Discrete & Continuous Audiovisual Recognition of Spontaneous Emotions

Mihalis A. Nicolaou

Department of Computing
Imperial College London

September 25, 2009

Outline

Introducing Emotion Recognition

Automatic emotion recognition is a multi-diverse research area

Machine Learning, Human-Computer Interaction, Computer Vision but also Psychology, Behavioural & Cognitive Sciences.

The ultimate goal

Is to generate robust, accurate and efficient emotion recognition systems operating in real-life scenarios.

Focus of Past Research

Traditionally, research focused on analysing posed emotional expressions, in controlled laboratory environments focusing on a set of discrete (basic) emotions (e.g. anger, happiness)

Questions w.r.t. goal

- Could the resulting systems be applied to real-life scenarios?
- Are these sets of basic emotions representative in everyday life situations?
- Can the systems trained on posed emotions also be used for spontaneous (i.e. naturalistic) emotion recognition?

State-of-the-art Research & Our Work

Traditionally, research focused on analysing posed emotional expressions, in controlled laboratory environments focusing on a set of discrete (basic) emotions (e.g. anger, happiness)

Recently, research in emotion recognition has shifted to analysing spontaneous emotion expressions in real-life scenarios, while continuous emotion recognition (i.e. instead of discrete categories, a set of latent dimensions) has been gaining interest.

The project follows this shift in the field and explores both discrete and continuous spontaneous emotion recognition

Description of Affect & Emotion

Three basic approaches to describing emotion

- Set of discrete emotions
- Number of latent dimensions
- Appraisal-based (context-dependent, person-dependent evaluations of events)

Will briefly discuss the first two approaches.

Categorical Approach (Discrete)

Common Approach: A Set of Discrete, Basic Emotions

- The approach is based on a set of emotions which are biologically hard-wired to humans and can be recognised independently of factors such as culture and race.
- Interesting to note that such theories have been around since the 3rd century B.C. (Stoics, the Li Chi encyclopedia).
- Modern theory accepts the following emotions as basic: Happiness, sadness, surprise, fear, anger and disgust.

\+ Intuitive mapping for humans.

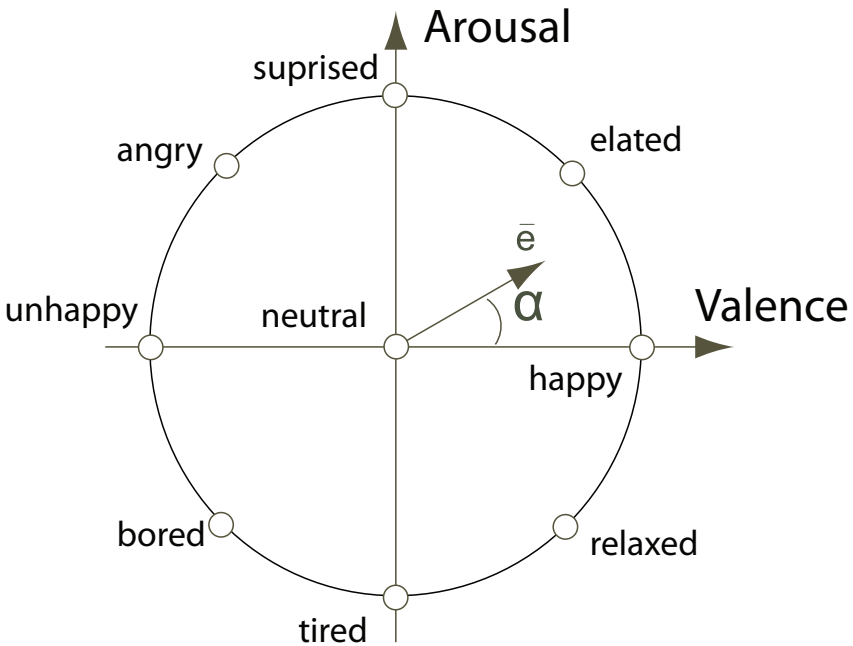
– Criticism on whether this model can capture complex emotional states. Researchers argue that these basic emotions are a small fragment of more complex & subtle emotions expressed in every-day life (e.g. boredom, tiredness).

Dimensional Approach

Much more recent approach, initiated by Wilhelm Wundt (1897)

Russel's valence/arousal space

- Valence: pleasantness (positive) or unpleasantness (negative).
- Arousal: relaxed / passive vs. aroused / active.



Dimensional Approach (2)

Much more recent approach, initiated by Wilhelm Wundt (1897)

Russel's valence/arousal space

- Valence: pleasantness (positive) or unpleasantness (negative).
- Arousal: relaxed / passive vs. aroused / active.

\+ Much more expressive than discrete labels, allows the expression of continuous emotional states.

– Certain emotions can become degenerately indistinguishable.

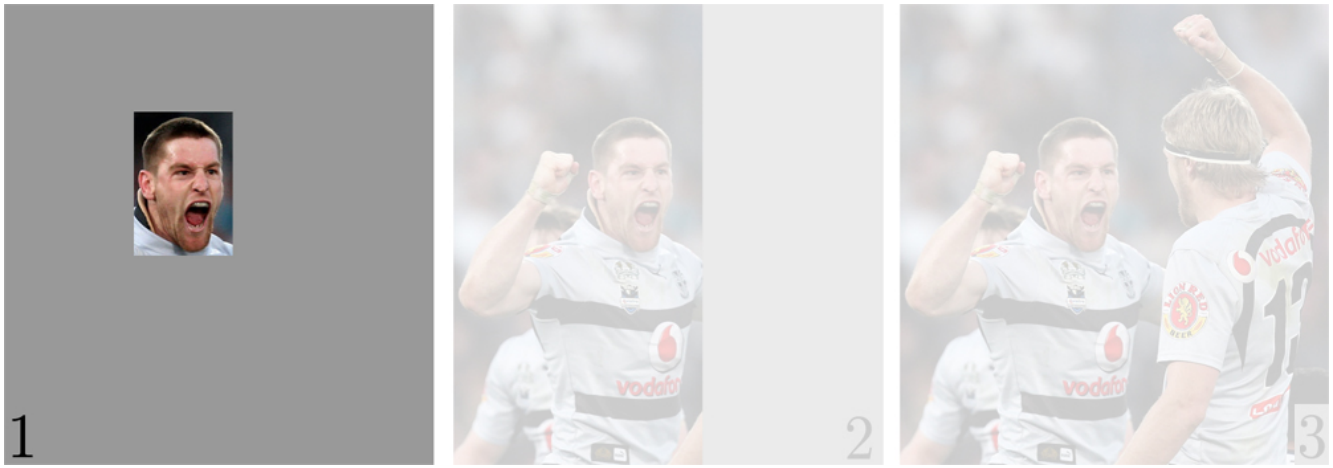
– A set of emotions falls out of this dimensional space.

Emotion Perception in Humans

- Affective information in human-to-human interaction is communicated over a set of modalities (audio, visual, tactile), corresponding to human senses. Each modality has a set of related cues, e.g. facial expressions and body gestures for the visual.
- Fusing sets of cues/modalities can resolve ambiguities, and is more naturalistic w.r.t. human emotion perception (see figure below).

Emotion Perception in Humans

- Affective information in human-to-human interaction is communicated over a set of modalities (audio, visual, tactile), corresponding to human senses. Each modality has a set of related cues, e.g. facial expressions and body gestures for the visual.
- Fusing sets of cues/modalities can resolve ambiguities, and is more naturalistic w.r.t. human emotion perception (see figure below).



Emotion Perception in Humans

- Affective information in human-to-human interaction is communicated over a set of modalities (audio, visual, tactile), corresponding to human senses. Each modality has a set of related cues, e.g. facial expressions and body gestures for the visual.
- Fusing sets of cues/modalities can resolve ambiguities, and is more naturalistic w.r.t. human emotion perception (see figure below).

Emotion Perception in Humans

- Affective information in human-to-human interaction is communicated over a set of modalities (audio, visual, tactile), corresponding to human senses. Each modality has a set of related cues, e.g. facial expressions and body gestures for the visual.
- Fusing sets of cues/modalities can resolve ambiguities, and is more naturalistic w.r.t. human emotion perception (see figure below).

Defining the Project

We use the SAL database: 4 subjects interacting with a virtual avatar. Each audiovisual session is annotated manually by 3 or 4 trained humans (the coders) in the valence/arousal emotion space.

We extract two sets of audiovisual segments from the database: the set of negative and positive emotion expressions, along with the valence/arousal ground truth.

- We perform audiovisual emotion recognition by fusing the audio, shoulder and facial expression cues.
- For discrete emotion recognition, we focus on classifying our audiovisual data into two coarse classes: Positive vs. Negative.
- The scenario relating to continuous emotion recognition (which is our focus) relies on predicting the valence/arousal distribution of values.

Learning Techniques

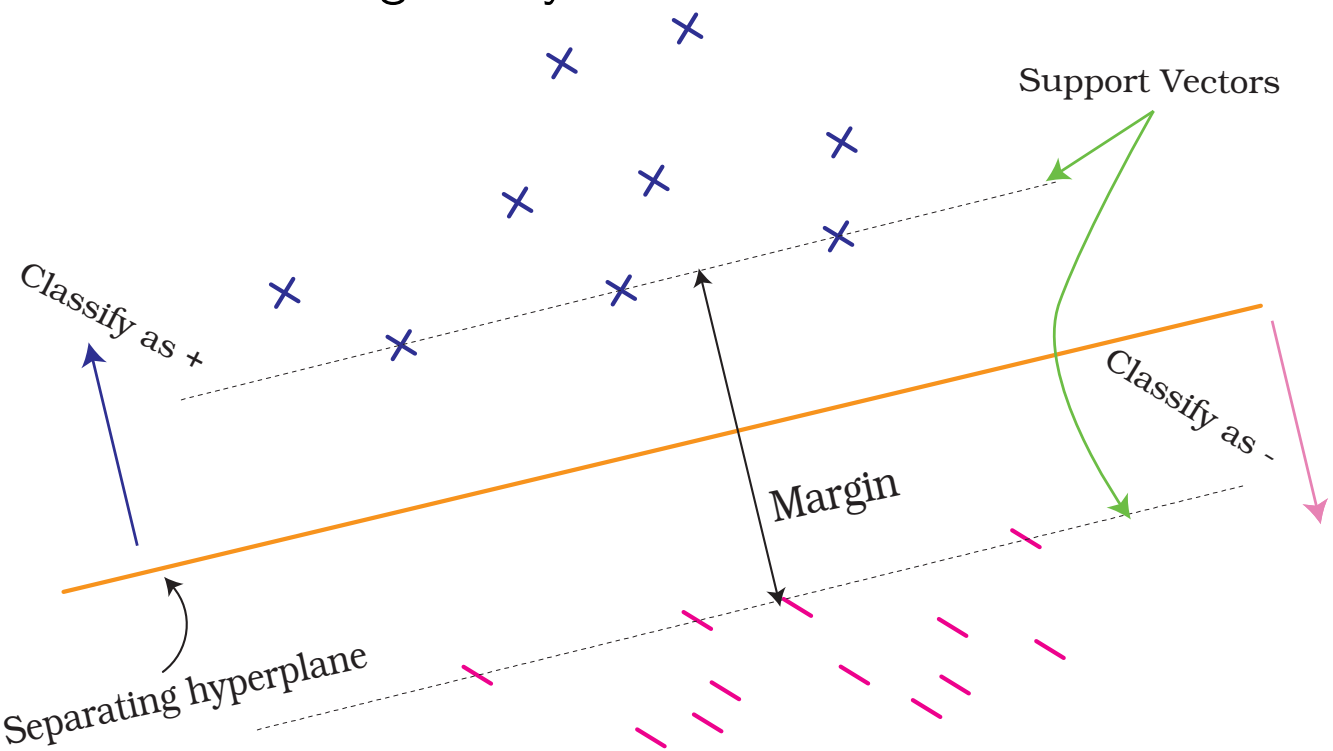
- Discrete: Coupled Hidden Markov Models (CHMMs), Support Vector Machines (SVM).
- Continuous: Support Vector Machines for Regression (SVR), Long Short-Term Memory (LSTM) Recurrent Neural Networks.

Our focus lies generally in methods for continuous emotion recognition. We will now briefly describe some of them.

Support Vector Machines

For a binary classification problem, given a set of training input vectors $\mathbf{x_i} \in \mathbb{R}^n$ and corresponding labels $y_i \in \{-1, 1\}$.

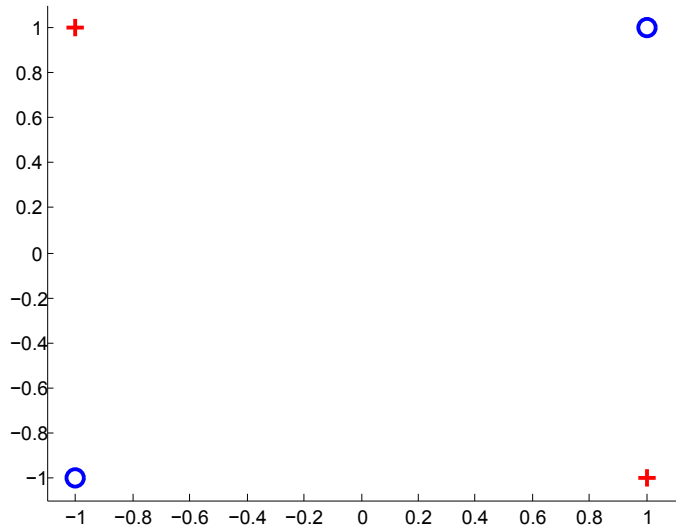
- A SVM will return the separating hyperplane which generates the maximal margin, thus minimising the generalisation error according to statistical learning theory.

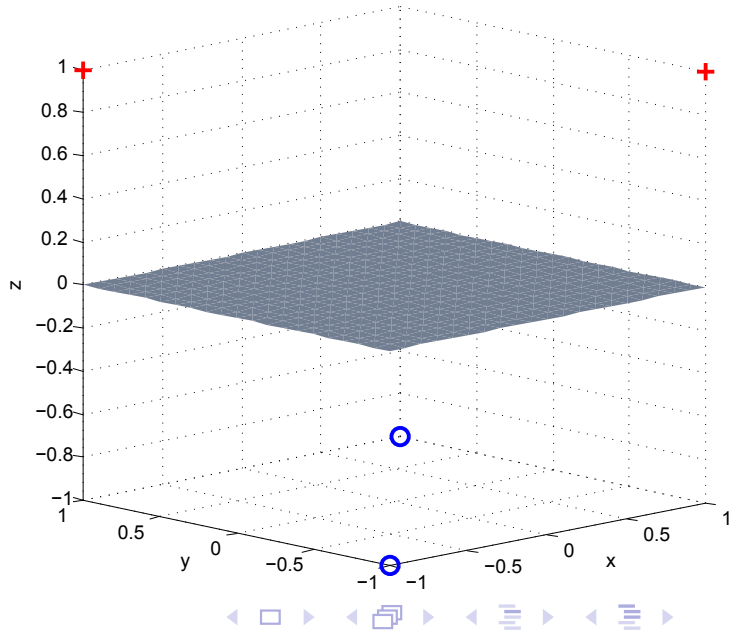


Non Linearly-separable Datasets

From statistical learning theory: The generalisation error does not depend on the dimensionality.

- The $\mathbf{x_i}$'s of the input problem define the *input space*. Map into a (higher) dimensionality space (the *feature* space) using a kernel to replace the dot product in the quadratic optimisation equations of SVM. In the figure, this solves the XOR problem by mapping to 3D.





Soft Margins and Slack Variables

- Previously, no points were allowed to fall in the margin (hard margin constraint).
- The model would then have to fit the noise in the training set.

Remedy: use a set of slack variables ξ in order to allow some points to fall inside the margin.

- Assuming Φ_m is the function that maximises the margin when minimised, the optimisation problem for SVMs becomes:

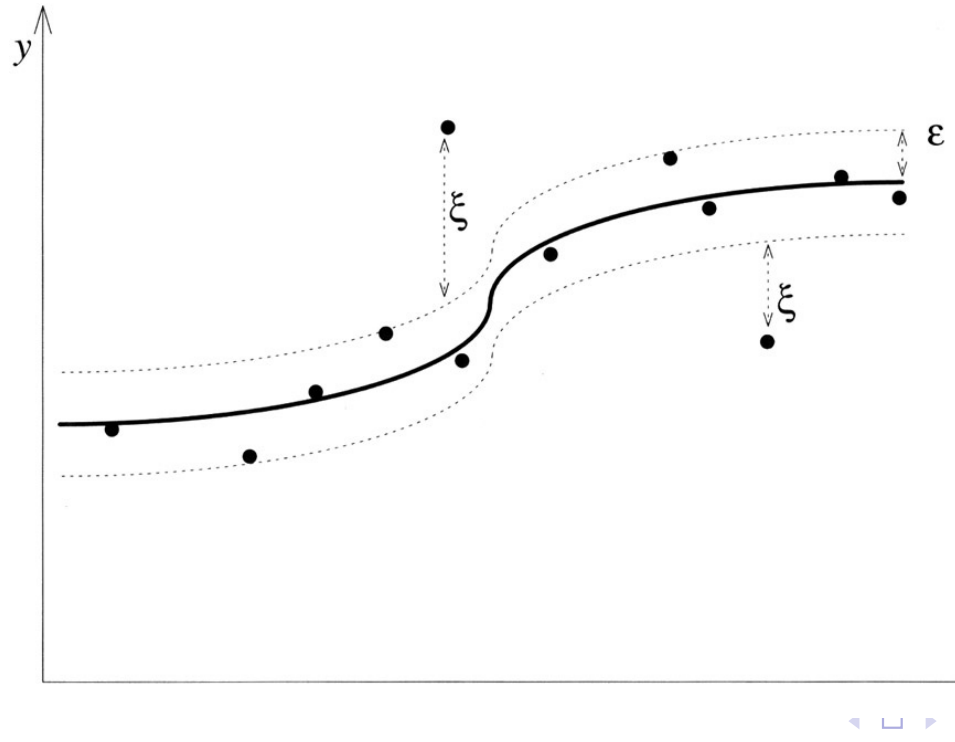
$$min[\Phi_m + C\sum_{i=1}^{m}\xi_i]$$

- The error penalty is the constant $C > 0$, which balances the tradeoff between maximising the margin and tolerating training errors. The higher it is the less errors.

Support Vector Machines for Regression

Use a loss function to weight the error of the point with respect to the distance from the correct prediction. Using ϵ insensitive regression, there is no charge for a band ϵ from the estimation to the actual training instance.

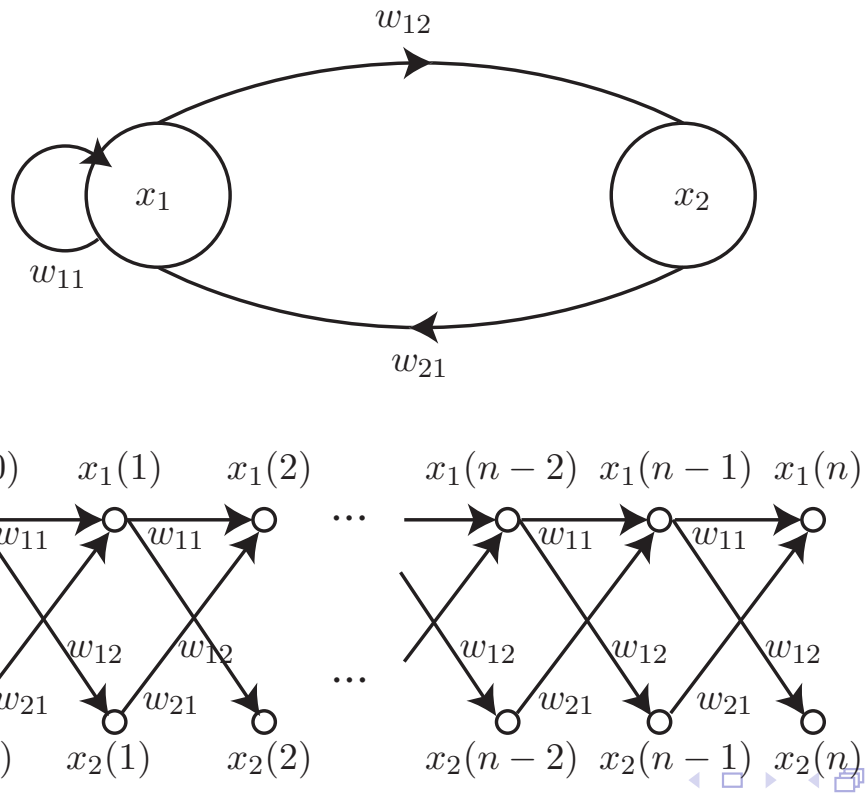
- Other concepts generalise directly from SVM.



Recurrent Neural Networks (RNNs)

Recurrent neural networks, in contrast to feedforward neural networks allow feedback connections which implicitly maintain an internal state of the network representing past inputs.

- A typical training algorithm is Back-Propagation Through Time (BPTT), where the network is unfolded for n time steps (figure).



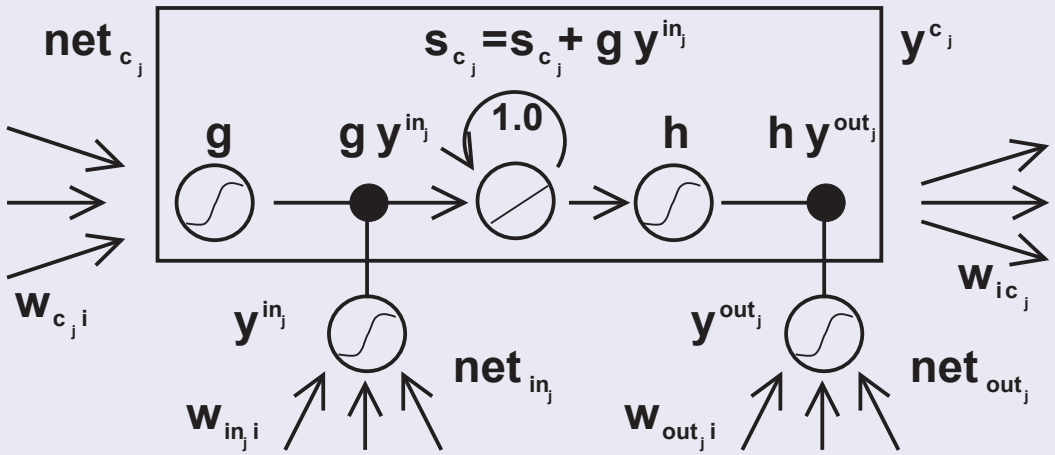
RNNs vs Long Short-Term Memory (LSTM) RNNs

Hochreiter (1991) shows that in traditional RNNs, the error flowing backwards either vanishes or grows exponentially. Thus, long-range dependencies can not be learnt.

- LSTMs offer a solution:
  - Use a basic unit called Constant Error Carousel (CEC): Essentially a node with one feedback connection to itself, with a weight of 1 and a linear activation function. This keeps the local error flow constant.
  - A memory cell contains a CEC and the input/output gates, which are multiplicatively involved in determining the activation and the output of the cell. They are supposed to learn when to propagate the error backwards and when to allow activations to enter/exit the cell.
  - This way LSTMs can learn long-range dependencies in the input data.

The Memory Cell

An LSTM memory cell along with the input gates, the CEC and the nonlinear squashing/scaling functions. The cell state is s_c.



- Some extensions:
  - Forget Gates: Avoid the states s_c growing unboundedly by replacing the CEC weight with the activation of a gate (Learn to forget).
  - Peephole Connections: Connect the CEC with the gates.
  - Bidirectional: Train two networks, one with the sequence forwards and the other in backwards, in order to learn dependencies not only w.r.t. the past but also w.r.t. the future.

LSTMs vs SVM(R)

A comparison of basic characteristics also taking into account issues specific to emotion recognition:

- LSTMs are dynamic learning techniques, SVM(R) are static. Dynamic learning allows us to capture "emotional history" - LSTMs can capture long-range dependencies. Crucial w.r.t. the temporal dynamics of human emotion expression.
- Both methods can capture non-linear correlations due to using non-linear kernel functions (kernels for SVM(R), squashing/scaling functions for LSTMs).
- SVM(R)s optimise a convex function; they have no problems of getting stuck in local optima, unlike methods for learning in neural networks.
- SVM(R)s training does not only rely on the training error for stopping, as the goal is to maximise the margin. Neural networks are prone to overfitting.

Pre-processing & Segmentation

We have a set of audiovisual sessions, annotated by 3-4 coders in the valence/arousal emotion space.

The goal for our segmentation is to produce audiovisual segments belonging to the set of positive/negative emotions, with a set of valence/arousal values to be used as ground truth.

We also attempt to capture a *baseline* - a condition against which the classifier can learn to compare against.

The experimentation was an iterative procedure:

- Determine normalisation/segmentation procedure.
- Segment according to above, generate ground truth.
- Evaluate ground truth, inspect video segments.

Pre-processing: Our final selection

To ensure a common time scale

We bin the annotations per video frame (0.4 seconds, 25 fps)

To suppress inter-coder variances

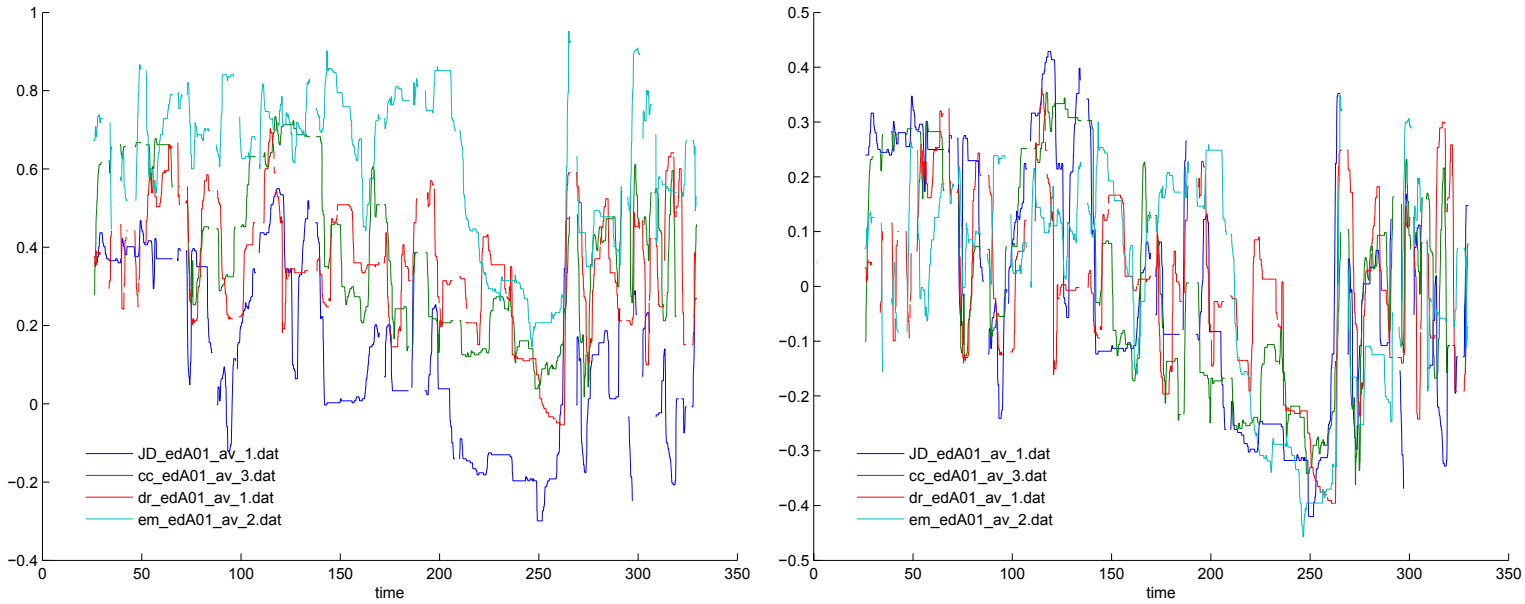
We normalise each set of annotations to have a zero mean locally.

To fill-in missing values in annotations

We perform cubic interpolation, maintaining the shape and the maxima of the distributions.

Pre-processing: Normalisation Example

Normalising to zero mean locally reduces the inter-coder MSE from 0.083 to 0.055. Figure (b) is normalised.



(a) (b)

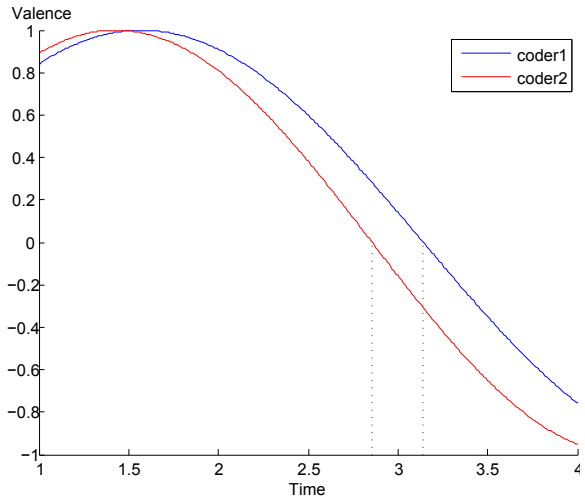
Segmentation

We will provide the segmentation procedure on an intuitive level, for extracting a positive emotion expression (the negative case is similar)

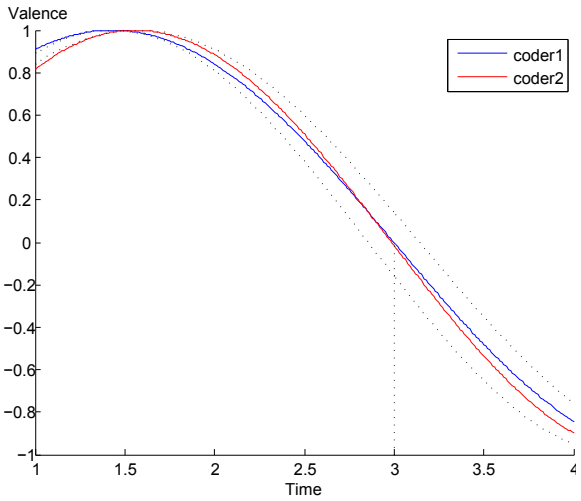
- For all coders c which annotate each session, find each set of transitions to a positive emotional state which are within a predefined offset of 0.5 seconds.
- Match this set (giving precedence to the number of agreeing coders) and time-shift to produce final averaged values for frames, valence and arousal.
- Where there are only two coders, weight by using their correlation w.r.t. the rest of the coders in order to propagate information from all the coders to the agreeing subset.

The transition to a positive emotional state is determined by a sign change of the valence values turning positive. We detect the turning back to a non-positive emotional state when the sign changes again.

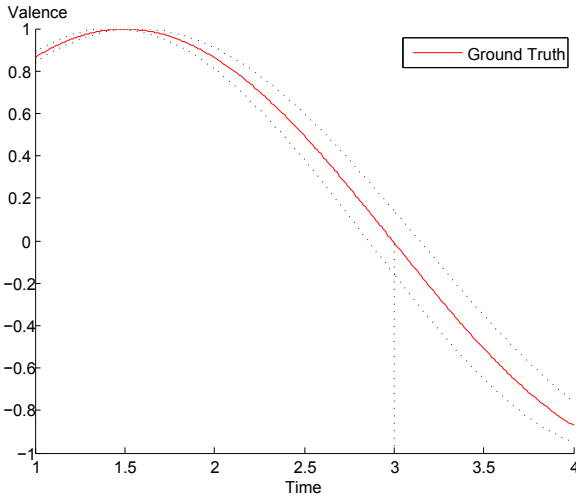
Segmentation: Time Shifting



(a)



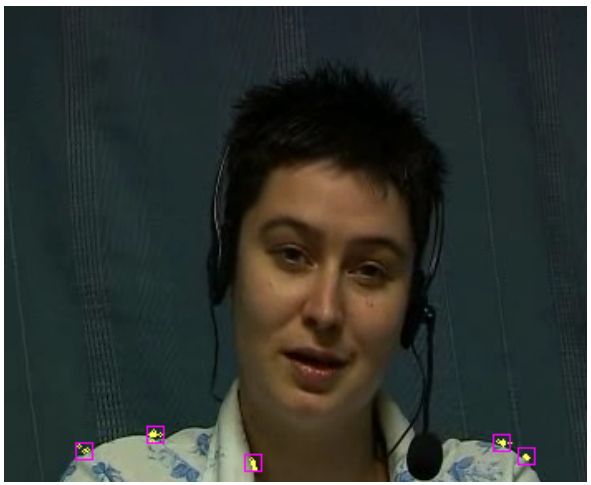
(b)



Feature Extraction

We extract the following features from the segments produced:

- Audio Cues: We extract the MFC coefficients and prosody features related to pitch, energy, leading to a set of 15 features. We removed the background speech and suppressed the noise in the clips.
- Facial Expression Cues: Tracking 20 2D points, generating 40 features.
- Shoulder Cues: Five 2D points are tracked, producing a 10 dimension feature vector.



Discrete Emotion Recognition

Classify resulting segments into positive/negative

- Subject-Independent: Leave one subject at each fold.
- Subject-Dependent: Perform typical 10-fold cross validation.

- Results show that subject-independent is difficult with just 4 subjects.
- We gained 10% accuracy by noise reduction/background speech removal for the audio cues.

We present the results attained from using CHMMs for subject dependent (which can also be theoretically justified)

F	S	A	FS	SA	FA	FSA
73.13%	73.88%	61.19%	78.36%	68.66%	70.9%	79.1%

Classification in the Likelihood Space (1)

- (C)HMMs are generative models. We have one model for each of the two classes that we perform recognition on.
- Given the sequence of features, the observations, each HMM returns a likelihood for that specific model having produced the specific sequence.

- The maximum likelihood (ML) principle assumes that the correct network is the one with the maximum likelihood; this assumes that the learnt distribution is the true distribution, something that is not always the case.
- Let us consider the 2D space of points (l_1, l_2), with each dimension corresponding to the likelihood of each network.
- The ML classification can be imagined as a line that bisects the space.

Classification in the Likelihood Space (2)

- We shift the line to better fit the training data. This does not produce consistent results, leading us to conclude that the training data is not always linearly separable or that more complex functions would represent the distribution better.

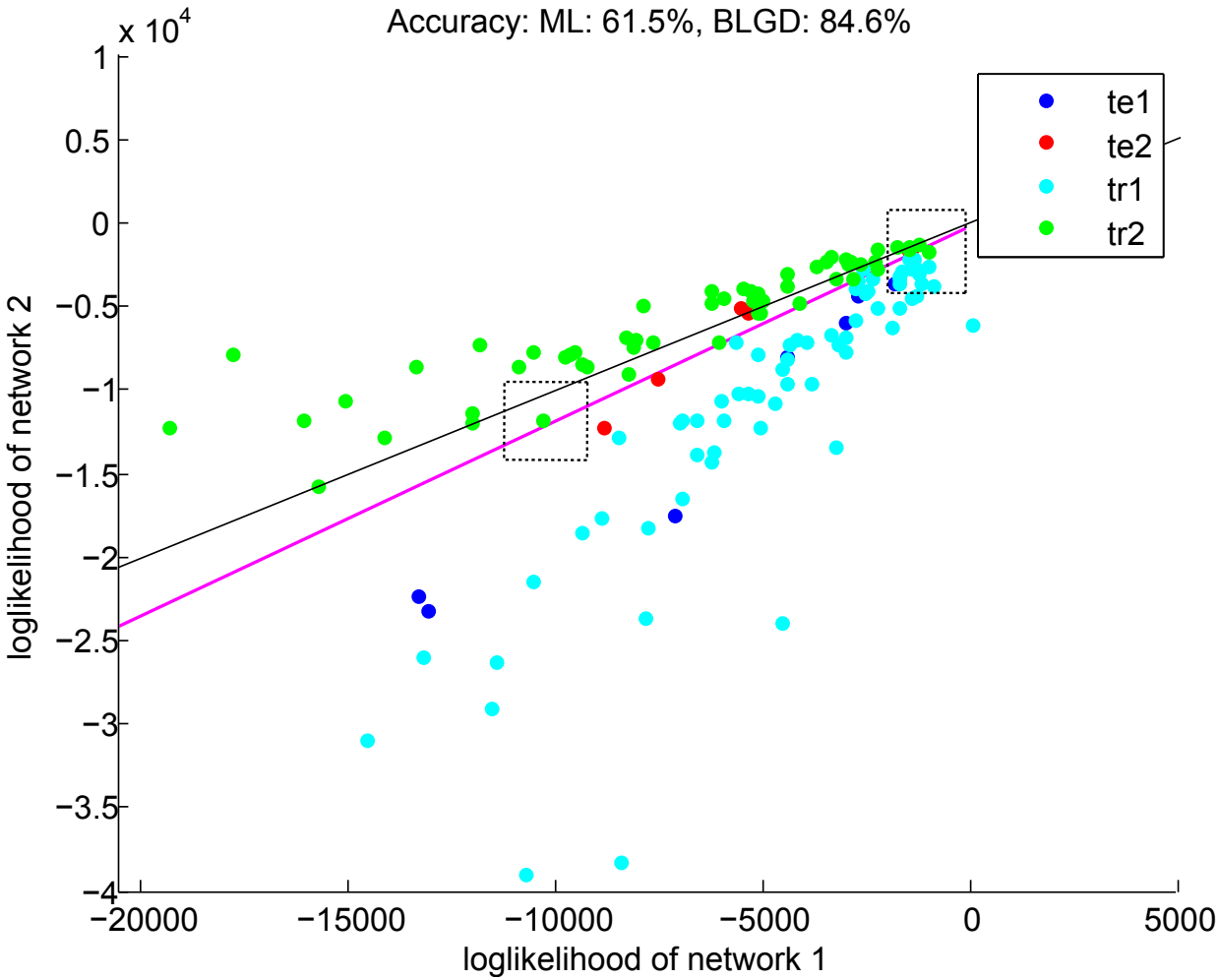
- We apply SVM classification on the 2D points that belong to the likelihood space, by using a radial basis function (RBF) kernel:

$$K(\mathbf{x}, \mathbf{x}') = e^{(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)}$$

- Where γ determines the width of the function; the bigger the gamma the smaller the width, the closer the function follows the distribution - and more prone it is to overfitting.
- Audio cues $\gamma = 8.65$, for all other cues: $\gamma < 0.3$.

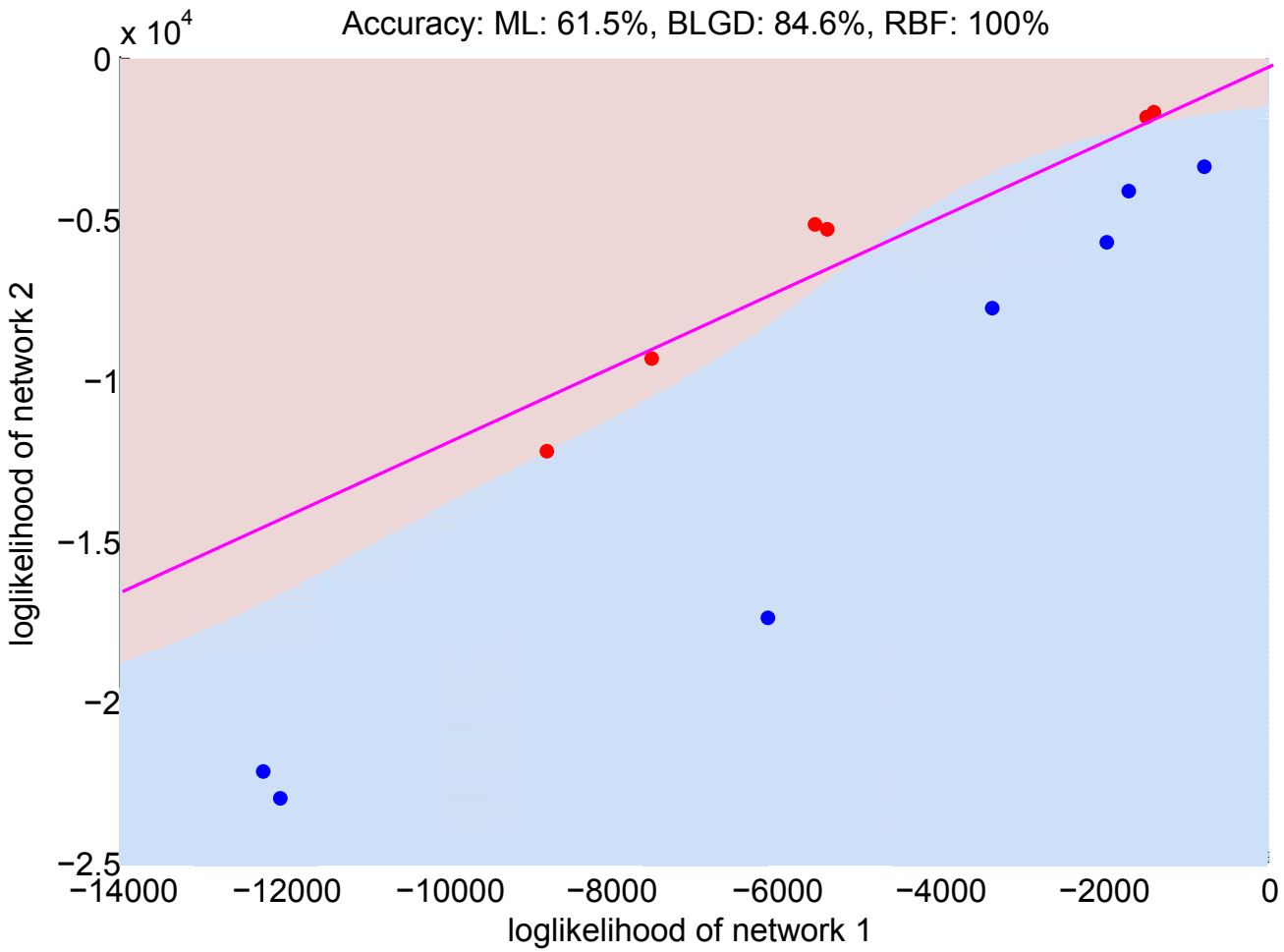
The Distribution of Likelihood points

...for the training data is not always linearly separable:



The RBF Decision Surface

...accurately classifies the testing data:



SVR-RBF vs. ML

	F	S	A	FS	SA	FA	FSA
RBF	**87.53%**	**79.07%**	**66.92%**	**85.05%**	**74.07%**	**86.65%**	**88.19%**
ML	73.13%	73.88%	61.19%	78.36%	68.66%	70.90%	79.10%

The more subtle spontaneous emotion expressions produce a distribution better classified by RBF kernels. In 4 cases performance is over 85%, reaching 88.2%.

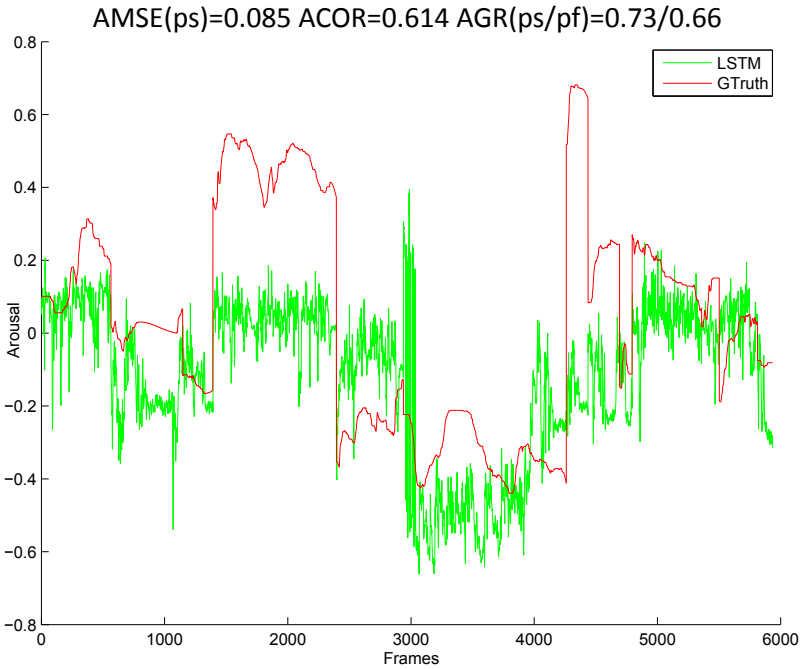
Continuous Emotion Recognition

Where we do not only have the audiovisual segments and the classification, but also continuous values for valence & arousal.

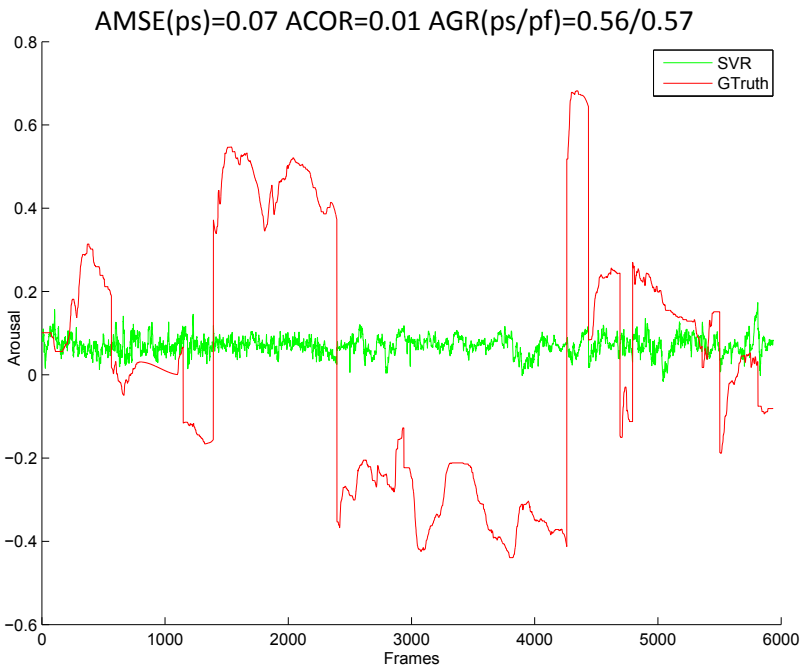
- Metrics we use include the mean squared error (MSE) ($\frac{1}{n}\sum_{i=1}^{n} err^2$), the correlation coefficient (COR), as well as the agreement (AGR).
- The agreement metric essentially translates to the percentage of each sequence where the predication w.r.t. the sign of the ground truth was correct. For valence this translates to predicting the correct emotion in terms of positive/negative.
- Where not otherwise mentioned, we use feature-level fusion, i.e. merge all features into one vector and train.
- We will present an overview of our results/conclusions...

Mean Squared Error Evaluation

First experiments show the LSTMs overperforming SVR-RBF and SVR-P on average. Very good MSE results for SVRs with polynomial kernel, but we should not only trust the MSE:



(a)



(b)

The optimal evaluation metric(s) are still an open research issue.

MSE, Correlation & Agreement (1)

For the rest of the presentation, we will focus on the correlation performance, ignoring very small variations in the MSE (≈ 0.01).

- On average, LSTMs overperform SVR, achieving correlations of up to 0.511 while the maximum correlation attained by SVRs is 0.376.
- Compared to facial expressions, the audio cues perform better for both arousal (0.511 vs. 0.237) and valence (0.397 vs. 0.205).
- Audio features are extracted at the double video frame rate. If we extract both cues at the same frame rate, the facial expressions overperform the audio cues for valence, as theoretically expected (audio drops from 0.4 to 0.08), while are still better for arousal. One justification is that the longer sequences for audio generate more long-range dependencies captured by LSTMs.

MSE, Correlation & Agreement (2)

In general, the shoulder cues seem to perform bad (correlation of 0.06-0.02). We assume that the variations in shoulder movements are not continuous/enough to be mapped to real, continuous values.

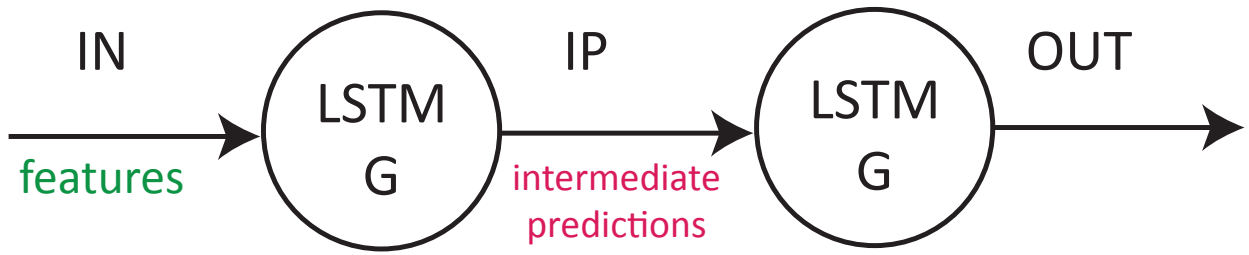
Experiments with[a] dimensionality reduction and PCA have shown the shoulder cues performance to increase when only the 4 dimensions with the greatest variances are used. It is a sign that not all feature dimensions we extract are useful for continuous emotion recognition. A relevant increase was also observed in the fusion of the facial expression/shoulder cues.

[a] It is noted that from now and on we will experiment with LSTMs

Capturing Temporal Patterns

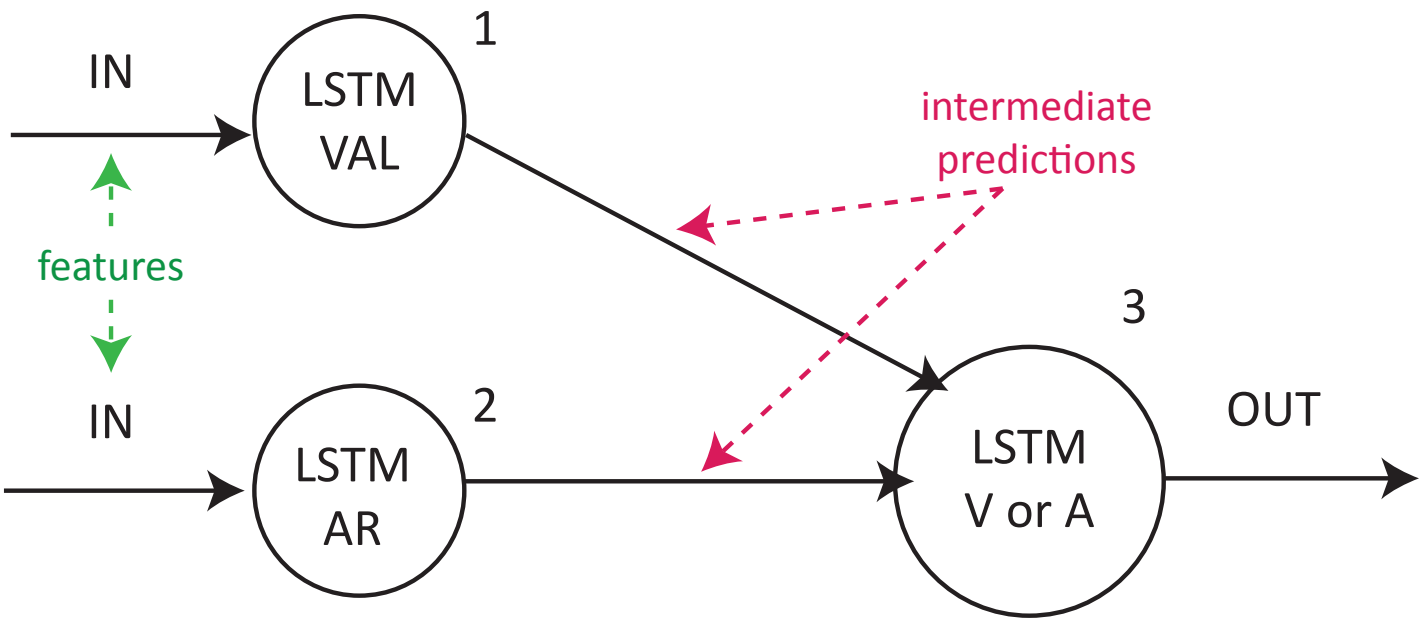
It is reasonable to assume that the valence/arousal predictions exhibit temporal patterns also found in the ground truth.

- How can we enforce the presence of such patterns?
- How can we suppress the false patterns emerging?



Capturing Correlations Between Valence/Arousal

It is also logical to assume that there exist correlations of co-occuring patterns between the valence/arousal estimations. Again, we want to detect such correlations which correspond to events in the ground truth and ignore the false ones:



Capturing Correlations/Temporal Patterns: Conclusion

In cases where the correlation was already good enough (0.2 and above, no less than 0.04), an improvement was observed.

E.g. for audio/arousal:

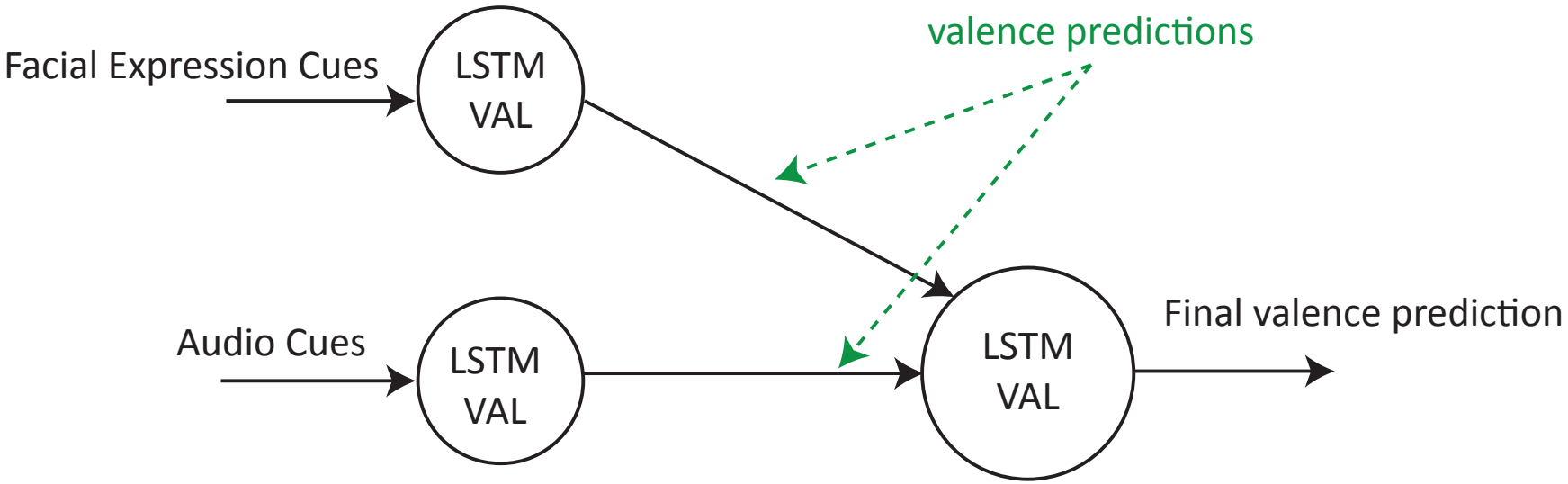
- Original Prediction: 0.51
- Capturing Temporal Patterns: 0.53
- Capturing Temporal Patterns & Correlations: 0.58

But let's explore this notion at decision-level fusion...

Decision-level fusion

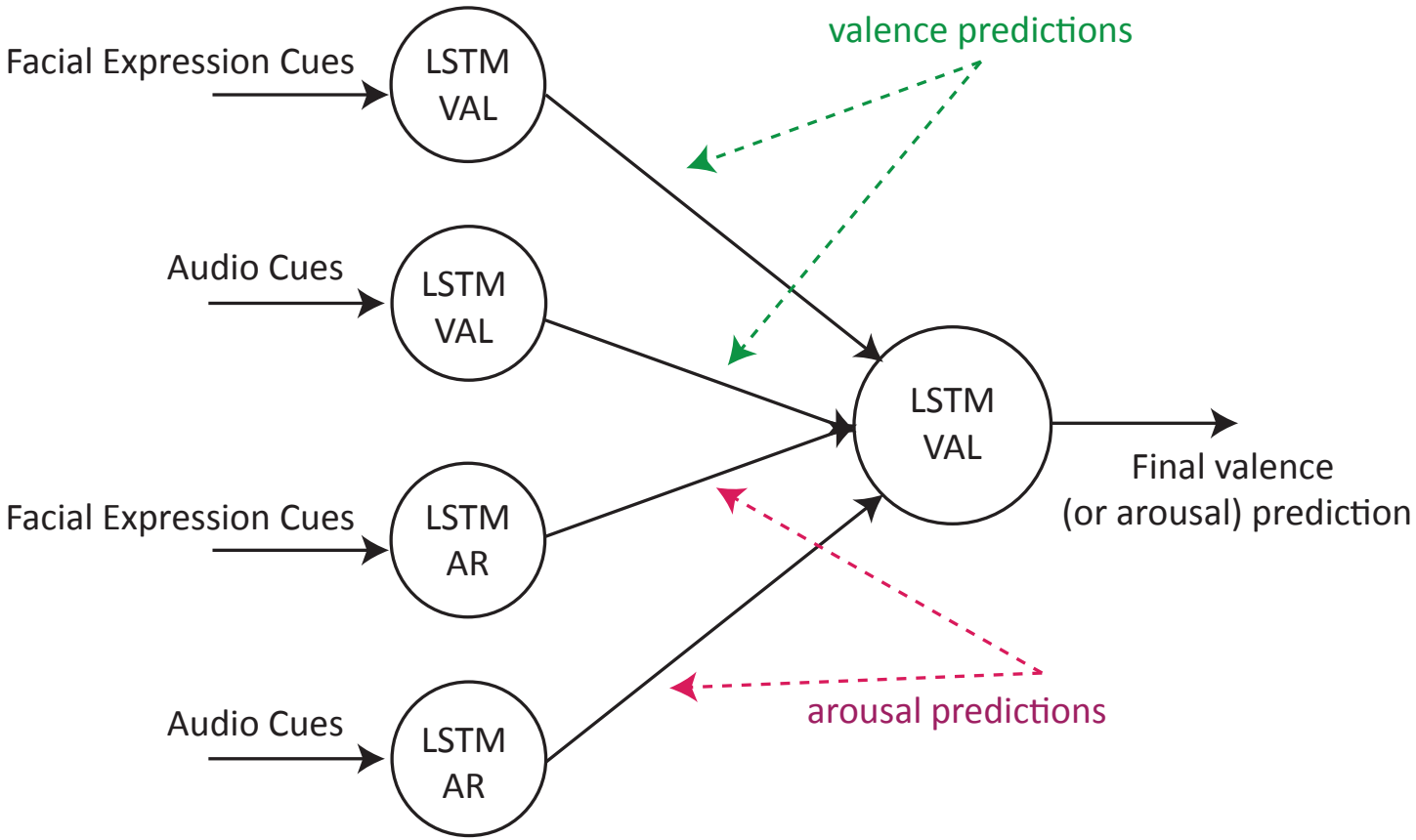
Our previous experiments applied feature level fusion: Merge all features from available cues into one feature vector and train.

With decision-level fusion we train one classifier for each set of cues and then fuse their estimation.



Decision-level fusion and Capturing Correlations

As with single cues, we attempt to capture and enforce correlations between valence and arousal by using both of the values as inputs at decision-level.



Decision-level fusion/Correlations Results

Improvement of Dec-Lev & Dec-Lev VACorr w.r.t:

	Dec-Lev			Dec-Lev VACorr		
	Valence	**Arousal**	**AVG**	**Valence**	**Arousal**	**AVG**
PCA	22.83%	21.55%	22.19%	22.91%	25.65%	24.28%
fLSVR	23.11%	27.40%	25.26%	23.19%	31.50%	27.35%
FeatLEV	21.59%	12.20%	16.90%	21.68%	16.30%	18.99%

f-level SVR $<$ **f-level LSTM** $<$ **d-level LSTM** $<$ **d-level/cor LSTM**

For discrete emotion recognition, decision-level fusion has been criticised for being unable to learn correlations between cues.

For continuous emotion recognition, each of the classifiers fused at decision-level produce a real number and not just a label. We believe that this could implicitly propagate enough information to capture inter-cue (and inter-estimation) correlations and patterns.

Continuous-to-Discrete Emotion Recognition

A final experiment...

The idea is to go from the continuous valence estimation, which measures the positiveness/negativeness of an emotional state, into the discrete positive vs. negative classification we performed with CHMMs. We experimented with the audio valence estimation.

We separated the estimation into n windows, extracted features such as the standard deviation and the mean value. We fed these features in a SVM. The results slightly overperformed the ML HMM performance (61.4% - 61.2%).

The approach seems promising for future work with more cues and can be a means to evaluate the estimation itself.

Conclusions

- We have already provided a description of our experiments in an observation/conclusion style.
- For more details please refer to the report.
- Continuous emotion recognition is still at its infancy, with just 3 papers for previous work, all in the past year (2008-2009).

It should be stressed that discrete and continuous emotion recognition are two vastly different problems.

Sequence (dynamic) learning techniques appear to be very promising. We have approximated the average coder correlation for valence and the maximum coder correlation for arousal by 3%, while the system performs 8% better than the average coder correlation for arousal.

Future Work

- Further experimentation with continuous to discrete emotion recognition.
- Model individual coders with learning techniques instead of the produced ground truth.
- Longer sequences / higher video frame rate to enforce sequence learning techniques.
- Is a baseline required for continuous emotion recognition?
- Experiment with other techniques (e.g. Conditional Random Fields) & other fusion methods (e.g. linear combinations etc.)
- Improve our normalisation/segmentation procedure.

Thank You. Questions?

Comment: Aphasic Patients

In The President's Speech from the book The Man Who Mistook his Wife for a hat by Oliver Sacks, the writer refers to how aphasic patients who were incapable of understanding words as such, could not realise that they were aphasic, because when they were addressed they grasped most of the meaning from the cues communicated by natural human interaction.

One had to remove all extraverbal cues - tone of voice, intonation, suggestive emphasis as well as all visual cues (expressions, gestures, posture) in order to make the patient sure of their aphasia.

An indication of how with no context - similarly to how the majority of emotion recognition systems works - human to human communication conveys a most significant amount of information.