Predicting users' task difficulty using Social Signals: a Preliminary Model

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

Humans communicate social intentions through patterns of nonverbal language, using posture, gestures and body motion. This social signalling is present in human to human interaction as well as in human-computer interaction. Our daily dependence on computers emphasizes the need and importance for good interaction quality. While humans have an innate ability to recognize and respond to social signalling, machines don't. Our work aims to develop a Social Signal Processing model based on features extracted using simple video processing techniques, applied in a real context and running in real-time, to predict interaction's difficulties and problems. In this study we report a preliminary model where features extracted from user motion within 60 seconds of video recordings can predict 46,6% of variance in task difficulty.

Nonverbal Behaviour; Social Signals; Thin Slices; Video Coding; Video Processing.

1. 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 behaviours and cues (Sanghvi et al., 2011). 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.

Based on Curhan and Pentland (2007) methodology, the present work presents the results of the preliminary model that used features extracted from motion, Emphasis and Activity, to infer the user's experienced difficulty towards the task on a photocopier.

2. OUR APPROACH

Our observations from initial trials (N. Branco et al., 2011) 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 Kapoor and Picard (2005) and Sanghvi et al. (2011).

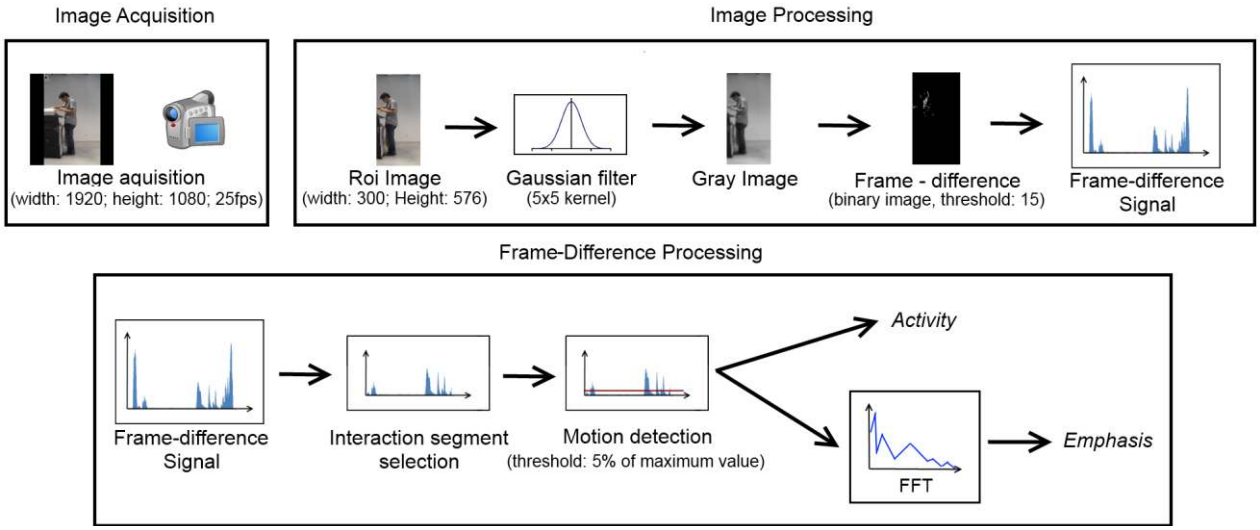

Figure 1. *Image acquisition, image processing and frame-difference processing schematic.*

Based on the four measures of the speech signal presented on Curhan & Pentland (2007): *activity*, *engagement*, *mirroring*, and *emphasis*, and following the proposed computational model of social signalling that those same four measures can be applied to video data (A. Pentland, 2006), we selected activity and emphasis to analyse 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 Curhan and Pentland (2007) *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 Curhan and Pentland (2007)). 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" behaviour, as described in Pentland (2008), and is associated with emotionality and stress. This measure on the speech signal is measured in Curhan and Pentland (2007) 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 behaviour.

3. 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 analysing 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.

4. 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.

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.

5. 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.

6. RESULTS

We recorded 24 volunteers, each performing three tasks with three different levels of difficulty, totalling 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 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 ($r_s = .646$, $p < .001$).
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 coefficient.*

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.

8. CONCLUSIONS

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 analysing 60 second time intervals follow previous results on thin slices of behavioural 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 signalling.

9. REFERENCES

Branco, N., Ferreira, J. P., Noronha e Sousa, M., Branco, P., Otero, N., Zagalo, N., & Ferreira, M. J. (2011). Blink: observing thin slices of behavior to determine users' expectation towards task difficulty. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (p. 2299). New York, New York, USA: ACM Press.

Curhan, J. R., & Pentland, A. (2007). Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, *92*(3), 802-811.

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 677-682). New York, NY, USA: ACM.

Pentland, A. (2006). A Computational Model of Social Signalin. *18th International Conference on Pattern Recognition (2006)*, *1*(c), 1080-1083. IEEE.

Pentland, Alex. (2008). *Honest Signals: How They Shape Our World*. *Technology Review* (Vol. 133, p. 205). The MIT Press.

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. *Proceedings of the 6th international conference on Human-robot interaction* (pp. 305-312). New York, NY, USA: ACM.