Layered representations for human activity recognition

Patrick Stahlberg

Fakultät für Informatik
Universität Karlsruhe (TH)

7th July 2004

Outline of this talk

Outline of this talk

1. What is our goal?
2. Which probabilistic model to choose?
3. Implementation
4. Experiments
5. Conclusions

Outline of this talk

Outline of this talk

1. What is our goal?
2. Which probabilistic model to choose?
3. Implementation
4. Experiments
5. Conclusions

Outline of this talk

Introduction

- Project by Nuria Oliver and Eric Horvitz at Microsoft Research, Redmond.
- Part of the Attentional User Interface (AUI) project.
- Presented at the International Joint Conference on Artificial Intelligence (IJCAI), Seattle in August 2001.

Why human activity recognition?

Automated surveillance

- elderly and ill persons
- children

Improve (or introduce) “context-awareness” of computers

context = identity, location, intentions, recent activities

Could enable:

- more natural communication
- notion of interruptability

Why human activity recognition?

Automated surveillance

- elderly and ill persons
- children

Improve (or introduce) "context-awareness" of computers

context = identity, location, intentions, recent activities

Could enable:

- more natural communication
- notion of interruptability

What do we want to recognize?

- Not simple movements (like waving a hand or a pointing gesture) but more complex activities (like talking on the phone, having a face-to-face conversation).
- We want that in realtime.
- Focus on office situations.

What do we want to recognize?

- Not simple movements (like waving a hand or a pointing gesture) but more complex activities (like talking on the phone, having a face-to-face conversation).
- We want that in realtime.
- Focus on office situations.

Classify office situations into the following:

classes

- Nobody is present
- Phone conversation
- Face to face conversation
- An ongoing presentation
- A distant conversation
- A user is present and engaged in some other activity

(Proposed earlier as indicators for a person's availability.)

Classify office situations into the following:

classes

- Nobody is present
- Phone conversation
- Face to face conversation
- An ongoing presentation
- A distant conversation
- A user is present and engaged in some other activity

(Proposed earlier as indicators for a person's availability.)

Experiment Setup (Hardware)

Multimodal approach

Audio Two Microphones

- Capture ambient noise, used for sound classification and localization.

Video USB Camera

- 30fps, used to determine the number of persons present in the scene.

Traditional input devices Keyboard and Mouse

- Keep a history of events during the last 5 seconds.

Experiment Setup (Hardware)

Multimodal approach

Audio Two Microphones

- Capture ambient noise, used for sound classification and localization.

Video USB Camera

- 30fps, used to determine the number of persons present in the scene.

Traditional input devices Keyboard and Mouse

- Keep a history of events during the last 5 seconds.

Experiment Setup (Hardware)

Multimodal approach

Audio Two Microphones

- Capture ambient noise, used for sound classification and localization.

Video USB Camera

- 30fps, used to determine the number of persons present in the scene.

Traditional input devices Keyboard and Mouse

- Keep a history of events during the last 5 seconds.

Experiment Setup (Hardware)

Multimodal approach

Audio Two Microphones

- Capture ambient noise, used for sound classification and localization.

Video USB Camera

- 30fps, used to determine the number of persons present in the scene.

Traditional input devices Keyboard and Mouse

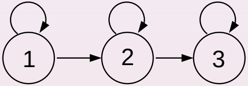
- Keep a history of events during the last 5 seconds.

Which kind of probabilistic model?

Many of the past works successfully used Hidden Markov Models (*HMMs*) or extensions. See American Sign Language, earlier in this seminar.

Other probabilistic models have been used, such as probabilistic finite-state automatons or Bayesian networks.

Hidden Markov Model



Tupel:
(S, A, V, B)

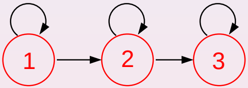
Description

- set of states
- state transition probability distribution
- set of observation symbols
- observation probability distribution for each state

Formal

$S = \{S_1, \ldots, S_N\}$, State at time t: q_t
$A = \{a_{ij}\}, \quad a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$
$V = \{v_1, \ldots, v_M\}$
$B = \{b_j(k)\}, \quad b_j(k) = P(v_k \text{ at } t | q_t = S_j)$

Hidden Markov Model



Tupel:
(S, A, V, B)

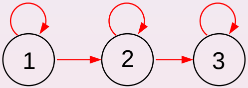
Description

- set of states
- state transition probability distribution
- set of observation symbols
- observation probability distribution for each state

Formal

$S = \{S_1, \ldots, S_N\}$, State at time t: q_t
$A = \{a_{ij}\}, \quad a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$
$V = \{v_1, \ldots, v_M\}$
$B = \{b_j(k)\}, \quad b_j(k) = P(v_k \text{ at } t | q_t = S_j)$

Hidden Markov Model



Tupel:
(S, A, V, B)

Description

- set of states
- state transition probability distribution
- set of observation symbols
- observation probability distribution for each state

Formal

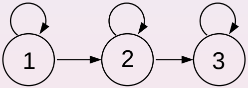
$S = \{S_1, \ldots, S_N\}$, State at time t: q_t

$A = \{a_{ij}\}, \quad a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$

$V = \{v_1, \ldots, v_M\}$

$B = \{b_j(k)\}, \quad b_j(k) = P(v_k \text{ at } t | q_t = S_j)$

Hidden Markov Model



Tupel:
(S, A, V, B)

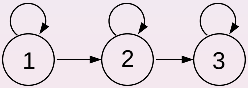
Description

- set of states
- state transition probability distribution
- set of observation symbols
- observation probability distribution for each state

Formal

$S = \{S_1, \ldots, S_N\}$, State at time t: q_t
$A = \{a_{ij}\}, \quad a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$
$V = \{v_1, \ldots, v_M\}$
$B = \{b_j(k)\}, \quad b_j(k) = P(v_k \text{ at } t | q_t = S_j)$

Hidden Markov Model



Tupel:
(S, A, V, B)

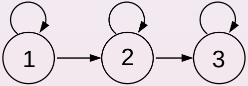
Description

- set of states
- state transition probability distribution
- set of observation symbols
- observation probability distribution for each state

Formal

$S = \{S_1, \dots, S_N\}$, State at time t: q_t
$A = \{a_{ij}\}, \quad a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$
$V = \{v_1, \dots, v_M\}$
$B = \{b_j(k)\}, \quad b_j(k) = P(v_k \text{ at } t | q_t = S_j)$

Hidden Markov Model



Tupel:
(S, A, V, B)

Description

- set of states
- state transition probability distribution
- set of observation symbols
- observation probability distribution for each state

Formal

$S = \{S_1, \ldots, S_N\}$, State at time t: q_t
$A = \{a_{ij}\}, \quad a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$
$V = \{v_1, \ldots, v_M\}$
$B = \{b_j(k)\}, \quad b_j(k) = P(v_k \text{ at } t | q_t = S_j)$

Hidden Markov Model (cont.)

Use it to:

- **Evaluate** Compute the probability that a certain observation sequence was generated by this HMM (Viterbi algorithm or Forward algorithm).
- **Decode** Compute the most probable state sequence for a given observation (Viterbi algorithm).
- **Train** Change the parameters of an HMM to better reflect real observations (Baum-Welch algorithm).

Good paper about HMMs: [Rabiner, 1989].

Hidden Markov Model (cont.)

Use it to:

Evaluate Compute the probability that a certain observation sequence was generated by this HMM (Viterbi algorithm or Forward algorithm).

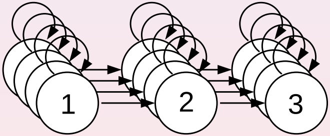
Decode Compute the most probable state sequence for a given observation (Viterbi algorithm).

Train Change the parameters of an HMM to better reflect real observations (Baum-Welch algorithm).

Good paper about HMMs: [Rabiner, 1989].

Hidden Markov Model (cont.)

Typical use of HMMs:

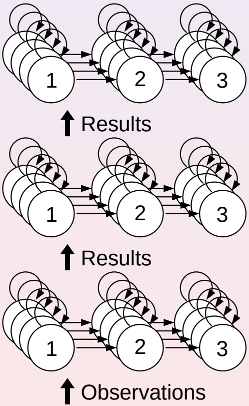


- Multiple HMMs, each is trained to accept one class.
- On an observation, each of the HMMs are evaluated in parallel.
- The HMM with the highest probability wins.

Drawbacks of HMMs

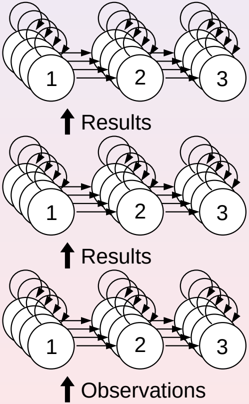
- Lack of structure.
  $\longrightarrow$ Search for a representation that is structured more like the problem (psychologists have found that many human behaviors are hierarchically structured).
- New training required when moving the system to another place.
  $\longrightarrow$ Need robustness to changes of lighting and acoustics.
- $=>$ Multilevel representation needed, for explanations at multiple temporal granularities.

Layered HMMs



- Layer architecture where each layer consists of a set of HMMs.
- Each layer is connected to the next one via its inferential results.
- Every layer operates on a different temporal granularity.
- Each layer can be trained and infered independently – lowest layer can be retrained when moving to a new office.

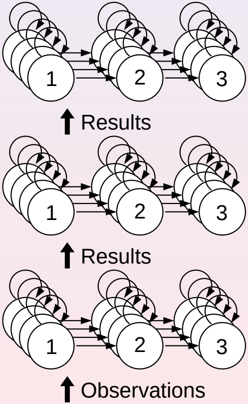
Layered HMMs



Decomposition per temporal granularity

- Layers generate one observation every n time intervals.
- Lowest level gets the features extracted from the raw sensor data, any other level gets results from previous level.
- n for each layer is determined by intuition. (Example: sensor signals: 100 miliseconds; outputs of first layer: less than one second; second layer: 5-10 seconds)

Layered HMMs



inference with LHMMs

Two approaches:

- Maxbelief: Pass the number of the HMM with the highest probability to the next layer.
- Distributional: Pass the full probability distribution of the HMMs to the next layer.

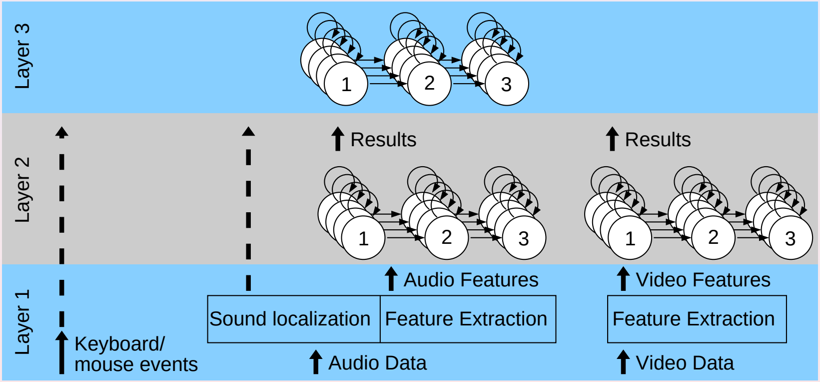
→ Maxbelief is used here, because the Distributional approach didn't improve results.

Implementation

In a system called *SEER*.

A two-layer HMM implementation. Three processing layers.

Architecture of SEER



Low layer preprocessing (Feature Extraction)

Audio

- Compute linear predictive coding coefficients, use the 7 principal coefficients.
- Locate source of sound using the Time Delay of Arrival method.
- Also: other features (like energy).

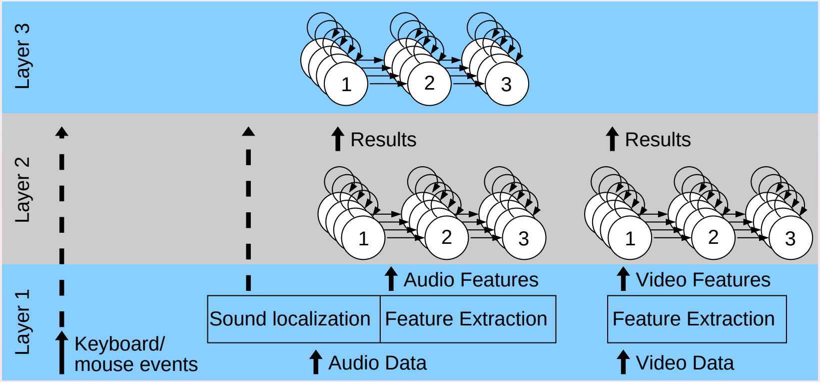
Video

- Density of skin color.
- Density of motion.
- Density of foreground pixels.
- Density of face pixels (using a realtime face detector).

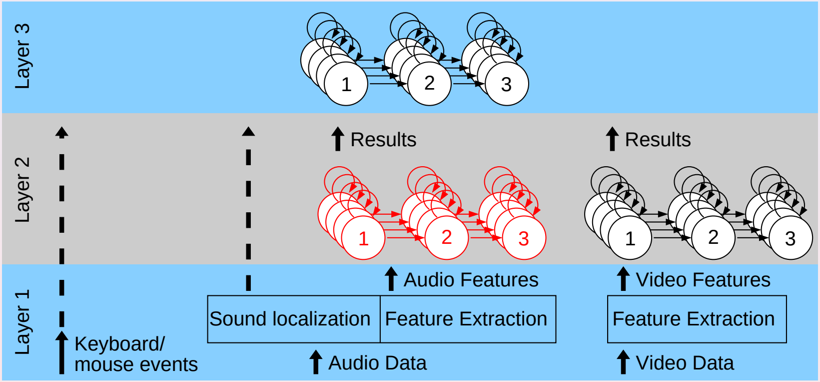
Mouse/Keyboard

- Keep last 5 seconds of mouse and keyboard events.

Architecture of SEER (cont.)

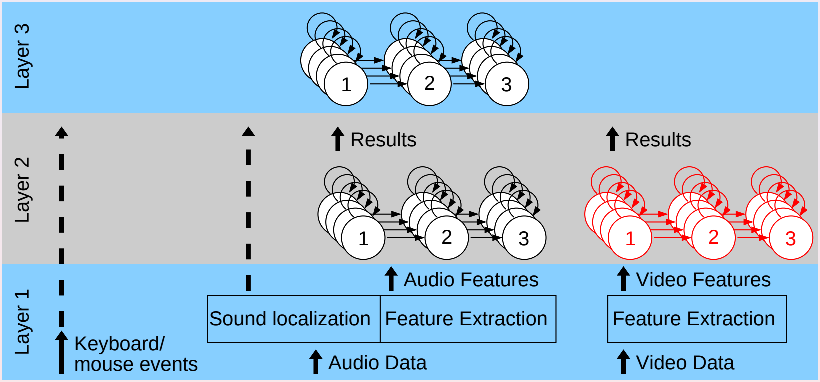


Architecture of SEER (cont.)



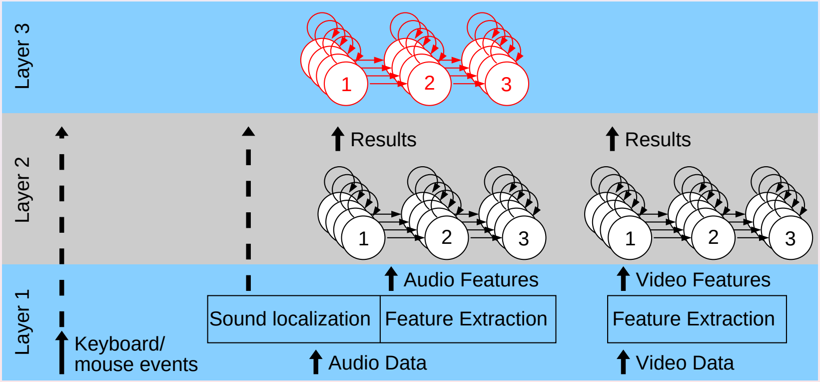
Audio HMMs: human speech; music; silence; ambient noise; phone ringing; keyboard typing.

Architecture of SEER (cont.)



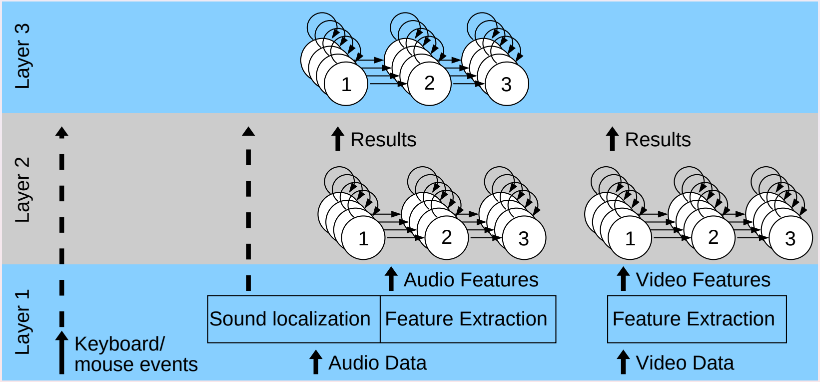
Video HMMs: nobody present; one person present (semi-static); one active person present; multiple people present

Architecture of SEER (cont.)



Top-Layer HMMs: phone conversation; face to face conversation; presentation; other activity; nobody around; distant conversation.

Architecture of SEER (cont.)



Learning SEER

Each set of HMMs is trained individually.

Experiment: comparison between LHMM and HMM

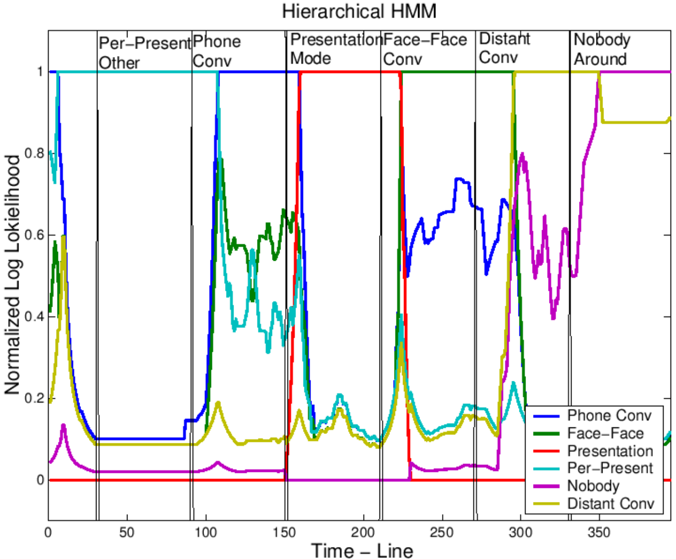
Layered Hidden Markov Models

Tested in multiple offices, with different users, for several weeks.

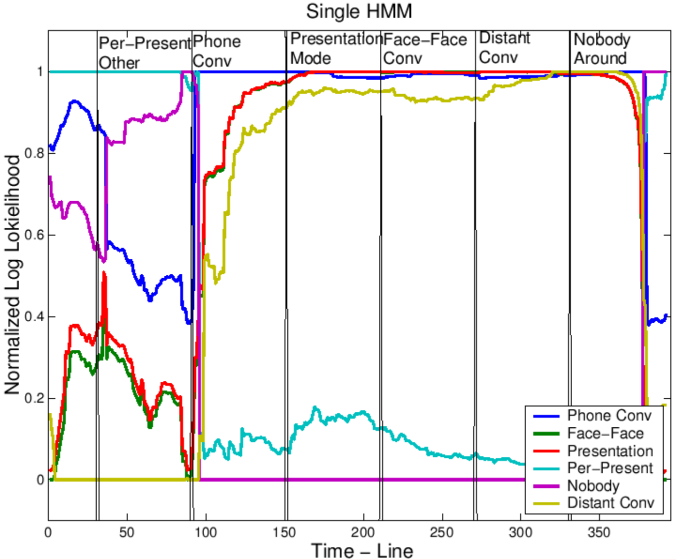
Standard Hidden Markov Models

Concatenate all the feature vector data to a new, large feature vector, which is input to a single set of discriminative HMMs.

Results: Layered HMMs



Results: Single HMMs



Experiments: comparison to standard HMMs

- High-level layers of SEER are relatively robust to changes in the environment, because inputs to each level are more stable in LHMMs.
- Encoding prior knowledge about the problem in the structure of the models decomposes the problem and reduces the dimensionality of the overall problem.
- For the same amount of training data, LHMMs have superior performance
- It's not considerably more difficult to determine the structure of LHMMs versus that of HMMs.

One additional set of experiments

- test LHMMs against HMMs on 60 minutes of recorded office activity (10 minutes per activity, 6 activities, 3 users)
- use 50 percent of data for training, 50 percent for testing.

One additional set of experiments

LHMMs:

	PC	FFC	P	O	NA	DC
PC	1.0	0.0	0.0	0.0	0.0	0.0
FFC	0.0	1.0	0.0	0.0	0.0	0.0
P	0.0	0.0	1.0	0.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0	0.0
NA	0.0	0.0	0.0	0.0	1.0	0.0
DC	0.0	0.0	0.0	0.0	0.0034	0.9966

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

LHMMs:

	PC	FFC	P	O	NA	DC
PC	1.0	0.0	0.0	0.0	0.0	0.0
FFC	0.0	1.0	0.0	0.0	0.0	0.0
P	0.0	0.0	1.0	0.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0	0.0
NA	0.0	0.0	0.0	0.0	1.0	0.0
DC	0.0	0.0	0.0	0.0	0.0034	0.9966

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

LHMMs:

	PC	FFC	P	O	NA	DC
PC	1.0	0.0	0.0	0.0	0.0	0.0
FFC	0.0	1.0	0.0	0.0	0.0	0.0
P	0.0	0.0	1.0	0.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0	0.0
NA	0.0	0.0	0.0	0.0	1.0	0.0
DC	0.0	0.0	0.0	0.0	0.0034	0.9966

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

LHMMs:

	PC	FFC	P	O	NA	DC
PC	1.0	0.0	0.0	0.0	0.0	0.0
FFC	0.0	1.0	0.0	0.0	0.0	0.0
P	0.0	0.0	1.0	0.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0	0.0
NA	0.0	0.0	0.0	0.0	1.0	0.0
DC	0.0	0.0	0.0	0.0	0.0034	0.9966

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

LHMMs:

	PC	FFC	P	O	NA	DC
PC	1.0	0.0	0.0	0.0	0.0	0.0
FFC	0.0	1.0	0.0	0.0	0.0	0.0
P	0.0	0.0	1.0	0.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0	0.0
NA	0.0	0.0	0.0	0.0	1.0	0.0
DC	0.0	0.0	0.0	0.0	0.0034	0.9966

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

LHMMs:

	PC	FFC	P	O	NA	DC
PC	1.0	0.0	0.0	0.0	0.0	0.0
FFC	0.0	1.0	0.0	0.0	0.0	0.0
P	0.0	0.0	1.0	0.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0	0.0
NA	0.0	0.0	0.0	0.0	1.0	0.0
DC	0.0	0.0	0.0	0.0	0.0034	0.9966

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

LHMMs:

	PC	FFC	P	O	NA	DC
PC	1.0	0.0	0.0	0.0	0.0	0.0
FFC	0.0	1.0	0.0	0.0	0.0	0.0
P	0.0	0.0	1.0	0.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0	0.0
NA	0.0	0.0	0.0	0.0	1.0	0.0
DC	0.0	0.0	0.0	0.0	0.0034	0.9966

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

LHMMs:

	PC	FFC	P	O	NA	DC
PC	1.0	0.0	0.0	0.0	0.0	0.0
FFC	0.0	1.0	0.0	0.0	0.0	0.0
P	0.0	0.0	1.0	0.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0	0.0
NA	0.0	0.0	0.0	0.0	1.0	0.0
DC	0.0	0.0	0.0	0.0	0.0034	0.9966

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

HMMs:

	PC	FFC	P	O	NA	DC
PC	0.8145	0.0679	0.0676	0.0	0.0	0.05
FFC	0.0014	0.9986	0.0	0.0	0.0	0.0
P	0.0	0.0052	0.9948	0.0	0.0	0.0
O	0.0345	0.0041	0.003	0.9610	0.0	0.0
NA	0.0341	0.0038	0.0010	0.2524	0.7086	0.0
DC	0.0076	0.0059	0.0065	0.0	0.0	0.98

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

HMMs:

	PC	FFC	P	O	NA	DC
PC	0.8145	0.0679	0.0676	0.0	0.0	0.05
FFC	0.0014	0.9986	0.0	0.0	0.0	0.0
P	0.0	0.0052	0.9948	0.0	0.0	0.0
O	0.0345	0.0041	0.003	0.9610	0.0	0.0
NA	0.0341	0.0038	0.0010	0.2524	0.7086	0.0
DC	0.0076	0.0059	0.0065	0.0	0.0	0.98

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

HMMs:

	PC	FFC	P	O	NA	DC
PC	0.8145	0.0679	0.0676	0.0	0.0	0.05
FFC	0.0014	0.9986	0.0	0.0	0.0	0.0
P	0.0	0.0052	0.9948	0.0	0.0	0.0
O	0.0345	0.0041	0.003	0.9610	0.0	0.0
NA	0.0341	0.0038	0.0010	0.2524	0.7086	0.0
DC	0.0076	0.0059	0.0065	0.0	0.0	0.98

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

HMMs:

	PC	FFC	P	O	NA	DC
PC	0.8145	0.0679	0.0676	0.0	0.0	0.05
FFC	0.0014	0.9986	0.0	0.0	0.0	0.0
P	0.0	0.0052	0.9948	0.0	0.0	0.0
O	0.0345	0.0041	0.003	0.9610	0.0	0.0
NA	0.0341	0.0038	0.0010	0.2524	0.7086	0.0
DC	0.0076	0.0059	0.0065	0.0	0.0	0.98

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

HMMs:

	PC	FFC	P	O	NA	DC
PC	0.8145	0.0679	0.0676	0.0	0.0	0.05
FFC	0.0014	0.9986	0.0	0.0	0.0	0.0
P	0.0	0.0052	0.9948	0.0	0.0	0.0
O	0.0345	0.0041	0.003	0.9610	0.0	0.0
NA	0.0341	0.0038	0.0010	0.2524	0.7086	0.0
DC	0.0076	0.0059	0.0065	0.0	0.0	0.98

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

HMMs:

	PC	FFC	P	O	NA	DC
PC	0.8145	0.0679	0.0676	0.0	0.0	0.05
FFC	0.0014	0.9986	0.0	0.0	0.0	0.0
P	0.0	0.0052	0.9948	0.0	0.0	0.0
O	0.0345	0.0041	0.003	0.9610	0.0	0.0
NA	0.0341	0.0038	0.0010	0.2524	0.7086	0.0
DC	0.0076	0.0059	0.0065	0.0	0.0	0.98

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

HMMs:

	PC	FFC	P	O	NA	DC
PC	0.8145	0.0679	0.0676	0.0	0.0	0.05
FFC	0.0014	0.9986	0.0	0.0	0.0	0.0
P	0.0	0.0052	0.9948	0.0	0.0	0.0
O	0.0345	0.0041	0.003	0.9610	0.0	0.0
NA	0.0341	0.0038	0.0010	0.2524	0.7086	0.0
DC	0.0076	0.0059	0.0065	0.0	0.0	0.98

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

One additional set of experiments

HMMs:

	PC	FFC	P	O	NA	DC
PC	0.8145	0.0679	0.0676	0.0	0.0	0.05
FFC	0.0014	0.9986	0.0	0.0	0.0	0.0
P	0.0	0.0052	0.9948	0.0	0.0	0.0
O	0.0345	0.0041	0.003	0.9610	0.0	0.0
NA	0.0341	0.0038	0.0010	0.2524	0.7086	0.0
DC	0.0076	0.0059	0.0065	0.0	0.0	0.98

legend

PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

Conclusions

1. For the same amount of training data, the accuracy of LHMMs is significantly higher than that of HMMs.
2. LHMMs are more robust to changes in the environment than HMMs.
3. The discriminative power of LHMMs is notably higher than that of HMMs.

Summary

We have...

- Presented a real-time, multimodal approach for human activity recognition in office environments.
- Analysed Layered HMMs and compared them to HMMs.

We found...

- LHMMs work better because they better reflect the hierarchical structure of the problem.
- LHMMs need less training data.
- LHMMs are more robust to changes in the environment.
- LHMMs are not significantly more difficult to design.

References / For Further Reading

N. Oliver and E. Horvitz
Layered Representations for Human Activity Recognition
IJCAI, Seattle. 2001.
http://research.microsoft.com/~nuria/papers/icmi2002.pdf, http://research.microsoft.com/~horvitz/seer.HTM

L. R. Rabiner
A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
Proceedings of the IEEE, Vol. 77, No. 2, pages 257–286. 1989

E. Horvitz, C.M. Kadie, T. Paek and D. Hovel
Models of Attention in Computing and Communication: From Principles to Applications
Communications of the ACM 46(3):52-59. 2003.
http://research.microsoft.com/~horvitz/cacm-attention.htm