Multimodal Meeting Segmentation

Marc Al-Hames
Institute for Human-Machine Communication
Technische Universität München, Germany
alh@mmk.ei.tum.de

Augmented Multi-party Interaction (AMI)
http://www.amiproject.org

Wednesday, 13th July 2005
10.00 – 11.00
Outline

Augmented Multi-party Interaction (AMI)

Multimodal meeting segmentation

Graphical models and multimodal meeting segmentation

Summary
Augmented Multi-party Interaction

Integrated Project
under the European Commission's
Sixth Framework Programme

Jointly managed by
Prof. Herve Bourlard (IDIAP) and
Prof. Steve Renals (University of Edinburgh)

www.amiproject.org
AMI General Goals

“AMI is concerned with new multimodal technologies to support human interaction, in the context of smart meeting rooms and remote meeting assistants.

The project aims to enhance the value of multimodal meeting recordings and to make human interaction more effective in real time.

These goals are being achieved by developing new tools for computer supported cooperative work and by designing new ways to search and browse meetings as part of an integrated multimodal group communication, captured from a wide range of devices.”
AMI at a Glance

• 3 instrumented “smart” meeting rooms

• 100 hours of multimodal data

• Development of meeting browsers and remote meeting assistants for instrumented meeting rooms and required component technologies

• R&D themes
  – Group dynamics
  – Audio/visual/multimodal processing
  – Content abstraction
  – Human-computer interaction
IDIAP Smart Meeting Room

- Whiteboard
- Projector Screen
- Microphone array
- Participants
- Lapel microphones
- Cameras
AMI Camera Output

Left

Centre

Right

Close-up 1

Close-up 2

Close-up 3

Close-up 4
AMI Meeting Example

Centre

Close-up 3

Right
Meeting Browser

P. Wellner, M. Flynn, and M. Guillemot, Browsing Recorded Meetings with Ferret, IDIAP-RR 04-32, 2004
Augmented Multi-party Interaction (AMI)

Multimodal meeting segmentation
  Introduction
  Features
  Models

Graphical models and multimodal meeting segmentation

Summary
Meetings can be modelled as a sequence of events, group, or individual actions from a set

\[ A = \{A_1, A_2, A_3, \ldots, A_N\} \]
Meetings can be modelled as a sequence of events, group, or individual actions from a set

\[ A = \{A_1, A_2, A_3, \ldots, A_N\} \]

Different sets represent different meeting views
Structuring of Meetings

Group Level:
Phase: Discussion
Task: Decision making
Interest Level: High

Individual #1:
Action: Speaking
Interest Level: High

Individual #3:
Action: Idle
Interest Level: Low

Meetings: A multimodal problem!
Meeting Views

### Group

<table>
<thead>
<tr>
<th>Discussion Phase</th>
<th>Task</th>
<th>Interest Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion</td>
<td>Brainstorming</td>
<td>High</td>
</tr>
<tr>
<td>Monologue (with ID)</td>
<td>Decision making</td>
<td>Neutral</td>
</tr>
<tr>
<td>Note-taking</td>
<td>Information sharing</td>
<td>Low</td>
</tr>
<tr>
<td>Presentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whiteboard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monologue (ID) + Note-taking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presentation + Note-taking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whiteboard + Note-taking</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Individual

<table>
<thead>
<tr>
<th>Actions</th>
<th>Interest Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking</td>
<td>High</td>
</tr>
<tr>
<td>Writing</td>
<td>Neutral</td>
</tr>
<tr>
<td>Idle</td>
<td>Low</td>
</tr>
</tbody>
</table>

Further sets could represent the **current agenda item** or the **topic discussed**, but may require higher semantic knowledge for an automatic analysis.
**Action Lexicon**

**Group**

**Discussion Phase**
- Discussion
- Monologue (with ID)
- Note-taking
- Presentation
- Whiteboard
- Monologue (ID) + Note-taking
- Presentation + Note-taking
- Whiteboard + Note-taking

**Action lexicon I**
Only single actions → 8 action classes

**Action lexicon II**
Single actions and combinations of parallel actions → 14 action classes
The output can then be used as **direct input to a meeting browser**, or as **input for a higher semantic structuring**
Outline

Augmented Multi-party Interaction (AMI)

Multimodal meeting segmentation
  Introduction
  Features
  Models

Graphical models and multimodal meeting segmentation

Summary
Audio Features

Microphone array
For each position

• For each seat (4), whiteboard, and projector screen
• SRP-PHAT measure to estimate a speech activity
• And a speech and silence segmentation

Lapel microphones
For each person

• For each speech segment
• Energy
• Pitch (SIFT algorithm)
• Speaking rate (combination of estimators)
• MFC-Coefficients
Lexical Features

For each person

• Speech transcription (or output of ASR)

• Gesture transcription (or output of an automatic gesture recognizer)

• Gesture inventory
  – Writing
  – Pointing
  – Standing up
  – Sitting down
  – Nodding
  – Shaking head
Visual Features

**Global motion**
*For each position*

- Centre of motion
- Changes in motion (dynamics)
- Mean absolute deviation
- Intensity of motion

**Skin colour blobs (GMM)**
*For each person*

- Head: orientation and motion
- Hands: Position, size, orientation, and motion
- Moving blobs from background subtraction
Global Motion Features

- For each location $L$ a difference image sequence is calculated: $I_d^L(x, y, t)$

- Centre of motion for the x- and y-direction:

  $$m_x^L(t) = \frac{\sum_{(x,y)} x \cdot |I_d^L(x,y,t)|}{\sum_{(x,y)} I_d^L(x,y,t)}$$
  $$m_y^L(t) = \frac{\sum_{(x,y)} y \cdot |I_d^L(x,y,t)|}{\sum_{(x,y)} I_d^L(x,y,t)}$$

- Changes in motion express the dynamic of the movements:

  $$\Delta m_x^L(t) = m_x^L(t) - m_x^L(t-1)$$
  $$\Delta m_y^L(t) = m_y^L(t) - m_y^L(t-1)$$

- Mean absolute deviation relative to the centre of motion:

  $$\sigma_x^L(t) = \frac{\sum_{(x,y)} |I_d^L(x,y,t)| \cdot (x - m_x^L(t))}{\sum_{(x,y)} I_d^L(x,y,t)}$$
  $$\sigma_y^L(t) = \frac{\sum_{(x,y)} |I_d^L(x,y,t)| \cdot (y - m_y^L(t))}{\sum_{(x,y)} I_d^L(x,y,t)}$$

- Intensity of motion:

  $$i^L(t) = \frac{\sum_{(x,y)} I_d^L(x,y,t)}{x \cdot y}$$
Global Motions

M. Zobl et al., *Action Recognition in Meeting Scenarios using Global Motion features*, PETS-ICVS, 2003
Global Motions

M. Zobl et al., Action Recognition in Meeting Scenarios using Global Motion features, PETS-ICVS, 2003
Global Motions
Outline

Augmented Multi-party Interaction (AMI)

Multimodal meeting segmentation
  Introduction
  Features
  Models

Graphical models and multimodal meeting segmentation

Summary
BIC-like approach

- Strategy similar to the Bayesian information criterion (BIC)
- Segmentation and classification in one step

- Two windows with variable length shifted over the time scale
- Inner border is shifted from left to right
- Classify each window: different results for left and right window
  - inner border is considered as a boundary of a meeting event
- No boundary is detected -> enlarge the whole window
- Repeat until right border reaches end of meeting
BIC-like approach

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Insertion-rate [%]</th>
<th>Deletion-rate [%]</th>
<th>Accuracy [s]</th>
<th>Rec-Error [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>14.7</td>
<td>6.22</td>
<td>7.93</td>
<td>39.0</td>
</tr>
<tr>
<td>GMM</td>
<td>24.7</td>
<td>2.33</td>
<td>10.8</td>
<td>41.4</td>
</tr>
<tr>
<td>MLP</td>
<td>8.61</td>
<td>1.67</td>
<td>6.33</td>
<td>32.4</td>
</tr>
<tr>
<td>RBF</td>
<td>6.89</td>
<td>3.00</td>
<td>5.66</td>
<td>31.6</td>
</tr>
<tr>
<td>SVM</td>
<td>17.7</td>
<td>0.83</td>
<td>9.08</td>
<td>35.7</td>
</tr>
</tbody>
</table>

Experimental setup: 57 meetings using “Lexicon 1”
Multi-layer Hidden Markov Model

- Individual action layer: **I-HMM**
  - Speaking, Writing, Idle

- Group action layer: **G-HMM**
  - Discussion, Monologue, …

- Each layer trained independently

- Simultaneous segmentation and recognition

Multi-layer Hidden Markov Model

- Smaller observation space -> stable for limited training data
- I-HMMs are person independent -> much more training data from different persons available
- G-HMM less sensitive to small changes in the low-level features
- Two layers are trained independently -> different HMM combinations can be explored
### Multi-layer Hidden Markov Model

#### Table: AER (%) for Single-layer and Multi-layer HMMs

<table>
<thead>
<tr>
<th>Method</th>
<th>AER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-layer HMM</strong></td>
<td></td>
</tr>
<tr>
<td>Visual only</td>
<td>48.2</td>
</tr>
<tr>
<td>Audio only</td>
<td>36.7</td>
</tr>
<tr>
<td>Early integration</td>
<td>23.7</td>
</tr>
<tr>
<td>Multi-stream</td>
<td>23.1</td>
</tr>
<tr>
<td>Asynchronous</td>
<td>22.2</td>
</tr>
<tr>
<td><strong>Multi-layer HMM</strong></td>
<td></td>
</tr>
<tr>
<td>Visual only</td>
<td>42.4</td>
</tr>
<tr>
<td>Audio only</td>
<td>32.3</td>
</tr>
<tr>
<td>Early Integration</td>
<td>16.5</td>
</tr>
<tr>
<td>Multi-stream</td>
<td>15.8</td>
</tr>
<tr>
<td>Asynchronous</td>
<td>15.1</td>
</tr>
</tbody>
</table>

Experimental setup: 59 meetings using “Lexicon 2”
Outline

Augmented Multi-party Interaction (AMI)

Multimodal meeting segmentation

Graphical models and multimodal meeting segmentation

Summary
Graphical Models in a Nutshell

• Graphical models are a marriage between probability and graph theory

• They provide an intuitive way of modelling data and dependencies

• They can deal with two common problems in pattern recognition: Complexity and uncertainty
Bayesian Networks

- Variables are modelled with nodes:
  - Discrete: \( P(A = \bar{x}) = \sum_{i=1}^{N} w_i \cdot \delta(\bar{x} - \bar{\mu}_i) \)
  - Continuous: \( P(B = \bar{y}) = N(\bar{y}, \bar{\mu}, \Sigma) \)

- Dependencies between them with arcs:
  \[ P(A, B) = P(B \mid A) \cdot P(A) \]
Bayesian Networks

- Each node can have several parent and child nodes

![Bayesian Network Diagram]

- The joint probability of the model then becomes:

\[
P(X_1 \ldots X_N) = \prod_{i=1}^{N} P(X_i \mid pa(X_i))
\]
Will Your Paper Be Accepted

Accept

- Work
  - Scope
  - Results
  - Interest
  - Writing

- Style
  - Layout
  - Referene.

- Soft
  - Known
  - Board
Introducing Time: Dynamic BN

- Use a BN to model one “time slice”

- Repeat ("unroll") the BN for each time step of the observation

- Define connections between subsequent time-slices
Some Models Represented as DBN

Discrete HMM

HMM

HMM with GM

IOHMM

LDS
More about Graphical Models

Boring  Too hard  Too easy  Need a try

Time


K. Murphy’s Bayes Net Toolbox for Matlab (BNT): Free, open source toolbox; perfect to start with

Interesting topics, such as DBNs or approximate inference are addressed in: M. Jordan (editor), Learning in Graphical Models, MIT Press, 1996
Multi-Stream DBN Model

- **Counter structure**
  - Improves state-duration modelling
  - Action counter C is incremented after each “action transition”

- **Hierarchical approach**

- **State-space decomposition**:
  - Two feature related sub-action nodes ($S_1^1$ and $S_2^2$)
  - Each sub-state corresponds to a cluster of feature vectors
  - Unsupervised training

- **Independent modality processing**

---

Marc Al-Hames, Technische Universität München, Germany

*Multimodal Meeting Segmentation*

---

A. Dieball and S. Renals, *Dynamic Bayesian networks for meeting structuring*, ICASSP, 2004
## Experimental Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>6x6</td>
<td>7x7</td>
<td>6x6</td>
</tr>
<tr>
<td>UEDIN (Turn.+Pros.)</td>
<td>48.4</td>
<td>17.1</td>
<td>11.0</td>
<td>18.9</td>
</tr>
<tr>
<td>IDIAP (Audio + Video)</td>
<td>61.6</td>
<td>26.7</td>
<td>16.7</td>
<td>24.9</td>
</tr>
<tr>
<td>TUM (Beamf. + Video)</td>
<td>92.9</td>
<td>21.4</td>
<td>21.4</td>
<td>21.4</td>
</tr>
</tbody>
</table>

Experimental setup: cross-validation over 53 meetings using “Lexicon 1”
Disturbed Channels

- Multi-modal features always outperform single-modal features
- Use as much channels as possible
- But, what if one channel is disturbed or occluded?
  - We certainly don’t want the disturbed channel to ruin our recognition results!
  - Use all streams to improve a disturbed one (even without knowing it is disturbed)
  - Thus squeeze the last bit of information out of each channel
Disturbed Meeting Data
Couple an HMM with a LDS

- Couple a multi-stream HMM with a linear dynamical system (LDS)
- HMM is driving input for the LDS
- With the HMM as driving input the visual channel is improved
LDS with state-space equations:

\[ \begin{align*}
\ddot{x}_0 &= B\ddot{u}_0 + \ddot{v}_0 \\
\ddot{x}_t &= A\ddot{x}_{t-1} + B\ddot{u}_t + \ddot{v}_t \\
\ddot{o}_t &= C\ddot{x}_t + \ddot{w}_t
\end{align*} \]

LDS with probability distributions:

\[ \begin{align*}
P(\ddot{x}_0 | \ddot{u}_0) &= N(B\ddot{u}_0, \ddot{\mu}, \Sigma) \\
P(\ddot{x}_t | \ddot{x}_{t-1}, \ddot{u}_t) &= N(\ddot{x}_t - A\ddot{x}_{t-1} - B\ddot{u}_t, \ddot{\mu}, \Sigma) \\
P(\ddot{o}_t | \ddot{x}_t) &= N(\ddot{o}_t - C\ddot{x}_t, \ddot{\mu}, \Sigma)
\end{align*} \]

x: hidden state, u: driving input, o: observation, A: state transition, B: input, C: observation matrix
Multi-Stream Mixed-State DBN

Multi-Stream HMM

LDS

Micro. array
Lapel microphones
Visual

Ob, H, O, M, X, U, V

Discrete
Gaussian
Observed
Joint Probability

The joint probability of the model, given the unknown observation $O$ is

$$P(O, E_j) = P(H_0^B) \prod_{t=1}^{T-1} P(H_t^B | H_{t-1}^B, H_{t-1}^M) \prod_{t=0}^{T-1} P(O_t^B | H_t^B)$$

$$P(H_0^M) \prod_{t=1}^{T-1} P(H_t^M | H_{t-1}^M, H_{t-1}^B, H_{t-1}^V)$$

$$\prod_{t=0}^{T-1} P(O_t^M | M_t^M, H_t^M)P(M_t^M | H_t^M)$$

$$P(H_0^V) \prod_{t=1}^{T-1} P(H_t^V | H_{t-1}^V, H_{t-1}^M) \prod_{t=0}^{T-1} P(U_t^V | H_t^V)$$

$$P(X_0^V | U_0^V) \prod_{t=1}^{T-1} P(X_t^V | X_{t-1}^V, U_{t-1}^V) \prod_{t=0}^{T-1} P(O_t^V | X_t^V)$$
Multi-Stream Mixed-State DBN

- The model is parameterised by $E_j = \{H^B, H^M, M^M, H^V, U^V, X^V\}$
- Train parameters for each class with approximate EM algorithm
- Classify with: $\text{argmax } P(O|E_j)$
## Experimental Results

### Recognition Results (%)

<table>
<thead>
<tr>
<th></th>
<th>Single modal HMMs</th>
<th>Multi-modal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Audio</td>
<td>Array</td>
</tr>
<tr>
<td>I: Clear test data</td>
<td>83.1</td>
<td>83.6</td>
</tr>
<tr>
<td>II: Lapel disturbed</td>
<td>61.1</td>
<td></td>
</tr>
<tr>
<td>III: Array disturbed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV: Video disturbed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V: All streams disturbed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The model uses much more information from the visual channel
- Thus it increases the robustness against disturbances in the channels
Summary

- Augmented Multi-party Interaction (AMI)
- Meeting segmentation
- Multimodal features
- Graphical Models
- Models for automatic meeting segmentation
References

EU Multimodal (Meeting) Projects
• AMI, Augmented Multi-party Interaction, http://www.amiproject.org
• M4, Multimodal meeting manager, http://www.m4project.org, (finished)
• CHIL, Computers In the Human Interaction Loop, http://chil.server.de

Smart Meeting Room

Meeting Browser

The AMI website, contains a wealth of information about the project.
Global Motion Features:


Graphical Models:

- J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publisher, 1988

Both Jensen books: very good tutorials on static BNs.

“The” fundamental book

Practical book, covers dynamic BNs as well

Open source Matlab BN Toolbox. Good examples. Good to start with. Yet, no Viterbi!
Multimodal Meeting Segmentation:

- D. Zhang et al., *Multimodal Group Action Clustering in Meetings*, ACM Multimedia, 2004

The first paper introducing the action classes
References

Multimodal Meeting Segmentation:


• M. Al-Hames and G. Rigoll, *A Multi-Modal Mixed-State DBN for Robust Meeting Event Recognition from Disturbed Data*, ICME, 2005

• M. Al-Hames and G. Rigoll, *A Multi-modal Graphical Model for Robust Recognition of Group Actions in Meetings from Disturbed Videos*, ICIP, 2005

• S. Reiter and G. Rigoll, *Segmentation and Classification of Meeting Events using Multiple Classifier Fusion and Dynamic Programming*, ICPR, 2004


• S. Reiter and G. Rigoll, *A Neural-Field-like Approach for Modeling Human Group Actions in Meetings*, ICME, 2005