



The University of  
Nottingham

# HEAD TALK

Issues in multi-modal communication analysis

# COMMUNICATION & HEAD NODS



The University of  
Nottingham

- Beyond language: Communication as a ‘complex network’ of ‘semiotic channels’ (Brown, 1986: 409)
- These channels are multimodal and can be indirect or direct
- There are many possible, ‘different, independent, pragmatic and semantic functions’ of signs making them specific to their (Argyle, 1975)

**- Type**

**- Function**

**- Context of use**

- Effective communication relies upon the receiver successfully detecting, processing and understanding these interactive ‘signs’ in its given context of use.

# Discussion outline: SYSTEM ARCHITECTURE



The University of  
Nottingham

## ➤ RECORD

- METHODS of recording visual and verbal data
- SYSTEM ARCHITECTURE- physical layout of recording process

## ➤ REPRESENT & REPLAY

- CODING & CLASSIFICATION methods used to synthesise the data
- PRESENTATION of data in an interactive interface

# RECORD: An example

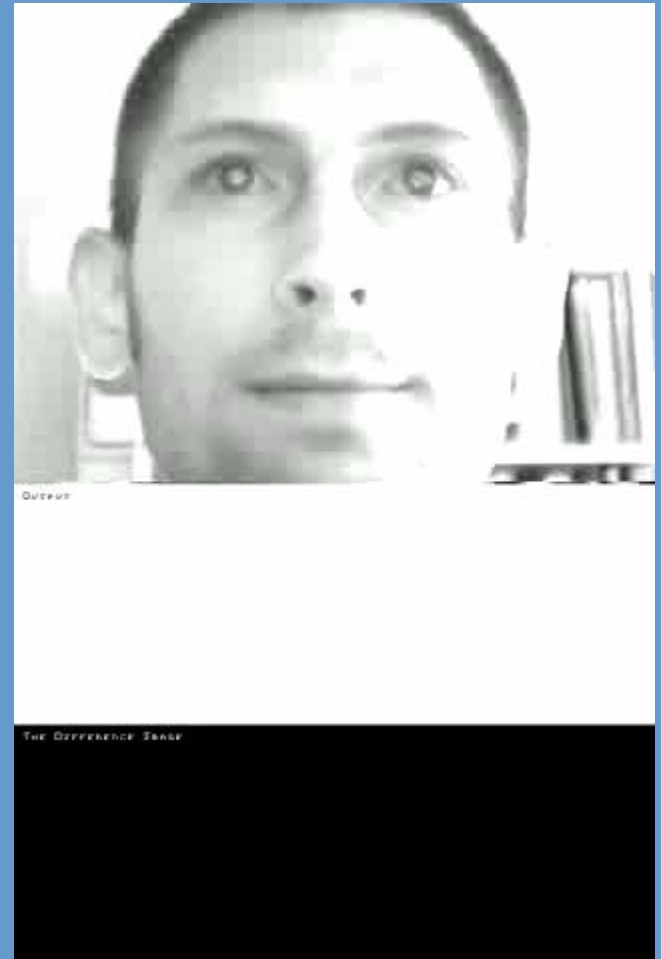


The University of  
Nottingham

## 'A Real-Time Head nod and shake detector' (Kapoor & Picard, 2001)

This is a system based upon tracking movements of the eyes as the basis for determining head movements. It uses a single infrared sensitive digital camera with concentric rings of LEDs with the ability to film at 30fps.

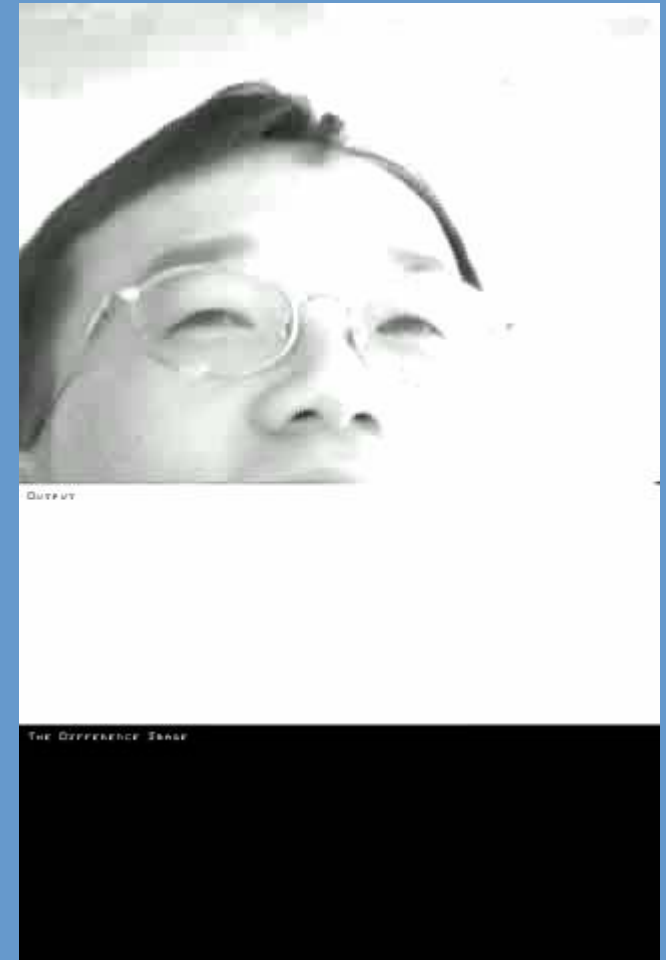
➤ This system is close-up and accurate, with no blind spots in the immediate Field Of Vision



# RECORD: Kapoor & Picard- Limitations



The University of  
Nottingham



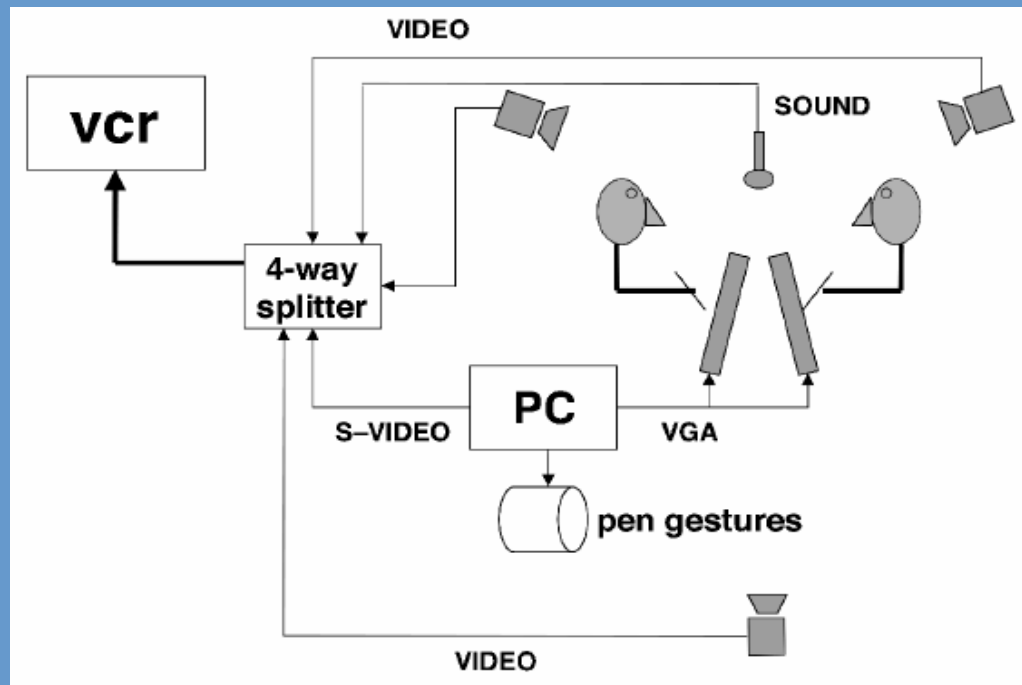
- The Camera is **restricted** in its **positioning**, only enables the filming of an **artificial HCI context**
- Narrow Field Of View can **miss** data
- 'Obstructions' (e.g. glasses) may cause the eye-tracker to **fail**
- Does not record verbal data
- Only reliable for slow movement

# RECORD: Another example



The University of  
Nottingham

## 'SLOT: Spatial Logistics Task System' (De Reuter et al., 2003)



- An example of **HH** interaction.
- Uses 3 cameras 'to provide different position, perspectives and depth' (2003: 411).
- Simultaneous recording of verbal, visual and written data which is compressed into a 4-fold (2x2) split screen image.

# RECORD: SLOT (Spatial Logistics Task System)



The University of  
Nottingham

## Advantages:

Solves problems of **perspective**, allowing us to record dyadic communication.

Facilitates an exploration of **multimodal** data.

## Limitations:

Although it focuses upon HH interaction, it is **not** strictly a '**naturalistic**' setting as we require.

Sound data is not a high priority

# RECORD: Summary/Requirements



The University of  
Nottingham

- \* To record multiple modes of communication in a natural context.
- \* To record both the individual & synchronised patterns of speech / head movements simultaneously, within the same frame of reference.
- \* Recordings to be accurate and able to be replayed & annotated in the future.

# REPRESENT & REPLAY



The University of  
Nottingham

The manual or automatic  
**coding, classification** and  
**representation** of recorded  
data.

# Automatic coding: (FACS) Facial Annotation Coding System



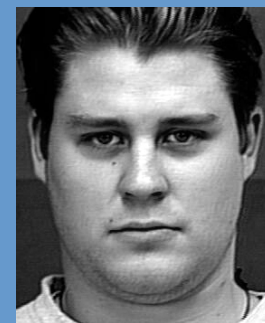
The University of  
Nottingham

- Developed by Ekman et al, 1976.
- Divides face into motion reference points- AUs (Action Units) for encoding movement.
- 46 different AUs account for facial expression, 12 account for head orientation and gaze.



**AU 53**

Head Up



**AU 54**

Head Down

**AU53+54 = Head Nod**

# Using HMMs- Hidden Markov Models



- When defined automatically, AUs are used as input into a HMM 'pattern analyser' in order to statistically define a given sequence of observations, providing a 'measure of confidence' for such definitions.
- It compares past, present and future states (A) of a sequence to define a specific movement (B).

**1:** 'The probability Transition Matrix': Establishes the probability distribution between each state. (where  $A = \{a_{ij}\}$ )

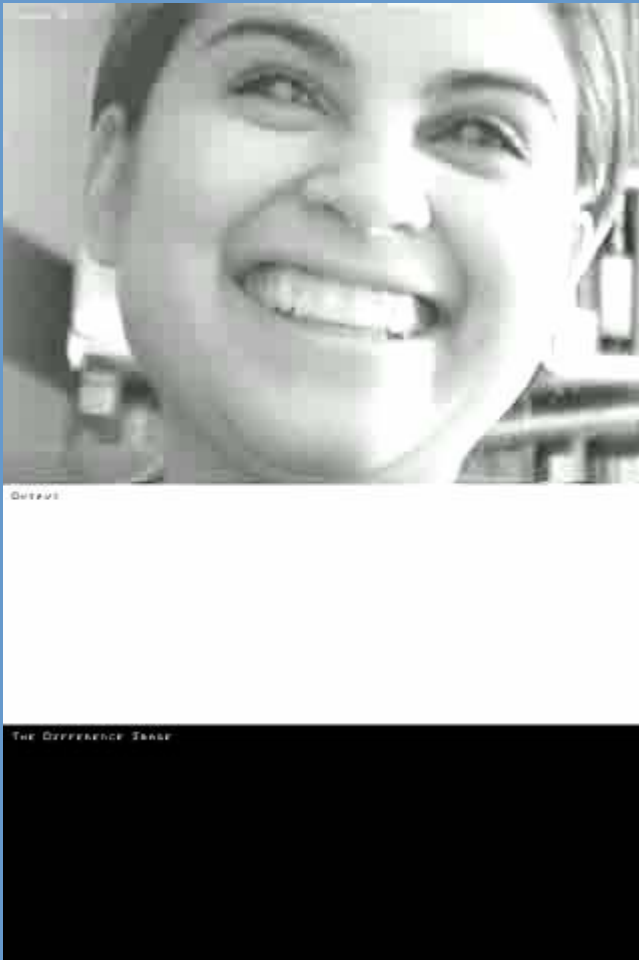
**2:** Establishes the 'output probability matrix' (where  $B = \{b_i(k)\}$ )

**3:** Provides the 'special initial probability vector'

# Limitations- AUs & HMMs



The University of  
Nottingham



AUs 55,  
56, 57, 57

- × 'Irregular' Movements- altering the AU53/A54 pattern may cause problems in the classification of head nods.
- × Process is mainly automatic, little provision for manual coding
- × Codes and classifies movement only, not speech

# Manual Coding: ICODE

<http://www.2.cs.cmu.edu/~face/index2.htm>

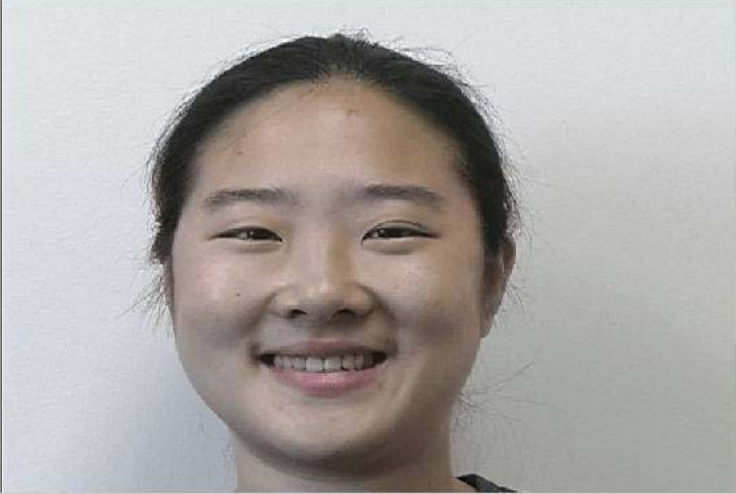


The University of  
Nottingham

ICODE

File ICODE

Open index Load labels Save labels Save Matrix True size image Index: F2A\_s\_Ha\_0001.idx Coder1: za VITC: 00:00:00  
FPS: 30



ICODE, Image sequence coder  
(c) Carnegie Mellon University, University of Pittsburgh  
\$Revision: 1.15 \$

Coder1 Coder2

```
00015: 6+12+25
00016: 6+12+25
00017: 6+12+25
00018: 6+12+25
00019: 6+12+25
00020: 6+12+25
00021: 6+12+25
00022: 6+12+25
00023: 6+12+25
00024: 6+12+25
00025: 6+12+25
00026: 6+12+25
00027: 6+12+25
00028: 6+12+25
00029: 6+12+25
00030: 6+12+25
00031: 6+12+25
00032: 6+12+25
00033: 6+12+25
00034: 6+12+25
00035: 6+12+25
00036: 6+12+25
00037: 6+12+25
00038: 6+12+25
00039: 6+12+25
00040: 6+12+25
00041: 6+12+25
00042: 6+12+25
00043: 6+12+25
00044: 6+12+25
00045: 6+12+25
00046: 6+12+25
00047: 6+12+25
00048: 6+12+25
00049: 6+12+25
00050: 6+12+25
00051: 6+12+25
```

First Last Reverse Stop Play << Step Step >> -10 +10 Preview

Frame: 20 Label: 6+12+25 +

End Frame: 56  
N = 37 Append N Remove N Append -> Remove -> Undo

# Limitations of manual coding systems



The University of  
Nottingham

✘ Consistency of coding can be difficult to maintain- leading to questions of reliability

✘ Statistically unsound?

✘ Transferability of results & techniques across the given context may be difficult- may be study-specific.

# Observation system: DIVER

<http://diver.stanford.edu>



The University of  
Nottingham

The screenshot shows a web browser window titled "WebDiver". The address bar contains "http://diver.stanford.edu". The browser's toolbar shows various bookmarks like "Live Home Page", "Apple", "Apple Support", "Apple Store", "iTools", "Mac OS X", "Micro", "Mac OS X", and "soft MacTopia".

The main content area is titled "WebDiver" and includes a video player showing a man pointing at a whiteboard. Below the video is a search box and a "Search" button. The search options are "Search in Comments" and "Search in Dive Annotations", both of which are checked.

The right-hand side of the page displays a "Welcome, Kenneth Dauber" message with a "Sign Out" link. Below this is the title "Effects of Space, Gesture and Age in Math Problem Solving" and the author "Michael Mills".

The content is organized into sections:

- 1: Introduction**: A video clip (01:34 - 02:25) titled "Visualization..." is shown. The text describes how students at different age levels use facilities like huddle boards and electronic media in their scientific visualizations.
- 2: Guided noticing through gesture**: A video clip (02:01 - 02:24) titled "Visualization..." is shown. The text notes that two groups were "democratic" in sharing, while a third group had a clearly defined leader.
- Comment 1. Communication among students**: A comment by "Harvey Keck, Jr." dated 01/25/03 at 12:44 PM. The message notes that CK is giving AL credit for the main solution by looking in his direction several times.
- 3: Guided noticing through language**: A video clip (10:23 - 11:25) titled "Visualization..." is shown. The text notes that in contrast to JK, LP, and HR, very little body movement and hand gesturing were observed during their dialogue.

The browser's status bar at the bottom indicates "Internet zone".

# WHERE NEXT?



- **MULTI-MODAL:** Allow for the analysis and exploration of data from a variety of multimedia (sound and visual data) simultaneously both within a single frame and a combined frame of reference when desired.
- **FLEXIBLE:** To allow the exploration of specific frames or sequences of data, as well as allowing the exploration of specific modes of data.
- **SYSTEMATIC:** Accurate and systematic analysis of verbal and visual records, achieved either through automatic or manual coding.
- **PROFICIENT:** To be able to synthesise, tag, code and transcribe large quantities of multimodal (sound and visual) data.
- **ACCESSIBLE:** Produce a user-friendly interface to access and search specific frames or sequences of frames.

# Towards a video-corpus interface: an example



The University of  
Nottingham

## Paradigmatic representation (based on linguistic form)

<\$1> Right. <\$2> Yeah we did have some forms whe	Sound	Video
whatever. <\\$O1> Yeah. Okay. <\$1> Thanks very	Sound	Video
now then? <\$2> Yeah if you could. If you could pla	Sound	Video
you there? <\$2> Yeah.	Sound	Video
<\$O2> You are yeah. <\\$O2> Yeah. Yeah. <\$2	Sound	Video
<\$2> <\$O2> Yeah. <\\$O2> Mm. Right. Actually	Sound	Video
an hour? <\$2> Yeah okay. <\$1> Okay. Bye.	Sound	Video

# Towards a video-corpus interface: another example



The University of  
Nottingham

## Syntagmatic representation (based on linguistic form)

S1: The way this was done was a Scottish lady who lived across the road from us.

S2: Yeah.

S1: And she would soak some grey wool. A length of grey wool in a saucer with olive oil.

S2: Yeah.

S1: And then she'd tread it through an extremely large darning needle.

S2: Yeah

S1: Then there was a cork held together. It was a perfectly clean cork a new cork held behind your earlobe and she just treaded the needle with the wool [...]

**Sound**   **Video**

