Thin Slices of Interaction: Predicting Users' Task Difficulty within 60 sec.

João Pedro Ferreira
engageLab
University of Minho, Portugal
jpferreira@engagelab.org

Marta Noronha e Sousa
engageLab / Dep. Com. Sciences
University of Minho, Portugal
msousa@engagelab.org

Nuno Branco
School of Technology and Management of Felgueiras / engageLab
University of Minho, Portugal
nuno@engagelab.org

Manuel João Ferreira
engageLab / Dep. Industrial Elec.
University of Minho, Portugal

Nuno Otero
Center for Learn. and Know. Tech.
Linnaeus University, Sweden; and
University of Minho, Portugal

Nelson Zagalo
engageLab / Dep. Com. Sciences
University of Minho, Portugal

Pedro Branco
engageLab / Dep. Inf. Systems
University of Minho, Portugal
pbranco@dsi.uminho.pt

CHI'12, May 5–10, 2012, Austin, Texas, USA.

Abstract
We report on an exploratory study where the first 60 seconds of the video recording of a user interaction are used to predict the user's experienced task difficulty. This approach builds on previous work on "thin slices" of human-human behavior, and applies it to human-computer interaction. In the scenario of interacting with a photocopy machine, automated video coding showed that the *Activity* and *Emphasis* predicted 46.6% of the variance of task difficulty. This result closely follows reported results on predicting negotiation outcomes from conversational dynamics using similar variables on the speech signal.

Keywords
Nonverbal Behavior; Social Signals; Thin Slices; Video Coding

ACM Classification Keywords
H.5.0 Information interfaces and presentation: General.

General Terms
Human Factors, Experimentation.

Introduction
Machines are increasingly present in our everyday lives. Even the simplest tasks, such as paying for groceries,

buying a train ticket, or paying for the car parking, may involve dealing with technological devices, often without anyone's help. However, not all of us feel equally comfortable when dealing with machines, and common machines are still not smart enough to deal with our doubts and inadequacies, at our personal pace and respecting our own likes and dislikes.

The present work is part of a project that aims precisely at improving the interaction between humans and public space utility machines. The overall goal is to learn, through a set of observational studies, which social signals could express the user's level of experience, the quality of the interaction and any interaction incident. By social signals, we mean signals that are the expression of a person's attitude towards social interactions, conveyed through a variety of nonverbal behaviors and cues [21]. We believe that the ability to detect these social cues could then lead to systems that are better designed to assess the quality of the interaction and provide more effective responses.

The study presented here follows the methodology of analyzing thin slices of behavioral data. This methodology has been shown to predict a broad range of interaction outcomes [1, 3]. We are relying on the user's "social signaling" towards the machine, conveyed through movement, to infer the user's experienced difficulty towards the task.

Social Signals Processing

Social Signal Processing (SSP) refers to the analysis of human nonlinguistic behavior (e.g., body language, facial expressions, and tone of voice) to make inferences on social relations and roles, to predict the behavioral outcomes of a particular social situation, and to reveal attitudes and relevant social information. The term was coined to denominate the body of seminal work presented by Pentland and colleagues [22] on the study and analysis of social signals.

Though pervasively present in our everyday lives, social signals work in somewhat complex ways. Most social nonverbal signaling is processed at an unintentional and unconscious level, and yet, it is extremely effective [7, 12]. Humans seem to be "hardwired" both to read other people's expressive behaviors (decoding), and to naturally express them (encoding) [16]. Even with minimal amounts of information, we are able to make rather accurate judgments [1].

In one example of that work [3], it is demonstrated that four metrics derived from the conversational dynamics occurring within the first five minutes of a two-party, simulated employment negotiation, predict the outcomes of that negotiation. Activity level, conversational engagement, prosodic emphasis, and mirroring predicted 30% of the variance in/of individual outcomes.

Cues from Nonverbal Language in HCI

The first applications for automatically monitoring users' nonverbal signals within HCI emerged from the need to closely and frequently check the operators' alertness, attention, and cognitive load in critical applications, such as air traffic control or military applications, for a review see [20].

The pioneering work of Picard [17] that lead to the establishment of the Affective Computing field launched a new era of interest in user's emotional aspects. Many projects were developed to infer the user's emotions,

levels of well-being, attention, interest or confusion while interacting with a computer system from physiological data, like respiration, heart rate, skin conductance, and muscle activity [e.g. 8, 18]. Other researchers have turned their focus onto visible and audible nonverbal signals such as facial expressions and vocal quality [for a review see 14]. In the past few years, other implicit modalities are being more frequently used, such as body movement, gestures and posture [6, 9, 10, 11, 21].

Besides the choice of modalities, most HCI studies use those cues to infer the users' emotions or affective states (hence, *affective* computing). The standard method is to borrow predefined models from psychology to understand and organize the expressive data collected. Many of the existing applications use Ekman's basic emotions model [4]. However, "pure", clean-cut emotions are seldom experienced by machine users; they often display mixed or confusing emotions, which makes strict affective categories hard to apply to natural interaction contexts [2, 5, 17]. Furthermore, while interacting with a machine, the user also experiences other rather important mental states [5, 9], such as attention, cognitive processing, interest, engagement, confusion, boredom, frustration, etc. These are not proper emotions, but cognitive states, not appropriately tackled by the typical affective model approach. Some projects, however, have lately been designed to approach such high level mental states [9, 11, 21].

The main difficulty is that most studies are still typically conducted on a laboratory, well controlled environment, mostly monitoring the user for relatively short periods of time [17, 22]. In a natural scenario, though, contextual variables may influence both the interaction and the meaning of the displayed cues, and no existing device is yet able to collect contextual information [22: 1062]. Systems based on physiologic measures require physical contact with the user, limiting its applicability. Time of day, tiredness or even coffee consumption can also bias the user's physiologic responses [19].

Nonverbal behavior analysis systems are working increasingly well when the users are in a fixed position, but, in situations where they are able to move around freely, it is still problematic to robustly track the data [14]. Moreover, the fact that there are contextual, cultural and individual differences in the way emotions and attitudes are exhibited [4, 7] introduces further challenges.

A New Approach

An interesting approach to the problem is presented by the SSP domain, originally concerned with human-to-human interaction. This framework functions as an alternative to the affective approach: the question is no longer to infer the subjects' emotions, but their attitude towards social interaction, in a context-sensitive fashion [22: 1062]. These attitudes are not inferred from isolated nonverbal or physiologic signals, observed in only one of the interactants, but from signals that refer to what one interactant is doing in relation to the other [22]. How one interlocutor positions him/herself from the other, action/response patterns between the two parts, or the amplitude and frequency of prosodic and gestural activities are good examples of social signals. This type of observable behavior is both even less conscious and more stable, and thus reliable, than nonverbal cues, because it is influenced by universal

biologic, rather than cultural or individual, determinations [16].

With this framework in mind, Pentland and colleagues [e.g. 3, 15] have coded four measures of social signaling (activity, engagement, emphasis and mirroring) to make quite accurate predictions of the outcomes of human interaction. The premise behind that approach is that humans are generally able to accurately predict interaction outcomes from the observation of just a *thin slice* (brief moment) of expressive behavior [1].

In this project we mean to call these findings onto a HCI context, while trying to avoid the limitations of previous similar studies. In the first place, our work aims to be usable outside the laboratory. Since it is not practicable to have users wear sensing devices, we relied on the video recording of the users' behavior. Secondly, our approach was not to model the user emotional phenomenon, but more the "sense of difficulty" by capturing signals that would illustrate the quality of the experience. Therefore we use variables such as the ones suggested in [3] to infer on the experienced task difficulty as self-reported by the user.

Study design

In the experiment we are reporting, participants were asked to perform three tasks on a photocopier, while being recorded on video. Each task had a distinct level of difficulty: make a single page copy (easy), make a front and back copy (intermediate), and make a front and back copy with two pages per side (difficult). The order of the tasks was assigned randomly to each participant. Participants had different degrees of experience in using photocopiers, ranging from seldom using any photocopying machine to using this particular model several times a day. Half of the participants had used this photocopier machine (or a similar one) before.

Before each task, participants were instructed on what they were expected to do and filled a form indicating the expected level of difficulty on a 5-point Likert scale ranging from 1 (easy) to 5 (difficult). They would then approach the photocopier to execute the task. Upon completion, the participant would return to the seat and indicate the experienced level of difficulty on an identical scale. In the results reported we are just analyzing this last variable, the difficulty level indicated after performing the task.

A total of 24 participants took part in this experiment. On average, each task took 3m:14s with a standard deviation of 3m:36s. The shortest lasted 18 seconds and the longest lasted 12m:51s.

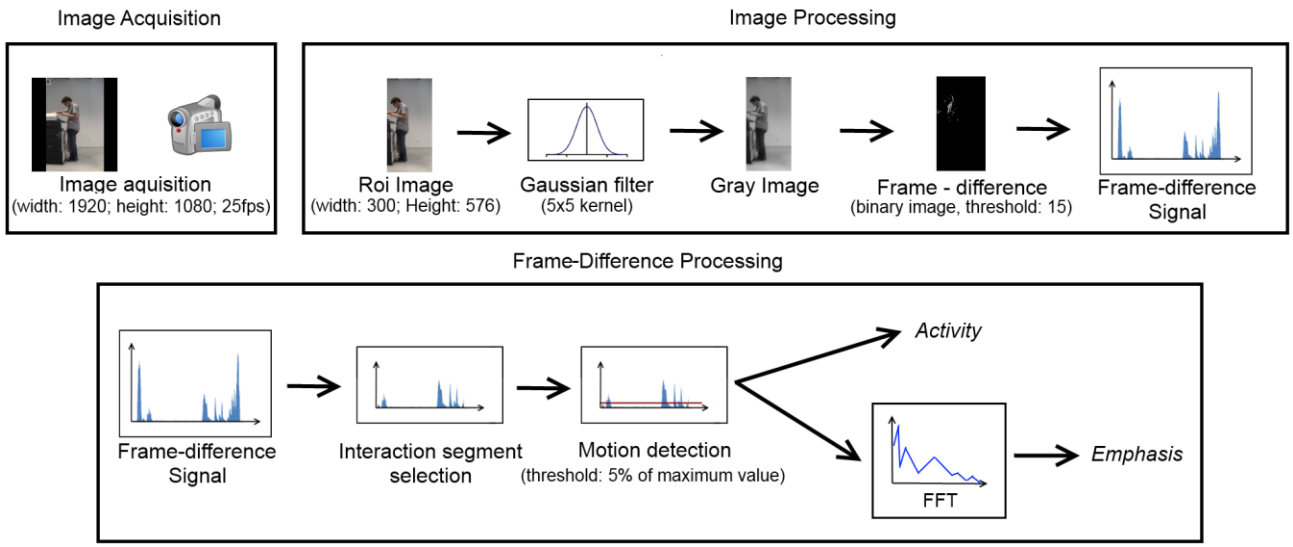

figure 1. Image acquisition, image processing and frame-difference processing schematic.

Video Processing

The interaction task was recorded with 3 cameras capturing different angles, a general view, a face view and a profile view. In this study only the profile view recordings were used. These were recorded at a 1920x1080 image size at 25 frames per second.

The authors decided to use the first 15, 30 and 60 second time slices of the video for the analysis. In this study we are reporting the results for the first 60 seconds of video.

The image processing phase (figure 1) starts with the selection of the image's region of interest, corresponding approximately to the user location (Roi Image). To remove video noise, a low pass filter (Gaussian filter) is applied to the recording and the image converted to grayscale (Gray Image). The difference between consecutive frames is then used to compute the movement on the video. From this *frame-difference* signal, the amplitude and the frequency of the motion can be identified.

For each video of the task, we remove the volunteers' entrance in the scene by detecting a maximum peak in the frame-difference signal. If the video is shorter than the time window used (60s) we also remove the exit from the scene in a similar manner. We are left with a time interval corresponding to the users' interaction segment.

From this interaction interval we computed two measures that we will introduce next.

Variables extracted

Our observations from initial trials suggested that body movement might be one of the most telling social signals in the present interaction context, namely the amplitude and pace of movement and posture changes [also suggested by 10, 21].

Based on the four measures of the speech signal presented on [3]: activity, engagement, mirroring, and emphasis, and following the proposed computational model of social signaling that those same four measures can be applied to video data [15], we selected activity and emphasis to analyze movement from a video signal. Mirroring and engagement are hardly applicable in this context, since they depend on the presence of a human interlocutor.

In [3] activity is the fraction of time a person is speaking and is known to be correlated with interest levels and extraversion (for a review see [3]). In the current study, *Activity* is defined as the fraction of time the volunteer is moving, and measured through the frame-difference signal.

Emphasis represents "jerky, unevenly accented and paced" behavior, as described in [16: 4], and is associated with emotionality and stress. This measure on the speech signal is measured in [3] by variation in speech prosody – pitch and volume. In our experiment, emphasis means that the user displays an uneven rhythm of movements, either moving slowly, with low amplitude gestures, or even stopping, and then suddenly increasing the pace and gesturing more amply. Low emphasis (consistency), on the other hand, is observed when, either presenting low or high activity levels, the user maintains a steady motor behavior.

Computed Variables

Activity: this variable is calculated as the fraction between the number of motion frames and the number of total frames of interaction time. Motion frames are considered to be those where frame-difference is greater than a threshold value, defined as 5% of the maximum movement for all tasks.

Hypothesis 1: *Activity* is correlated with experienced difficulty.

Emphasis: A fast Fourier transform was applied to the frame-difference signal of the motion segments to compute the frequencies' weighted standard deviations and the signal's energy standard deviation. The *Emphasis* is the sum of these two measures. In other words, *Emphasis* measures the variation of motion's energy and frequency.

Hypothesis 2: *Emphasis* is correlated with experienced difficulty.

Results

We recorded 24 volunteers, each performing three tasks with three different levels of difficulty, totaling 72 videos. Two volunteers were excluded since the instructions were not followed correctly. A single recording of another volunteer was also dismissed for the same reasons. Another volunteer's recordings were dismissed due to a camera failure during the session. In total, 62 video recordings were considered.

Table 1 indicates the correlations among all variables. The low level of interdependence between *Activity* and *Emphasis* variables suggests these variables are measuring different features of the signal (r_s = *.187*, *n.s.*).

The results of Pearson[1] correlation tests between all variables are presented in table 1. The correlation between experienced difficulty (*M = 3.08*, *SD = 1.61*), *Activity* (*M = .358*, *SD = .216*), and *Emphasis (M = 107.01*, *SD = 57.79*) was tested.

table 1. Pearson correlations among variables.

	Variables	1	2	3
1	experienced difficulty	-	-.384**	.627***
2	*Activity*		-	-.184
3	*Emphasis*			-

NOTE: ** *p<.01*. *** *p<.001*. (All two-tailed tested)

Hypothesis 1 is confirmed, since *Activity* is negatively correlated with the difficulty level of the task (r_s = -.381, *p < .01*). *Activity* levels decrease as experienced difficulty increases.

Hypothesis 2 is also confirmed as *Emphasis* is positively correlated with the experienced difficulty (r_s = .646, *p < .001*).

[1] We considered the distance between different levels of experienced difficulty (a Likert scale of five points) to be well defined. The experienced difficulty is therefore used as an interval variable.

Multiple regression standardized coefficients (β) are presented in table 2. This model takes *Activity* and *Emphasis* to justify the experienced difficulty.

table 2. Multiple regression standardized coefficients (β).

Variables	experienced difficulty
Activity	-.277**
Emphasis	.574***
R^2	.466

NOTE: ***p<.01*. *** *p<.001*. (All two-tailed tested)

Comparing both standardized coefficients, *Activity* is less important than *Emphasis* to justify the experienced difficulty. These two variables justify 46.6% of the variance of task difficulty.

In a summary, motion tends to be lower (*Activity*) and more irregular (*Emphasis*) with the increase in task difficulty.

Discussion

Previous studies on time slice analysis of behavior have consistently revealed the predictive power of the signaling in respect to the outcomes of social interactions. Those studies look at the phenomenon of social interaction and the signals that emerge in that context. In [3] it is presented a set of variables computed from the interlocutor's voice to predict the outcome of negotiation discussions. The work that we present uses two of those same variables computed from the movement of the user, *Activity* and *Emphasis* to explore the hypotheses that they could predict the user difficulty with a computer task.

Analyzing the user movement to predict the task difficulty was to our knowledge never approached before. The effect size of those two variables .683 is slightly higher than the same level of magnitude reported in the work by [3].

Something should be said about the nature of the task. The interaction with the photocopier was chosen because it allows a good view of the body movement of the user, while simulating a situation where the machine is used by a wide range of people with different skill levels. The sort of movements required as part of the normal task execution is finger movement, to interact with the touch panel, and placing and removing paper from the paper tray. Overall body movements are not an integral part of the task and therefore its presence or absence might very well serve as signaling as the results suggest. For tasks of different nature one should reflect on the context of the movements within the task and which might make sense, or not, to analyze.

From the 62 analyzed videos there are 19 where the time for completion was less than 60 seconds. For those we should not talk about "time-slice" since the task was all contained within the 60 seconds. We decided not to remove those videos from the analysis since the vast majority corresponded to easy or intermediate tasks and removing them would unbalance the number of tasks for each difficulty level. In any case the overall goal of using brief segments of time to infer on the task difficulty remains valid for those videos even if we cannot technically call them "time-slices".

One shortcoming of the study is the lack of validation of the variables extracted from the video. The variables were automatically calculated by the computer from the video signal and there is no attempt to attest if the *Activity* and *Emphasis* correspond to the intended movement dynamics. By construction we did try to minimize perturbations to the users' movement signal that could directly result from the increase in task difficulty. The movements strictly required for completing any of the tasks regardless of its difficulty was the same: place a page on the photocopier, interact with the touch panel and remove the paper. Still, it could be argued that users that experienced higher difficulty with the task would need more trials and therefore the variables that we are measuring are just an effect of the number of trials, for instance the movement of placing and removing paper from the photocopier. In fact just 3 of the 62 analyzed videos contained more than one trial within the 60 seconds, for the rest there was only one trial within that time. The region of interest from video chosen to compute the user movement is centered around the user and it remains fixed during the video. The parts of the machine that could be moved as part of the interaction do not overlap, or they have a minimum overlap over that rectangle. The exceptions are the paper drawers that are not necessary for a successful execution of the task but some users still did use it; in any case those represented 5 videos out of the 62.

More advanced video processing techniques that better segment the user's body and differentiate the different body parts could help improve variable measurement and the model's robustness.

Also, self-reported measures of experienced task difficulty might profit from a more objective validation. That goal could be achieved by comparing the classifications given by the users with less subjective data, such as task duration, number of attempts and success or failure in completion.

Conclusion

The methodology of applying social signals derived from body movement to the study of human-computer interaction is a relatively new and unexplored approach. Other studies have considered motion or posture to infer user states and engagement in computing systems and game applications, but none, to our knowledge, has focused on the quality of the HCI.

The results of analyzing 60 second time intervals follow previous results on thin slices of behavioral data, shown to predict a broad range of interaction outcomes. Specifically, this study suggests higher levels of task difficulty can origin changes in motion amplitude and frequency: Motion tends to be lower (*Activity*) and more irregular (*Emphasis*).

The results here discussed, though preliminary, suggest that video-based sensing systems could be developed that are capable of inferring the users' task difficulty from a thin time-slice of the interaction. The recent appearance of commercially available 3D range cameras that are capable of tracking the user body in real-time indicates that the application of those results in generic interactive systems could be possible in a not so distance future. Questions are then raised if and how those systems could be design to respond to that social signaling.

Acknowledgements

The project is funded by FEDER (Fundo Europeu de Desenvolvimento Regional) through COMPETE (Programa Operacional Factores de Competitividade) and by National funds through FCT (Fundação para a Ciência e a Tecnologia) in the context of the project PTDC/EIA-EIA/098634/2008.

References

[1] Ambady, N., and Rosenthal, R. Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: a meta-analysis. *Psychological Bulletin* III, 2 (1992), 256-274.

[2] Castellano, G., Aylett, R., Paiva, A., and McOwan P. Affect Recognition for Interactive Companions. *Proc. ICMI'08* (2008).

[3] Curhan, J., and Pentland, A. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first five minutes. *Journal of Applied Psychology* 92, 3 (2007), 802–811.

[4] Ekman, P. Universals and cultural differences in facial expressions of emotion. Cole, J. ed. *Nebraska symposium on motivation, 1971*, 19, Lincoln University of Nebraska Press, Lincoln, USA, 1972, 207-283.

[5] Gunes, H., and Pantic, M. Automatic, Dimensional and Continuous Emotion Recognition, *International Journal of Synthetic Emotions* 1, 1 (2010), 68-99.

[6] Gunes, H. Piccardi, M., and Jan, T. Face and Body Gesture Analysis for Multimodal HCI. *Computer Science* 3101 (2004), 583-58.

[7] Hall, E. T. *The Hidden Dimension / A Dimensão Oculta*. Relógio d'Água, Lisbon, PT, 1966.

[8] Hazelett, R. L. Measurement of User Frustration: A biologic approach. *Ext. Abstracts CHI 2003*, ACM Press (2003), 734–735.

[9] el Kaliouby, R., and Robinson, P. Real-time inference of complex mental states from facial

expressions and head gestures. *Proc. CVPRW '04*, IEEE (2005), 181-200.

[10] Kapoor A., and Picard, R. W. Multimodal affect recognition in learning environments. *Proc 13th annual ACM international conference on Multimedia*, ACM (2005), 677-682.

[11] Kapoor, A., Burleson, W., and Picard, R. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65 (2007), 724–736.

[12] Knapp, M., and Hall, J. *Non-Verbal Communication in Human Interaction*. Thompson Wadsworth, Toronto, CA, 2006.

[13] Mehrabian, A. *Nonverbal Communication*. Aldine Transaction, New Brunswick, USA, 1972.

[14] Pantic, M. Affective Computing, 2005 http://pubs.doc.ic.ac.uk/Pantic-Encyclopedia05-2/Pantic-Encyclopedia05-2.pdf.

[15] Pentland, A. A Computational Model of Social Signaling. *Proc. ICPR'06*, IEEE (2006).

[16] Pentland, A. *Honest Signals. How they shape our world*. MIT Press, Cambridge, 2008.

[17] Picard, R. *Affective Computing*. MIT Press, Cambridge, 1997.

[18] Picard, R.W., Vyzas, E., and Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10 (2001), 1175-1191.

[19] Quinlan, P., Lane, J., Aspinall, L. Effects of hot tea, coffee and water ingestion on physiological responses and mood: the role of caffeine, water and beverage type. *Psychopharmacology* 134, 2 (1997), 164-173.

[20] Rowe, D., Sibert, J., and Irwin, D., Heart Rate Variability: Indicator of User State as an Aid to Human-Computer Interaction. *Proc. CHI 1998*, ACM Press (1998), 480–487.

[21] Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. Automatic analysis of affective postures and body motion to detect engagement with a game companion. *Proc. 6th Intern. Conf. on Human-Robot Interaction*, (2011), 305-312.

[22] Vinciarelli, A., Pantic, M., Bourlard, H., and Pentland, A. Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain. *MM'08*, ACM Press (2008), 1061-1070.