Layered Representations for Human Activity Recognition

Nuria Oliver
Ashutosh Garg
Eric Horvitz

Adaptive Systems & Interaction
Microsoft Research

Outline

- Motivation & Approach:
  - Human Behavior Perception & Understanding
- Perceptual Input:
  - Vision, Speech, Keyboard-mouse
- Layered HMMs for Behavior Understanding
- SEER: A real-time multimodal office awareness system
- Experiments
- Future Work & Discussion

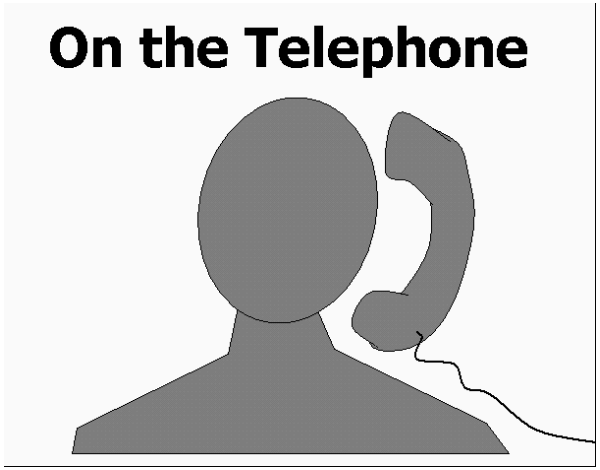
Introduction

- GOAL: Automatic recognition of human behavior from sensory observations
- Relevant in many domains:
  - visual surveillance, office awareness
  - multimodal human-computer interaction
  - accessibility, medical applications
- CONTEXT is critical:
  - Endowing computers with richer notions of context can enhance the communication between humans and computers

Context-Sensitive Systems

- Location and Identity have been the most common properties considered as ‘context’
- Richer notions of context include the user’s intentions and past/present actions
- Multimodal systems allow for richer context
- A key component is the fusion of low-level streams of sensor data with higher-level assessments of activity

Example of Multi-scale Activity Recognition



Time

Audio

Phone Ring | Speech | Silence | Speech | Silence

Video

One Person Present

High Level Interpretation

On the Telephone

Desirable Properties

- Hierarchical statistical approach for detecting and recognizing human activities
- Multimodal, real-time
- Multiple levels of abstraction, time granularities
- Robustness to changes in the environment: lighting, users, room
- Trainable and easy to train

Previous Work

- Computer Vision
  - HMMs and Extensions for gestures/activity recognition (Starner&Pentland ISCV'95, Wilson&Bobick ICCV'98, Galata et al IJCV-01,Brand et al CVPR'96, Yivanov&Bobick ICCV'98)
  - Bayesian Networks for complex activity recognition in a particular domain (Intille&Bobick AAAI'99, Madabhushi&Aggarwal'99, Buxton et al ICCV'95)
  - Multimodal Wearable System (Clarkson&Pentland'99)
- AAAI/CHI
  - Inference of the user's context and intentions (Horvitz et al UAI'99, Salber et al CHI'99, Horvitz et al UAI'98)
  - Office Awareness (Johnson&Greenberg'99, Pedersen&Soloker CHI'97, Gellersen et al HUC'99)

Tractable and Robust Context Sensing

- We have developed a *probabilistic representation* based on a tiered formulation of dynamic graphical models that we refer to as Layered Hidden Markov Models (LHMMs)

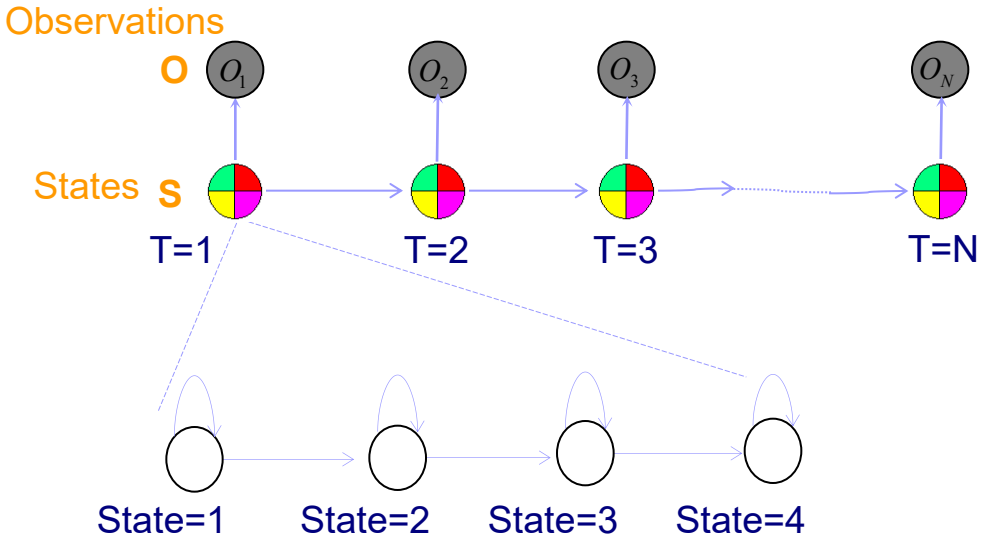
Multimodal Input

- Vision: one static USB camera sampled at 15 fps
- Audio: two binaural mini-microphones (20-16000 Hz, SNR 58 dB), sampled at 44100 KHz
- Keyboard and mouse: history of the activity during the last 5 sec

HMMs for Behavior Recognition

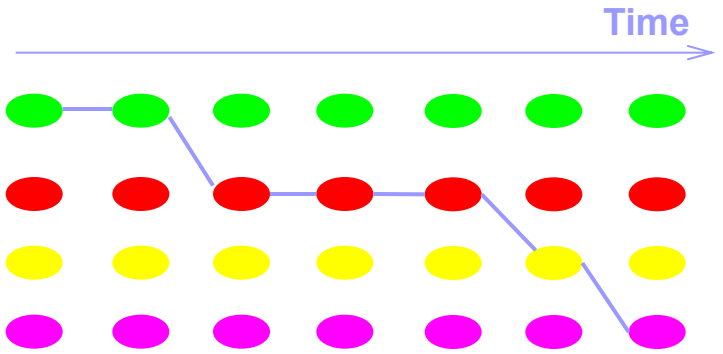
4 state HMM

Graphical Model



State Trellis

Most likely Path ⟶ Viterbi



$$P[S \mid O] \propto P[S]P[O \mid S] = \left(P_{s_1} \prod_{t=2}^{T} P_{s_t \mid s_{t-1}} \right) \left(\prod_{t=1}^{T} P_{s_t}(o_t) \right)$$

$$\hat{S} \mid O = \arg\max_{S} \left(P_{s_1} p_{s_1}(o_1) \prod_{t=2}^{T} P_{s_t \mid s_{t-1}} p_{s_t}(o_t) \right)$$

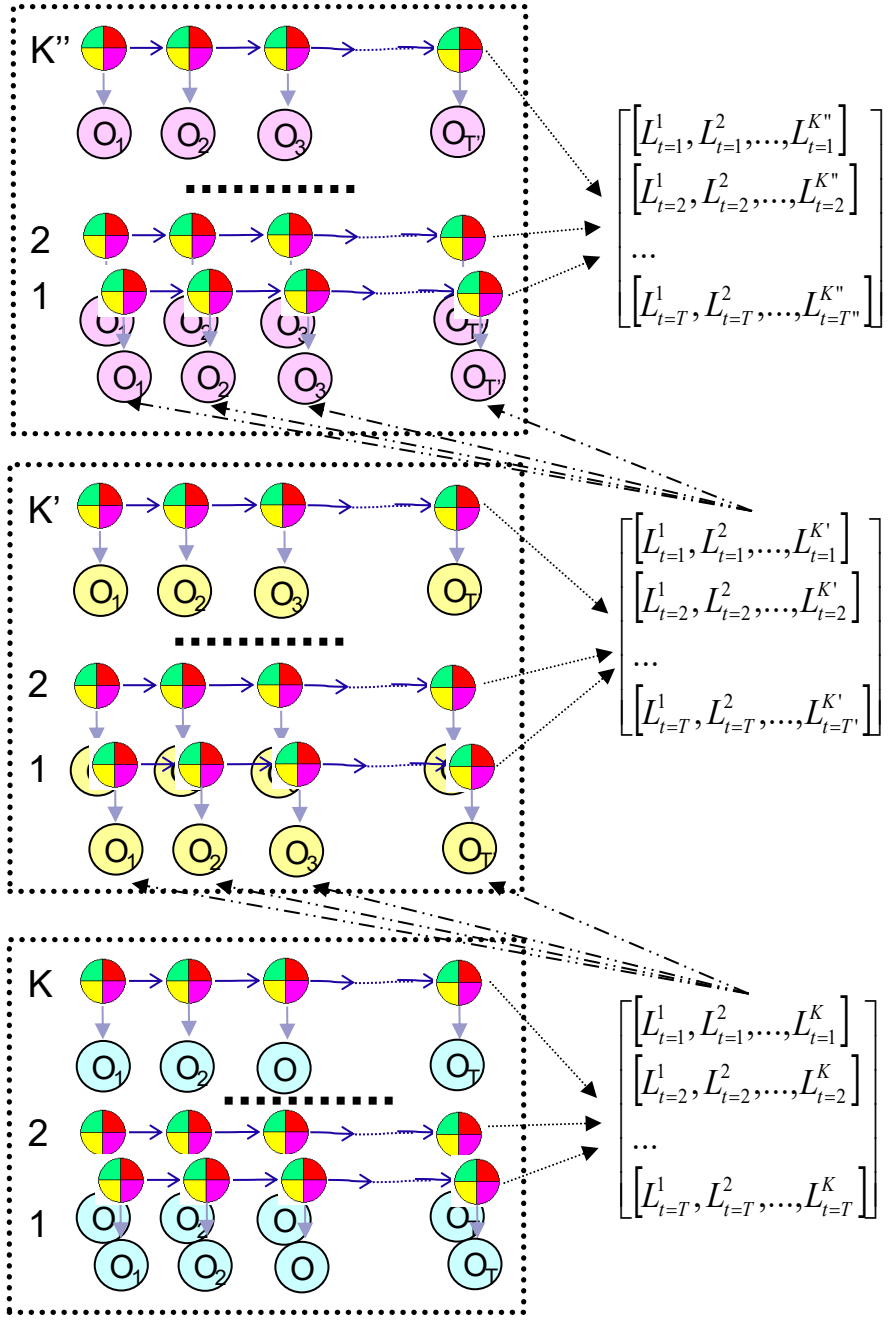
$$\hat{S}_{i,t} \mid O = \arg\max_{j} \left(p_i(o_t) P_{i \mid j} \hat{S}_{j,t-1} \mid O \right)$$

HMMs: Limitations

- First-order Markov assumption may not handle *long-term dependencies* and *multiple time granularities*
- Based on single process dynamics. BUT, signals may be generated by multiple processes. (This is not well modeled by the unstructured transition matrix of HMMs)
- Context limited to a single state variable. If multiple processes (e.g. signals coming from multiple sensors), the Cartesian product HMM becomes intractable very quickly.

LHMMs: Layered HMMs

- Goal: Decompose the parameter space in a way to enhance ROBUSTNESS of the system by reducing training and tuning requirements
- Approach: Segment the problem into distinct layers that operate at different temporal granularities
- Consequence: The data can be explained at different levels of abstraction from pointwise observations at particular times into explanations at varying temporal intervals



$$\begin{bmatrix} [L^1_{t=1}, L^2_{t=1}, \ldots, L^{K''}_{t=1}] \\ [L^1_{t=2}, L^2_{t=2}, \ldots, L^{K''}_{t=2}] \\ \ldots \\ [L^1_{t=T}, L^2_{t=T}, \ldots, L^{K''}_{t=T''}] \end{bmatrix}$$

$$\begin{bmatrix} [L^1_{t=1}, L^2_{t=1}, \ldots, L^{K'}_{t=1}] \\ [L^1_{t=2}, L^2_{t=2}, \ldots, L^{K'}_{t=2}] \\ \ldots \\ [L^1_{t=T}, L^2_{t=T}, \ldots, L^{K'}_{t=T'}] \end{bmatrix}$$

$$\begin{bmatrix} [L^1_{t=1}, L^2_{t=1}, \ldots, L^{K}_{t=1}] \\ [L^1_{t=2}, L^2_{t=2}, \ldots, L^{K}_{t=2}] \\ \ldots \\ [L^1_{t=T}, L^2_{t=T}, \ldots, L^{K}_{t=T}] \end{bmatrix}$$

Layer 3

Bank of K”
-Level 3 HMM
Classifiers with
time granularity T”

Layer 2

Bank of K’
-Level 2 HMM
Classifiers with
time granularity T’

Layer 1

Bank of K
-Level 1 HMM
Classifiers with
time granularity T

LHMMs: Layered HMMs

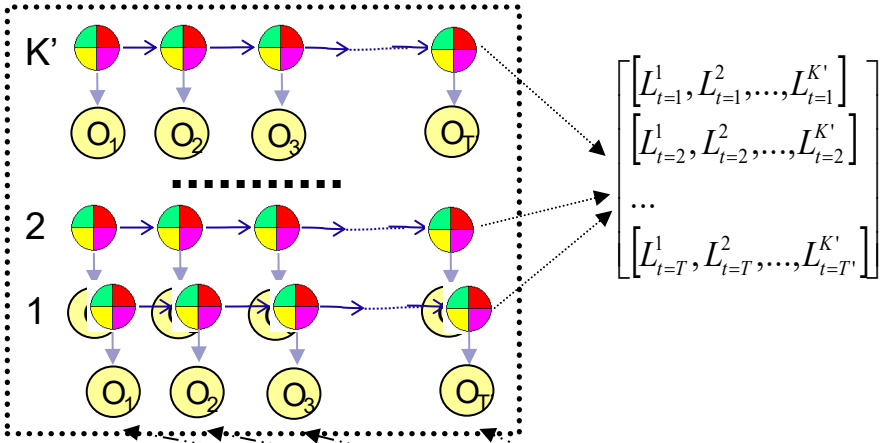
- Each layer handles a higher time granularity than the previous layer
- The inferential results of one layer are the inputs to the next layer
  - MaxBelief: The index of the maximum likelihood model is passed to the next level
  - Distributional: The full pdf over the models is passed to the next level
- Only the lowest layer (leaves) ‘sees’ the observations

Layered HMMs

- Inference is possible at any layer, i.e. at different time granularities
- Learning is performed independently on each layer using standard HMM Baum-Welch Algorithm
- Possible retraining of the lower layers without retraining the higher layers

Layered HMMs

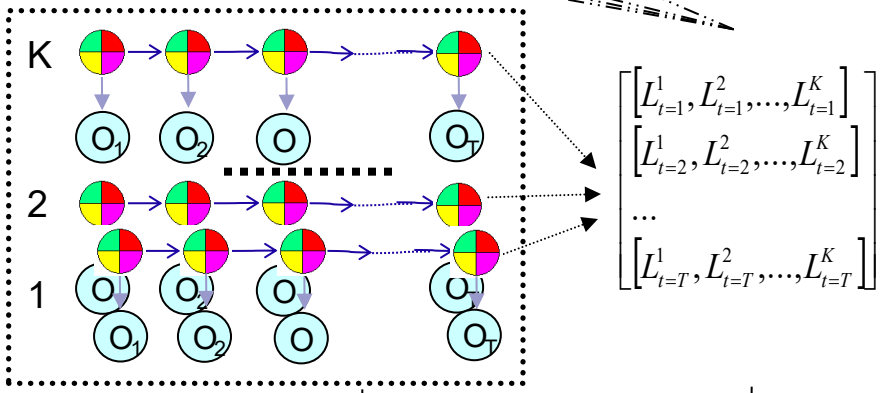
Phone Conversation
Face to Face Conversation
Working on the Computer
Nobody Around
Distant Conversation
Presentation



Layer 2

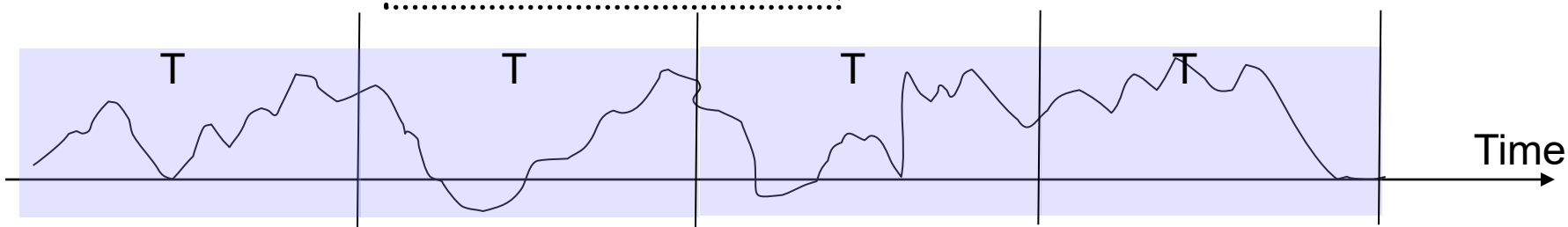
Bank of K'
-Level 2 HMM Classifiers with time granularity T'

Phone Ringing
Human Speech
Music
Keyboard Typing
Office Noise

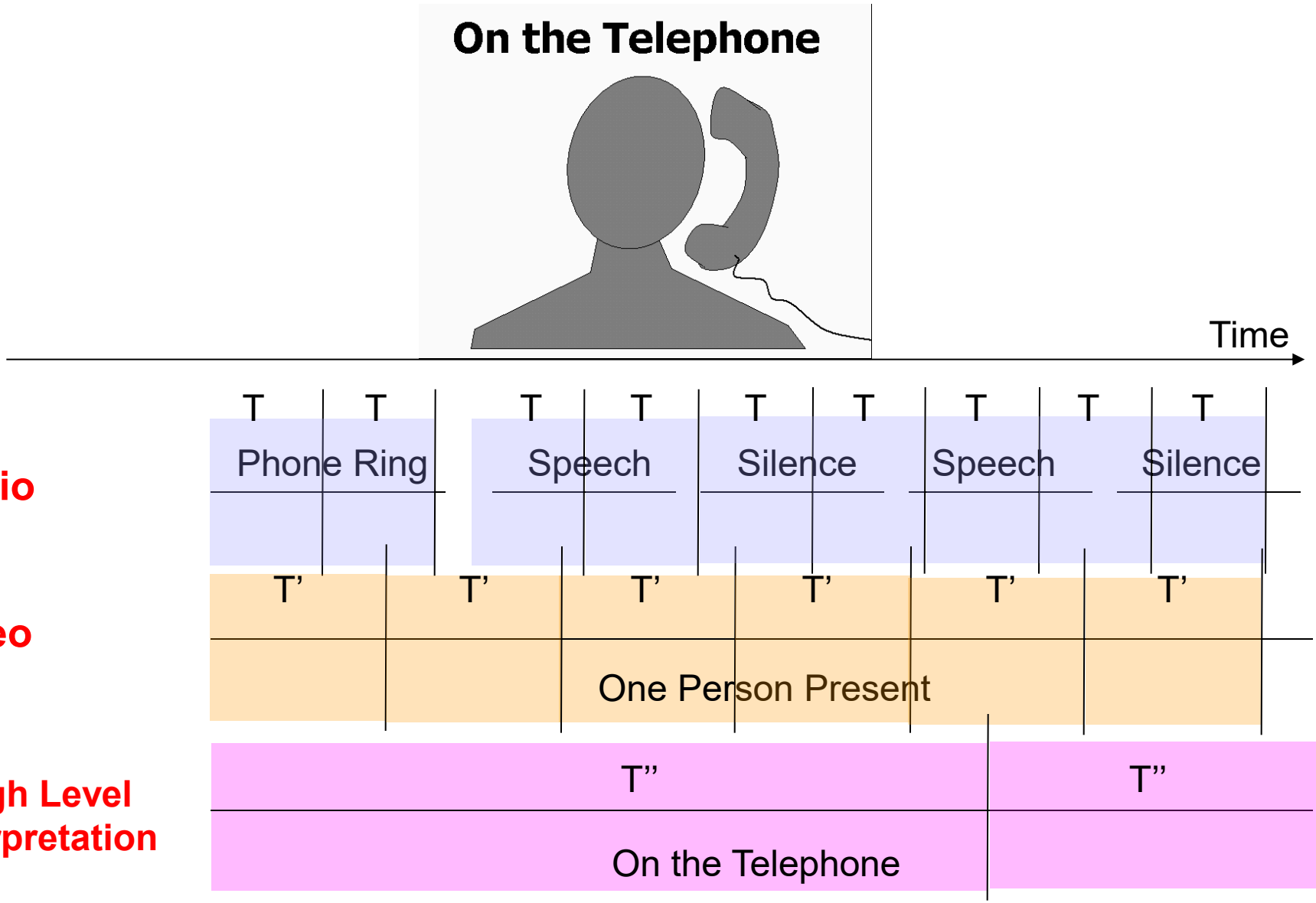


Layer 1

Bank of K
-Level 1 HMM Classifiers with time granularity T



Interpretations at Multiple Levels of Granularity

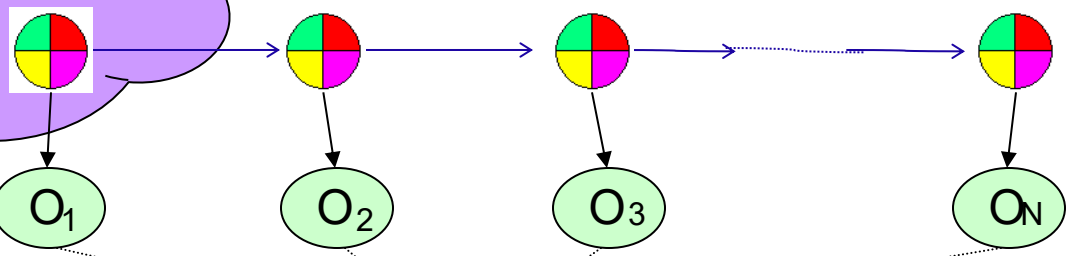


Seer: A Multimodal Office Awareness System

- Seer employs a two-layer HMM architecture
- Multimodal Input:
  - Vision
  - Audio
  - Keyboard/mouse activities
- Automatic, real-time recognition of six typical office activities:
  - Phone Conversation
  - Face to Face Conversation
  - Working on the Computer
  - Presentation
  - Nobody Around
  - Distant Conversation

Face to Face Conversation
Phone Conversation
Presentation
Person Present/Other activity
Distant Conversation
Nobody

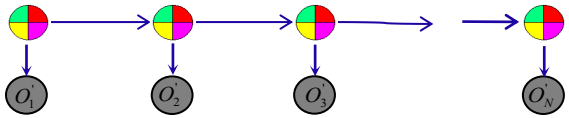
Behavior Models



Sound Localization

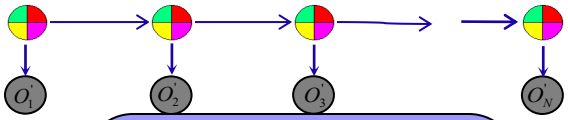
Feature Vector

Sound Classification



AUDIO PROCESSING:
PCA on LPC,
Discriminitive Features

Person Detection



VIDEO PROCESSING:
Skin color density,
Motion density,
Foreground/Background Segmentation,
Face Detection

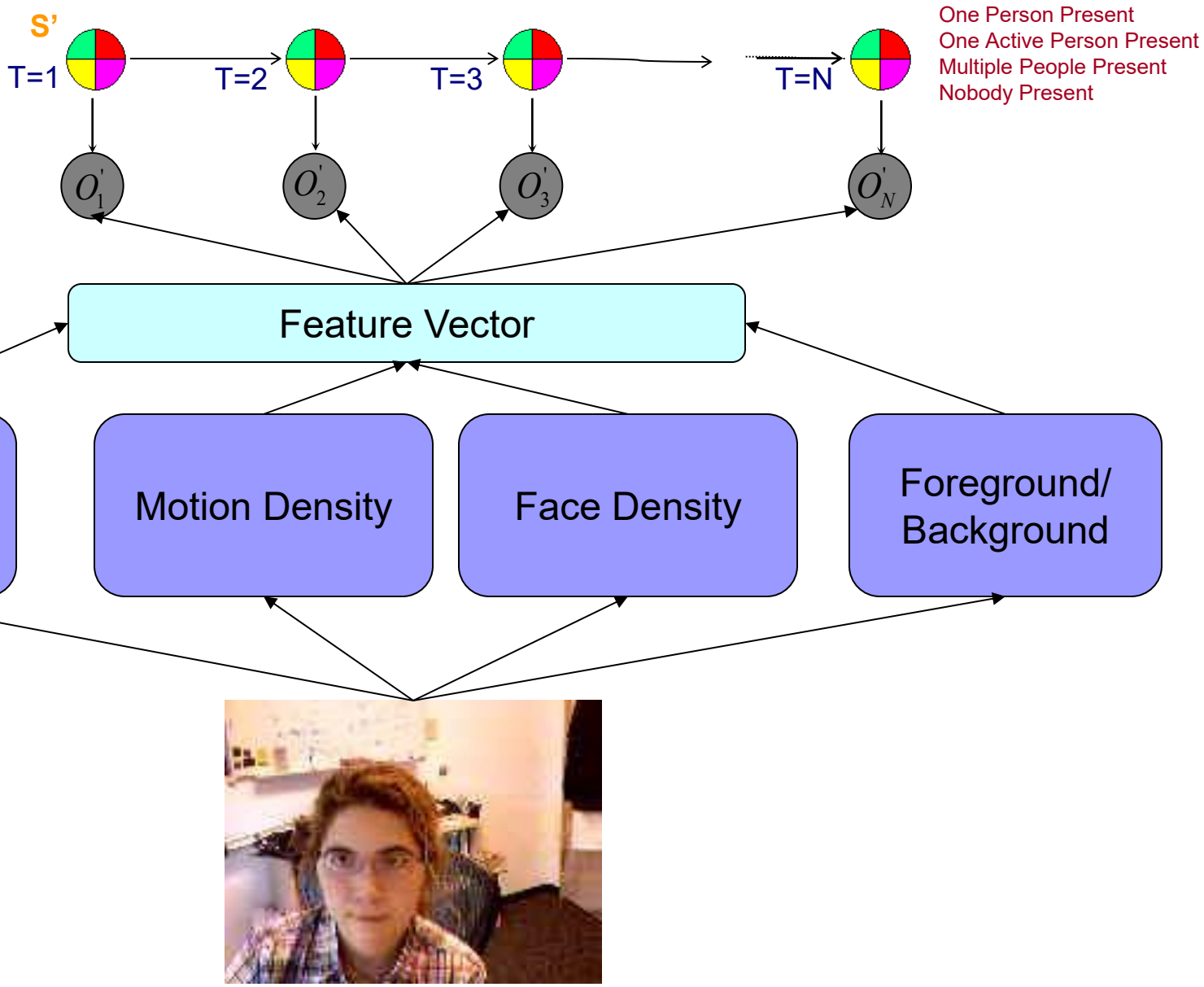
Computer Activity

Vision Processing

- Person Detection based on:
  - Color: Discriminative histogram-based detection in HS space
  - Motion: Motion density using image differences
  - Foreground/background segmentation: Background learned at initialization time.
  - Face Detection based on the MSR Beijing's Face Detector
- The four vision-based features are the input to the vision HMMs for determining the number of users present in the office

Vision Processing

'Vision' HMMs for Detecting the Number of Persons Present



Audio Processing

- Audio signals coming from 2 binaural mini-microphones
- Feature selection of the Linear Predictive Coding (LPC) Coefficients using PCA
  - 95% variability of the data is represented
  - Normally no more than 7 features

Audio Processing

- Other features:
  - Zero Crossing Rate (ZCR)

$$Z_s(m) = \frac{1}{N} \sum_{n=m-N+1}^{m} \frac{\left| sign(s(n)) - sign(s(n-1)) \right|}{2} \bullet w(m-n)$$

with ‘m’ the frame number, ‘N’ the frame length, ‘w’ window function, ‘s(n)’ digitized audio signal at frame ‘n’, and sign(s(n))={+1, s(n)>=0, -1, s(n)<0}

  - Energy, mean and variance of the fundamental frequency over a time window

Audio Processing

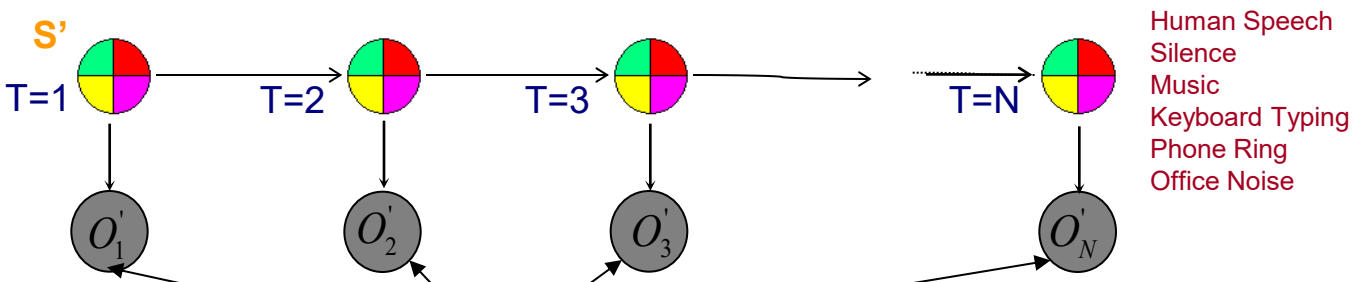
- Sound Localization based on Time Delay of Arrival (TDOA)
- Assume no reverberation
- Find the peak in the time correlation of the signals $x_i(n)$ coming from both microphones

$$x_i(n) = a_i s(n - t_i) + b_i(n)$$

- where s(n) is the source signal, $x_i(n)$ is the i-th sensor received signal, a_i is the attenuation factor and b_i is additive noise

Audio Processing

'Audio' HMMs for Classifying the Sound



Human Speech
Silence
Music
Keyboard Typing
Phone Ring
Office Noise

Feature Vector

PCA on LPC Coeff

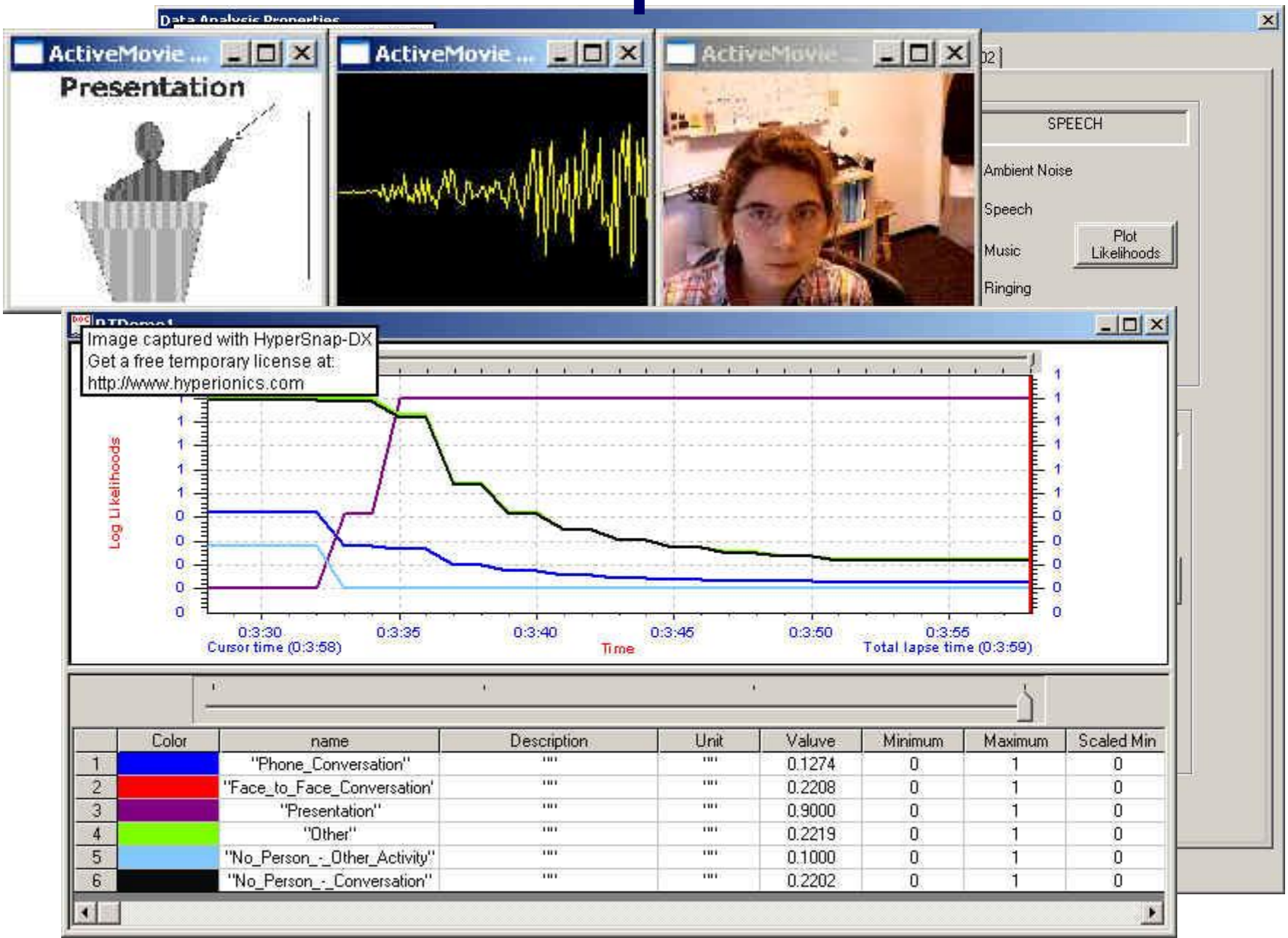
Energy, Mean & Variance of Fundamental Freq

Zero Crossing Rate (ZCR)

Seer: A Multimodal Office Awareness System

- Seer employs a two-layer HMM architecture
- Automatic, real-time recognition of six typical office activities:
  - Phone Conversation
  - Face to Face Conversation
  - Working on the Computer
  - Presentation
  - Nobody Around
  - Distant Conversation

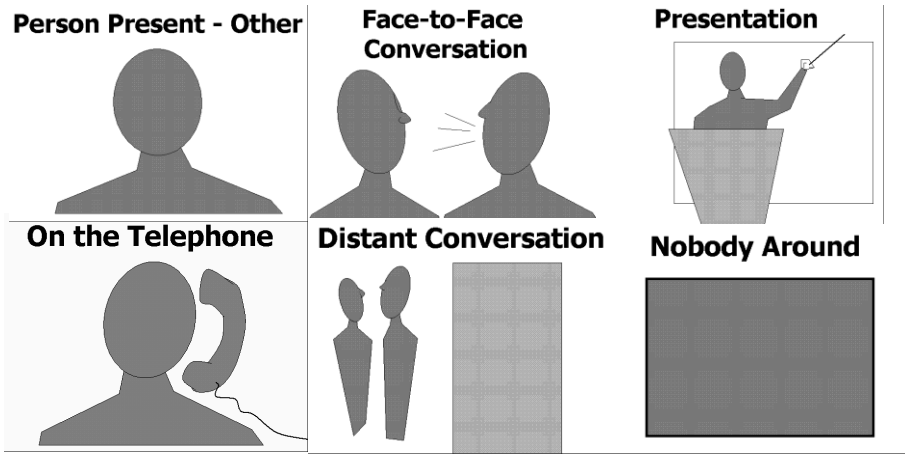
Interface Example: Presentation



IJCAI Real-time Demo

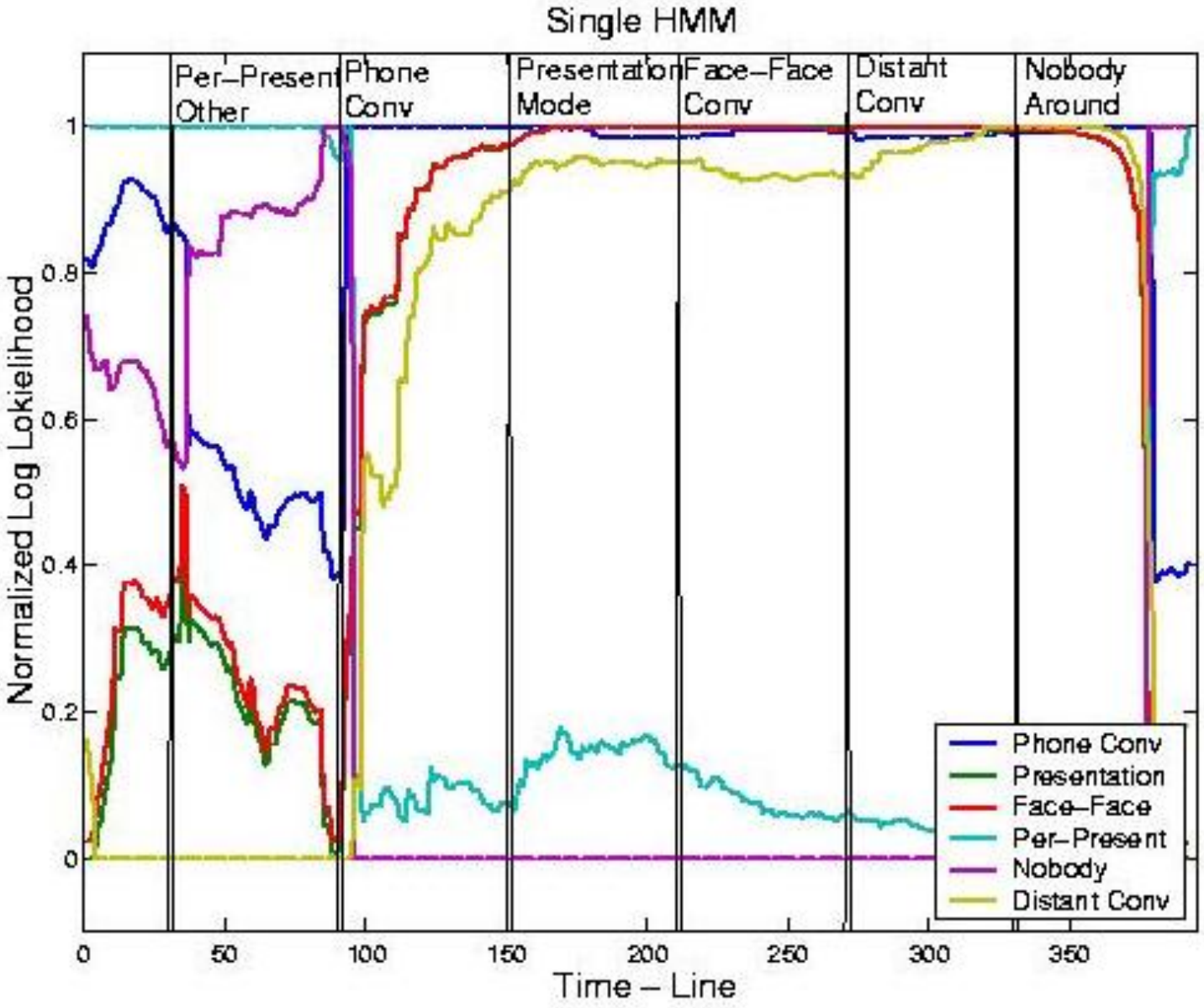
Experiments

- Comparison between Cartesian Product (CP) HMMs and LHMMs
- 60 minutes of office activity data (10min/activity;3 users)
- 50% of data for training and 50% for testing
- 6 office activities recognized:
  - Phone Conversation
  - Face to Face Conversation
  - Working on the computer
  - Distant conversation
  - Presentation
  - Nobody Around

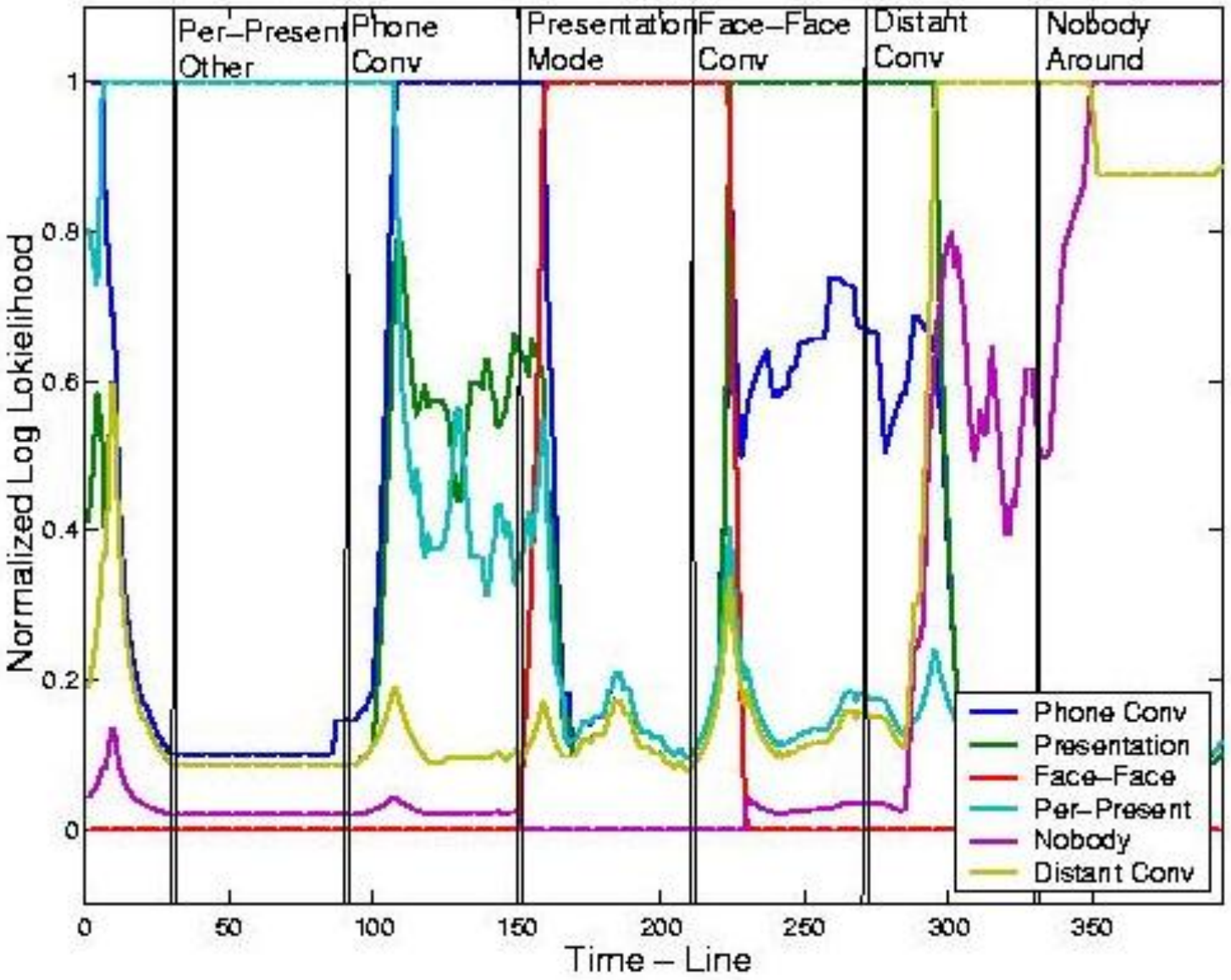


	HMMs	LHMMs
Accuracy	**72.7%**	**99.7%**
#param	**1360**	**670**

CP Hidden Markov Models



Layered Hidden Markov Models



Conclusions

- For same amount of training data, LHMMs are significantly **more accurate** than CP HMMs
  - The number of parameters of CP HMMs is higher than that of LHMMs
  - CP HMMs are more prone to overfitting and worse generalization
- LHMMs **more robust** to changes in the environment than CP HMMs
  - CP HMMs carry out high level inferences directly from the raw sensor signals, whereas LHMMs isolate the sensor signals in different sub-HMMs for each input modality and time granularity
- HHMMs have **higher discriminative** power than HMMs, i.e. distance between the log-likelihood of the two most likely models

Future Work/Discussion

- Deeper understanding of the influence of layered decompositions on the size of the parameter space, and the resulting effects on
  - Learning requirements
  - Accuracy of inference for different amount of training
- Intelligent selection of which sensors to use and when
- New representations, feature selection
- Unsupervised and semi-supervised methods for training one or more layers in LHMMs without explicit training effort
- User Interface Issues: What to do with the recognized activity information

Thank You!