Affective and Pleasurable Design, Vol. 71, 2023, 1–10
https://doi.org/10.54941/ahfe1003238

Quantifying Interest From Facial Images, and the Role of Video in Effective Business Calls

Ryosuke Katsuki[1], Keiichi Watanuki[2,3], Suguru Mashima[2], and Yusuke Osawa[2]

[1]IgnitusAI, Citus LLP, Shubuya-ku, Tokyo 150-0043, Japan
[2]Graduate School of Science amd Engineering, Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338-8570, Japan
[3]Advanced Institute of Innovative Technology, Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338-8570, Japan

ABSTRACT

It is said, to succeed in business, put the interest of customers ahead of your own. This is great advice to businesses, racing to deploy video calls to provide COVID-safe meetings and/or benefit from general conveniences and gain productivity. If 55% of the communication is truly nonverbal as suggested by Prof. Albert Mehrabian from University of California in Los Angeles, putting the interest of the customer ahead means listening carefully to the nonverbal channels. In video calls, one must try hardest to read, understand and manage the customer interest expressed in the video images. In this study, our primary objective was to analyse and evaluate the practicality of a project to develop a machine-learning model that can predict and quantify a level of interest of a video caller from his facial images. As such, our secondary objective was to explore facial and behavioural expressions, highly correlated to or triggered by interests that are confirmed by imaged human subjects themselves, as well as explore areas and methods to capture them as to generate machine-learning training data. The result suggested the project to develop a machine-learning model would be practical. The finding included that a model based on static facial features visible in a video frame could be possible, but a model based on moving facial features estimated from sequence of consecutive video frames could do better, especially those with acute focus on eye movements.

Keywords: Video call, Video conference, Online meetings, Remote work, Interest, Nonverbal communication, Machine learning, Facial images, Facial expressions, Facial emotions, Eye, Blink, Saccade

INTRODUCTION

The one of the must haves in a business relationship is a common interest. Hence, understanding and responding to customer interests have almost always been fundamental steps to start a business relationship. At least a slightest interest or a trigger of a kind is almost a requirement to start a business conversation. That is why we often hear, to succeed in business, put the interest of customers ahead of your own.

This is great advice to businesses, racing to deploy video calls to provide COVID-safe meetings and/or benefit from general conveniences and gain productivity. If 55% of the communication is truly nonverbal as suggested by Prof. Albert Mehrabian from University of California in Los Angeles, putting the interest of the customer ahead means listening carefully to the nonverbal channels. In video calls, one must try hardest to read, understand and manage the customer interest expressed in the video images.

As the world returns to normal from COVID 19 and lockdowns, businesses can ride on the current accelerated rate of the video call adoption, seek not only to expand pandemic-safe meeting capabilities for business continuity, but also to exploit the videos to start and build relationships with customers at a different level and more confidently on the basis that there exist foundational interests.

OBJECTIVE

Our primary objective was to analyse and evaluate the practicality of a project to develop a machine-learning model that can predict and quantify a level of interest of a video caller from his facial images.

Our feasibility theory of such a machine-learning model was based on the existence of successful Facial Emotion Recognition, which may be just a few steps short of predicting and quantifying a level of interest of a person in the image. Our instinct was that some facial emotions would accompany the presence of some interest, and conversely no facial emotion in its absence. Thus, our theory was that the probability of neutral state predicted by Facial Emotion Recognition could also be just a few steps short of a proxy to the probability of the absence of interest, and one minus the probability of neutral state of a proxy to the probability of the presence of interest.

As such, our secondary objective was to explore facial and behavioural expressions, highly correlated to or triggered by interests that are confirmed by imaged human subjects themselves, as well as explore areas and methods to capture them as to generate machine-learning training data.

For the purpose of this study, interest was defined as any interest including deep interest, defined as long lasting ones strongly associated with personal affinity, and shallow interest defined as short lived interest more related to the present circumstances, including a quick glance to an annoyance and an interest of a kind nonetheless.

METHODOLOGY

To test our instincts and theories, as well as to collect facial behaviours associated with interests, we ran the following laboratory experiment. During the span of May 19, 2022 and June 3, 2022, 30 human subjects, randomly picked from the pool of volunteers and undergraduate students from the Saitama University campus, participated in the experiment as shown in Fig. 1. Of 30 subjects, 17 were male and 13 female. The laboratory was set up with a monitor of 24 inches, 1920×1080 pixels, and 93 DPI resolution, replaying the stimulus video clips. Attached to the same monitor was a Logitech

Figure 1: Laboratory setup.

C270n web camera, recording participating human subjects' facial images. Each of them were asked to sit on a desk chair, facing the monitor at his eye level, simulating a typical ergonomic environment in which one would take a business video call.

Each subject participated sequentially, one at a time. Each subject was first asked to participate in training runs on the interest level feedback application on an iPad device, followed by steps to calibrate the biometric measurement devices. Then, the subject was asked to watch randomly sequenced twelve 30-second stimulus video clips and 30-second rest screen (See Figure 2) in between the clips, and simultaneously provide interest level feedback via the application on the iPad device (realtime interest level). As stimulus video clips played, the web camera attached to the monitor recorded the human subject's facial images. At the end of the series of the clips, the subject was asked to rate, from 1 (lowest) to 7 (highest), the level of deep interest and the level of shallow interest each clip triggered (deep interest level and shallow interest level).

The web camera captured facial images were processed through our naive machine-learning model, predicting the subject's interest level communicated nonverbally thought facial expressions. Our naive model was based on the Face Expression Recognition feature of justadudewhohacks/face-api.js. The output value of 0 represented the probability that the emotional state is absolutely neutral, and the value of 1 represented the probability of absolute absence of the neutral emotional state. Intuitively, the spectrum also represented a rough proxy to the level of interest, 0 representing a full absence of interest and 1 a full presence of interest.

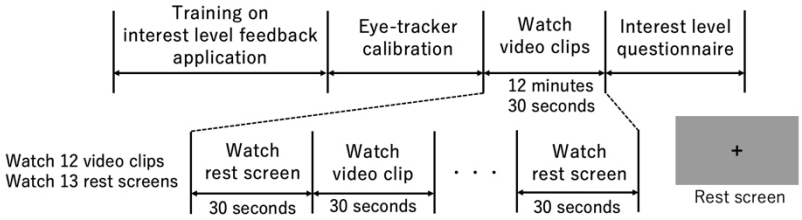

Figure 2: Experiment sequence per subject.

Also, for the purpose of further explorations and potential triangulations, human subjects' biometrics data was collected as they watched the clips. Eye behaviours were tracked by eye tracking system Tobbi and electrocardiogram, heart rate and precipitation by physiological signals acquisition system NeXus-10 MK II.

The iPad application is used to record the realtime interest level. The iPad application was designed to render a 355 pixel (9.39 centimetre) radius circle on the touch screen, representing an empty template of Russell's Circular Model of Emotions, where the horizontal centre line represented the pleasant-unpleasant axis and the vertical centre line the arousal-non-arousal axis, and developed to track and record the location of the subject's index finger placed on it. The location was recorded continuously at 10 Hz and in terms of x and y coordinates, each ranging between from −177.5 to 177.5, and 0 and 0 being the centre of the circle as shown in Fig. 3.

Operation training were provided to each subject before the actual experiment. Subjects were instructed to place their index fingers at the centre at the start, as well as during the rest between the clips, and when the interest level is absolute zero while watching any of the stimulus video clips. They were instructed, otherwise, to move their index fingers, without lifting it off of the pad, away from the circle with rising interest, such that the distance between the centre and x/y coordinate of the finger represented the interest level. In terms of directions, they were instructed to move their index fingers towards the top-right when interest is associated with excitement, bottom-left with depression, top-left with stressfulness, and bottom-right with calmness. Sealed in the center was a tiny nail-size tape, rough on the surface, so that the subjects could find the center easily without taking his eyes off of the stimulus video clips.

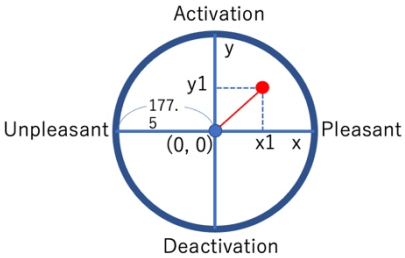

Figure 3: The application interface and Russell's circular model of emotions.

THE NAIVE ML MODEL PERFORMANCE

The laboratory experiment collected data about 360 cases (30 human subjects X 12 stimulus video clips). The min-max feature scaled naive model prediction average in each case was compared to the respective min-max feature scaled average realtime interest level and min-max feature scaled deep and shallow interest levels from the questionnaire.

To analyse how closely the naïve ML model tracked the actual interest level submitted by the subject, realtime interest levels, deep interest levels

and shallow interest levels were plotted against the respective naive model predictions. but no correlation was obvious. However, in plotting only the substantiated relative interests and substantiated relative disinterests, the correlation was observed.

The substantiated relative interests were cases in which the subject's realtime interest level, deep interest level and shallow interest level were all above 0.5. Conversely, substantiated relative disinterests were cases in which the subject's realtime interest level, deep interest level and shallow interest level were all below 0.5.

The model performed remarkably predicting the substantiated relative disinterest as shown by the orange dots being concentrated at the bottom half (See Figure 4). This supports that subjects tend to show little to no facial emotion that are visible in web camera images when they feel little to no interest. This also supports that a machine-learning facial image model for predicting low to no interest is potentially feasible with more research, in particular in reducing the false positive.

Where the correlation is short is drawn by the blue dots inside the red circle in the graph above (See Figure 4). These represent cases where substantiated relative interest were accompanied with a low amount of facial emotion. In these cases, the naive model predicted low to no interest but the subject substantiated his interest. This suggested facial emotions do not always accompany high interests, but something else might. Perhaps there are behaviours that are not being picked up by the naive model, which estimates interest level per static video image or video frame, independent of the previous or following frame, ignoring movements.

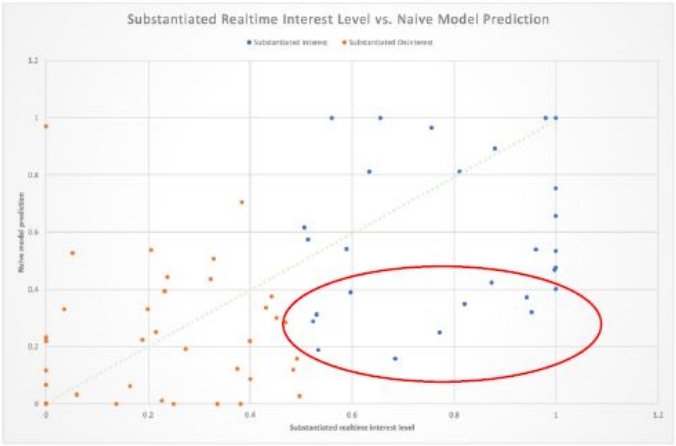

Figure 4: Substantiated realtime interest level vs Naïve model prediction.

AREAS FOR IMPROVEMENTS

Our exploration and analysis with Tobii data surfaced promising two means to improve on the naïve model performance, and especially to address the very shortcoming noted in the previous section. One was to collect movement

patterns and build the prediction model based on them. The other was to focus more on the eyes where small and quick movements are potentially more telling. These represented precisely what the naïve model was not designed to do.

The subject responses to two clips that invited the most realtime interest level changes, No. 8 honda optical illusion and No. 9 pink graphic art, and one clip that invited the least realtime interest level changes, No. 6 leaky faucet, were analyzed. Out of many, two explorations surfaced repeating patterns, correlating with strong interest levels.

For one of the exploration, each of the above clips and for each of the subjects for which Tobii successfully collected data, a chart was created, plotting the amount of line of sight changes and realtime interest levels on the y axis against clip replay seconds elapsed on the x axis as shown in Figs. 5a-5c. Z-scores was computed for amount of line of sight changes and realtime interest levels, so to remove the influence of individual differences between subjects and to bring both into the same range for easier comparison (Press et al, 1989).

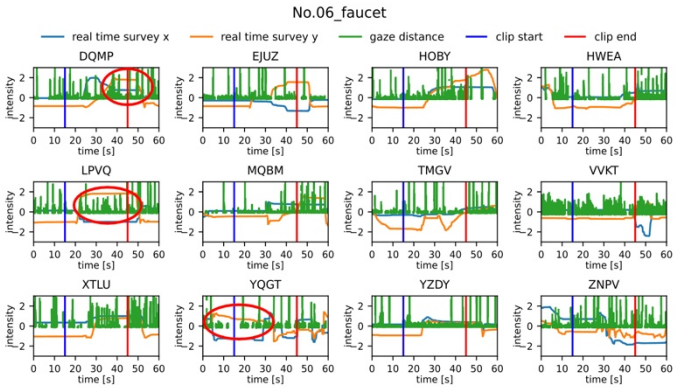

Figure 5a: Changes in gaze distances and realtime interest level over time (No 6).

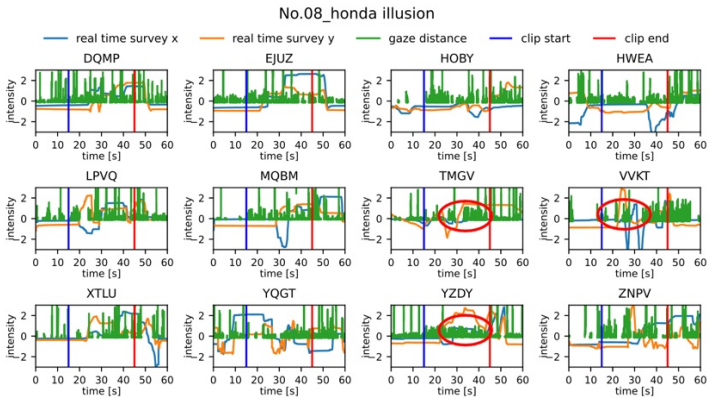

Figure 5b: Changes in gaze distances and realtime interest level over time (No 8).

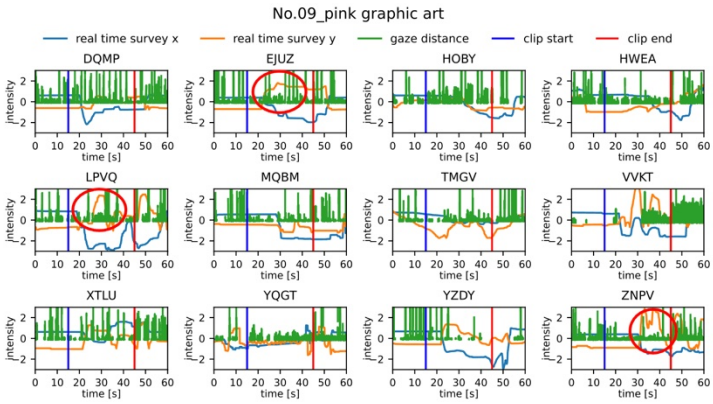

Figure 5c: Changes in gaze distances and realtime interest level over time (No 9).

It was found that the y component of realtime interest level tends to be large when Tobii detected a small amount of line of sight movements, or sign that the subject is excited and gazing, and in effect potentially having fixated and potentially with neutral expressions, suggesting reasons that the naive model could not pick up these interests.

This particular pattern can be interpreted as a tendency for one to gaze when aroused. Preferential gazes are known to occur when a person's gaze is directed more toward the desirable item compared to the less desirable one (Fantz, 1961). Moreover, in a separate study in which two T-shirt images were displayed and people were asked which T-shirt they wanted to buy, 80% of the people picked the T-shirt they gazed at for a longer time (Tagawa et al., 2014). It is also said that an initial fixation can become a longer one because one would typically not only pay attention to the point of interest, but also to obtain information from their interest.

These studies also support that the fixation occurs with interests, and longer one with increasing interests. The gaze and gaze time are related to interest, and are powerful facial features with which to construct a model for estimating the degree of interest.

In relation to the naive model prediction performance analysis, two cases, among others, highlighted the modeling improvement opportunities. Both subject TMGV and subject YZDY expressed their strong interest in the stimulus video clip the no 8 optical illusion in all three accounts – realtime, deep and shallow interest. However, the naive model that did not digest blinks or any movements predicted rather unconvincing interest levels, 0.4773 and 0.5004 respectively.

Similarly, charts were created, plotting the blink rates and realtime interest levels on the y axis against clip replay seconds elapsed on the x axis as shown in Figs. 6a–6c. Z-scores was computed for the Y axis values for the same reason. Each failure to detect a pupil was defined as one blink, and the number of blinks per second was defined as the blink rate (Hershman et al., 2018).

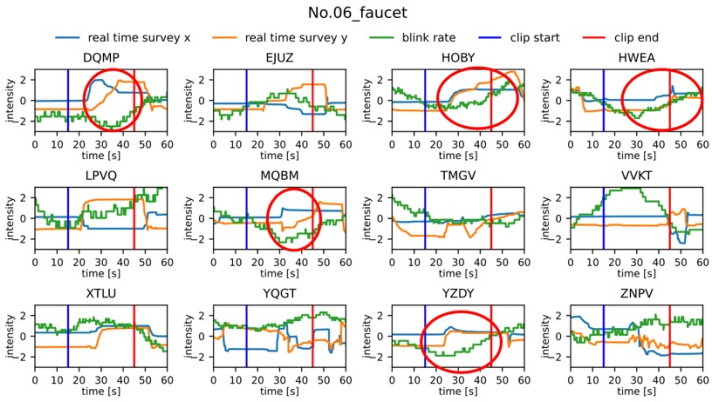

Figure 6a: Changes in blink rate and realtime interest level over time (No 6).

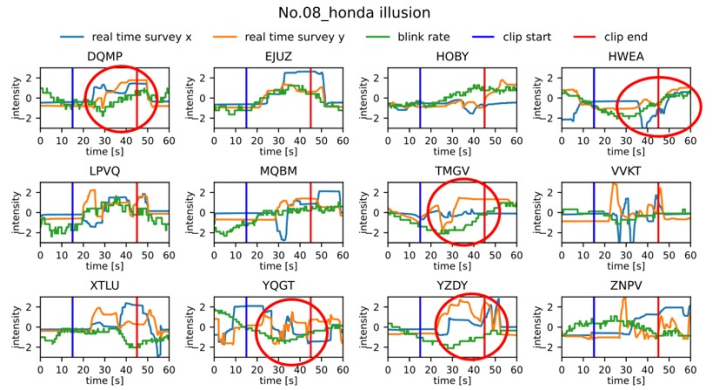

Figure 6b: Changes in blink rate and realtime interest level over time (No 8).

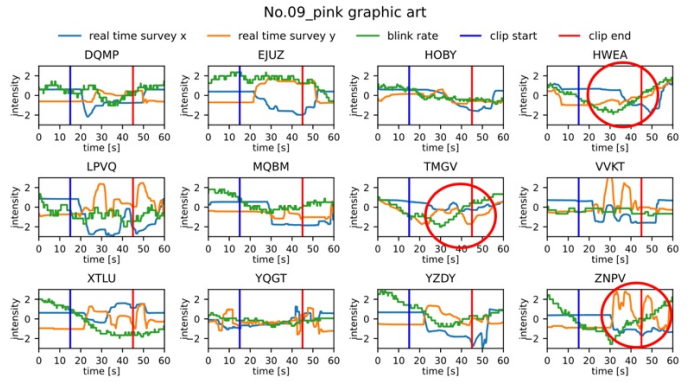

Figure 6c: Changes in blink rate and realtime interest level over time (No 9).

In the 2002 study "Evaluating Emotions from Blinking" showed that interest decreases blinking rate (Yamada, 2002). While this was observed, it was also apparent that a low and increasing blink rate accompanied increasing realtime interest level.

Blinks are classified into spontaneous blinks, voluntary blinks, and reflex blinks. Spontaneous blinks are blinks performed unconsciously, voluntary blinks are blinks performed consciously, and reflex blinks are blinks triggered by an external stimulus (Suyama et al., 2014). In this experiment, the percentage of reflective blinks is considered to be small, so it is thought that the percentages of voluntary blinks. It has been reported (Tada et al., 1990) that the blink rate decreases as the arousal level decreases. As such, it is thought that when there is no interest, the viewer feels that the video is boring, hence the blink rate decreases. However, it has been reported that the blink rate is higher when watching boring and unpleasant TV programs than when watching high-interest TV (Hideoki, 1986). When the blink rate increased, the x component (pleasure-discomfort) of the realtime interest level did not decrease. In addition, the reason that the blink rate increased from the second half of the video is that interest began in the first half of the video, the arousal level increased, the blink rate increased, and then the subjects reflected their interests in the realtime interest levels though the application on the device.

These previous studies clarified the observed relationship between the blink rate, the arousal state and interest, confirming that the blink rate is an important facial feature with which to construct a model to estimate interest level.

In relation to the naive model prediction performance analysis, the same two cases, among others, highlighted the modeling improvement opportunities. Both subject TMGV and subject YZDY expressed their strong interest in the stimulus video clip the no 8 optical illusion in all three accounts – realtime, deep and shallow interest. However, the naive model that did not digest blinks or any movements predicted rather unconvincing interest levels, 0.4773 and 0.5004 respectively.

CONCLUSION

The result suggested the project to develop a machine-learning model would be practical. The finding included that a model based on static facial features visible in a video frame could be possible, but a model based on moving facial features estimated from sequences of consecutive video frames could do better, especially those with acute focus on eye movements, such as blinking and gazing.

A more accurate interest predicting model, a realtime Interest Index, on video calls is a game changer, especially for teams orchestrating medium to large scale programs. The Interest Index can act as a proxy to video call performance. It does not measure outcomes such as conversion, sales or revenue, which are important but can be too short sighted as means to improve communications and relationships with customers. Instead, the index would measure customer interests, thereby measuring where one is in fundamental steps of communication and business. Team supervisors and call agents can use the Interest Index to gauge how well the current video calls are being conducted in managing customer interests, then restructure the approach, including team resources, workflow, assignments, scripts,

and/or slide, less likely compromising the customer focus fundamental steps of communication, and measure the effectiveness of the restructure.

The obvious extension to the advertising sales team study is to use the data to attempt to find interest patterns that are highly correlated to sales conversions, for example an average interest above certain level, a prolonged interest above certain level, an increasing interest level between phases, or some repeating simple to complex interest patterns relating to sales conversions. This will lead to better prediction of conversion. The next step is to analyze those high conversion patterns, and identify repeatable triggers, paths, and steps that one can learn to enact and drive the same high conversion patterns.

On the individual level, putting the interest of customers ahead means becoming tactically prepared. The Interest Index can be made available in realtime as one talks to another on a video call. It can potentially help agents be more adroit when unexpected interest observed, and more prudent when interest seem absent, and especially so with customers whose expressions are harder to read with the naked eyes, whether that's a technical or setting issues, cultural differences, or accessibility challenged situations with elderly, or autism or other mental illness patients cases.

REFERENCES

Fantz, R. L., The Origin of form perception, Scientific American, Vol. 204 (1961), pp. 66–72.

Hershman, R., Henik, A. and Cohen, N., "A novel blink detection methodbased on pupillometry noise," Behav. Res. Methods, Vol. 50, No. 1 (2018), pp. 107–114.

Hideoki, T., "Eyeblink rates as a function of the interest value ofvideo stimuli", Tohoku Psychologica Folia, Vol. 45, (1986), pp. 107–113.

Press, W. H., and George, B. R., "Fast algorithm for spectral analysisof unevenly sampled data", The Astrophysical Journal, Vol. 338 (1989), pp. 277–280.

Suyama, M., Kato, T., Takano, H., Nakamura, K., Development of Discrimination Method Between Voluntary and Spontaneous Eye Blinks, Proc. of FIT 2014, Vol. 13, No. 2 (2014), pp. 377–378. (in Japanese)

Tada, H., Yamada, F., and Hariu, T., "Changes of eye-blink activities during hypnotic state. Perceptual and motor skills", Vol. 71, No. 3(1990), pp. 832–834.

Tagawa, R., Kato, T., Sudo, K., Taniguchi, Y., Estimation of Dominant Factors in Shopping Based on Gaze Time Detection, IPSJ SIG Technical Report, Vol. 2014-EC-31 (2014), pp. 1–4. (in Japanese)

Yamada, F., "Evaluating Emotion by Blinking: Emotional Evaluation by Startle Blink Reflex and Spontaneous Blinking". Psychological Review, *45*(1), (2002), pp. 20–32.