Hierarchical Representations for Learning and Inferring Office Activity from Multiple Streams of Information

Nuria Oliver Eric Horvitz

Adaptive Systems & Interaction
Microsoft Research
Redmond, WA 98052

Ashutosh Garg

Computer Science Dept.
Univ. Urbana-Champaign
Champaign, IL

Abstract

We present the use of hierarchical probabilistic representations for modeling the activities of people, and describe how we use the representation to do sensing, learning, and inference at multiple levels of temporal granularity and abstraction. The approach centers on the use of a Hierarchy of Hidden Markov Models (HHMMs), using parameters that are learned from data. HHMMs provide a promising means for modeling diverse human activities. We illustrate the application of HHMMs within an office-awareness situation. We describe the ability to correctly classify in real-time typical office activities such as talking on the phone, being in a meeting with someone else, giving a presentation, or just performing work within an office setting. We believe that systems like ours constitute an important step towards richer human-computer interaction and computer-mediated communication.

1 Introduction

Researchers and application developers have long been interested in the promise of performing automatic and semi-automatic recognition of human behavior from observations. Successful recognition of human behavior is critical in a number of compelling applications, including automated visual surveillance and multimodal human–computer interaction (HCI)—user interfaces that consider multiple streams of information about a user's behavior and the context of a situation. We shall focus in this paper on human activity modeling and recognition employed in a real-time multimodal system for identifying activities in and around an office.

We tackle the challenge of performing inferences that take as inputs raw signals coming from multiple sensors and that yield high-level abstract descriptions of the human activities. The task of moving from low-level signals to abstract hypotheses about activity brings into focus a consideration of a broad spectrum of approaches. The variety of potentially valuable methods include template matching, context-free grammars, and various statistical approaches. We introduce a hierarchical statistical method for detecting and recognizing human actitivities in a multimodal system. Our method is based on a formulation of dynamic graphical models which we refer to as a Hierarchy of Hidden Markov Models (HHMMs).

We shall review relevant prior work in Section 2. This discussion will be useful for understanding background research and the nature of our contribution. In Section 3, we describe the system's architecture and the perceptual inputs. Then, in Section 4, we review the hierarchical behavior of the models. Experimental results are described in section 5. Finally, section 6 summarizes our work and highlights several conclusions.

2 Previous Work

Having access to a user's context is valuable in developing systems that could provide more intuitive, natural user interaction experiences. Important aspects of a user's context include the user's current and past activities and intentions. Recent work on probabilistic models for reasoning about a user's context and intentions have highlighted opportunities for building new kinds of applications and services [13, 14].

Most of the previous work on using perceptual information to recognize human activity has been for identifying a specific type of activity in a particular scenario. Many of these techniques are targeted at recognizing single, simple events, e.g. 'waving the hand' or 'sitting on a chair'. Much less has been done on methods for identifying more complex patterns of human activities, extending over longer periods of time.

Dynamic models of periodic patterns of people's movements are used in [8] to model the periodicity of activities such as walking. Other approaches to the recognition of human activity employ graphical models. A significant portion of work in this arena has harnessed Hidden Markov Models (HMMs). In [19], an HMM is used for recognizing hand movements used to relay symbols in American Sign Language. More complex models, such as Parametrized-HMM (PHMM) [20], Entropic-HMM [3], Variable-length HMM (VHMM)

[11] and Coupled-HMM (CHMM) [4, 17], have been used to recognize more complex activities such as the interaction between two people. In [2], a stochastic context-free grammar is used to compute the probability of a temporally consistent sequence of primitive actions recognized by HMMs. Clarkson and Pentland model events and scenes from audiovisual information in [7]. They have developed a wearable computer system that uses a hierarchy of HMMs for recognizing the user's location, *e.g.*, in the office, at the bank, etc. In [3], an entropic-HMM approach is used to organize the observed video activities (office activity and outdoor traffic) into meaningful states. Finally, in [12], a probabilistic finite-state automaton (a variation of structured HMMs) is used for recognizing different scenarios, such as monitoring pedestrians or cars on a freeway. Although HMMs appear to be robust to changes in the temporal segmentation of observations, they suffer from a lack of structure, an excess of parameters, and an associated over-fitting of data when they are applied to reason about long and complex temporal sequences with inadequate training data. Finally, in recent years, more complex Bayesian networks have also been adopted for the modeling and recognition of human activities [1, 6, 14, 15].

To date, however, there has been little research on multimodal systems for human-computer interaction that use statistical methods to model typical human activities in a hierarchial manner. The methods and working system described in this paper focus on this representation. We show how the methods can learn on the fly the most common actions that users perform in office settings.

3 System Architecture

We have built a system named Seer, that harnesses the HHMM representation to detect activities in an office. As it is shown in Figure 1, Seer is composed of three layers:

(1) At the lowest level, the sensor signals (described in section 3.1), *i.e.*, images, audio and keyboard and mouse activity, are captured and processed, resulting in a feature vector for each modality.

(2) In the middle layer, the sound is classified using discriminative Hidden Markov Models. Typical office sounds are trained in real-time, such as human speech, music, silence, ambient noise, phone ringing and keyboard typing. The source of the sound is also localized using a technique based on the Time Delay of Arrival (TDOA) [5]. The video signals are classified using discriminative HMMs to implement a person detector. At this level, the system detects whether one person is present in the room (semi-static), one active person is present, or multiple people are present. The inferential results from this layer (audio and video classifiers), the derivative of the sound localization component, and the history of keyboard and mouse activities, are passed to the next higher layer in the hierarchy.

(3) The third layer handles concepts that have longer temporal extent. Such concepts include the user's high-level activities in or near an office. The behavior models are HMMs whose inputs are the inferential outputs of the previous level. The office activities that are recognized by Seer include PHONE CONVERSATION, FACE TO FACE CONVERSATION, PRESENTATION, DISTANT CONVERSATION, NOBODY IN THE OFFICE and USER PRESENT, ENGAGED IN SOME OTHER ACTIVITY.

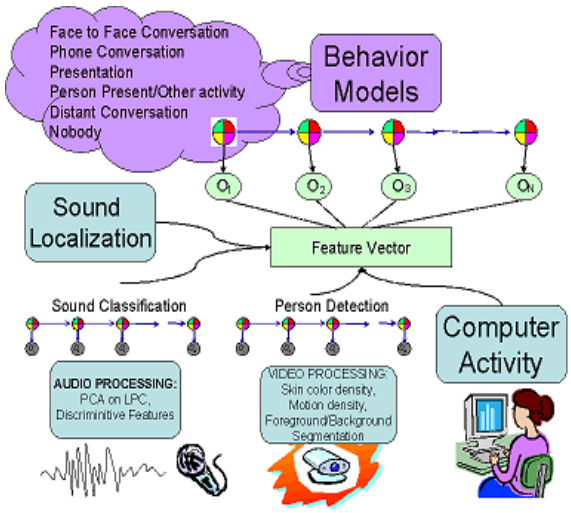


Figure 1: Architecture of the multimodal Seer system.

3.1 Perceptual Input

Seer is multimodal, accessing perceptual information from the following sources:

(1) **Binaural microphones:** Two mini-microphones (20 – 16000 Hz, SNR 58 dB) are being used. The audio signal is sampled at 44100 KHz. The microphones are used for sound classification and localization;

(2) **USB camera:** The video signal is obtained via a standard USB camera (Intel), sampled at 30 f.p.s. The video input is used to determine the number of persons present in the scene;

(3) **Keyboard and mouse:** The system keeps a history of keyboard and mouse activities.

3.2 Feature Extraction and Selection

The raw sensor signals are preprocessed to obtain feature vectors (*i.e.* observations) for the lowest-level HMMs. On the audio side, Linear Predictive Coding (LPC) coefficients are computed. Feature selection is applied on these coefficients by means of principal component analysis (PCA). The number of features is selected such that more than 95% of the variability in the data is maintained, which is typically achieved with

no more than 7 features. Other higher-level features are also extracted from the audio signal, such as the zero crossing rate (ZCR), the energy and the mean and variance of the fundamental frequency over a time window. On the video side, three features are computed: the density of skin color in the image (obtained via a discriminative histogram between skin and non-skin in HSV space), the density of motion in the image (obtained by image differences), and the density of foreground pixels in the image (obtained by background subtraction, after having learned the background). Finally, a history of the last 5, 60 and 600 seconds of mouse and keyboard activities is logged.

The source of the sound is localized using the Time Delay of Arrival (TDOA) [5] method. In TDOA, the measure in question is not the acoustic data received by the sensors, but rather the time delays between the signals coming from each sensor. Typically, TDOA-based approaches have two steps: the time delay estimation (TDE) and the sound localization (SSL). Let $s(n)$ be the source signal and be $x_i(n)$ the i-th sensor received signal. If we assume no reverberation, we have $x_i(n) = a_i s(n - t_i) + b_i(n)$. To model reverberation, we would add the non-linear reverberation function: $x_i(n) = g_i * s(n - t_i) + b_i(n)$, where a_i is the attenuation factor, b_i is additive noise and g_i is the response between the source and the sensor. Seer includes multiple approaches for estimating the time delay of arrival between the left and right audio signals. We have obtained the best performance by estimating the peak of the time cross-correlation function between the left and right audio signals over a finite time window $[N_1, N_2]$, i.e.: $r_{lr}(d) = \sum_{n=N_1}^{n=N_2} l(n)r(n - d)$.

4 Activity Modeling using HHMMs

We now turn to the learning of the representation and overall architecture that allows for temporal abstraction from pointwise observations at particular times into explanations over varying temporal intervals. We have explored hierarchical representations for reasons of robustness, naturalness, and training efficacy.

In building Seer, we faced the challenge of processing continuous flows of data coming from multiple sensors in a reliable manner. One of our design goals was developing methods that are robust to typical variations of lighting and acoustics within office environments. Beyond robustness in a single office, we desired a representation that would allow the models to perform well when transferred to new office spaces with minimal tuning through retraining. We also sought a representation that would map naturally onto the problem space. Human behaviors appear to be hierarchically structured in many cases [16]. We pursued a representation that could capture such hierarchical properties in an elegant manner.

We converged on the use of a multilevel representation of observations that allows for explanations at different granularities of time, or levels of detail. For example, in the domain of office awareness, one level of description is the analysis and classification of the raw sensor signals. This level corresponds to a fine time granularity on the order of milliseconds. Another level of description is the detection of the user's presence, by use of audio, keyboard, mouse, and video. In this case, the time granularity is of several seconds. At another level, one could describe what the user has done in the last N minutes. Finally, one could provide an explanation of the user's activities during the day.

Our hierarchical approach to learning human activities from data employs directed acyclic graphs (DAGs), also referred to as dynamic Bayesian networks. Statistical DAGs [9] can provide a computationally efficient and sufficiently expressive solution to the problem of human activity modeling and recognition. HMMs and their extensions, e.g. CHMMS, PHMMs, VHMMs, or the architecture proposed in this paper, HHMMs, can be viewed as a particular case of temporal DAGs.

DAGs present several important advantages that are relevant to the problem of human behavior modeling from multiple sensors. In addition, employing a hierarchical structure provides several valuable properties. A hierarchical formulation makes it feasible to decouple different levels of analysis for training and inference. As it is further explained in Section 4.1.1, each level of our hierarchy is trained independently, with different feature vectors and time granularities. Once the system has been trained, inference can be carried out at any level of the hierarchy. One could retrain the lowest (most sensitive to variations in the environment) level, for example, without having to retrain any other level in the hierarchy.

4.1 Hierarchical Hidden Markov Models

The structure of HHMMs for human activity recognition is displayed in Figure 2.

4.1.1 Learning

A formulation for hierarchical HMMs was first proposed in [10] in work that extended the standard Baum-Welch procedure and presented an efficient estimation procedure of model parameters from unlabeled data. A trained model was applied to an automatic hierarchical parsing of an observation sequence as a dendrogram. Because of the computational complexity of the original algorithm, the authors suggest an efficient approximation to the full estimation scheme. The approximation can further be used to construct models that adapt both their topology and parameters. The authors briefly illustrate the performance of their models in natural written English text and in English handwriting recognition.

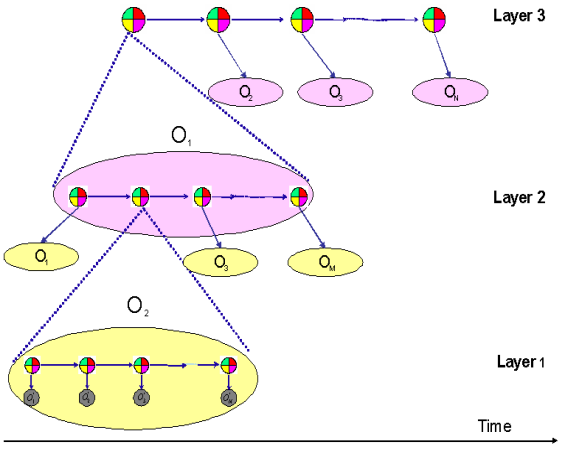


Figure 2: Hierarchical HMMs structure for human behavior recognition at different levels of temporal granularity

HHMMs differ in fundamental ways from the hierarchical approach explored in [10]. In our case, each layer of the architecture is connected to the next layer via the inferential results (log likelihoods). In contrast, in [10], each state of the architecture is another HHMM, and therefore a time sequence of the raw signals. Moreover, rather than training all the levels at the same time, the parameters for each level of our HHMMs can be trained independently by means of the Baum-Welch algorithm [18]. The inputs of each level are the outputs of the previous level. At the lowest level, the observations (the leaves of the tree) are the feature vectors extracted directly from sensor signals.

4.1.2 Inference

Inference can be performed at any level of the hierarchy by means of the Viterbi algorithm. Each level encodes a different temporal abstraction, going from the finest time granularity, at the leaves, to the highest time granularity (highest abstraction level), at the root of the tree.

Focusing more specifically on our target application of office awareness, we employed HHMMs composed of a three-layer structure. At the bottom, the raw sensor signals are processed with time windows of duration less than 100 miliseconds. Next, the audio and video information is classified with a time granularity of less than 3 seconds. Finally, typical office activities are represented at the highest level (root), corresponding to a time granularity of about $10-15$ seconds. The activities modeled in this setting are: (1) PHONE CONVERSATION; (2) PRESENTATION; (3) FACE-TO-FACE CONVERSATION; (4) USER PRESENT, ENGAGED IN SOME OTHER ACTIVITY; (5) DISTANT CONVERSATION (outside the field of view); (6) NOBODY PRESENT.

5 Experiments

As we mentioned earlier, a basic motivation for using a hierarchical representation is the intrinsic hierarchical structure of many human activities [16] and the importance of reasoning and providing explanations at different levels of abstraction. From a more practical viewpoint, there are other several important advantages of our hierarchical framework. First is the robustness of the real-time inference capabilities in light of environmental and subject variation. An additional goal is to minimize the amount of training required to adapt Seer to a new environment. Thus, we are motivated to decompose the model into distinct layers in accordance with the sensitivity of layers of analysis to environmental data. We found that the temporally fine-grained signal-detection level of analysis typically required retraining in new environments. However, temporally coarser, higher level layers of the model could be reused. Note that with our HHMMs once the system has been trained, inference can be carried out at any level of the hierarchy. Moreover, one can retrain the lowest (most sensitive to variations in the environment) levels without having to retrain any other level in the hierarchy.

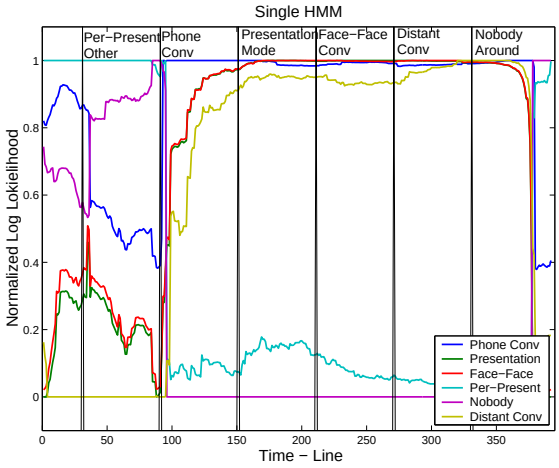
We have been running Seer in multiple offices, with different users and respective environments for several weeks. In our preliminary tests, we have found that the high-level layers of Seer are relatively robust to changes in the environment. In all the cases, when we moved Seer from one office to another, we obtained nearly perfect performance *without* the need for retraining the higher levels of the hierarchy. Only some of the lowest-level features (such as the ambient noise, the background image, and the skin color threshold) required some re-training to tune the lowest level to the new conditions. In summary, the hierarchical structure greatly contributes to the overall robustness of the system given changes in the environment.

In a more quantitative study, we compared the performance of our model and single, standard HMMs. The feature vector in the latter case results from the concatenation of the audio, video and keyboard/mouse activities features in one long feature vector. We refer to these HMMs as the Cartesian product HMMs. Note that 5-state HMM with single Gaussian observations of dimensionality 16 would have $5 * 16 * (16 + 1) = 1360$ parameters to estimate. An equivalent HHMM with 2 levels, two 5-state HMMs at the lowest level (audio and video, with dimensionalities 10 and 3 respectively) and one 5-state HMM at the highest level (of dimensionality 3), would have $5*10*(10+1)+5*3*(3+1)+5*3*(3+1) = 670$ parameters. Note that encoding prior knowledge about the problem in the structure of the models significantly reduces the dimensionality of the problem. Therefore, for the same amount of training data, it is expected for HHMMs to have superior performance than

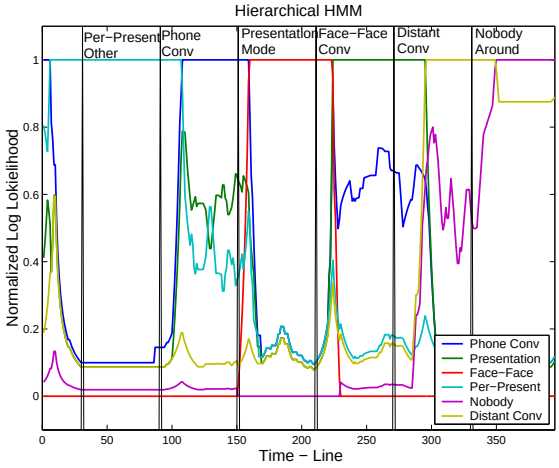
HMMs. Our experimental results corroborate such expectation.

Figure 3 illustrates the per-frame normalized likelihoods on testing in *real-time* both HMMs and HHMMs with the different office activities. By 'normalized' likelihoods, we denote the likelihoods whose values have been bounded between 0 and 1. They are given by: $NormLike_i = \frac{Like_i - min_j(Like_j)}{max_j(Like_j) - min_j(Like_j)}$, for $i = 1, ..., N$, $j = 1, ...N$, and N models. Note that we only plot the likelihoods for the last half of the testing data to avoid instabilities in the transitions.

The accuracies of both HMMs and HHMMs when tested on 8 real-time sequences of each class were of 72.68% (STD 8.15) and 99.5% (STD 0.95) respectively.



(a) Single HMMs



(b) HHMMs

Figure 3: Log Likelihoods for each of the activity models over time when tested in real time

Finally, we compared the performance on 30 minutes of office activity data (5 minutes per activity and 6 activities) of HHMMs and HMMs. The results are summarized in table 1. Note that the HMMs were specifically tuned to the particular testing data. Their performance was otherwise so poor that we could not make any meaningful comparison with the equivalent HHMMs. On the other hand, the HHMMs had been trained many days before, under different office conditions than that of testing.

Confusion Matrix for highly-tuned 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
Confusion Matrix for generic HHMMs						
	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

Table 1: Confusion matrix for highly-tuned HMMs and generic HHMMs on 30 min of real data, where PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NA=Nobody Around; DC=Distant Conversation.

Some important observations are:

(1) *For the same amount of data, the accuracy of HHMMs is significantly higher than that of HMMs.* There are several reasons for the better performance of HHMMs when compared to HMMs: (1) The number of parameters of HMMs is approximately double that of HHMMs for the office activities being modeled in our experiments. As a consequence, for the same amount of training data, HMMs have many more parameters to estimate than HHMMs. Therefore HMMs are more prone to overfitting and worse generalization than HHMMs, especially when trained with limited amounts of data. (2) HMMs carry out high-level inferences about the user's activity, directly from the raw sensor signals. HHMMs, on the other hand, isolate the sensor signals in different sub-HMM models for each input modality. The inferential results of these models feed the next layer HMMs that characterize the office activities. Due to its hierarchical structure, HHMMs are more robust to noise in the sensor signals and have therefore better generalization performance than HMMs.

(2) *HHMMs are more robust to changes in the environment than HMMs.* We could not obtain any reasonable performance on HMMs had they not been *highly tuned* to the particular testing environment and conditions. We had to retrain the HMMs every time we needed to test them on some particular data. On the contrary, our HHMMs did *not* require retraining, de-

spite the changes in office conditions. HMMs were more sensitive to changes in the environment than HHMMs. HMMs required significant retraining when changing the conditions.

(3) *The discriminative power of HHMMs is notably higher than that of HMMs.* By *discriminative power*, we mean the distance between the log-likelihood of the two most likely models. The log likelihoods for the HMMs tend to be much closer to each other, making them prone to instability and errors in the classification. Note in Figure 3 how the normalized likelihoods between the two best models in HMMs are much closer than that in HHMMs. This phenomenon is particularly noticeable in the PRESENTATION, FACE TO FACE CONVERSATION, DISTANT CONVERSATION and NOBODY AROUND activities.

6 Summary and Conclusions

We have described a real time, multimodal framework for human activity recognition in an office environment. We have introduced a new hierarchical HMM approach (HHMM) that has the ability to capture different levels of abstraction and corresponding time granularities. The approach appears to be well matched to the decomposition of signals and hypotheses for discriminating a set of activities in an office setting. Our models are learned from data. Some important characteristics of HHMMs when compared to HMMs are: (1) HHMMs encode the hierarchical temporal structure of the discrimination problem; thus, the dimensionality of the state space that needs to be learned from data is much smaller than that of their corresponding Cartesian Product HMMs; (2) HHMMs are easier to interpret, and, thus, to refine and improve, than the corresponding Cartesian Product HMMs, thanks to their hierarchical structure; (3) HHMMs can encode different levels of abstraction and time granularities that can be linked to different levels of representation for human behaviors; (4) the modularity of HHMMs allows the selective retraining of the levels that are most sensitive to environmental or sensor variation, minimizing the burden of training during transfer among different environments.

We have demonstrated the performance of the HHMM representation in Seer, a real-time system for recognizing typical office activities. Seer can accurately recognize when a user is engaged in a phone conversation, giving a presentation, involved in a face-to-face conversation, or doing some other work in the office—or when a distant conversation is occurring in the corridor. We believe that our framework can be harnessed to enhance multimodal solutions on the path to more natural human-computer interaction.

References

[1] John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213, 1997.

[2] A. Bobick and Y. Ivanov. Action recognition using probabilistic parsing, 1998.

[3] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *PAMI, August 2000.*

[4] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR'97.*

[5] M. Brandstein and H. Silverman. A practical methodology for speech source localization with microphone arrays, 1997.

[6] Hilary Buxton and Shaogang Gong. Advanced Visual Surveillance using Bayesian Networks. In *International Conference on Computer Vision*, Cambridge, Massachusetts, June 1995.

[7] B. Clarkson and A. Pentland. Unsupervised clustering of ambulatory audio and video. In *In ICASSP'99. 1999.*

[8] L. Davis, R. Chellapa, Y. Yacoob, and Q. Zheng. Visual surveillance and monitoring of human and vehicular activity, 1997.

[9] T. Dean and K. Kanazawa. Probabilistic temporal reasoning. In *AAAI'88, pp. 524–529.*

[10] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. In *Machine Learning 32(1)*, 1998.

[11] A. Galata, N. Johnson, and D. Hogg. Learning variable length markov models of behaviour. 81(3), IJCV-2001, pp. 398–413.

[12] S. Hongeng, F. Bremond, and R. Nevatia. Representation and optimal recognition of human activities. CVPR'00.

[13] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proc. of CHI'99, pp. 159–166.*

[14] E. Horvitz, A. Jacobs, and D. Hovel. Attention-sensitive alerting. In *Proc. of UAI'99, pp. 305–313.*

[15] S. S. Intille and A. F. Bobick. A framework for recognizing multi-agent action from visual evidence. In *AAAI'99, 518-525.*

[16] Zacks J. and Tversky B. Event structure in perception and cognition. *Psychological Bulletin*, 127(1), 3-21, 2001.

[17] N. Oliver. *Towards Perceptual Intelligence: Statistical Modeling of Human Individual and Interactive Behaviors.* PhD thesis, MIT, 2000.

[18] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE*, 77(2):257–286, February 1989.

[19] T. Starner and A. Pentland. Real-time ameerican sign language recognition from video using hidden markov models, ISCV'95.

[20] A. Wilson and A. Bobick. Recognition and interpretation of parametric gesture, 1998.