Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 503–511
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Human dynamics: computation for organizations

Alex Pentland *, Tanzeem Choudhury, Nathan Eagle, Push Singh

Room E15-387, MIT Media Laboratory, 20 Ames St, Cambridge, MA 02139, USA

Available online 28 September 2004

Abstract

The human dynamics group at the MIT Media Laboratory proposes that active pattern analysis of face-to-face interactions within the workplace can radically improve the functioning of the organization. There are several different types of information inherent in such conversations: interaction features, participants, context, and content. By aggregating this information, high-potential collaborations and expertise within the organization can be identified, and information efficiently distributed. Examples of using wearable machine perception to characterize face-to-face interactions and using the results to initiate productive connections are described, and privacy concerns are addressed.

1. Introduction

Studies of office interactions have discovered that 35–80% of work time is spent in spoken conversation, 14–93% of work time is spent in opportunistic communication, and 7–82% of work time is spent in meetings (Allen, 1997). Senior managers represent the high end of these scales. Clearly, face-to-face interaction within the worksplace is highly important, and critical pieces of information are often transmitted by word of mouth in a serendipitous fashion. The money and time spent on business travel and conferences further underscores the value of face-to-face interactions.

Given the importance of face-to-face interaction, it is notable that the majority of working professionals already carry around a computer, sensors, and network connection in the form of a cellular phone or personal digital assistant (PDA). Many of these devices have computational power similar to those found in desktop computers only a few years ago.

Our proposal is to harness the wearable computing and sensing power provided by mobile information devices to provide a 'social intelligence' that improves the functioning of work groups. We believe it will provide oganizations with an extraordinary resource for collaboration, team formation, knowledge management, and social network analysis.

To explain this vision we will first discuss the information that can be obtained from many

* Corresponding author.
E-mail address: pentland@media.mit.edu (A. Pentland).
URL: http://hd.media.mit.edu

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.08.012

streams of sensor data using an unobtrusive wearable computation platform. These interactions will then be mathematically modeled using probabilistic graphical models, allowing global analysis of the dynamics of the organization's information flow and potentially allowing far more effective management of the organization. We will then describe how this information can be combined with knowledge about human networks and commonsense knowledge about topics of conversation to characterize the semantics and the function of the interactions, and suggest some of the applications this might enable. Finally, we will discuss the privacy implications of such a system. Papers containing technical details can be found at http://hd.media.mit.edu.

2. Characterizing face-to-face interactions

The first step we take toward understanding face-to-face interactions is characterize them using inconspicuous wearable sensors. To accomplish this we have developed the MIThril architecture to add functionality to existing commercial mobile platforms, as shown in Fig. 1 (DeVaul et al., 2003).

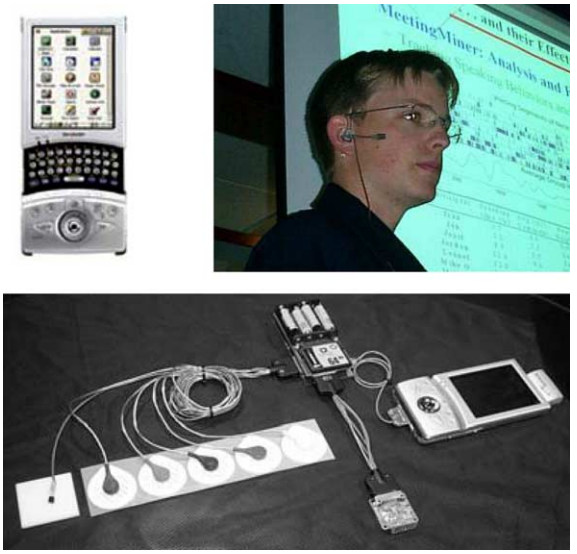

Fig. 1. The PDA-based computing engine, the audio interface, a full MIThril system with biosensors.

This architecture allows us to customize the mobile hardware and software architecture for distributed applications in classroom and collaborative settings.

The MIThril hardware architecture consists of flexible, modular networking protocols and a unified multi-wired power/data bus for sensors and peripherals. The computing core is a commerical PDA or mobile telephone, which provides users with an audio and graphical interface. The software architecture supports extremely flexible streaming and processing of sensor data among large groups of users.

Fig. 2 shows a self-contained, badge-like version called the Sociometer, which senses ambient audio, acceleration to characterize body language, and uses infrared (IR) beacons to establish the wearers' location and proximity to other people, as described in (Basu, 2002; Choudhury and Pentland, 2004; Choudhury, 2003).

There are four types of information inherent within streams of sensor data that are easily recorded from individuals of a common social network: features of the interaction (body language, speech prosody, etc.), participant identity, context, and content.

2.1. Features

Interaction features can be extracted from audio (e.g., speaking pitch, talking/not talking, and stress) and tiny, body-worn accelerometers (e.g., body language). In (Basu et al., 2001; Basu, 2002), we analyzed features such as speech energy, duration, and speaker transitions in conversations. We showed that speaker transitions could be

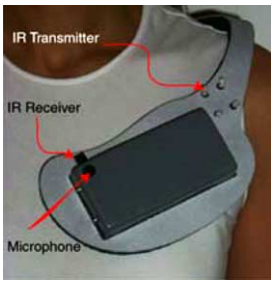

Fig. 2. The shoulder mounted sociometer.

learned and predicted given enough training data. We have since been able to show that these conversation features are indicative of the type of interaction. Conversations that are dominated by a single person who leaves no time for interjections are quite different from those conversations that have lower speech energy and regular speaker transitions. Conversation analysis allows group dynamics to be quantified and provides individual behavior profiles that can provide information to assist in assembling a project team.

2.2. Participants

There are many ways to establish the identity of people in nearby proximity. These include IR tags, radio frequency (RF) tags, and Bluetooth beacons. However these sensing methods only establish proximity, and not who is actually participating in a conversation. We have found, however, that we can reliably determine who is participating in a conversation (as opposed to merely being nearby) by calculating the mutual information between streams of audio conversations (Basu et al., 2001; Basu, 2002; Choudhury and Pentland, 2004; Choudhury, 2003; Clarkson et al., 1998). People who are conversing take turns, so that their pattern of speaking vs not-speaking is (approximately) complementary. We have also been able to reliably map the topology of wearer's social networks using this passive sensing methodology (DeVaul et al., 2003; Basu et al., 2001; Choudhury and Pentland, 2004).

2.3. Context

IR tags, RF tags, and Bluetooth beacons provide one way to establish the location of a face-to-face interaction. Additional context comes from how the user employs electronic aids. For example, if the user enters an appointment into his electronic calendar, the time, date, and key words entered can help identify the context of the conversation.

We can also use audio and video analysis to establish context. In (Clarkson and Pentland, 2000; Clarkson, 2002) we showed that analysis of either ambient 'background' noises or low-resolution 'peripheral' video can provide a reliable method of establishing both location and contextual situation.

In one experiment Clarkson used a wearable computer to collect 'peripheral' video over a period of 100 days, annotating the moment-by-moment context by hand. Annotations were by situation and not by location; they included when he was in any restaurant, any part of the laboratory, any classroom, any meeting, anywhere in his office, using public transportation, or walking to and from home (among other categories). We found that we could analyze very low-resolution version of this peripherial video (32×32 pixels, one frame per second) and still classify the context with 97% accuracy (Clarkson, 2002). Interestingly the processing required was minimal: it could be accomplished in real-time on todays' camera-equiped mobile telephones.

2.4. Content

Today's speech recognition engines typically have accuracy rates that vary widely, and unless great care is taken to control the ambient audio environment the overall word recognition accuracy will usually be quite low. Despite this lackluster performance, unique keywords can usually be identified and, as we have shown in (Jebara et al., 2000; Eagle et al., 2003), these noisy keyword inputs are sufficient for spotting topics within spoken conversation.

Keyword vocabulary also leverages the conversational feature information, and may lend insight into difficult questions such the nature of the conversation (e.g., is this a question or a response?). Such information is important to distinguishing relationship types, expertise areas, and conversation relevancy to enable appropriate collaborations.

3. The influence model

Once we have characterized face-to-face conversations using wearable sensors, the next challenge is to build a computational model that can be used to predict the dynamics of the individuals and their

interactions. The learnability and interpretability of a model greatly depends on its parameterization. The requirement for a minimal parameterization motivated our development of Coupled Hidden Markov Models (CHMMs) to describe interactions between two people, where the interaction parameters are limited to the inner products of the individual Markov chains (Oliver et al., 2000).

The "influence model" is a generalization of this approach, and describes the connections between many Markov chains as a network of convex combinations of the chains (Asavathiratham, 2000). This allows a simple parameterization in terms of the "influence" each chain has on the others. As described in (Asavathiratham, 2000), complex phenomena involving interactions between large numbers of chains can be analyzed by use of this simplified model.

The influence model is a tractable framework for understanding the recurrent classes of the global system and its steady state behavior by doing eigenstructure analysis of the "influence matrix" that describes the network of connections between the various constituent Markov chains. This representation makes the analysis of global behavior possible, which otherwise would become intractable with increasing number of individuals or agents. To apply the influence model to human networks we have extended the original formulation to include hidden states and develop a mechanism for learning the parameters of the model from observations (Choudhury et al., 2003).

The graphical model we use is identical to that of the generalized N-chain coupled HMM, but there is one very important simplification. Instead of keeping the entire $P(S_t^i|S_{t-1}^1,\ldots,S_{t-1}^N)$ for each of Q states, we only keep $P(S_t^i|S_{t-1}^j)$ and approximate the former with N^2 coupling or 'influence' parameters α_{ij}:

$$P(S_t^i|S_{t-1}^1,\ldots,S_{t-1}^N)=\sum_j \alpha_{ij}P(S_t^i|S_{t-1}^j)$$

In other words, we form our probability for the next state by taking a convex combination of the pairwise conditional probabilities for our next state given our previous state and the neighbors' previous state. As a result, we only have $NQ\times Q$ tables and $N\alpha$ parameters per chain, resulting in a total of NQ^2+N^2 transition parameters—far fewer parameters than any of the above models.

This simplification seems reasonable for the domain of human interactions and potentially for many other domains. Furthermore, it gives us a small set of interpretable parameters, the α values, which summarize the interactions between the chains. By estimating these parameters, we can gain an understanding of how much the chains influence each other.

To estimate the influence parameters we maximize the per-chain likelihood by gradient ascent using the derivative w.r.t. α_{ij}:

$$\frac{\partial}{\partial\alpha_{ij}}(\cdot)=\sum_t \frac{P(S_t^i|S_{t-1}^j)}{\sum_k \alpha_{ik}P(S_t^i|S_{t-1}^k)}=\sum_t \frac{P(S_t^i|S_{t-1}^j)}{\langle\alpha_i,B_t^i\rangle}$$

$$\text{where } \alpha=\begin{bmatrix}\alpha_{i0}\\ \vdots\\ \alpha_{iN}\end{bmatrix} \text{ and } B_t^i=\begin{bmatrix}P(S_t^i|S_{t-1}^0)\\ \vdots\\ P(S_t^i|S_{t-1}^N)\end{bmatrix}.$$

Typically no more than 20 iterations are required to ensure convergence.

3.1. An example: information propagation

One of the most important features of group interaction is the direction of information flow. Unfortunately, speech recognition and understanding technology cannot yet reliably determine the direction of information flow. However we can obtain a useful estimate of the direction of flow from 'verbal body language': we have discovered that the person driving the *dynamics* of the conversational turn-taking (e.g., the pattern of who speaks when) is usually the person eliciting new information (Madan et al., 2004).

We estimate directionality by calculating the influence each speaker has on the turn-taking dynamics of the conversation. Beginning with a classification of who is speaking at each instant in time, we characterize the dynamics of the individuals speaking vs not-speaking behavior using individual Hidden Markov Models (HMMs). We then estimate the influence parameters to describe interactions between pairs of people. This allows a simple parameterization in terms of the *influence* each person has on the other.

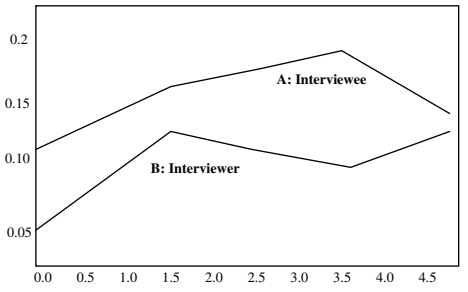

Fig. 3. Influence experienced by participants.

Fig. 3 shows the typical evolution of influence parameters for an interview conversation, where speaker A is interviewed by speaker B. As can be seen, the interviewee—the person providing the information—experiences much higher influence values. This is consistent with our previous research: in (Choudhury and Pentland, 2004; Choudhury, 2003) this measure of influence was shown to have an extremely high correlation with the 'betweeness centrality' of the information provider, and thus (according to common interpretation) with the directionality of information propagation. Assuming that these findings hold generally, then the social behavior that relates influence on conversational dynamics to direction of information flow gives us a way to characterize information flow within an organization without needing to do full speech understanding.

4. Using commonsense knowledge

Understanding the dynamics of group interactions using the influence model is only the start. For many applications we must combine interaction information from wearable sensors with higher-level, propositional knowledge about day-to-day problems and work patterns. We have found that this task can be greatly aided by using commonsense knowledge (Singh, 2002). For instance, by combining information from the audio stream and from other available contextual cues with commonsense knowledge about the activities people engage in and the topics they care about, we can infer a clearer picture of the content of conversations and the context of their participants (Eagle et al., 2003; Lieberman and Liu, 2002).

We make use of OMCSNet, a large-scale semantic network built at the Media Lab by aggregating and normalizing the contributions from over 10,000 people from across the web (Singh, 2002). It presently consists of over 250,000 commonsensical semantic relationships of the form 'a printer is often found in an office', 'going to a movie requires buying a ticket', and so forth. OMCSNet contains a wide variety of knowledge about many aspects of everyday life: typical objects and their properties, the effects of ordinary actions, the kinds of things people like and dislike, the structure of typical activities and events, and many other things. OMCSNet has been used in a variety of applications to date.

OMCSNet uses a hybrid knowledge representation strategy where individual concepts are represented linguistically (lexically and phrasally), and are related by a small set of about twenty specific semantic relationships such as LocationOf, SubeventOf, HasEffect, and so on.

4.1. Fine-grained topics: Gists

Our system's goal is to infer the 'fine grained topic', or gist, of the conversation. A gist is the class of event that most accurately summarizes the current subject of the conversation. For example:

- buying a ticket to a baseball game,
- looking for a restaurant,
- scheduling a meeting,
- canceling a meeting.

These gists are represented within OMCSNet as simple verb phrases. For our set of target gists, we use the 700 most richly defined situational aspects within OMCSNet (those for which at least 10 facts are asserted).

One feature that distinguishes commonsense reasoning from other forms of reasoning is that it involves making inferences using many different kinds of knowledge: about objects, events, goals, locations, and so forth. Accordingly, our system uses a probabilistic model that incorporates different types of knowledge, as well as contextual data from the mobile devices.

4.2. Inference in OMCSNet

Inference over the OMCS network can be done with varying levels of complexity, ranging from simple network analysis metrics to probabilistic modeling using Bayesian networks.

Before the inference, the transcriptions are preprocessed to reduce the noise of the speech recognition engine and improve inference performance. The transcriptions are first lemmatized and filtered for stop words (such as 'like', 'the', 'a', etc.). A second filtering process is then performed using a clustering metric to reduce the number of weakly connected words. These outliers, words with very sparse links to the rest of the transcription, are removed from the data set.

By flattening the networks of the different relationship types, a bipartite network can be formed to incorporate all ties from words to gists. The probability of a specific gist can be modeled as proportional to the gist's links to the selected words:

$$P(g_i|k) \propto \frac{k_i}{\sum_{i=1}^{G} k_i}$$

where k_i is the number of links between a gist, g_i, and the observed transcript, and G is the number of potential gists (approximately 700). This method is capable of identifying a small group of potential gists, frequently with the 'correct' one dominating the others.

Once the probable topics of conversation have been identified and ranked, contextual information about the conversation is incorporated into the model. In many instances, information such as location or participant identity can identify the gist from the small subsection of topics. In our initial tests we incremented a gist's score for each of its links to a keyword related to the given context. A more sophisticated model is being developed to incorporate the conditional probability distributions determined by training data.

4.3. Experiments

We ran a series of 20 interaction experiments on speech segments ranging from 50 to 150 words covering a wide range of topics. Conversational context was limited to location; the 802.11b network was used to give a general sense of location such as in an office or cafeteria.

Using the method described above, a ranking of the top ten gists for each interaction was created. The model gave a correct gist the number 1 ranking in 40% of the tests. In 70% of the tests, a correct gist was one of the top ranking three. However in 25% of the tests, a correct gist was ranked outside the top ten.

As an example to illustrate the functioning of the system, in one test we captured conversations from the student center cafeteria streaming data to an access point mapped as 'restaurant'. Using words along produced a useful result, but using words along with the contextual information to condition the model greatly improved our results:

Actual situation: Deciding what to get for lunch in the cafeteria.

Automatic transcription: Store going to stop and listen to type of its cellular and fries he backed a bill in the one everyone get a guess but that some of the past like a salad bar and some offense militias cambers the site fast food them and the styrofoam large chicken nuggets son is a pretty pleased even guess I as long as can't you don't have to wait too long its complicity sunrise against NAFTA pact if for lunch.

Automatically selected keywords: Wait type store stop salad past lunch long long listen large fry food fast chicken cellular bill big bar back (see Table 1).

5. Applications

Synergistic collaborations, real-time expertise, and redundant work can be identified by clustering people based on profiles generated from an aggregate of conversation, email, location, and web data. Additionally, by leveraging recent advances in machine learning, robust computational models can be built to simulate the effects of organizational disruptions in the existing social networks, such as relocating a group to a different location or merging two departments. Indeed, such a data-driven model offers the potential to transcend

Table 1
Gist classification with and without locational context; the numerical score is the number of links to the gist

Without location context		With location context	
5	Talk with someone far away	27	Eat in fast food restaurant
5	Buy beer	21	Eat in restaurant
5	Eat in restaurant	18	Wait on table
5	Eat in fast food restaurant	16	You would go to restaurant because you
5	Buy hamburger	16	Wait table
4	Go to hairdresser	16	Go to restaurant
4	Wait in line	15	Know how much you owe restaurant
4	Howl with laughter	12	Store food for people to purchase
4	Eat healthily	11	Sitting down while place order at bar
4	Play harp	11	Cook food

the traditional org-chart, perhaps by drawing parallels to ad-hoc network optimization. Forming groups based on inherent communication behavior rather than rigid hierarchy or formal education may yield significant improvements to the organizations' performance. The following sections suggest some ways in which this might be accomplished.

5.1. Expert and collaborator locator

Leveraging the techniques described here it may be possible to automatically generate profiles of each individuals expertise. By querying these profiles, a manager can form a team that has synergistic skills and social behavior. Clusters of people working on similar projects within a large organization can be identified to instigate collaboration and avoid redundant work. Experts can be identified similarly from conversation features, keywords and participants.

5.2. Collaboration tools

The emerging area of knowledge management attempts to capture and visualize the collective knowledge of an organization from individuals' posted profiles and web documents. Although existing corporate information repositories can be easily analyzed using standard data mining operations, the output reflects a severely limited and static view of an organization's human and social capital. Augmenting knowledge management and traditional social network analysis with information gathered by unobtrusive wearable sensors has enormous potential benefit for organizational collaboration. Querying this database for interests, skills, or simply recent vocabulary would be an efficient way to instigate collaboration, and should provide sufficient data to make expertise classifications (Foner, 1996).

5.3. Ad hoc conversation patching

There is no technical obstacle from prohibiting the system described above from operating in real-time. As conversations occur throughout an office space, they can be automatically streamed to a server that spots keywords that are relevant to his or her profile. These keywords of running conversations can be detected and clustered into topic categories using the common sense database.

If a public brainstorming session occurs and a specific technology is repeatedly mentioned, another employee who is interested and experienced with it could be automatically patched into the conversation.

6. Privacy concerns

Continually recording, transcribing, and archiving all conversations within an organization may seem unreasonable, and if misused, could be potentially dangerous. In an attempt to assuage some of these legitimate concerns, several methods of collecting this data will be discussed.

6.1. Conversation posting

One method of giving users control over their data is to have all the data stored locally on the individual's machine. At the end of the week a topic-spotting algorithm would be used to summarize each conversation, allowing the user to get a list of the week's conversations, the participants involved, and the duration. By each conversation there would be box to check if the conversation is private, public, or should be deleted.

Types of environments where this sort of system would flourish could be places where individuals need to keep careful track of how they spend every minute of their day. For billing purposes, law firms could use such a system to account for their lawyers' time.

6.2. Ten minute delete/mute button

Simply put, the device recording the audio could also be equipped with a button to delete the last 10 min of the recording, or mute the audio for ten minutes into the future. In this way, employees could have a private conversation while at work with a push of the button.

6.3. Demanding environments

In some instances, the environmental demands may supersede privacy concerns. Environments such as these have minimal private conversations, and the needs for all available information is so great, that many of the privacy concerns may not be relevant. Other testing grounds for such a system could be emergency response teams, airport environments, or military applications.

7. Conclusion

We have proposed that active analysis of interactions within the workplace can radically improve the functioning of the organization. There are several different types of information in face-to-face interactions that are measurable by inconspicuous wearable devices: prosidic and body language features, participants identity, context, and conversational content. By aggregating this information, interpreting it in terms of work tasks, and modeling the dynamics of the interactions, we hope to be better able to understand and manage complex organizations.

References

Allen, T., 1997. Architecture and communication among product development engineers. MIT Press, Cambridge, MA, pp. 1–35.

Asavathiratham, C., 2000. The influence model: A tractable representation for the dynamics of networked Markov chains, department of EECS, MIT.

Basu, S., 2002. Conversation scene analysis, Ph.D. thesis in department of EECS. MIT.

Basu, S., Choudhury, T., Clarkson, B., Pentland, A., 2001. Towards measuring human interactions in conversational settings. IEEE Int. Workshop on Cues in Communication (CUES 2001) at CVPR 2001. Kauai, Hawaii.

Choudhury, T., 2003. Sensing and modeling human networks, Ph.D. thesis, department of MAS, MIT.

Choudhury, T., Pentland, A., 2004. Characterizing social network using the sociometer, In: Proc. NAACOS, Pittsbugh, PA, June 17–19. Kluwer.

Choudhury, T., Basu, S., Clarkson, B., Pentland, A., 2003. Learning communities: Connectivity and dynamics of interactive agents, IJCNN, special session on autonomous mental development, IEEE Press October, pp. 2797–2802.

Clarkson, B., Pentland, A., 2000. Framing through peripheral perception, international Conference on Image Processing (ICIP-2000), September 10–13, Vancouver, BC.

Clarkson, B., Sawhney, N., Pentland, A., 1998. Auditory context awareness via wearable computing, In: Proc. of the Perceptual User Interfaces Workshop, San Francisco, CA. September.

Clarkson, B., 2002. Life Patterns: Structure from Wearable Sensors. Ph.D. thesis in MAS, MIT.

DeVaul, R., Sung, M., Gips, J., Pentland, A., 2003. MIThril 2003: Applications and architecture, Proc. ISWC '03, White Plains, NY, October, pp. 4–11.

Eagle, N., Singh, P., Pentland, A., 2003. Common Sense Conversations: Understanding Casual Conversation using a Common Sense Database, Mobile Computing Workshop, IJCAI, Acapulco, MX. August.

Foner, L., 1996. A Multi-Agent System for Matchmaking. In First Intl. Conf. Practical Application of Intelligent Agents and Multi-Agent Tech, Practical Application Company.

Jebara, T., Ivanov, Y., Rahimi, A., Pentland, A., 2000. Tracking conversational context for machine mediation of human discourse. In AAAI Fall 2000 Symposium—-Socially Intelligent Agents—The Human in the Loop. November.

Lieberman, H., and Liu, H., 2002. Adaptive linking between text and photos using common sense reasoning, Conference

on Adaptive Hypermedia and Adaptive Web Systems, Malaga, Spain, May.
Madan, A., Caneel, R., Pentland, A., 2004. GroupMedia: Using wearable devices to understand social context, MIT Media Laboratory Technical Report 582.
Oliver, N., Rosario, B., Pentland, A., 2000. A Bayesian computer vision system for modeling human interaction. IEEE Trans. Pattern Anal. Machine Intell. 22 (8), 831–843.
Singh, P., 2002. Open Mind Common Sense Database: http://commonsense.media.mit.edu/.