Towards Activity Databases: Using Sensors and Statistical Models to Summarize People's Lives

Tanzeem Choudhury[1,2], Matthai Philipose,[1] Danny Wyatt,[2] Jonathan Lester[1,2]
[1]Intel Research Seattle, 1100 NE 45th Street (6th Floor), Seattle, WA 98105.
[2]Univ. of Washington, Seattle, WA 98195
{tanzeem.choudhury,matthai.philipose}@intel.com
danny@cs.washington.edu, jlester@ee.washington.edu

Abstract

Automated reasoning about human behavior is a central goal of artificial intelligence. In order to engage and intervene in a meaningful way, an intelligent system must be able to understand what humans are doing, their goals and intentions. Furthermore, as social animals, people's interactions with each other underlie many aspects of their lives: how they learn, how they work, how they play and how they affect the broader community. Understanding people's interactions and their social networks will play an important role in designing technology and applications that are "socially-aware". This paper introduces some of the current approaches in activity recognition which use a variety of different sensors to collect data about users' activities, and probabilistic models and relational information that are used to transform the raw sensor data into higher-level descriptions of people's behaviors and interactions. The end result of these methods is a richly structured dataset describing people's daily patterns of activities and their evolving social networks. The potential applications of such datasets include mapping patterns of information-flow within an organization, predicting the spread of disease within a community, monitoring the health and activity-levels of elderly patients as well as healthy adults, and allowing "smart environments" to respond proactively to the needs and intentions of their users.

1 Introduction

For computers to become increasingly useful and capable of independently assisting human beings, they need to be given a richer understanding of how humans behave "in the world." The more a computer knows about the environment in which its user exists, the better it will be able to respond to and meet a user's needs. Example uses of such new understanding cover a wide range of applications, from a messaging application that does not interrupt its user when she is a giving talk, to a surgical assistant application that follows a doctor's motions and suggest diagnoses and actions.

Even if a system cannot fully model a user's beliefs, desires, and intentions, it can still be useful if it can simply recognize her activities. The recognition of human activities is becoming a central component to a many of the pervasive computing usage models and applications, such as activity-aware actuation in smart environments, embedded health assessment, assistive technologies for elder-care, task monitoring and prompting in the workplace, enhancing workplace efficiency and information flow, surveillance and anomaly detection, etc.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

For these applications to be practical, the underlying activity recognition module often needs to detect a wide variety of activities (people may routinely perform dozens to hundreds of relevant activities a day, for instance) performed in many different manners, under many different environmental conditions, and across many different individuals. The particular aspects of the activity that are of interest also vary widely across applications (e.g. user motion, whom the user interacts with, task progress, object usage or space usage). Hence, robust recognition across a variety of activities and individuals and their variations has proved to be difficult to engineer.

The current methods available for tracking activities are time and resource consuming manual tasks, relying on either paid trained observers (i.e. a job coach who periodically monitors an individual performing their job or a nurse monitoring an elderly patient) or on self-reporting, namely, having people complete an activity report at the end of the day. However, these methods have significant deficiencies in cost, accuracy, scope, coverage, and obtrusiveness. Paid observers such as job coaches and nurses must typically split their time among several clients at different locations. Also, extensive observation causes fatigue in observers and resentment in those being observed; in addition the constant involvement of humans makes the process very expensive. Self-reporting is often inaccurate and of limited usefulness due to patient forgetfulness and both unintentional and intentional misreporting, such as a patient reporting more fitness activities than they actually completed.

An automatic activity recognition system would help not only to reduce the errors that arise from self-reporting and sparse observational sampling, but also to improve the quality of service that coaches and caregivers can provide, as they would spend less of their time performing bookkeeping duties. In addition, unobtrusive monitoring enables people to go about their daily lives in an unimpeded manner, while providing their caregivers with a more accurate assessment of their real life activities, rather than of a small sample. An accurate automated system does has another clear benefit over existing methods such as surveys, in that it provides a continuous activity log along with times and durations for a wide range of activities.

Activity recognition is also an important component for modeling group-level behavior and social dynamics. Large businesses have long been interested in the flow of information within their organization, as the difference between success and bankruptcy can depend on how well information flows between different groups of employees. Although people heavily rely on email, telephone and other virtual means of communication, highly complex information is primarily exchanged through face-to-face interactions [1]. An understanding of these face-to-face interactions and the social networks in which they take place would enable businesses to determine bottlenecks and breakdowns in communication before they become serious problems. Another real-world problem in which social networks play a central role is the spread of disease. An infectious outbreak in a self-contained village community would exhibit a completely different propagation pattern than an outbreak in a busy metropolitan city. Knowing the social networks in these communities can have enormous practical benefits, from predicting the rate of propagation of a given disease to determining where it will spread to next. This information would enable doctors to curb the further spread of a disease and begin treatment of those likely to be infected, long before they might be aware of their illness. Wearable sensing combined with statistical reasoning techniques can play an important role in discovering and modeling face-to-face interactions.

2 Building an Activity Recognition System

An activity recognition system typically has three main subcomponents: (i) A low-level sensing module that gathers relevant information about activities, e.g., camera, microphone, acceleration, RFID etc. (ii) A feature-processing and feature-selection module that processes the raw sensor data into features that can help discriminate between activities. Features can be low-level information such frequency content or correlation coefficients, or higher level information such as objects detected or the number of people present in a scene. The third subcomponent, (iii), is a computational model that uses these various features to infer the activity that an individual or a group of individuals are engaged in, e.g. walking, talking, making tea, having a conversation etc.

Because human activities are complex and sensor signals have varying amounts of noise, these activity models are almost always probabilistic.

One has also to consider and specify the requirements for an activity recognition system that may determine the choice of sensors, form-factor, and the complexity of the models needed for inference. The aspects to consider are (i) functionality: whether the system will be used for logging and classifying activities (e.g. for a doctor or to understand the usage of space) or for taking an action based on inference (e.g. an application that reminds someone to take their medicine), (ii) speed: real-time inference is necessary for prompting but not necessary for logging, (iii) resolution and timescale, e.g. whether the system needs to detect number of steps a person or takes or how long a person spends at work, and (iv) accuracy: how well the inference system has to perform in order to be useful will depend on the application (e.g. a trade off might exist between allowing more false alarms but preventing more potentially harmful false negatives in medical domains). For an activity recognition system to be widely deployable and useable, it will need to support queries that are meaningful to the user of the system, and to provide the user with the right level of summarization of a person's life.

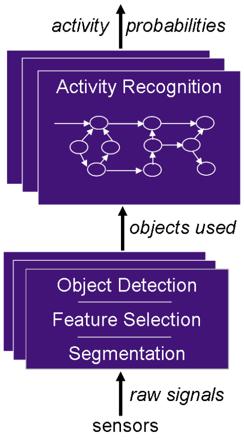

Figure 1: A typical activity recognition system

3 Sensing

The most common sensing approach is to use a few very rich sensors (typically one per room or user) such as cameras and microphones, which can record very large quantities of data about the user and their environment. For example, originally most of the research in activity recognition was done using vision and audio sensors [2, 3]. Although in principle the data captured by these sensors should be as useful as that captured by the key human senses of sight and hearing, in practice the task of extracting features from rich low-level representations has proved to be challenging in unstructured environments [4, 5].

An increasingly popular alternative approach is to use personalized sensors (one set of sensors per user) such as accelerometers and location beacons to get precise information about a particular small set of features related to the user, such as limb-movement and user location. The majority of research using wearable devices has concentrated on using multiple sensors of a single modality, typically accelerometers on several locations on the body [6, 7]. The placement of sensors in multiple pre-defined locations can be quite obtrusive and is one of the limitations of such an approach. As a result, a single sensing device that can be integrated into existing mobile platforms, such as a cell phone, would be more appealing to users and is likely to garner greater user acceptance. In our work, we have shown that incorporating multiple sensor modalities (e.g. accelerometer, audio, light, barometric pressure, humidity, temperature, and compass) will offset the information lost by using a single sensing device. Furthermore, multiple modalities will be better suited to record the rich perceptual cues that are present in the environment, cues that a single modality often fails to capture.

Recent advancements in miniaturization and wireless communication have seen the emergence of a third approach to sensing that may be termed *dense* sensing. In this approach, sensors are directly attached to many objects of interest. These sensors are either battery-free wireless stickers called Radio Frequency Identification (RFID) tags [8] or small wireless sensor nodes powered by batteries [9]. The sensors transmit to ambient readers the usage of the objects they are attached to by detecting either motion or hand-proximity to the object. Since each sensor has a unique identifier, fixed metadata about the object (such as its color, weight or even ownership), which would conventionally have to be discerned by sensors, can be easily associated in a directly machine-readable way with the object. The reliable sensing of detailed object-use that is enabled by dense sensing has several advantages: (i) for many day-to-day activities, the objects used serve as a good indicator of the activity being performed (ii) objects used remain fairly invariant across the different manners of performing these activities (iii) since the sensors detect the features quite well regardless of most environmental conditions, activity recognition can be robust to changes in the environment or individual. Finally, knowing the class of objects being used can serve as a powerful cue to constrain the search space of possible activities.

4 Feature Extraction

For an automated system to recognize people's behavior accurately, the choice of features is critical. The usefulness of certain features will depend on the application and the activities that need to be inferred. For example, frequency information from acceleration is important in determining activities such as walking, running and related gait. The periodicity of the auditory signal is useful in determining speech and whether someone is talking or not. The overall visual shape of objects appearing in an image can be used to detect the presence of a person. Some of the features may be deterministic transformations of the raw sensor data (e.g. frequency content), while others can be a probability measure (e.g. the likelihood that an image contains a human shaped blob or likelihood of a person being in a certain location). The time-scale at which features are computed also impacts recognition, e.g. human speech is usually analyzed at millisecond resolution whereas a variety of physical activity models use features computed at 0.1 to 10Hz, and contextual information about behavior is often computed over minutes or even hours.

It is conceivable that in the near future many people will be logging information about their activities and interactions continuously, for a variety of different purposes. The need for generating reliable databases that store features and support various types of queries over time, space and other sensor attributes will be increasingly important. Example queries may be of the form, "*get audio features from all the people who were in the computer science building at time t*" or "*get camera information from a specific location when there are more than 5 people present with at least 80% certainty*."

5 Models

The two main approaches that are used for classification in machine learning are: (i) generative techniques that model the underlying joint probability distribution P(X,Y) of the classes/activities (Y) and features (X), e.g. Naïve Bayesian models, Hidden Markov models, Dynamic Bayesian networks etc. and (ii) discriminative techniques that focus on learning the class boundaries [10] or only the class posterior probability P(Y|X), e.g. support vector machines, logistic regression and conditional random fields. Both of these approaches have been used extensively for recognizing various human behaviors and activities. Although discriminative techniques sometimes outperform generative approaches in classification tasks, generative models are necessary for synthesis (e.g. if a robot has to perform an instance of an activity), in anomaly detection, or in circumstances where the discovery of the underlying process is a goal. Another recently developed class of models, called probabilistic relational models and relational Markov networks, incorporate relational structure within the probabilistic framework [11, 12]. In relational models, the properties of a certain entity can depend probabilistically on the properties of other entities (e.g. a person's role can depend on the roles of other individuals in his social network as well as his own attributes). As a result these models have been successfully applied to activity recognition and social network modeling tasks. Another dimension along which techniques vary is the manner in which the models are learned. A conventional approach is to label traces of sensor data collected during the performance of a set of activities one wants to recognize, and use the labeled examples to learn the structure and parameters of the model; this is referred to as supervised learning. The other approach is unsupervised, where the underlying structure and associated parameters are learned automatically given only the sensor traces[13]. Semi-supervised techniques use sparsely labeled data to seed the parameters of unsupervised models [14].

No matter what kinds of model and learning techniques are used, sensor data has significant variability across instances and individuals, and it is nearly impossible for automated activity classification systems to recognize every event with absolute certainty. Thus, most of the time classifiers are probabilistic or have confidence values associated with them, especially when the activities being modeled are complex or the number of activities being recognized is large. Consequently, activity databases will also need to support probabilistic entries and queries. For example, a typical query may be, "*what is the expected time spent walking for person A on a weekday?*."

6 Sensing and Modeling Activities at Different Granularities

Human activities naturally fall into two categories: (i) activities that an individual does by himself (e.g. brushing teeth) and (ii) activities that he engages in with others (e.g. conversation). When it comes to probabilistic representation, joint activities require explicit modeling of the relationships between individuals and how they affect each other. Another aspect of activities that influences the choice of models is the time-scale, we currently break them into (a) short time-scale activities, where the pattern or regularity in the sensor data is present within a short time window (on the order of seconds, e.g. walking) and (b) long time-scale activities which have regularities at a longer time window (on the order of minutes and hours, e.g. attending a meeting). For short-time scale activities, static models are often sufficient (i.e. no temporal constraints), whereas for longer time-scale events temporal models are usually required. Most approaches to constructing models suffer from what may be termed as the *model completeness* problem: models have observations that are either missing or that have inappropriate probabilities. Incomplete models can, of course, result in faulty inference. For example a model for making tea may have probabilities of various objects being used, e.g. "kettle", "teabag", teacup" and "sugar", but may mention neither "coffee cup" nor "honey". Similarly, given the inconvenience of generating labeled examples of all (or most) possible ways to execute an activity, it is likely that probabilities associated with certain observations will be under-represented. Below we give a brief overview of the work done by our group in the following areas: (i) representation of individual-level activities using multi-modal sensing (ii) automated approaches to handling model incompleteness (iii) multi-person sensing and modeling of group interactions.

Representation of individual-level activities: Advances in the development of multi-modal wearable sensors enable us to gather rich datasets of human activities. However, the problem of automatically identifying the most useful features for modeling such activities remains largely unsolved. We have developed a discriminative approach based on boosting [15] to select the most useful features and learn a weighted set of static classifiers (decision stumps) that can be additively combined to recognize different activities. During training we provide a set of labeled examples to the system, which it uses to learn the most discriminative features and the model parameters. A trained system will then use the selected features to output a probability score that a given data point or data sequence is generated by a specific activity. To capture the temporal smoothness of activities, a first-order probabilistic Markov-model (a hidden Markov model, or HMM [16]) is trained using the output of the static classifiers, where the observation sequence of the HMM consists of the probabilities from the static classification step. This combination leverages the good discriminative qualities of the decision stumps with the temporal smoothness of the HMM. By automatically inferring the features that were most useful, we discovered that two modalities in particular (out of

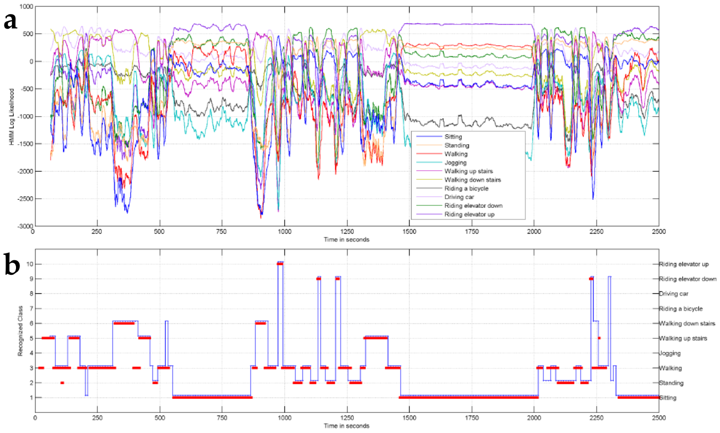

Figure 2: For a forty minute segment of data (a) the likelihood of the different activities over time and (b) the final output of the activity classification system in blue (based on the class that has maximum likelihood) and the ground truth in red.

seven) yielded the most discriminative information for our activities: the audio and accelerometer sensors. These two modalities provide complementary information about the environment and the wearer. The audio captures the sounds produced during the various activities, whereas the accelerometer data is sensitive to the movement of the body. In addition, other sensors yielded more specific types of information for certain activities, such as the barometric pressure sensor being sensitive enough to detect the activity of riding in an elevator[17, 18].

Automated approaches to handling model incompleteness: Learning from data requires labeling, and since the amount of data is large it is impractical to expect an appreciable portion of it to be labeled. However, although activities are varied and idiosyncratic, they have common features that most people recognize, i.e. they have a generic "common sense" aspect that often suffices to recognize them. Furthermore, many daily activities are performed using objects that can be easily recognized if they have RFID tags on them. We have developed techniques for mining from the web simple but useful discriminative models of numerous object-based activities, which can be applied to segment and label object-use traces (thereby avoiding the need for hand-labeling). These segments can then be used to effectively bootstrap the learning of better model parameters for activities.

Given a set of activities A, we mine from the web a set of objects O used for each activity a in A and their associated usage probabilities p(o ∈ O | a ∈ A). The mining process proceeds in four distinct steps. First, for each activity in A, we identify web pages that describe that activity being performed. Second, having identified these pages, we extract phrases from them that describe the objects used during the performance of the activity. Third, once the set of pages and phrases have been found, co-occurrence statistics for the pages and phrases are used to estimate the object-use probabilities. Finally, we use the mined information to assemble a Hidden Markov Model (HMM) capable of recognizing activities in traces of object data; the hidden states of the HMM correspond to the various activities, and the observation probabilities of the HMM are the object-use probabilities. Now, given a set E of unlabeled traces (a trace is a sequence of sensed objects), we use the mined models as a basis for learning an improved or more customized model. To train this customized model from the generic mined model, we first apply the most probable labeling for the traces E (using the Viterbi algorithm [16]) given the model. We then re-estimate the model parameters according to the labeled trace. If certain part of the model are not observed then their parameters are not changed, and remain set to the mined probabilities [8].

The use of objects as the underlying features being modeled suggests another simple approach to countering models with missing information. Intuitively, we can exploit common-sense information about which objects are functionally similar. If the model ascribes very different probabilities to two very similar objects, we can "smooth" these probabilities into more similar values. As a degenerate case, if the model omits an object while incorporating very similar ones, we can postulate that the omitted object is likely to be observed in the model. We have developed a completely unsupervised approach to realizing this idea. By using auxiliary information, called an ontology, about the functional similarities between objects, we mitigate the problem of model incompleteness. The similarity information is extracted automatically from WordNet, a large, relevant ontology of lexical reference system for the English language, and incorporated into our models by using a statistical smoothing technique, called shrinkage [19].

Multi-person sensing and modeling of group interactions: The structure and dynamics of face-to-face social networks are of critical importance to many social phenomena, ranging from organizational efficiency to the spread of knowledge and disease. Research in face-to-face networks has an abundance of interesting and important questions, but has been faced with a paucity of data rich enough to answer many of these questions. We believe better models of social network and organizational dynamics will facilitate efficient means of collaboration and information propagation. Virtually all of the datasets are collected manually by human observers or via surveys, which are very labor intensive and yield only a small number of observations, sparsely spread over time. In our work, we have demonstrated the feasibility of learning social interactions from raw sensor data. We collected a large repository of wearable sensor data that includes auditory features (raw audio signal is not stored for privacy reasons) for several hours everyday over several weeks from multiple individuals (more than twenty people). We have developed a framework for automatic modeling of face-to-face interactions, starting from data processing and going all the way up to capturing the structure and dynamics of social networks, by analyzing whom we talk to and how we talk to them. The micro-level inter-relationship between individuals is modeled via a coupled probabilistic model of turn-taking during conversations. The coupled model allows us to estimate how much influence an individual has in the overall turn-taking that occurs during conversations. Furthermore, we were able to show how this measure of "influence" correlates significantly with betweenness centrality [20], an independent measure of an individual's importance in a social network. This result suggests that micro-level measures such as conversational influence can be predictive of more macro-level social influence [21, 22].

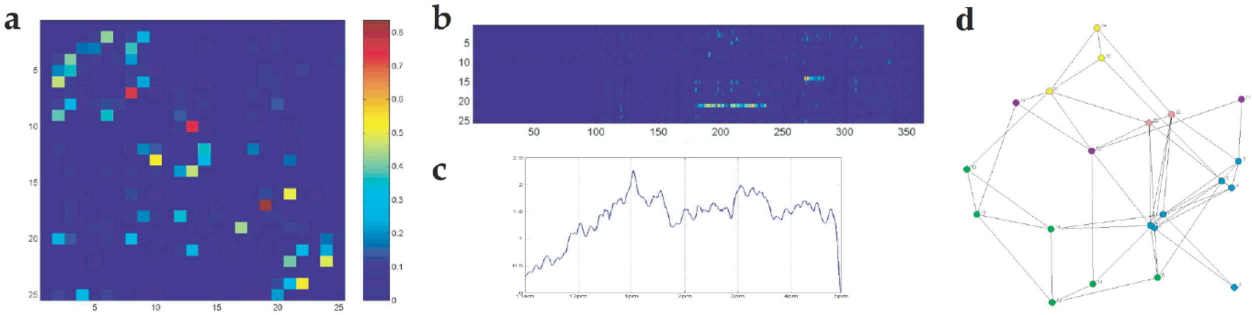

Figure 3: Some information of interest about social networks. (a) Interaction matrix, I, of a face-to-face network. Each row corresponds to a different person. I(i,j) is the fraction of person i's total interaction with person j. (b) A given person's interaction likelihood with other people in the network. The x-axis is time in minutes (6 hours) and the y-axis numbers are the IDs of people in the network. (c) Speech activity over the course of the day, averaged over all subjects. (d) Interaction network diagram, based on multi-dimensional scaling of geodesic distances. Node numbers represent the subject IDs.

7 Conclusion

This paper provides a brief introduction to the sensing and statistical reasoning techniques used in building systems that reason about human activities and interactions. The approaches outlined here are common to many systems, although the illustrative examples given in the paper have been drawn mostly from our own work.

Databases for storing the output of activity inference systems need to meet several challenges. They must be able to to support a variety of activity-based queries, while also protecting raw sensor data and sensitive private information. It is important that different users can be given different levels of access privilege to specific types of query. For example, the raw sensor data should be accessible only to a minimal set of individuals, whereas a broader set of users may be able to compute deterministic features or issue probabilistic queries. Privacy protection is even more important when answers to a query require access to the data from multiple individuals (e.g. "Did A and B have a conversation today?"). Social network information or any relational data can often destroy anonymity, so queries need to support varying levels of anonymity.

An activity database will certainly be probabilistic, as both entries and answer to queries will often be probabilistic. Such databases will also need to be able to deal with sporadically missing data, and with combining data from sensors that record at varying rates — the data from real-world activity recognition systems are all too rarely uniform or complete. The statistical techniques used in activity-recognition modeling already offer potential methods for handling missing information. Such tools may be applicable in ranking queries and in dealing with inconsistent data in probabilistic databases. Finally, these databases will also need to deal with temporal queries about peoples behavior. For example, a query may ask not what an individual was doing at a particular instant, but instead what their pattern of behavior was over the course of a day.

Sensors are being embedded more and more into the everyday objects around us: phones and watches contain cameras, microphones and GPS, and objects in shops, factories and hospitals are being tagged with RFID. Powerful computer processors are being incorporated into previously "dumb" consumer products. However, such technology will do little to improve usability if it is not sensitive to people's needs, and these needs vary as a function of the activities that people are engaged in. We therefore believe that the need for activity recognition, and for the management and retrieval of information about activities, will present important new research challenges for a long time to come.

8 References

1. Allen, T., Architecture and communication among product development engineers. 1997, Sloan School of Management, MIT: Cambridge. p. 1-35.
2. Pentland, A., *Smart Rooms.* Scientific American, 1996. **274**(4): p. 68-76.
3. Gavrila, G., *The visual analysis of human movement: A survey.* Computer Vision & Image Understanding, 1999, **75**(1).
4. Moore, D.J., I. Essa, and M.H. Hayes. Exploiting object context for recognition tasks. In *Proceedings of the Conference on Computer Vision*. 1999: IEEE Press.
5. Davis, J.W. and A.F. Bobick. The representation and recognition of action using temporal templates. In the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1997: IEEE Press.
6. Bao, L. and S. Intille. Activity recognition from user-annotated acceleration data. In *Pervasive* 2004.
7. Kern, N., B. Schiele, and A. Schmidt. Multi-sensor activity context detection for wearable computing. in the *Proc. EUSAI, LNCS*. 2003. Eindhoven, The Netherlands.
8. Wyatt, D., M. Philipose, and T. Choudhury. Unsupervised activity recognition using automatically Mined common sense. In *the Proc. of Twentieth National Conference on Artificial Intelligence (AAAI)*. 2005.
9. Munguia-Tapia, E., et al., MITes: Wireless portable sensors for studying behavior. In the *Proceedings of Extended Abstracts Ubicomp 2004: Ubiquitous Computing*. 2004: Vienna, Austria.
10. Rubinstein, Y.D. and T. Hastie. Discriminative vs. informative learning. In *the Proceedings of Knowledge Discovery and Data Mining*. 1997.
11. Getoor, L., et al., *Learning probabilistic models of relational structure.* Journal of Machine Learning Research, 2002. **3**: p. 679-707.
12. Taskar, B., P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Uncertainty in Artificial Intelligence (UAI)*. 2002.
13. Duda, R., P. Hart, and D. Stork, *Pattern Classification*. 2000: Wiley-Interscience.
14. Zhu, J., Semi-supervised learning literature survey. CS TR 1530, Univ. of Wisconsin-Madison, 2006.
15. Schapire, R.E. A brief introduction to boosting. In *16th Intl. Joint Conf. on Artificial Intelligence*. 1999.
16. Rabiner, L., A tutorial on hidden Markov models and selected applications in speech recognition. In the *Proceedings of IEEE*, 1989. **77**(2): p. 257-286.
17. Lester, J., Choudhury, T., et al. A hybrid discriminative-generative approach for modeling human activities. In *the Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*. 2005.
18. Lester, J., T. Choudhury, and G. Borriello. A practical approach to recognizing physical activities. In *Pervasive 2006*. Dublin, Ireland.
19. Munguia-Tapia, E., T. Choudhury, and M. Philipose. Building reliable activity models using hierarchical shrinkage and mined ontology. In *Pervasive 2006*. Dublin, Ireland.
20. Freeman, L.C., *A set of measures of centrality based on betweenness.* Sociometry, 1977. **40**: p. 35-41.
21. Choudhury, T., Sensing and Modeling Human Networks, *Doctoral Thesis*. 2003, MIT, Cambridge, MA.
22. Choudhury, T. and S. Basu. Modeling Conversational Dynamics as a Mixed Memory Markov Process.In *Neural Information Processing Systems (NIPS)*. 2004. Vancouver, BC.