

# Beyond the Text: Construction and Analysis of Multi- Modal Linguistic Corpora

Dawn Knight, Sahar Bayoumi, Steve Mills, Andy Crabtree,  
Svenja Adolphs, Tony Pridmore, Ronald Carter

# Introducing Corpus Linguistics (CL):

---

- 1) Corpus Linguistics (CL) is an empirically based methodological approach, involving three main processes: extraction of language data, processing of the output and interpretation of the output.
- 2) Language data, known as corpora, are stored as a corpus, i.e. a 'large and principled collection of texts' (Biber et al, 1998: 4). These can include written and/or spoken corpora.
- 3) New and different perspectives to language description are possible when using corpora, which prove useful beyond the field of Linguistics.

# Spoken Corpora: The Limitations

---

- However such corpora are unable to represent language ‘beyond the word’.
- This is problematic as communication exists as a ‘complex network’ of ‘semiotic channels’ (Brown, 1986: 409). These channels are multi-modal and specific to their form, function and context of use in discourse.
- To develop the scope of the kind of evidence we can extract from a spoken corpus a more multi-modal approach to discourse needs to be developed to facilitate explorations of communication beyond the text.

# The HeadTalk Project:

---

This project seeks to provide a rubric for the development of a multi-modal, multi-media corpus tool which can be utilised to explore gesture-in-talk, and a specific subset of these: head nods, in more detail.

# Backchannels:

---

- Head nods are a form of backchannel, signalling active listenership in conversation.
- A backchannel item is a short response token that does not take over a speaker turn and is not a response to a question.
- Backchannels are realised through a finite set of linguistic forms which can be extracted from corpora.
- Backchannelling also includes non-verbal response tokens, non-vocalised kinesic signals, and proxemic movement as a means for hearers to register and evaluate what is being said (i.e. Head nods).

# Coding Backchannels: A Linguistic Approach

---

- Backchannels adopt a variety of functions in discourse which both ‘complement and regulate each other’ in order to generate meaning’. (Brown, 1986: 409)
- In order to gain a better understanding of head nods we need to develop out abilities of ‘reading’ them in terms of their:
  - Specific function
  - When and where they occur in the co-text of discourse
  - When and where they occur in the context of discourse

# Automated Analysis of Conversation Gesture:

---

- Data on the role of head nods is provided by video records of natural conversations
- Computer vision techniques are needed that can recognise and extract descriptions of head nods (and other non-verbal response tokens)
- Work in the vision community to date has focussed on HCI applications involving direct communication with the device
  - the view of the head is restricted
  - simple image features are used to identify simple gestures

# Automated Analysis of Conversation Gesture:

---

- Given video records of natural conversations, how can we:
  - provide the ability to quickly search video for (likely) gestures of interest?
  - analyse these gestures to form meaningful descriptors of the motion involved?
  - identify common classes of gesture, on the basis of these descriptors?
  - combine gestural information with other modalities within the corpus?

# Gesture Detection: Discrete Wavelet Transform

---

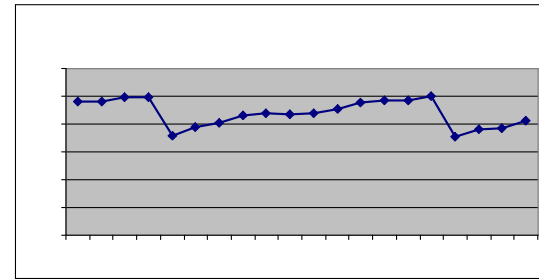
- Gesture detection and feature extraction need not be distinct, but gesture detection may need less detailed information
- Wavelet transforms decompose the video into four components – the 2<sup>nd</sup> and 3<sup>rd</sup> represent vertical and horizontal frequency respectively
- Variations in the moments of these components can provide indicators of head nods
- The method only detects motion, but it can rapidly detect candidate sections of video for more detailed analysis

# Gesture Detection: Discrete Wavelet Transform

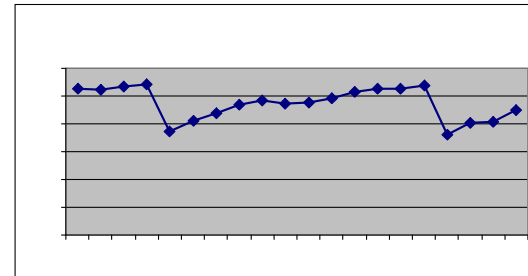


The first 3 moments of the vertical component of the wavelet transform

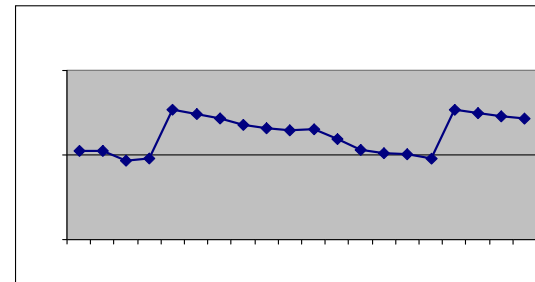
1st



2nd



3rd



frame

# Gesture Detection: Tracking Image Features

---

- Tracking user-specified targets both provides an alternative approach to detecting head gestures, and allows the user more control over the process

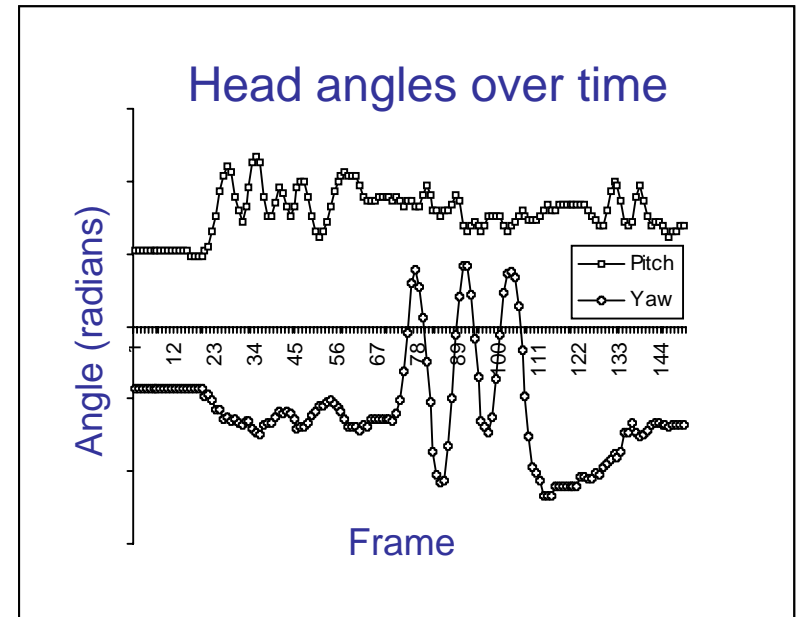
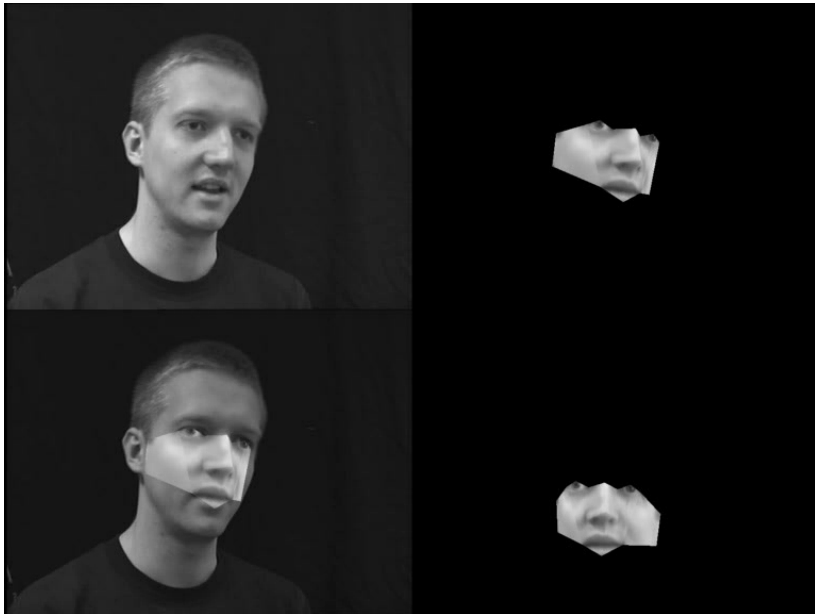


# Feature Extraction: A 3D Head Model

---

- We use a 3D head model to track the face through a video sequence, describe and analyse gestures
- The model is built from training data comprising stereo image pairs using a principle components analysis method capable of dealing with missing data
  - 6 parameters control the position & orientation of the model in 3D space
  - 15 parameters control changes in appearance (colour) and shape of the model
- Analysis of the orientation parameters allows us to characterise head nods and shakes

# Feature Extraction: A 3D Head Model



# Clustering & Linguistic Interpretation:

---

- Final step is to relate gestures, each described by a small number of features, to linguistic categories
- We combine automated clustering with expert linguistic classification, three approaches present themselves
  - Automated clustering first, analyse clusters to determine linguistic significance
  - Expert clustering first, then learn the features that distinguish between groups
  - Combine automatic and expert clustering in an iterative manner

# Clustering & Linguistic Interpretation:

- To examine the feasibility of the methodology we have examined a very simple feature - duration
  - an initial clustering was made using k-means,  $k = 4$
  - expert analysis showed that the clusters were not individually meaningful, but the two with the longest duration contained more non-verbal gestures
  - revised clustering,  $k = 2$

	Short	Long
Verbal	122	66
Non-verbal	47	39

$\chi^2 = 2.62$  close to the 10% level of confidence

There may be a relationship between duration & verbalisation

# Clustering & Linguistic Interpretation:

---

- Within the verbal gestures it is possible to distinguish gestures that act as a backchannel

	Short	Long
Backchannel	113	50
Non-backchannel	9	16

- $\chi^2 = 10.57$ , which is significant at a 99% confidence level
- Gestures that act as a backchannel tend to be longer than those that do not

# Conclusion:

---

- Construction and analysis of multi-modal corpora is an important and open research challenge
- Tools are required, and must incorporate computer vision techniques; no directly applicable system exists, though many valuable components are available
- We have identified a number of techniques which may support gesture detection and feature extraction and proposed a methodology for clustering and linguistic interpretation
- Initial results have been presented,  
but further work is required