# S-SEER: Selective Perception in a Multimodal Office Activity Recognition System

Nuria Oliver & Eric Horvitz

Adaptive Systems & Interaction
Microsoft Research
Redmond, WA USA
{nuria,horvitz}@microsoft.com

**Abstract.** The computation required for sensing and processing perceptual information can impose significant burdens on personal computer systems. We explore several policies for selective perception in SEER, a multimodal system for recognizing office activity that relies on a cascade of Hidden Markov Models (HMMs) named Layered Hidden Markov Model (LHMMs). We use LHMMs to diagnose states of a user's activity based on real-time streams of evidence from video, audio and computer (keyboard and mouse) interactions. We review our efforts to employ expected-value-of-information (EVI) to limit sensing and analysis in a context-sensitive manner. We discuss an implementation of a greedy EVI analysis and compare the results of using this analysis with a heuristic sensing policy that makes observations at different frequencies. Both policies are then compared to a random perception policy, where sensors are selected at random. Finally, we discuss the sensitivity of ideal perceptual actions to preferences encoded in utility models about information value and the cost of sensing.

## 1 Introduction

Investigators have long pursued the dream of building systems with the ability to perform automatic recognition of human behavior and intentions from observations. Successful recognition of human behavior enables compelling services and applications, including the provision of appropriate help and assistance, automated visual surveillance and multimodal user interfaces —user interfaces that allow human-computer interaction via perceptual channels such as acoustical and visual analyses. Such systems can employ representations of a user's *context* and reason about the most appropriate control and services in different settings. There has been progress on multiple fronts in recognizing human behavior and intentions. However, challenges remain for developing machinery that can provide rich, human-centric notions of context in a tractable manner.

We address in this paper the computational burden associated with perceptual analysis. Computation for visual and acoustical analyses has typically required a large portion, if not nearly all, of the total computational resources of personal computers that make use of such perceptual inferences. Thus, we have pursued principled strategies for limiting in an automated manner the computational load of perceptual systems.

For a testbed, we have considered the allocation of perceptual resources in SEER, a multimodal, probabilistic reasoning system that provides real-time interpretations of human activity in and around an office [1]. We have explored different strategies for sensor selection and sensor data processing in SEER. The result is a new system named S-SEER, or *Selective* SEER.

This paper is organized as follows: We first provide background on multimodal systems and principles for guiding perception in these systems in Section 2. In Section 3 we describe the challenge of understanding human activity in an office setting and review the perceptual inputs that are used. We also provide background on the legacy SEER system, focusing on our work to extend a single-layer implementation of HMMs into a more effective cascade of HMMs, that we refer to as Layered Hidden Markov Models (LHMMs). Section 4 describes the three selective perception strategies that we have studied in experiments: EVI-based, rate-based and random-based perception. In Section 5 we review the implementation of S-SEER. Experimental results with the use of S-SEER are presented in Section 6. Finally, we summarize our work in Section 7.

## 2 Prior Related Work

**Human Activity Recognition** Most of the prior work on leveraging perceptual information to recognize human activities has centered on the identification of a specific type of activity in a particular scenario. Many of these techniques are targeted at recognizing single events, *e.g.,* “waving the hand” or “sitting on a chair”. Over the past few years, there has been increasing work on methods for identifying more complex patterns of human behavior, including patterns extending over increasingly long periods of time. A significant portion of work in this arena has harnessed Hidden Markov Models (HMMs) [2] and extensions. Starner and Pentland in [3] use HMMs for recognizing hand movements used to relay symbols in American Sign Language. More complex models, such as Parameterized-HMMs [4], Entropic-HMMs [5], Variable-length HMMs [6], Coupled-HMMs [7], structured HMMs [8] and context-free grammars [9] have been used to recognize more complex activities such as the interaction between two people or cars on a freeway.

Moving beyond the HMM representation and solution paradigm, researchers have investigated more general temporal dependency models, such as dynamic Bayesian networks. Dynamic Bayesian networks have been adopted by several researchers for the modeling and recognition of human activities [10–15].

We have explored the use of a layering of probabilistic models at different levels of temporal abstraction. We have shown that this representation allows a system to learn and recognize in real-time common situations in office settings [1]. Although the methods have performed well, a great deal of perceptual processing has been required by the system, consuming most of the resources available by personal computers. We have thus been motivated to explore strategies for selecting on-the-fly the most informative features, starting with the integration of decision-theoretic approaches to information value for guiding perception.

**Principles for Guiding Perception** Decision theory centers on representations and principles for deciding among alternative courses of action under uncertainty. At the heart of decision theory are the axioms of utility, desiderata about preferences under uncertainty. The axioms imply the Principle of Maximum Expected Utility (MEU), which asserts that the best action to take is the one associated with the highest expected utility. The engineering discipline of Decision Analysis has developed a rich set of tools, methods and practices on top of the theoretical foundations of Decision Theory. Expected Value of Information (EVI) refers to the expected value of making observations under uncertainty, taking into consideration the probability distribution over values that would be seen should an observation be made. In practice, EVI computations can be used to identify MEU decisions with regards to information gathering actions.

The connection between decision theory and perception received some attention by AI researchers studying computer vision tasks in the mid-70's, but interest faded for nearly a decade. Decision theory was used to model the behavior of vision modules [16], to score plans of perceptual actions [17] and plans involving physical manipulation with the option of performing simple visual tests [18]. This early work introduced decision-theoretic techniques to the perceptual computing community.

Following this early research, was a second wave of interest in applying decision theory in perceptual applications in the early 90's, largely for computer vision systems [19] and in particular in the area of active vision search tasks [20].

## 3 Toward Robust Context Sensing

Before focusing on the control of perceptual actions, let us discuss in more detail the domain and original SEER office-awareness prototype. We shall turn to selective perception for HMM-centric multimodal systems in Section 4.

A key challenge in inferring human-centric notions of context from multiple sensors is the fusion of low-level streams of raw sensor data—for example, acoustic and visual cues—into higher-level assessments of activity. 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) [1]. For recognizing office situations, we have explored the challenge of fusing information from the following sensors:

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

**2. Camera:** A video signal is obtained via a standard Firewire camera, sampled at 30 f.p.s, that is used to determine the number of persons present in the scene.

3. **Keyboard and mouse:** SEER keeps a history of keyboard and mouse activities during the past 1, 5 and 60 seconds.

Figure 1 illustrates the hardware configuration used in the SEER system.

### 3.1 Hidden Markov Models (HMMs)

In early work on SEER we explored the use of single-layer hidden Markov models (HMMs) to reason about an overall office situation. Graphically, HMMs are

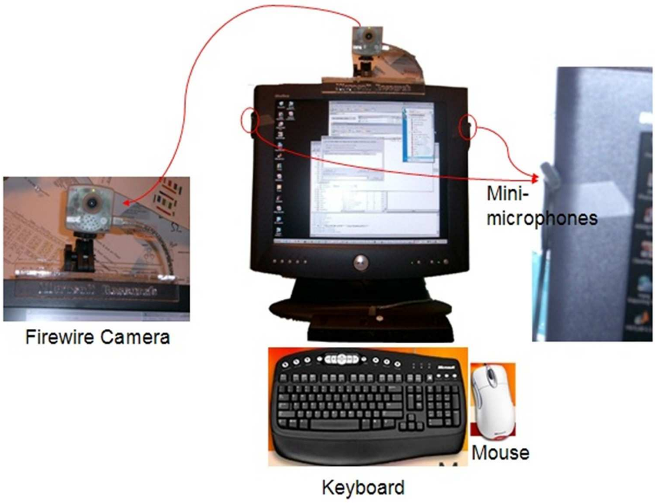

**Fig. 1.** Hardware utilized in the SEER system.

often depicted "rolled-out in time". We found that a single-layer HMM approach generated a large parameter space, requiring substantial amounts of training data for a particular office or user. The single-layer model did not perform well: the typical classification accuracies were not high enough for a real application. Also, when the system was moved to a new office, copious retraining was typically necessary to adapt the model to the specifics of the signals and/or user in the new setting. Thus, we sought a representation that would be robust to typical variations within office environments, such as changes of lighting and acoustics, and models that would allow the system to perform well when transferred to new office spaces with minimal tuning through retraining.

### 3.2 Layered Hidden Markov Models (LHMMs)

We converged on the use of a multilayer representation that reasons in parallel at multiple temporal granularities, by capturing different levels of temporal detail. We formulated a layered HMM (LHMM) representation that had the ability to decompose the parameter space in a manner that reduced the training and tuning requirements. In LHMMs, each layer of the architecture is connected to the next layer via its inferential results. The representation segments the problem into distinct layers that operate at different temporal granularities[1] —allowing for temporal abstractions from pointwise observations at particular times into explanations over varying temporal intervals. LHMMs can be regarded

[1] The "time granularity" in this context corresponds to the window size or vector length of the observation sequences in the HMMs.

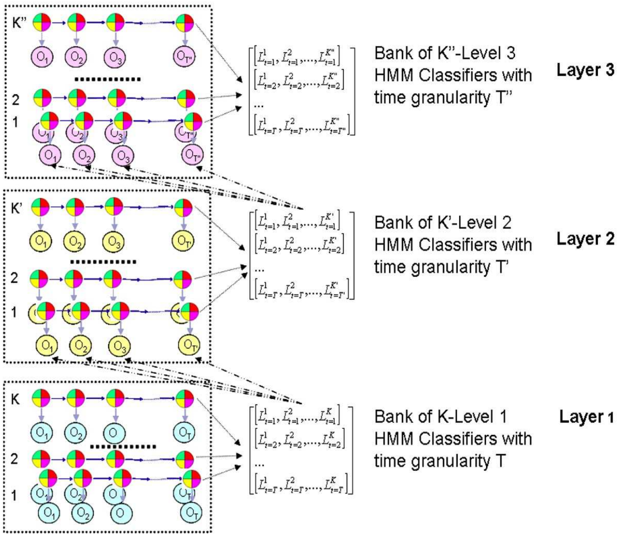
K''
2
1
Bank of K''-Level 3 HMM Classifiers with time granularity T''
Layer 3
K'
2
1
Bank of K'-Level 2 HMM Classifiers with time granularity T'
Layer 2
K
2
1
Bank of K-Level 1 HMM Classifiers with time granularity T
Layer 1

**Fig. 2.** Graphical representation of LHMMs with 3 different levels of temporal granularity.

as a cascade of HMMs. The structure of a three-layer LHMM is displayed in Figure 2.

The layered formulation of LHMMs makes it feasible to decouple different levels of analysis for training and inference. As we review in [1], each level of the hierarchy is trained independently, with different feature vectors and time granularities. In consequence, the lowest, signal-analysis layer, that is most sensitive to variations in the environment, can be retrained, while leaving the higher-level layers unchanged. Figure 2 highlights how we decompose the problem into layers with increasing time granularity.

## 4 Selective Perception Policies

Although the legacy SEER system performs well, it consumes a large portion of the available CPU time to process video and audio sensor information to make inferences. We integrated into SEER several methods for selecting features dynamically.

### 4.1 EVI for Selective Perception

We focused our efforts on implementing a principled, decision-theoretic approach for guiding perception. Thus, we worked to apply *expected value of information*

(EVI) to determine dynamically which features to extract from sensors in different contexts.

We compute the expected value of information for a perceptual system by considering the value of eliminating uncertainty about the state of the set of features $f_k, k = 1...K$, under consideration. For example, as illustrated in Figure 3, the features associated with the vision sensor (camera) are motion density, face density, foreground density and skin color density in the image[2]. There are $K = 16$ possible combinations of these features and we wish the system to determine in real-time which combination of features to compute, depending on the context[3].

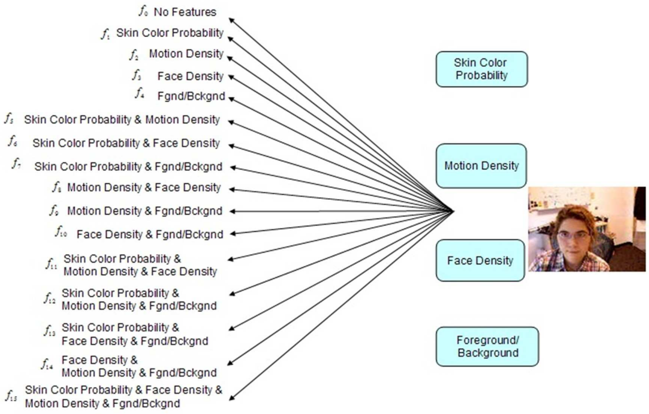
$f_0$ No Features
$f_1$ Skin Color Probability
$f_2$ Motion Density
$f_3$ Face Density
$f_4$ Fgnd/Bckgnd
$f_5$ Skin Color Probability & Motion Density
$f_6$ Skin Color Probability & Face Density
$f_7$ Skin Color Probability & Fgnd/Bckgnd
$f_8$ Motion Density & Face Density
$f_9$ Motion Density & Fgnd/Bckgnd
$f_{10}$ Face Density & Fgnd/Bckgnd
$f_{11}$ Skin Color Probability & Motion Density & Face Density
$f_{12}$ Skin Color Probability & Motion Density & Fgnd/Bckgnd
$f_{13}$ Skin Color Probability & Face Density & Fgnd/Bckgnd
$f_{14}$ Face Density & Motion Density & Fgnd/Bckgnd
$f_{15}$ Skin Color Probability & Face Density & Motion Density & Fgnd/Bckgnd
Skin Color Probability
Motion Density
Face Density
Foreground/ Background

**Fig. 3.** Possible feature combinations for the vision sensor.

**Perceptual Decisions Grounded in Models of Utility** We wish to guide the sensing actions with a consideration of their influence on the global expected utility of the system's performance under uncertainty. Thus, we need to endow the perceptual system with knowledge about the value of action in the world. In our initial work, we encoded utility as the cost of misdiagnosis by the system. We assess utilities, $U(M_i, M_j)$, as the value of asserting that the real-world activity $M_i$ is $M_j$. In any context, a maximal utility is associated with the accurate assessment of $M_j$ as $M_j$.

[2] By "density" we mean the number of pixels in the image that, for example, have motion above a certain threshold, divided by the total number of pixels in the image.

[3] In the following we will refer to features instead of sensors, because one can compute different features for each sensor input –e.g. skin density, face density, motion density, etc, for the camera sensor.

**Uncertainty About the Outcome of Observations** Let us take $f_k^m, m = 1...M$ to denote all possible values of the feature combination $f_k$, and $E$ to refer to all previous observational evidence. The expected value (EV) of computing the feature combination $f_k$ is,

$$EV(f_k) = \sum_m P(f_k^m|E) \max_i \sum_j P(M_j|E, f_k^m) U(M_i, M_j) \tag{1}$$

As we are uncertain about the value that the system will observe when it evaluates $f_k$, we consider the change in expected value associated with the system's overall output, given the current probability distribution of the different values $m$ that would be obtained if the features in $f_k$ would in fact be computed, $P(f_k^m|E)$.

The expected value (EVI) of evaluating a feature combination $f_k$ is the difference between the expected utility of the system's best action when observing the features in $f_k$ and not observing them, minus the cost of sensing and computing such features, $cost(f_k)$. If the net expected value is positive, then it is worth collecting the information and therefore computing the features.

$$EVI(f_k) = EV(f_k) - \max_i \sum_j P(M_j|E) U(M_i, M_j) - cost(f_k) \tag{2}$$

where $cost(f_k)$ is in our case the computational cost associated with computing feature combination $f_k$. Perceptual systems normally incur significant cost with the *computation* of the features from the sensors. Thus, we trade the information value of observations with the cost due to the analysis required to make the observations. Note that all the terms in Equation 2 should have the same units. Traditionally, EVI approaches convert all the terms to dollars. In our case, we use a scale factor for the $cost(f_k)$.

**EVI in HMMs** Our probabilistic modules are HMMs, with one HMM per class. In the case of HMMs, with continuous observation sequences $\{O_1, ..., O_t, O_{t+1}\}$ and an observation space of $M$ dimensions (after discretization[4]), the EVI of features $f_k$ is given by:

$$\begin{aligned} EVI \propto \sum_{m=1}^{M} \sum_n [\sum_s \alpha_t^n(s) \sum_l a_{sl}^n b_l^n(O_{t+1}^{f_k^m})] P(M_n) \\ \max_i \sum_j U(M_i, M_j) p(M_j) - \max_i \sum_j U(M_i, M_j) p(M_j) - cost(O_{t+1}^{f_k}) \end{aligned}$$

where $\alpha_t^n(s)$ is the alpha or forward variable at time $t$ and state $s$ in the standard Baum-Welch algorithm [21], $a_{sl}^n$ is the transition probability of going from state $s$ to state $l$, and $b_l^n(O_{t+1}^{f_k^m})$ is the probability of observing $O_{t+1}^{f_k^m}$ in state $l$, all of them in model $M_n$.

[4] In S-SEER $M$ is typically 10.

The computational overhead added to carry out the EVI analysis is –in the discrete case– $O(M * F * N^2 * J)$, where $M$ is the maximum cardinality of the features, $F$ is the number of feature combinations, $N$ is the maximum number of states in the HMMs and $J$ is the number of HMMs.

### 4.2 Alternative Perception Policies

In order to better understand the properties of the EVI approach, we have developed alternative methods for selective perception. We explored, in a second selective perception policy, a heuristic, rate-based approach. This policy consists of defining an observational frequency and duty cycle (*i.e.* amount of time during which the feature is computed) for each feature $f$. With this approach, each feature $f$ is computed periodically. The period between observations and the duty cycle of the observation is determined by means of cross-validation on a validation set of real-time data.

For another baseline policy, we developed a simple random-selection method, where features are selected randomly for use on a frame-by-frame basis. In this case, the average computational cost of the system is constant, independent of the current sensed activity, and lower than the cost of computing all of the features all the time.

## 5 Implementation of S-SEER

S-SEER operates the same way as its predecessor, SEER, except in the availability of several selection perception policies. Figure 4 illustrates S-SEER's architecture. For clarity, we shall include a brief summary of the core system and move onto the details of experiments with selective perception in Section 6.

### 5.1 Core Learning and Inference

SEER consists of a two-level LHMM architecture with three processing layers. For a more detailed description we direct the reader to [1].

The raw sensor signals are preprocessed to obtain feature vectors (*i.e.* observations) for the first layer of HMMs.

With respect to the audio analysis, Linear Predictive Coding coefficients [2] are computed. Feature selection is applied to these coefficients via principal component analysis. The number of features is selected such that at least 95% of the variability in the data is maintained, which is typically achieved with no more than 7 features. We also extract other higher-level features from the audio signal such as its energy, the mean and variance of the fundamental frequency over a time window, and the zero crossing rate [2]. The source of the sound is localized using the Time Delay of Arrival (TDOA) method.

Four features are extracted from the video signal: the density of skin color in the image (obtained by discriminating between skin and non-skin models, consisting of histograms in YUV color space), the density of motion in the image (obtained by image differences), the density of foreground pixels in the image (obtained by background subtraction, after having learned the background), and

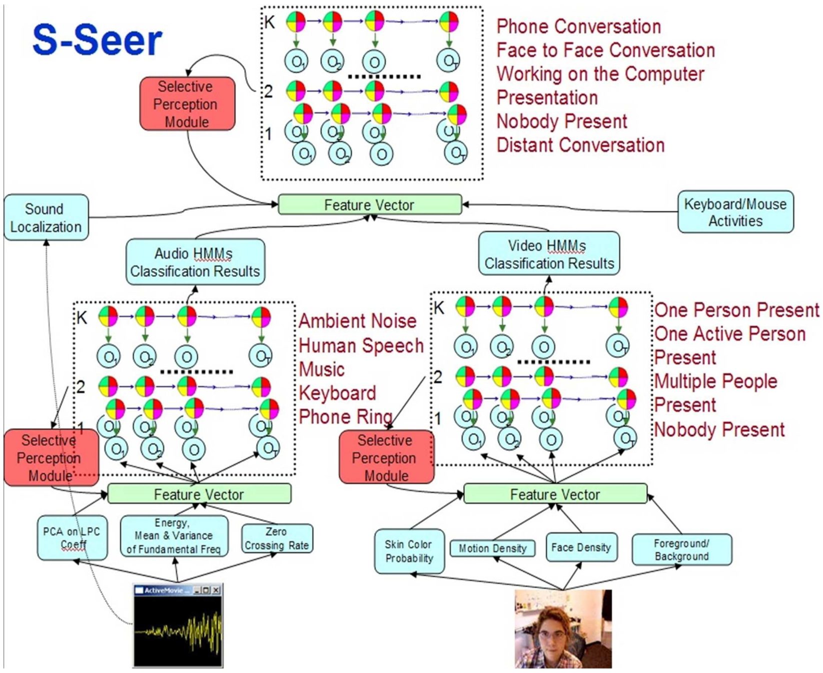

**Fig. 4.** Architecture of S-SEER.

the density of face pixels in the image (obtained by means of a real-time face detector).

Finally, a history of the last 1, 5 and 60 seconds of mouse and keyboard activities is logged.

**First Level HMMs** The first level of HMMs includes two banks of distinct HMMs for classifying the audio and video feature vectors. The structure for each of these HMMs is determined by means of cross-validation on a validation set of real-time data. On the audio side, we train one HMM for each of the following office sounds: *human speech, music, silence, ambient noise, phone ringing*, and the sounds of *keyboard typing*. In our architecture, all the HMMs are run in parallel. At each instant, the model with the highest likelihood is selected and the data –*e.g.* sound in the case of the audio HMMs– is classified correspondingly. We will refer to this kind of HMMs as *discriminative* HMMs. The video signals are classified using another bank of discriminative HMMs that implement a person detector. At this level, the system detects whether *nobody, one person (semi-static), one active person, or multiple people* are present in the office.

Each bank of HMMs can use any of the previously defined selective perception strategies to determine which features to use. For example, a typical scenario is one where the system uses EVI analysis to select in real-time the motion and skin density features when there is *one active person* in the office, and skin density and face detection when there are *multiple people* present.

**Second Level HMMs** The inferential results[5] from this layer (*i.e.* the outputs of the audio and video classifiers), the derivative of the sound localization component, and the history of keyboard and mouse activities constitute a feature vector that is passed to the next (third) and highest layer of analysis. This layer handles concepts with longer temporal extent. Such concepts include the user's typical activities in or near an office. In particular, the activities modeled 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. Some of these activities can be used in a variety of ways in services, such as those that identify a person's availability.

The models at this level are also discriminative HMMs and they can also use selective perception policies to determine which inputs from the previous layer to use.

### 5.2 Performance of SEER

We have tested S-SEER in multiple offices, with different users and respective environments for several weeks. In our tests, we have found that the high-level layers of S-SEER are relatively robust to changes in the environment. In all the cases, when we moved S-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 models required re-training to tune their parameters to the new conditions (such as different ambient noise, background image, and illumination) . The fundamental decomposability of the learning and inference of LHMMs makes it possible to reuse prior training of the higher-level models, allowing for the selective retraining of layers that are less robust to the variations present in different instances of similar environments. We direct the reader to [1] for a detailed description of the experiments comparing HMMs and LHMMs for office activity recognition as well as a review of an evaluation of the recognition accuracy of the system.

## 6 Experiments with Selective Perception

We performed a comparative evaluation of the S-SEER system when executing the EVI, rate-based, and random selective perception algorithms when applied at the highest level of the Layered HMMs architecture. Therefore, in the experiments that follow the selective perception policies will select any combination of four possible sensors: vision, audio, keyboard and mouse activities and sound localization.

### 6.1 Studies of Accuracy and Computation

In an initial set of studies, we considered diagnostic accuracy and the computational cost incurred by the system. The results are displayed in Tables 1 and 2. We use the abbreviations: PC=Phone Conversation; FFC=Face to Face Conversation; P=Presentation; O=Other Activity; NP=Nobody Present; DC=Distant Conversation.

[5] See [1] for a detailed description of how we use these inferential results.

Observations that can be noted from our experiements are: (1) At times the system does not use any features at all, as S-SEER is confident enough about the situation, and it selectively turns the features on only when necessary; (2) the system guided by EVI tends to have longer switching time (*i.e.* the time that it takes to the system to realize that a new activity is taking place) than when using all the features all the time. We found that the EVI computations trigger the use of features again only after the likelihoods of hypotheses have sufficiently decreased, *i.e.* none of the models is a good explanation of the data; (3) in most of our experiments, S-SEER never turned the *sound localization* feature on, due to its high computational cost versus the relatively low informational value this acoustical feature provides.

Tables 1 (a) and (b) compare the average recognition accuracy and average computational cost (measured as % of CPU usage) when testing S-SEER on 600 sequences of office activity (100 sequences/activity) with and without (first column, labeled "Nothing") selective perception. Note how S-SEER with selective perception achieved as high a level of accuracy as when evaluating all the features all the time, but with a significant reduction on the CPU usage. Given S-SEER's –without selective perception– perfect accuracy, one could think that the task is too easy for the model and that is the reason why the selective perception policies have reasonable accuracies as well. We would like to emphasize that the results reflected on the table correspond to a particular test set. In a real scenario, S-SEER's accuracy is on average 95% or higher, leaving still some room for improvement. We are also exploring more challenging scenarios for S-SEER, both in terms of the number of activities to classify from and their complexity.

**Table 1.** Average accuracies and computational costs for S-SEER with and without different selective perception strategies.

| **Recognition Accuracy (%)** | | | | |
|---|---|---|---|---|
| | Nothing | EVI | Rate-based | Random |
| PC | 100 | 100 | 29.7 | 78 |
| FFC | 100 | 100 | 86.9 | 90.2 |
| P | 100 | 97.8 | 100 | 91.2 |
| O | 100 | 100 | 100 | 96.7 |
| NP | 100 | 98.9 | 100 | 100 |
| DC | 100 | 100 | 100 | 100 |

(a)

| **Computational Costs (% of CPU time)** | | | | |
|---|---|---|---|---|
| | Nothing | EVI | Rate-based | Random |
| PC | 61.22 | 44.5 | 37.7 | 47.5 |
| FFC | 67.07 | 56.5 | 38.5 | 53.4 |
| P | 49.80 | 20.88 | 35.9 | 53.3 |
| O | 59 | 19.6 | 37.8 | 48.9 |
| NP | 44.33 | 35.7 | 39.4 | 41.9 |
| DC | 44.54 | 23.27 | 33.9 | 46.1 |

(b)

### 6.2 Richer Utility and Cost Models

The EVI-based approach experiments previously reported correspond to using an identity matrix as the system's utility model $U(M_i, M_j)$ and a measure of cost $cost(f_k)$, proportional to the percentage of CPU usage. However, we can assess more detailed models that capture a user's preferences about different misdiagnoses in various usage contexts and about latencies associated with computation for perception.

**Models of the Cost of Misdiagnosis** As an example, one can assess in dollars the cost to a user of misclassifying $M_i$ as $M_j$, $i, j = 1...N$ in a specific setting. In one assessment technique, for each actual office activity $M_i$, we seek the dollar amounts that users would be willing to pay to avoid having the activity misdiagnosed as $M_j$ by an automated system, for all $N-1$ possible misdiagnoses.

**Models of the Cost of Perceptual Analysis** In determining a real world measure of the expected value of computation, we also need to consider the deeper semantics of the computational costs associated with perceptual analysis. To make cost-benefit tradeoffs, we map the computational cost and the utility to the same currency. Thus, we can assess cost in terms of dollars that a user would be willing to pay to avoid latencies associated with a computer loaded with perceptual tasks.

We can introduce key contextual considerations into a cost-model. For example, we can condition cost models on the specific software application that has focus at any moment. We can also consider settings where a user is not explicitly interacting with a computer (or is not relying on the background execution of primary applications), versus cases where a user is interacting with a primary application, and thus, at risk of experiencing costly latencies.

We compared the impact of an activity-dependent cost model in the EVI-based perception approach. We run S-SEER on 900 sequences of office activity (150 seq/activity) with a fixed cost model (*i.e.* the computational cost) and an activity-dependent cost model. In the latter case, the cost of evaluating the features was penalized when the user was interacting with the computer (*e.g.* Presentation, Person Present-Other Activity), and it was reduced when there was no interaction (*e.g.* Nobody Present, Distant Conversation Overheard).

Table 2 summarizes our findings. It contains the percentage of time per activity that a particular feature was active both with constant costs and activity-dependent costs. Note how the system selects less frequently computationally expensive features (such as video and audio classification) when there is a person interacting with the computer (third and fourth columns in the table) while it uses them more frequently when there is nobody in front of the computer (last two columns in the table). Finally, the last row of each section of the table corresponds to the average accuracy of each approach.

**Table 2.** Impact of a variable cost model in EVI-based selective perception as measured in percentage of time that a particular feature was "ON".

| | PC | FFC | P | O | NP | DC |
|---|---|---|---|---|---|---|
| Video | 86.7 | 65.3 | 10 | 10 | 78.7 | 47.3 |
| Audio | 86.7 | 65.3 | 10 | 10 | 78.7 | 47.3 |
| Sound Loc | 0 | 0 | 0 | 0 | 0 | 0 |
| Kb/Mouse | 100 | 100 | 27.3 | 63.3 | 80.7 | 100 |
| Accuracy (%) | 100 | 100 | 97.8 | 100 | 98.9 | 100 |

(a) Constant Cost

| | PC | FFC | P | O | NP | DC |
|---|---|---|---|---|---|---|
| Video | 78 | 48.7 | 2 | 1.3 | 86 | 100 |
| Audio | 78 | 40.7 | 2 | 1.3 | 86 | 100 |
| Sound Loc | 14.7 | 0 | 2 | 1.3 | 86 | 100 |
| Kb/Mouse | 100 | 100 | 53.3 | 63.3 | 88 | 100 |
| Accuracy (%) | 82.27 | 100 | 97.7 | 87.02 | 98.47 | 100 |

(b) Variable Cost

**Volatility and Persistence of the Observed Data** We can extend our analysis by learning and harnessing inferences about the persistence versus volatility of observational states of the world. Rather than consider findings unobserved at a particular time slice if the corresponding sensory analyses have not been immediately performed, the growing error for each sensor (or feature computation), based on the previous evaluation of that sensor (or feature) and the time since the finding was last observed, is learned. The probability distribution of how each feature's uncertainty grows over time can be learned and then captured by functions of time. For example, the probability distribution of the skin color feature used in face detection that had been earlier directly observed in a previous time slice can be modeled by learning *via* training data. As faces do not disappear instantaneously –at least typically, approximations can be modeled and leveraged based on previously examined states. After learning distributions that capture a probabilistic model of the dynamics of the volatility versus persistence of observations, such distributions can be substituted and integrated over, or sampled from, in lieu of assuming "not observed" at each step. Thus, such probabilistic modeling of persistence can be leveraged in the computation of the expected value of information to guide the allocation of resources in perceptual systems.

We are currently working on learning the uncertainties for each sensor (feature) from data and applying this approach to our EVI analysis.

## 7 Summary

We have reviewed our efforts to endow a computationally intensive perceptual system for office activity recognition with selective perception policies. We have explored and compared the use of different selective perception policies for guiding perception in our models, emphasizing the balance between computation and recognition accuracy. In particular, we have compared *EVI-based* perception and *rate-based* perception techniques to a system evaluating all features all of the time all and a random feature selection approach. We have carried out experiments probing the performance of LHMMs in S-SEER, a multi-modal, real-time system for recognizing typical office activities.

Although the EVI analysis adds computational overhead to the system, we have shown that a utility-directed information-gathering policy can significantly reduce the computational cost of the system by selectively activating features, depending on the situation. When comparing the EVI analysis to the rate-based and random approaches, we found that EVI provides the best balance between computational cost and recognition accuracy. We believe that this approach can be used to enhance multimodal interaction in a variety of domains.

We have found that selective perception policies can significantly reduce the computation required by a multimodal behavior-recognition system. Selective perception policies show promise for enhancing the design and operation of multimodal systems–especially for systems that consume a great percentage of available computation on perceptual tasks.

## References

1. Oliver, N., Horvitz, E., Garg, A.: Layered representations for human activity recognition. In: Proc. of Int. Conf. on Multimodal Interfaces. (2002) 3–8
2. Rabiner, L., Huang, B.: Fundamentals of Speech Recognition. (1993)
3. Starner, T., Pentland, A.: Real-time american sign language recognition from video using hidden markov models. In: Proceed. of SCV'95. (1995) 265–270
4. Wilson, A., Bobick, A.: Recognition and interpretation of parametric gesture. In: Proc. of International Conference on Computer Vision, ICCV'98. (1998) 329–336
5. Brand, M., Kettnaker, V.: Discovery and segmentation of activities in video. IEEE Transactions on Pattern Analysis and Machine Intelligence **22(8)** (2000)
6. Galata, A., Johnson, N., Hogg, D.: Learning variable length markov models of behaviour. International Journal on Computer Vision, IJCV (2001) 398–413
7. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: Proc. of CVPR97. (1996) 994–999
8. S. Hongeng, F.B., Nevatia, R.: Representation and optimal recognition of human activities. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'00. (2000)
9. Ivanov, Y., Bobick, A.: Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. on Pattern Analysis and Machine Intelligence, TPAMI **22(8)** (2000) 852–872
10. Madabhushi, A., Aggarwal, J.: A bayesian approach to human activity recognition. In: In Proc. of the 2nd International Workshop on Visual Surveillance. (1999) 25–30
11. Hoey, J.: Hierarchical unsupervised learning of facial expression categories. In: Proc. ICCV Workshop on Detection and Recognition of Events in Video, Vancouver, Canada (2001)
12. Fernyhough, J., Cohn, A., Hogg, D.: Building qualitative event models automatically from visual input. In: ICCV'98. (1998) 350–355
13. Buxton, H., Gong, S.: Advanced Visual Surveillance using Bayesian Networks. In: International Conference on Computer Vision, Cambridge, Massachusetts (1995) 111–123
14. Intille, S.S., Bobick, A.F.: A framework for recognizing multi-agent action from visual evidence. In: AAAI/IAAI'99. (1999) 518–525
15. Forbes, J., Huang, T., Kanazawa, K., Russell, S.: The batmobile: Towards a bayesian automated taxi. In: Proc. Fourteenth International Joint Conference on Artificial Intelligence, IJCAI'95. (1995)
16. Bolles, R.: Verification vision for programmable assembly. In: Proc. IJCAI'77. (1977) 569–575
17. Garvey, J.: Perceptual strategies for purposive vision. Technical Report 117, SRI International (1976)
18. Feldman, J., Sproull, R.: Decision theory and artificial intelligence ii: The hungry monkey. Cognitive Science **1** (1977) 158–192
19. Wu, H., Cameron, A.: A bayesian decision theoretic approach for adaptive goal-directed sensing. ICCV **90** (1990) 563–567
20. Rimey, R.D.: Control of selective perception using bayes nets and decision theory. Technical Report TR468 (1993)
21. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceed. of the IEEE **77** (1989) 257–286