective phenomenon under study, ...



Event-based Fusion

• **Event Value** e(E). Can be dynamically calculated from the probabilities of a detected cue.

• **Decay Speed**. Determines the time it takes for the event's influence to decrease to zero and get discarded. Events that strongly indicate the fusion's target class can be given longer decay times, in order to prolong their influence on the result.

t [s]	Description	e(E)
5.4	vocal event indicating enjoyment	0.9
6.2	facial expression indicating enjoyment	1.0
8.0	vocal event indicating enjoyment	0.8

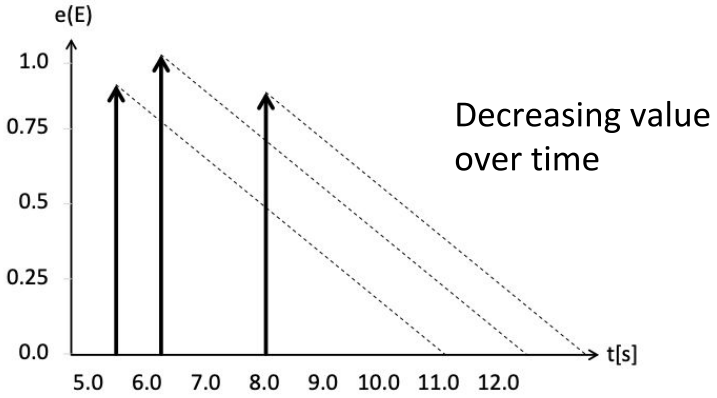


Figure 1: Multi-modal events mapped into the event space.

Event-based Fusion

t [s]	Description	e(E)
5.4	vocal event indicating enjoyment	0.9
6.2	facial expression indicating enjoyment	1.0
8.0	vocal event indicating enjoyment	0.8

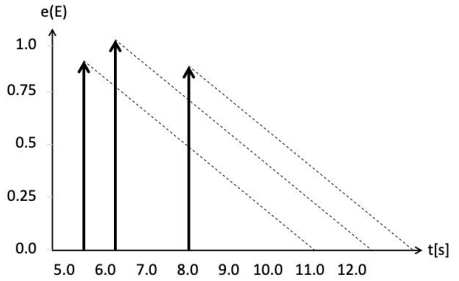


Figure 1: Multi-modal events mapped into the event space.

• **Vector Weight**. Defined per modality to emphasize more reliable or important information sources. For example, in case of a high noise level (or cultural difference), audio might be given less weight.

Likelihood of a given affect at time *t*: sum of all event-values present at *t* modified by their current influence (decreased weight), averaging over the number of active events.