

General Examination on Technical Area:
Recognizing human activity in audio/visual scenes

Nitin Sawhney, MIT Media Lab
August 7, 2000

Examiner: Trevor Darrell
Associate Professor, Dept. of EECS and MIT A.I. Laboratory

Here are the questions for your exam; you may choose to answer any two (or all 3 if you prefer).

1. Relate the literature in Sociology to the task of perception & representation of human activity by computers.

Consider an "always-on" audio/video service that allows people to stay in touch and maintain an on-going awareness of others in some manner; Goffman proposes that people have a sophisticated means for managing their "face" in the presence of others in everyday situations. What social challenges are posed by a system that mediates people's "face" in such a continuous and distributed manner?

Note: I chose not to answer Q.1 here, preferring instead to write an extended survey for Q.2. I realized that in my response to Mark Ackerman's exam, I extensively reviewed the literature in HCI and sociology in relation to shared media spaces, which is quite similar to the scenario posed here. Please refer there for details. I will address some of these issues in my response to Q.3.

2. Compare probabilistic models of human activity - Bayesian Networks and HMMs.

Compare and contrast these methods and their limitations, referring both to static and dynamic forms of Bayesian Networks. What class of applications are they generally better suited for and why?

Note: I provide a broad survey of Bayesian Networks, a framework which helps characterize what are generally considered HMMs, Kalman filters, and static and dynamic DBNs. There is much confusion in terminology in the literature; often these models are only slight structural variations of another, but their particular representational expressiveness and computational methods make them better suited for certain kinds of problems. I emphasize the distinctions and limitations as I understand them, and discuss known applications.

3. Propose perceptual sensing mechanisms to help to mediate communication and awareness.

Consider issues such as managing privacy, interruption and one's "face" in different situations with different people. Refer to prior work in HCI & scene understanding, and pose a means for analyzing audio/video that is meaningful for such settings. What features and techniques would one utilize and how? Perhaps propose a concrete (simplified) machine learning experiment for some aspect of such a system. Discuss how it would be modeled and the challenges.

Note: To answer this question I pose a domestic communication scenario. For this domain I chose to answer in an unconventionally frank manner, based on my intuitions, personal life and some recent field studies I'm involved in, rather than from a purely academic perspective. This is because I am struggling hard with such questions, and haven't managed to find a satisfactory framework in sociology or HCI (perhaps I'm just burnt-out on it). The response may point to areas for further social and perceptual research. I hope it will be useful.

2. Probabilistic Models of Human Activity

Most types of human activity inferred from auditory, visual or sensory data can generally be represented as time series models with stochastic and deterministic components. Consider the time-varying features of speech utterances and physical gestures; such expressions occur over an extended time period, vary slightly each time, and their accurate perception is influenced by context, prior actions and the environment. For example, making sense of an ambiguous spoken utterance over a phone connection requires that we recognize the context of what is being said, based on past utterances, the person's accent and the noise in the background. Hence to recognize the dynamic complexities in human speech, we need some form of signal processing for feature extraction, language models and statistical learning for incorporating prior knowledge, and modeling uncertainty to predict future outcomes. Here we focus on understanding statistical learning of causality in situations and temporal sequences using a Bayesian framework.

Overview of Bayesian Networks

The *Bayesian network formalism*, also referred to as probabilistic graphical models or belief networks, is a combination of probability theory and graph theory in which dependencies between random variables is expressed graphically. Hence a Bayesian network can be defined as a "graphical model for representing conditional independencies between a set of random variables" [Ghahramani97]. Let us consider an example from a tutorial by Ghahramani. Figure 1 shows a graphical representation of the joint probability $P(W,X,Y,Z)$ that can be factorized as a set of conditional independence relations, as follows:

$$P(W,X,Y,Z) = P(W) P(X) P(Y|W) P(Z|X,Y)$$

Given the values of X and Y , we can show that Z and W are independent.

$$P(Z,W|X,Y) = P(W|Y) P(Z|X,Y)$$

So the Bayesian network is a way of graphically representing a *particular factorization* of a joint distribution. This factorization implies a certain ordering of the random variables in a manner that defines a directed acyclic graph (DAG). Undirected graphical models are considered Markov networks, with a different set of semantics. In a DAG each node (variable) is conditionally independent from its non-descendants, given its parent nodes. For example, we can visually infer from the DAG that W is conditionally independent from X given the set $\{Y, Z\}$, but not necessarily from X given Z (cannot infer that from the graph). Here the set $\{Y, Z\}$ *d-separates* the disjoint nodes W and X .

The graph not only allows us to understand which variables affect others, but also serves as a means to efficiently compute marginal and conditional probabilities for inference and learning. For *singly connected networks*, in which the underlying undirected path has no more than one path between any two nodes (i.e. no loops), the general algorithm used is called *Belief Propagation*. For *multiply connected networks*, in which there can be more than one undirected path between any two nodes, a more general algorithm used is the *Junction Tree Algorithm*.

A Bayesian network can be constructed by combining *a priori* knowledge about conditional independencies between variables, either from an expert in a particular domain by asking questions about causality (as is often done in *Static Bayes nets*) or from observed temporal data (as modeled by *Dynamic Bayesian networks*).

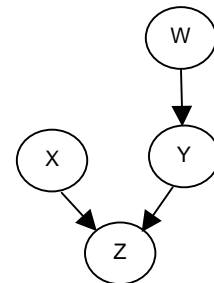


Figure1: A *directed acyclic graph* (DAG) representing the conditional independence between random variables in the joint probability distribution $P(W,X,Y,Z)$.

(Static) Bayesian Networks - Modeling Causality in Situations for Decision Theory

Bayesian networks allow one to model situations in which causality plays a role but our understanding of what is going on is incomplete [Charniak91]. Networks that do not explicitly model temporal sequences, but rather causality relations (not necessarily temporal) between random variables, are generally referred to as (static) Bayesian networks. These variables may have discrete or continuous values, although most algorithms developed for (static) Bayesian networks operate on discrete variables. Bayesian networks allow one to calculate the conditional probabilities of the nodes in the network given that values of some of the nodes have been observed. For example, the nodes W, X, Y, Z in Figure 1 could represent states such as family-out, dog-out, lights-on, and dog-barking, respectively. Here if we observe the *evidence* that the dog is barking and the lights are off, we can *evaluate* the conditional probability of the family being out. As additional evidence is observed, the conditional probability of all nodes changes.

Such Bayesian networks are extensively used for solving problems related to *decision theory*. Here one specifies the desirability of various outcomes (their utility) and the costs of various actions that might be performed to affect the outcomes. The goal of decision theory is to find the action or plan that maximizes the expected utility minus the costs [Charniak91]. Bayesian networks extended for decision theory, by incorporating decision nodes and value nodes, are called *influence diagrams*. In *Pathfinder* [Heckerman90], a physician can manually enter information regarding a patient's symptoms and get the conditional probabilities of the diseases (related to the lymph node) given the evidence so far. Here decision theory is used to choose which test to perform next, when current tests are not sufficient to make a diagnosis; however it is not used to make treatment decisions as it is sensitive to details of the utilities (e.g. how much pain to tolerate to decrease risk of death by $X\%$). Besides medical diagnosis, Bayesian networks are also used for conventional A.I. problems such as language (story) understanding [Charniak91], map learning and navigation, heuristic search, time-critical decision making [Horvitz95], intelligent software help assistants (Lumiere project at Microsoft) [Horvitz98] and most recently for email alerting based on inferred priorities and costs of interruption [Horvitz99].

One potential problem with such networks, for example in the Lumiere system is that the user may wish to violate the distribution of probabilities that the system is built upon, causing it not to anticipate novel situations. Another is the computational difficulty of updating all probabilities in exploring previously unknown network. Finally a Bayesian network is only as useful as the prior knowledge is reliable. An excessively optimistic or pessimistic expectation of the quality of these beliefs will distort the entire network and invalidate the results. Selecting the appropriate statistical distribution model to describe the data has a notable effect on the quality of the resulting network.

Dynamic Bayesian Networks - Modeling Temporal Events

For time-series modeling we can assume that an event can cause another event in the future, but not vice-versa. This simplifies the design of Bayesian networks allowing directed arcs to flow forward in time. The simplest causal model for such temporal data is a *first-order Markov model*; here each variable is directly influenced only by one in the previous time-step. Hence, having observed a sequence $\{Y_1, \dots, Y_t\}$, the model makes use of Y_t to predict Y_{t+1} . These models can be extended to allow higher-order interactions between variables. Alternatively, these Markov models can be extended by assuming that the observations are dependent on hidden variables $\{X_1, \dots, X_t\}$, called *states*; this sequence of states is called a *Markov process*. Such a state-space model with linear functions and gaussian variables is called a *Kalman filter* (described below).

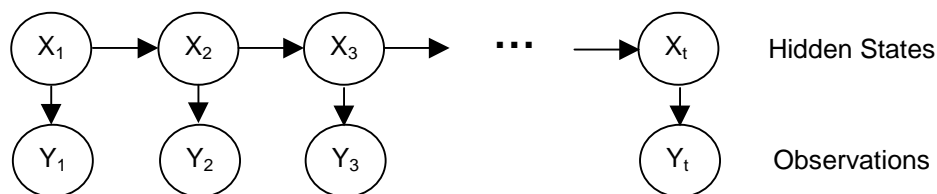


Figure 2: A Bayesian network showing a first order Markov process of hidden states $\{X_1, \dots, X_t\}$ generating observations $\{Y_1, \dots, Y_t\}$ at each time-step.

State-Space Models: Kalman Filters - continuous valued hidden states

In state-space models, a sequence of D-dimensional real-valued observations $\{Y_1, \dots, Y_T\}$ is modeled by assuming that each is generated by a K-dimensional real-valued state variable X_t . The state transition probabilities $P(X_t|X_{t-1})$ and observation probabilities $P(Y_t|X_t)$ are both decomposed into deterministic and stochastic components. If both the transition and output functions are linear and time-invariant, and the distribution of states and observation noise variables is Gaussian, we get a Linear-Gaussian state-space model [Ghahramani97]:

$$\begin{aligned} X_t &= A X_{t-1} + w_t & \text{- where } A \text{ is the state transition matrix and } w_t \text{ is a noise vector} \\ Y_t &= C X_t + v_t & \text{- where } C \text{ is the observation matrix and } v_t \text{ is a noise vector} \end{aligned}$$

Such models, also known as Kalman filters [Kalman61], are used extensively in control and signal processing applications. In contrast to HMMs, Kalman filters are most relevant when the hidden state is most naturally described by continuous variables (although they can be extended to problems with discrete-state components). They are also distinctive because they factor the hidden state into a combination of quantities, as indicated by the vector nature of the hidden variables [Zweig97]. Although different schemes for applying Kalman filtering to speech recognition have been tested, better results are generally obtained by using HMMs for that domain (described next). However, Kalman filters have been found to better model finer properties like smoothness and continuity in human driving behavior [Pentland99].

Hidden Markov Models (HMMs) - discrete hidden states and continuous density observations

In an HMM, the observation sequence is modeled by assuming each observation Y_t depends on a *discrete* hidden state S_t and the sequence of these hidden states $S = \{S_1, \dots, S_T\}$ is distributed according to a Markov process. The state can take on one of N discrete values, $S_t \in \{1, \dots, N\}$. So the state transition probabilities $P(S_t|S_{t-1})$, for a time-invariant HMM can be specified by a single $N \times N$ transition matrix A. If the output observations M per state are discrete symbols, the emission probabilities $P(Y_t|S_t)$ can be specified by a $N \times M$ observation matrix B. An initial state distribution π is also needed. Hence a HMM can be fully specified by the two model parameters N, M, the observation symbols, and 3 sets of probability measures A, B and π . Notationally, an HMM is typically written as $\lambda = \{A, B, \pi\}$ to indicate the parameter set of the model [Rabiner93].

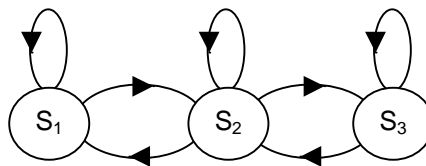


Figure 3: A conventional three state left-to-right HMM.

Given appropriate values for these parameters, an HMM can generate an observation sequence $Y = \{Y_1, \dots, Y_T\}$. Given this form of the HMM, three basic problems must be solved to make them useful for a real-world application [Rabiner93 and tutorial notes by Victor Zue]:

1. *Scoring*: Given an observation sequence Y and a model λ , efficiently compute the probability of observing the observations $P(Y|\lambda)$. This is solved by the *Forward-Backward Algorithm*.
2. *Matching*: Given an observation sequence Y , choose an optimal state sequence S that best explains the observations. This can be found by the *Viterbi Algorithm*.
3. *Training*: Adjust the model parameters $\lambda = \{A, B, \pi\}$ to maximize $P(Y|\lambda)$. This is solved by *Baum-Welch Reestimation Procedures*, similar to *Expectation Maximization (EM) Algorithm*.

Conventional HMMs can be extended to handle continuous (real-valued) observation vectors (such as speech) by replacing the output probabilities (matrix B) with continuous probability densities, usually represented as a mixture of Gaussians. Continuous-density HMMs currently yield the best performance for large-vocabulary speech recognition [Rabiner93] have been

applied extensively to problems in computational biology such as DNA sequencing [Baldi94] and protein modeling [Krogh94], and visual recognition of human gestures such as American Sign Language [Starnes95].

Several structural extensions and variations of HMMs have been explored recently. One approach of associating output distributions with transitions rather than states has been found to be equivalent. In many realistic applications the assumptions of linear state dynamics and Gaussian noise are not valid. In HMMs, unconstrained mixture models can be used to model any distribution given infinite mixture components [Ghahramani97]. An unconstrained state transition matrix can also model arbitrary non-linear dynamics. However, modeling any system that requires a large number of states is computationally inefficient and difficult to interpret. An unconstrained HMM with K^M states has K^{2M} parameters in the transition matrix; this would require a large amount of training data to capture all possible transitions. Several extensions have been proposed such as *Factorial HMMs* [Ghahramani97] (underlying state transitions are constrained), *Tree-Structured HMMs* (coupling state variables in each time step), and *Switching-State space models* (combining real-valued states of Kalman filters with discrete states of HMMs); however in these models most probabilities are intractable to compute. Hence structured approximations to these extensions are proposed to make the computations tractable.

Modeling structure both in space and time requires multiple interacting processes, and hence a compositional representation of 2 or more state variables in a HMM simultaneously. With a single-state variable, Markov models are ill suited to these problems. Extensions of the basic HMM to 3 or more chains is intractable and needs approximation techniques. However an exact solution exists for 2 interacting chains that influence each other through causal links. Such a structure called Coupled HMMs [Brand96] has been effectively used in modeling simple human behavior between 2 interacting pedestrians [Oliver99].

The literature on learning a model of human behavior over time-spans greater than a few seconds is sparse. Hogg et al. demonstrate how to learn characteristic motions of pedestrians in a plaza, by representing maps of pedestrian trajectories as non-parametric distributions over several hours [Johnson96]. Brand has developed a process of *entropic estimation* of parameter values to induce the structure of relations between hidden variables by trimming uninformative transitions and/or removing entire states [Brand97]. Such an entropic prior on parameter values and a solution to the MAP estimator has been used for learning a model of human activities over medium to long-term ambient video.

Differences in Modeling Temporal Sequences

Before describing *what is referred to as* DBNs, lets consider the differences and limitations of Kalman filters and (conventional) HMMs for temporal processing along these axes [Zweig97]:

Linearity - Kalman Filtering is fundamentally linear, although various schemes are developed to model non-linear systems, they tend to be complex and less applicable. HMMs are more suited to non-linear processes, due to arbitrary conditional probabilities associated with the transition and observation matrices.

Interpretability - Kalman filters are most interpretable since the matrices involved are usually designed by hand to reflect known physical processes. However the states of HMMs are not clearly interpretable, especially after training.

Factorization - Different modeling techniques have a wide variation in the degree to which they can be factorized, depending on how well the conventional model is modified. In Kalman filters the states and observation vectors are inherently factored, leading to reduced parameters. But conventional HMMs are fundamentally unfactored; i.e. if the state of a system consists of a combination of factors it cannot be concisely represented in this model. Although some schemes are proposed to represent such combinations (like speech and noise or articulatory features in HMMs), there is a parameter reduction but no corresponding computational savings.

Extensibility - Kalman filters are quite extensible as the state and observation variables are vectors; using higher dimensional vectors can increase system complexity. HMMs generally require additional states or chains to increase their modeling complexity.

Dynamic Bayesian Networks (DBNs) - Encoding model semantics with sub-models

In contrast to (static) Bayesian networks, a probabilistic network that models a system as it evolves over time is referred to as a Dynamic Bayesian Network (DBN) [Zweig97]. DBNs are time-invariant so that the topology of the network is a repeating structure, and its conditional probabilities do not change in each time-slice. When applied to an observation sequence of a given length, the DBN is "unrolled" to produce a network of the appropriate size to accommodate the observations.

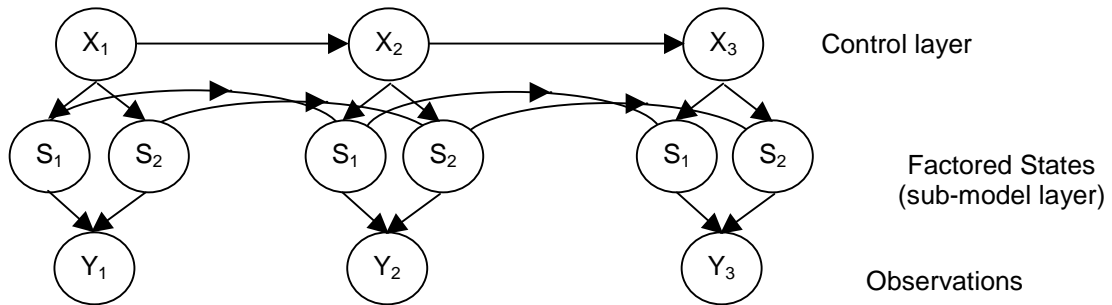


Figure 4: A Dynamic Bayesian network (DBN) showing control layer states $\{X_1...X_3\}$ factored into composite variable nodes $\{S_1, S_2\}$ chained together at each time-step.

Many temporal processes like speech recognition evolve through a number of distinct stages each best represented by a sub-model (such as for phonemes or articulatory dynamics). This composition achieves a combinatorial reduction in the number of models needed to be learned for a complex phenomenon. Such model composition requires constraining the manner in which sub-model sequences are represented in a Bayesian network. The sub-model state layer represents the hidden variables in the DBN sub-models. By conditioning state variables in control layer, a switching behavior between models is achieved. Hence an arbitrary set of variables encoding additional characteristics (such as duration of phone or type of speaker) can be associated with each time-slice, making a highly expressive representational framework [Zweig97].

In terms of the comparison axis mentioned earlier, DBNs have the following characteristics: DBNs can easily represent arbitrary non-linearities (via tabular representation of conditional probabilities), and allow efficient computation even for continuous state variables and Gaussian conditional probabilities. DBNs are interpretable as each variable (node) represents a specific concept. The joint distribution in a DBN can be factorized as much as possible. Compared to an unfactored HMM with equal number of states, a DBN with factored states and sparse connections between variables requires exponentially fewer parameters, leading to better computational efficiency. Finally DBNs are extensible as they can handle a large number of variables, given a sparse graph structure. In its simpler form the DBN merges into the HMM framework; conversely as implicit factorization is added to HMMs, they blend into the DBN methodology [Zweig97].

Experiments with application of DBNs to isolated word recognition, show that use of an auxiliary context variable improves recognition performance over a conventional HMM [Zweig99]. DBNs have been demonstrated for keyhole plan recognition [Albrecht97] to identify a user's next actions and goals in MUDs, text-based multi-player adventure games, based on training data of action/location probabilities over time. Finally DBNs have been recently applied towards visual sensor fusion, combining the output of several vision algorithms in an interactive kiosk to detect the presence of a speaker [Rehg99]. They have also been applied towards visual surveillance and analysis tasks, requiring the recognition of complex multi-agent actions in a scene [Intille98].

3. Perceptual mechanisms to mediate Communication and Awareness

Lately, I find myself more interested in understanding socially motivated problems on the "field", rather than in laboratory settings. Developing perceptual interfaces for real-world settings poses many social and technical challenges. I will attempt to discuss such issues in a domestic domain.

Supporting Communication Patterns in Domestic Life

Increasingly the distinction between people's work and homes lives, is becoming somewhat blurred. We need to maintain rich relationships with family and friends; when not in face-to-face contact, awareness and communication in such relations are largely mediated via everyday devices like phones, email, letters and faxes, and even via word of mouth from known others (your sister tells you that mom is ok). How can one maintain a richer awareness within such a domestic network and know when and how to communicate with others? Conversely how does one also maintain control over one's privacy and interruptability at certain times of day.

To answer these questions, we have undertaken an ethnographic field-study where we interview members of immigrant families and ask them about their domestic relations and communication patterns. As we are in early stages of this study (initiated late July '2000), we do not have results but we are beginning to recognize a few aspects. People's lives are not centered around any one device at all times, but many modalities serve complementary functions in different situations. People are increasingly mobile, and spend less time in any one physical location such as home or work. In particular, mobile-phones are increasingly being used in the U.S., even by members of lower-income families. Many people (like myself) who consciously choose not to carry mobile-phones have fairly consistent means for letting others know where they can generally be reachable (although part of the reason they don't carry phones is a desire to retain more control over such communication). Mobile-phone users have good strategies for maintaining contact with desired parties (via caller ID) and consciously leaving it on/off at certain times and places, as well as relying on voice/SMS messaging. Although clearly such devices tend to be used in a manner that can be quite disruptive to them and others (another reason why many others prefer not to use them). Awareness of others also comes via word-of-mouth from mutual friends and family.

Despite the vast array of literature on shared media spaces, we have not seen such systems (like video-conferencing) frequently being used in everyday situations. My response to Mark Ackerman's exam reviews the HCI and social issues involved, such as affordances of gaze and audio, peripheral awareness, interruption, ownership and so on in both work and domestic settings. Goffman's notion of "face" plays an important role in explaining why we don't see people using video-mediated communication more often. Although a phone connection is less expressive, it allows people to maintain a "face" that does not necessarily reveal their physical and emotional disposition at that moment. It allows people to invoke a limited interaction with another party, without all the social formalities associated with a face-to-face interaction. However, an audio channel does indeed allow others to infer many aspects of one's "face" from style of speaking, intonation in the voice, phrases spoken and so on. Email and instant messaging allow a form of low-bandwidth asynchronous communication, that many find appropriate in certain contexts. Hence people will often email or leave voice messages, even if their colleague is in an office around the corner. There has been much social and HCI research conducted on media spaces, but much less on audio-only spaces [Heath92, Watts96, Ackerman97] or asynchronous communication patterns. Some limited perceptual mechanisms have been used in media space applications, such as low-disturbance audio, shadow-views [Smith95], activity graphs and blurring filters [Zhao98, Crowley2000]. These techniques have not been evaluated extensively in social settings outside the lab; but more seriously, they assume highly simplified notions of people's expressive communication behaviors and complex social needs. Hence, there are greater social and technical challenges for developing perceptual mechanisms for everyday use in homes and on the street.

Trends indicate that mobile devices will be far more ubiquitous as wireless services become more affordable, and an array of richer information services on these devices become available. GSM

phones in Europe and I-mode phones in Japan already point towards the usage of many wireless, location-dependent and 'continuous-on' services. In addition, use of VoIP (voice over IP) via Internet devices and local phones is greatly expanding as the quality and accessibility of the services improves. Such devices & services could begin to incorporate sensing and processing to make perceptual mechanisms feasible. Despite my own proclaimed aversion to mobile-phones (and pagers/PDAs), I am beginning to recognize (from the limited field studies) that in conjunction with other methods they continues to serve as an important communication link in domestic life (it is clearly useful in business, but that's not my concern). However to understand and enhance domestic communication & awareness, I think we must also consider interaction in home and work life, where such devices may be turned off. Here regular phone use, email and even post-it notes are frequently used to mediate short messages and extended communication.

Key issue: How does one provide appropriate forms of awareness and interruption mechanisms that support a variety of practices and social behaviors in different contexts (home, work and commuting)?

My feeling is that despite the rich array of communication options available today, we still have difficulties staying in touch when needed, while retaining control over when we can be disrupted. With the general motivation I have outlined above, I'll now develop a scenario structured around hypothetical domestic life, as we don't have results from the ethnographic study yet. Here, I'll show the role of communication modalities used, problems/issues experienced and consider how to incorporate perceptual and learning mechanisms for improved awareness and communication.

Scenarios in Everyday Domestic Life

Joe is 28 and lives in Cambridge. He bikes to work everyday and uses desktop computing in his job. His sister Jane is an architect living in LA. She drives and carries a mobile-phone everywhere. His parents live abroad in South Africa, where long-distance calling can be expensive. Joe has several colleagues and friends in Cambridge who keep in touch via phone and voice messages. He lives with a roommate, but they tend to work on different time schedules, hence rarely meet. They tend to leave paper notes for each other and scribble messages on the fridge door. Without developing the scenario any further, we will make assumptions as we try to model the communication patterns.

In my mind, there are two key aspects of social interactions, that can benefit from methods grounded in perception and learning:

1. **Asynchronous Perceptual Awareness:** One goal is to model patterns in one's everyday lifestyle to provide an "abstracted form" of on-going awareness to "known" others about their daily routines. This may provide a better understanding of when people are better disposed for particular kinds of communication during the day. In our scenario, Joe would like to know when to call his sister, such that he does not disrupt her in a meeting or if sleeping earlier that day. If he knows something about her disposition, he may choose instead to email or leave a note for her (an SMS message or electronic note on her digital message pad at home). Similarly he prefers to speak with his mom when he's home cooking dinner, however she has no easy way to know when that occurs and ends up calling him at work when he 'd rather not to be disturbed. The manner in which a person facilitates their goals and actions, based on an awareness of the activities of others is referred to as 'social facilitation' [Ackerman95], and is found to be an important part of our social environments. So what is the role of perception and learning here? How can it be incorporated within the social affordances and concerns of such settings?

[Though the social impact of this scenario seems somewhat trivial, one can consider how it can be applied towards social facilitation among a team of nurses and emergency doctors who are on-call. Nurses frequently rely on pagers to contact doctors for critical events, without always recognizing their current disposition. This is a complex and life-critical scenario that can be better understood by doing ethnographic studies of their practices.]

2. **Perceptually Mediated Communication:** Let's consider a setting where several members of a group or domestic unit, maintain open audio channels at certain times of day. Say Joe (in his office around noon) likes to keep an open channel with his sister for 20-30 minutes while she drives to work in LA (around 9:00 AM), and his mom in South Africa is cooking dinner in the kitchen. They are all listening to a shared Internet radio broadcast of NPR news or Brazilian Jazz, while occasionally talking to each other over a shared VoIP (voice over IP) connection. Certainly in this scenario, they will all often be interrupted by phone calls, visitors and so on. In addition, although they would like to have an on-going awareness of the other, they may not wish to share all irrelevant or private conversations in their own space to every member of the audio space, including their uncontrolled or unintended acts (such as coughing). Actively managing such multi-party dialogue and spontaneous interruptions places great cognitive overhead and makes such an open audio space somewhat awkward to use. Can perceptual mechanisms help mediate real-time communication in this setting? What social issues arise?

[Again this scenario can be extended to complex/critical domains such as flight controllers sharing open voice channels [Watts96] or people monitoring 911 emergency calls.]

Lets now briefly consider some perception/learning approaches, within such contexts:

Recognizing Auditory Characteristics via Dynamic Bayesian Networks

A DBN can be trained on temporal sequences to distinguish voices of particular speakers, spoken background conversation, ambient sounds, driving, phones ringing and so on. Prior experiments have shown that HMMs can be used for this task [Clarkson and Sawhney, 98]. HMMs are trained on features such as Mel-scaled cepstral coefficients (which are good for recognizing speech), but poor at representing characteristics of everyday sounds. A major practical issue for such a system is to automatically acquire, segment and label different classes of sounds, especially when they are mixed in the background. These classifiers have to be frequently retrained for sounds in different locations, and tend to confuse many classes in new settings. One reason is that unlike humans, there is no rich context or prior knowledge built into the system. Hence even though we can expect to hear a phone ringing in an office, the system may confuse that with a high-pitched sound outdoors. Here prior knowledge via additional sub-models in a Dynamic Bayesian network or through some combination with a static Bayes net, may decrease the likelihood of such errors. This prior knowledge could be provided by acquiring data from people's natural interactions with other devices and from perceptual sensing. E.g. As I walk into my office and start using my desktop for email, the system can inform the mobile-phone that I'm indoors in a room-like environment, where it can select appropriate acoustic models. In an open-audio connection, auditory knowledge can allow the system to garble or lower the volume of conversations transmitted, or indicate the presence of people / context to the other party.

Modeling Availability and Interruptability via (Static) DBNs

A (static) Bayes net could be used to model causality between people's usage of devices and their location and availability. However, this model will rely on people's own assessment of their usage, availability and priorities in different situations. Horvitz et al. have reported on a system for managing email alerts based on such Bayesian inference [Horvitz99]. A rich understanding of one's auditory context, coupled with prior knowledge (in a static DBN) can guide a communication system to minimize interruption in situations or resolve ambiguities. For example, it may recognize that although my calendar indicates a meeting at this time, the auditory model shows no evidence of conversations in the last few minutes, and shows me logged into my desktop recently, increasing the likelihood that the meeting was cancelled or terminated early. In a social context with multiple individuals as part of a system, a much richer model of patterns could be derived. Socially available information, such as Joe's sister having just spoken to her mom at home, could be utilized in the system to update the likelihood of Joe finding mom at home for a voice conversation. In this way, a probabilistic model of availability/interruption patterns in family relations could be acquired via natural usage of (or lack thereof) multiple connected devices over time. An important issue is to allow people to easily view their model in a useful manner; however

probabilities are not easy to interpret, hence the system should show likely outcomes such as the chance of finding someone or being interrupted in different situations. E.g. Jane's mobile-phone could show that Joe is most likely busy but could be accessible by email or SMS at that moment.

In summary, these approaches rely on perceptual components and knowledge distributed across a network of devices, appliances and actual individuals. People's social interactions and everyday usage of devices allows the system to acquire usable models of their activity and availability over time. These models must be easily interpretable and modified to ensure trust, and the interfaces accessible and integrated in existing types of appliances to allow casual usage in domestic life. Extended ethnographic studies of people using such services will reveal their attitudes and social concerns.

References

Bayesian Networks Theory

- [Brand96] Brand, Matthew. 1996. Coupled Hidden Markov models for modeling interacting processes. Submitted to Neural Computation.
- [Charniak91] Charniak, Eugene. 1991. Bayesian Networks without Tears. *AI Magazine*, pp. 50-62.
- [Ghahramani97] Ghahramani, Zoubin. 1997. Learning Dynamic Bayesian Networks. *Adaptive Processing of Temporal Information*. Lecture Notes in Artificial Intelligence. Springer-Verlag. See related tutorial paper here - <http://www.cs.utoronto.ca/~zoubin/>
- [Heckerman96] Heckerman, David. 1996. A Tutorial on Learning with Bayesian Networks. Microsoft Research, Technical Report, MSR-TR-95-06.
- [Kalman61] Kalman, R., and Bucy, R. 1961. New Results in linear filtering and prediction theory. *Transaction ASME*, 83D, 95-108.
- [Rabiner93] Rabiner, L.R. and Juang, B.H. 1993. Ch. 6: Theory and Implementation of Hidden Markov Models. In *Fundamentals of Speech Recognition*.
- [Zweig97] Zweig, Geoffrey. 1997. Speech Recognition with Dynamic Bayesian Networks. Ph.D. Thesis, University of California, Berkeley. <http://www.cs.berkeley.edu/~zweig/>

Applications of Bayesian Networks

- [Albrecht97] Albrecht, D. W., Zukerman, I., Nicholson, A. E., Bud, A. 1997. In *User Modeling: Proceedings of the Sixth International Conference, UM97*. Vienna, New York. <http://um.org>
- [Baldi94] Baldi, P., Chauvin, Y., Hunkapiller, T., McClure, M.A. 1994. Markov models of biological primary sequence information. *Proc. Nat. Acad. Sci. (USA)*, 91(3):1059-1063.
- [Brand97] Brand, Matthew. 1997. Learning concise models of human activity from ambient video via a structure-inducing M-step estimator. Technical Report, Mitsubishi Electric Research Labs.
- [Clarkson98] Clarkson, B., Sawhney, N., Pentland, A. 1995. Auditory Context Awareness via Wearable Computing. 1998 Workshop on Perceptual User Interfaces (PUI98). Nov 4-6, 1998. pp. 37-43.
- [Heckerman90] Heckerman, David. 1990. Probabilistic Similarity Networks, Technical Report, STAN-CS-1316, Dept. of Computer Science and Medicine, Stanford Univ.
- [Horvitz95] Horvitz, E. and Barry, M. 1995. Display of Information for Time-Critical Decision Making. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal.
- [Horvitz98] Horvitz E. 1998. Lumiere Project: Bayesian Reasoning for Automated Assistance. Decision Theory and Adaptive Systems Group, Microsoft Research.
- [Horvitz99] Horvitz, E., Jacobs, A., Hovel, D. 1999. Attention-Sensitive Alerting. In *Proceedings of UAI '99, Conference on Uncertainty and Artificial Intelligence*, Stockholm, Sweden, pp. 305-313.

- [Intille98] Intille, S.S. and Bobick A. 1998. Representation and visual recognition of complex, multi-agent actions using belief networks. IN *CVPR '98 Workshop on Interpretations of Visual Motion*. Also see MIT Media Lab TR 454.
- [Johnson96] Johnson, N. and Hogg, D. 1996. Learning the distribution of object trajectories for event recognition. *IVC* 14(8):609-615.
- [Krogh94] Krough, A., Brown, M., Mian, I. S., Sjolander, K., Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501-1531.
- [Oliver99] Oliver, N., Rosario, B., Pentland, A. 1999. A Bayesian Computer Vision System for Modeling Human Interactions. *Proceedings of ICVS'99*, Gran Canaria, Spain.
- [Pentland99] Pentland, Alex and Liu, Andrew. 1999. Modeling and Prediction of Human Behavior. *Neural Computation*, 11, 229-242.
- [Rehg99] Rehg, J. M., Murphy, K. P., Fieguth, P. W. 1999. Vision-Based Speaker Detection Using Bayesian Networks. In *Proceedings of Computer Vision and Pattern Recognition*, pp. 110-116.
- [Starner95] Starner, Thad and Pentland, Alex. 1995. Visual Recognition of American Sign Language Using Hidden Markov Models. International Workshop on Automatic Face and Gesture Recognition (IWAAGR), Zurich, Switzerland.
- [Zweig99] Zweig, G. and Russell, S. 1999. Probabilistic Modeling with Bayesian Networks for Automatic Speech Recognition. *Australian Journal of Intelligent Information Processing Systems*, 5(4), 253-60.

Human-Computer Interaction

- [Ackerman95] Ackerman, M. and Starr, B. 1995. Social Activity Indicators: Interface Components for CSCW Systems. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST'95)*, pp. 159-168.
- [Ackerman97] Ackerman, M. S., Debby, H., Mainwaring, S. D. and Starr, B. 1997. Hanging on the 'Wire: A Field Study of an Audio-Only Media Space. *ACM Transactions on Computer-Human Interaction*, Vol.4, No.1, 39-66.
- [Crowley2000] Crowley, J. L., Coutaz, J. Berard, F. 2000, Things that See. *Communications of the ACM*. Vol.43, No.3, pp. 54-64.
- [Dourish96] Dourish, P. Adler, A. Bellotti, V., and Henderson, A. 1996. Your Place or Mine? Learning from Long-term Use of Audio-Video Communication. *Computer-Supported Cooperative Work*, 5(1), 33-62.
- [Heath92] Heath, Christian and Luff, Paul. 1992. Media Space and Communicative Asymmetries: Preliminary Observations of Video-Mediated Interaction. *Human-Computer Interaction*, 7(3), pp. 315-346.
- [Hudson96] Hudson, Scott E. and Smith, Ian. 1996. Techniques for Addressing Fundamental Privacy and Disruption: Tradeoffs in Awareness Support Systems. *Proceedings of CSCW '96*. pp. 238-247.
- [Smith95] Smith, I. and Hudson, S. E. 1995. Low-disturbance audio for awareness and privacy in media space applications. In *Proceedings of the ACM Conference on Multimedia*. ACM, New York, 91-97.
- [Watts96] Watts, J. C., Woods, D. D., Corban, J. M., Patterson, E. S.. 1996. Voice Loops as Cooperative Aids in Space Shuttle Mission Control. *Proceedings of CSCW '96*. pp. 48-56.
- [Zhao98] Zhao, Q. A. and Stasko, J. T. 1998. Evaluating image filtering based techniques in media space applications. *Proceedings of CSCW '98*.

Sociology

- [Goffman61] Goffman, Erving. 1961. *The Presentation of Self in Everyday Life*. Anchor-Doubleday, New York.

[O'Brien99] O'Brien, J. Rodeen, T. Rouncefield, M., Hughes, J. 1999. At Home with the Technology: An Ethnographic Study of a Set-Top-Box Trial. *ACM Transactions on Computer-Human Interaction*, Vol.6, No.3, 282-308.

[Whyte88] Whyte, William H. 1988. *City: Rediscovering the Center*. New York: Doubleday.