Comparison of Video-Based Pointing and Selection Techniques for Hands-Free Text Entry

Yulia Gizatdinova
Research Group for Emotions, Sociality and Computing, TAUCHI
University of Tampere, Finland
+358 (0)3 3551 4030
yulia.gizatdinova@uta.fi

Oleg Špakov
Visual Interaction Research Group, TAUCHI
University of Tampere, Finland
+358 (0)3 3551 8556
oleg.spakov@uta.fi

Veikko Surakka
Research Group for Emotions, Sociality and Computing, TAUCHI
University of Tampere, Finland
+358 (0)40 5573265
veikko.surakka@uta.fi

ABSTRACT

Video-based human-computer interaction has received increasing interest over the years. However, earlier research has been mainly focusing on technical characteristics of different methods rather than on user performance and experiences in using computer vision technology. This study aims to investigate performance characteristics of novice users and their subjective experiences in typing text with several video-based pointing and selection techniques. In Experiment 1, eye tracking and head tracking were applied for the task of pointing at the keys of a virtual keyboard. The results showed that gaze pointing was significantly faster but also more erroneous technique as compared with head pointing. Self-reported subjective ratings revealed that it was generally better, faster, more pleasant and efficient to type using gaze pointing than head pointing. In Experiment 2, *mouth open* and *brows up* facial gestures were utilized for confirming the selection of a given character. The results showed that text entry speed was approximately the same for both selection techniques, while mouth interaction caused significantly fewer errors than brow interaction. Subjective ratings did not reveal any significant differences between the techniques. Possibilities for design improvements are discussed.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Evaluation/methodology, prototyping, input devices and strategies*; I.2.10 [**Artificial Intelligence**] Vision and Scene Understanding – *Video analysis.*

General Terms

Human Factors, Design, Experimentation, Measurement, Performance, Reliability, Verification.

Keywords

Video-based interaction, text entry, virtual keyboard, eye tracking, computer vision, face detection, visual gesture.

1. INTRODUCTION

In video-based human-computer interaction, computer vision is applied to remotely detect visual cues from a user via camera, interpret and use them to control various applications or smart environments [15,19]. Hands-free text entry has been and still remains one of the important application areas of visual interfaces. Generally, hands-free text entry involves pointing at and selection of the graphical elements of an on-screen spelling application or virtual keyboard. For a long time eye tracking has been the primary method used for this purpose. Eye tracking technology utilizes cameras and infrared illumination to track eye/gaze and interprets it as a computer pointer.

Dwell time protocol has been intensively used in eye typing as a selection method which is based on placing a computer pointer over a graphical element for about 0.4-1 s in order to activate it. In typing with a virtual keyboard, a theoretical text entry limit of 22 words per minute (wpm, 1 word = 5 characters including spaces and punctuation marks) was suggested in [9], assuming 500 ms dwell time and 40 ms average saccade duration. In practice, dwell-based eye typing speed for novice users is not higher than 8-10 wpm (with somewhat longer dwell times), but experienced users may set shorter dwell times and reach the typing speed of 20 wpm [9].

Recently, head/face or facial feature tracking has been applied for pointing at the objects in a graphical user interface (aka camera mouse) [1,4,13,14]. Such systems utilize off-the-shelve hardware components and, therefore, do not require external equipment other than a PC and a webcam. They also offer non-contact, un-calibrated and self-initialized interaction. Non-intrusiveness is a highly desirable feature especially for disabled users (potentially the largest target user group for this technology) [1,11]. Many camera mouse applications for text entry utilize dwell time protocol as a selection method. Betke et al. [1] applied normalized correlation template tracking of eye, nose or lip in a spelling board application. The reported text entry speed was ~6 wpm with a dwell time of 0.5 s. Hansen et al. [6] reported similar results (~5,5 wpm) for typing with a dwell-based spelling application with letter/word prediction using a marker-based head pointing.

Despite the popularity of using dwell time protocol in visual interfaces, it has been shown [16] that dwell times less than a half of a second may be perceived by the users as too short to interact comfortably. For example, short dwell times may result in an unintentional selection of interactive elements when a user,

AVI '12, May 21-25, 2012, Capri Island, Italy.

for example, is investigating an interface. Text entry speed tends to increase with practice. To adjust for personal preferences in typing speed adaptive dwell times have been recently proposed [16,10].

However, dwell time protocol has another limitation as it can only substitute for one command. Several alternative methods such as directional gestures [21] and some indirect methods like graphical toolbars have been applied to execute and switch between different interface options. Recent eye/gaze-based dwell-time-free systems show potential (with advanced users) reaching about 11 wpm [20]. However, despite of a long search for accurate, fast and convenient selection/activation methods, dwell time protocol remains the most used one in text entry applications [9].

There are more than 40 muscles in the face which alone or in combinations produce visually detectable changes in facial appearance. These facial expressions or gestures can be automatically detected by computer vision methods and potentially replace or add to the functionality of standard computer devices and support the development of 'native' visual interfaces. Grauman et al. [5] utilized voluntary blinks and brow raises, both detected by motion analysis and normalized correlation template matching, as selection gestures in order to eliminate the use of dwell time and to emulate a single click functionality of a computer mouse. The interface was tested in a letter-scanning spelling application that required two selections to enter a single character. The reported typing speed (selection-only) was 1 wpm. De Silva et al., [2] applied template matching for head (nose) tracking and a hybrid approach to detect *mouth open* gesture. Although the interface supports both pointing and selection functionality it was tested in a point-only spelling application and the reported typing speed for two participants was 7.6 wpm.

The literature analysis revealed that the focus of research in this area has been mainly on a technological side, dealing with performance characteristics of various computer vision methods like their speed, robustness and accuracy. Computer vision-based interfaces for hands-free text entry are still rare and insufficiently studied, therefore their applicability and limitations for this task are not yet well understood. The proposed point- and select-only solutions have not yet utilized computer vision capabilities to their full potential. Especially little attention has been given to the investigation of efficiency and user satisfaction in using such systems [11].

We believe that apart from selection of reliable computer vision methods for visual processing, a proper selection of facial gestures to control a spelling application is important. However, this research question has received only a little attention in the past. Intuitively, a gesture set should satisfy a number of requirements and be (1) natural and emotionally neutral; (2) non-fatiguing in intensive and continuous use, (3) comfortable so that it does not entail muscle strain or pain, (4) detectable by computer vision methods and (5) unlikely to occur accidentally or unintentionally.

The present study is an ongoing work that aims at developing visual interface for hands-free text entry. The paper presents an empirical evaluation of the text entry performance of novice users and their subjective ratings while using several video-based pointing and selection techniques. The general aim was to find the pros and cons of the whole system and suggest future design improvements.

2. EXPERIMENT 1

The first experiment aimed to compare two pointing techniques, namely, gaze pointing and head pointing and investigate how they affect the performance and subjective experiences of novice users in text entry task.

2.1 Methods

2.1.1 Participants

Fifteen unpaid native Finnish speaking students of a local university (11 males, 4 females) between 20 and 35 years of age (M=25.1, SD=4.8) participated in the experiment. The participants had normal or corrected to normal vision; three participants had eye glasses. The participants were novice users regarding the tested techniques as they had no prior experience in typing with a virtual keyboard, gaze tracking or computer vision-based interfaces. All participants were highly experienced computer users and regularly used physical QWERTY keyboard.

2.1.2 Design

The participants typed text using eye or head movement for pointing at the keys of a virtual keyboard, and a spacebar of a physical keyboard for key selection. It was expected that the chosen manual method of key selection is the most robust and will help to eliminate errors and time required by other selection techniques: methods which are unknown to participants (e.g. dwell time protocol) could potentially affect the results in unpredictable way. This helped in measuring a 'pure' impact of pointing techniques on the users' text entry performance.

The experiment was a within subject 2 (pointing technique: gaze pointing and face pointing) × 3 (keyboard size: small, medium and large) repeated measures factorial design. The experiment consisted of two sessions (i.e. gaze pointing and head pointing) three blocks each (with small, medium and big keyboard). Each participant typed 5 random phrases (~150 characters) in native language within each block. The experiment was counterbalanced regarding a pointing method and a keyboard size. The total number of phrases typed was *15 participants × 2 pointing techniques × 3 keyboard sizes × 5 trials = 450.*

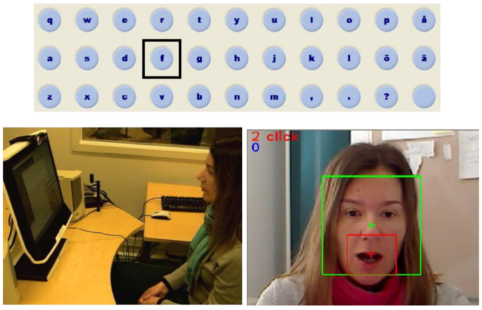

Figure 1: Top: The layout of a virtual keyboard. A black bounding box indicates the point-sensitive area of a key. Bottom-left: Positioning of the user in front of the camera. Bottom-right: Visual output of the face processing software.

2.1.3 Apparatus

The experiment was performed in controlled laboratory conditions. The distance from a participant's face to a monitor was about 50 cm. Hardware specifications included computer (Intel Core 2 quad, 2.66 GHz, 3 GB RAM), Tobii T60 eye tracker (60 Hz sampling rate), monitor (17", 1280x1024 pixel resolution) and web camera (Logitech Webcam Pro 9000, 320x240 pixel resolution, 25 frames per second capture rate).

The keys of 3×11 keyboard layout were visually presented as circles separated by a gap of 20 pixels; however, the point-sensitive areas of the keys were squares without any gap in between. A black bounding box in Figure 1 indicates the point-sensitive area of a key. The bounding box was not visible to the participants during the experiment. Small keyboard (825×225 pixels) was characterized by the keys of 75×75 pixels, medium keyboard (1045×285 pixels) had the keys of 95×95 pixels and finally big keyboard (1265×345 pixels) had the keys of 115×115 pixel size. The smallest key size was suggested by the earlier studies [12] and the biggest key size was limited by a size of the monitor. Finnish target phrases were taken from a corpus of approximately 500 phrases [18]. The QWERTY layout of keys (29 letters of Finnish alphabet plus comma, period, question mark and space) was used to minimize learning effects associated with memorizing positions of the keys in a new layout. A computer pointer was displayed as a dark blue circle of 10×10 pixels. Activation (or selection) of a key resulted in a 'click' sound, providing audio feedback to the participants during the experiment.

In gaze pointing tasks, the (*x,y*) coordinates of the computer pointer were calculated as averages of the left and right eye coordinates returned by the eye tracker. The accuracy of the eye tracker was about 0.5°-1° (~16-33 pixels on a 1280x1024 or 96 dpi monitor viewed at a distance of 50 cm), assuming nearly perfect calibration. Our camera mouse was supported by continuous head (face) detection from a video stream [3]. The camera mouse had pointing accuracy of approximately 5-10 pixels, assuming favorable illumination conditions. A moving average was applied to the five most recent head locations to remove a jitter from the detection output. In head pointing tasks, the image with detected head outline and its center of mass (green markings in the bottom-left image of Figure 1) helped the participants to adjust their position in front of the camera and, therefore, facilitate head detection.

User experiences were captured by seven bipolar rating scales: *general evaluation*, *pleasantness*, *dominance*, *quickness*, *accuracy*, *efficiency* and *difficulty*. The rating scales had nine points varying from –4 to +4 (the bigger the number the more positive is the rating, e.g. 4 in a difficulty scale means easy). The rating scales used in this experiment were inspired by [17]. A scale of *dominance*, not originally included in the earlier study, was added as it was recognized as an important metric for this particular experiment

2.1.4 Procedure

The gaze pointing session always started by calibrating the eye tracker until an acceptable level of calibration quality was recognized. The participants were instructed not to move their heads much during the session since large head movements have a negative effect on accuracy of the eye tracker. The head pointing session always started with adjusting the camera's tilt-and-zoom settings to achieve approximately same position and size of a face in the image for each participant.

Each session then continued with practicing in using the technology by typing the word "hello" and explanation of the task. The participants were instructed to memorize an appearing phrase before they start typing text, so that they would not spend time later looking at it again. However, verification of the target phrase as well as own typed text was allowed. The participants were instructed to type as fast and as accurate as they can. Corrections were not allowed (there was no backspace key). To continue to the next target phrase, the participants had to press the key "right" on the physical keyboard. The participants had a chance to rest between blocks of a session. The participants provided feedback regarding their experiences at the end of each session. The experiment, including self-reporting, lasted about 40-60 minutes. A history of actions produced by the participants during the experiment was recorded into a system log file including the target phrase, timestamp, action type (hover, leave or select) and its parameters (letter or action assigned to a key).

2.2 Results

A two-way ANOVA was used for data analysis with pointing technique and keyboard size as independent variables. The trials were treated as repetitions. The following dependent variables were analyzed: (1) *text entry speed* is a time required to type a target phrase in words per minute (wpm), (2) *relative error rate* calculated using Levenshtein string distance algorithm [7] as a ratio of erroneous or missed characters to a total number of characters in a phrase, and (3) *self-reported subjective ratings of the user experiences.* The data was first inspected for outliers and Grubb's test revealed 5 blocks as extreme outliers (i.e. they had average error rate larger than three standard deviations from the mean value across trials of all participants for a given keyboard size and a given pointing technique) which were excluded from the analysis.

2.2.1 Text entry speed

The effect of the keyboard size on the text entry speed is shown in Figure 2 for each pointing technique. In this figure and the following graphs the error bars show plus minus one standard deviation from the mean values. The ground mean of the text entry speed (averaged over all keyboard sizes) was *M=10.98*

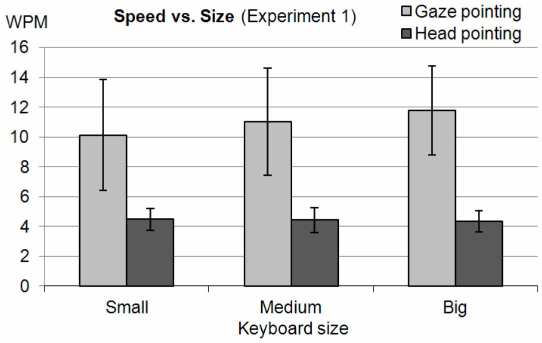

Figure 2: Mean text entry speed in words per minute.

wpm (*SD=0.39*) in case of gaze pointing and *M=4.42* wpm (*SD=0.06*) in case of head pointing. A two-way ANOVA showed a significant main effect of the pointing technique ($F_{1,79}=152.255$, $p<0.001$) indicating that gaze pointing resulted in significantly faster text entry speed than head pointing. There were no other significant main or interaction effects, although in case of gaze pointing the mean text entry speed gradually increased with increase of the keyboard size.

2.2.2 *Error rate*

The error analysis revealed that in average gaze pointing produced more errors than head pointing. The ground mean of the relative error rate (averaged over all keyboard sizes) was *M=8%* (*SD=1.77*) in case of gaze pointing and *M=3.8%* (*SD=0.3*) in case of head pointing. A two-way ANOVA revealed that there were significant differences in the mean values of the relative error rate between the pointing techniques ($F_{1,79}=16.512$, $p<0.001$). There were no other significant main or interaction effects, although the mean value of the relative error rate gradually decreased with the increase of the keyboard size when gaze pointing was used. Figure 3 shows a summarized effect of the keyboard size on the relative error rate for gaze and head pointing techniques.

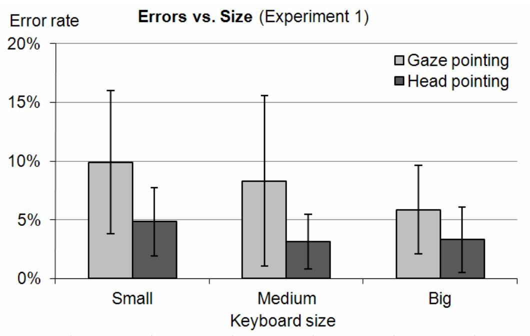

Figure 3: Mean relative error rate.

2.2.3 *Self-reported subjective ratings*

Figure 4 demonstrates the results of subjective evaluation of the pointing techniques by the participants. A Wilcoxon signed ranks test was applied for a pairwise comparison of the subjective ratings for different categories. The test indicated that the participants perceived gaze pointing technique as more pleasant, fast, efficient and generally better for the given task (refer to Table 1). All other differences in the ratings were not statistically significant.

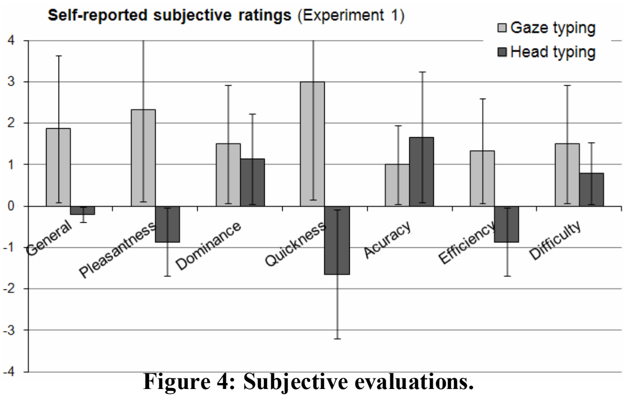

Figure 4: Subjective evaluations.

Table 1. Wilcoxon signed ranks test results with mean *M*, standard deviation *SD*, minimum *Min* and maximum *Max* scores, *Z* and *p* values.

Rating	**Pointin g**	**M**	**SD**	**Min**	**Max**	**Test**
General	gaze	1.87	2.07	-3	4	$Z=-2.180$, $p=0.029$
	head	-0.20	1.97	-3	2	
Pleasantness	gaze	2.33	1.40	-2	4	$Z=-3.085$, $p=0.002$
	head	-0.87	2.20	-4	4	
Dominance	gaze	1.50	1.79	-3	4	$Z=-0.537$, $p=0.592$
	head	1.14	2.18	-3	4	
Quickness	gaze	3.00	0.96	1	4	$Z=-3.306$, $p=0.001$
	head	-1.64	1.91	-4	3	
Accuracy	gaze	1.00	2.42	-4	4	$Z=-0.670$, $p=0.503$
	head	1.67	1.80	-2	4	
Efficiency	gaze	1.33	1.95	-2	3	$Z=-2.409$, $p=0.016$
	head	-0.87	1.92	-3	2	
Difficulty	gaze	1.50	1.95	-2	4	$Z=-0.773$, $p=0.439$
	head	0.79	2.08	-3	4	

3. EXPERIMENT 2

The second experiment aimed to study two selection techniques executed by *mouth open* and *brows up* facial gestures and investigate how they affect the performance and subjective experiences of novice users in text entry task.

3.1 Methods

3.1.1 *Participants*

Thirteen unpaid students of a local university (10 males, 3 females) between 19 and 35 years of age (M=25.6, SD=4.5) participated in the experiment. Eight participants had Finnish as native language and five were non-Finnish speaking students. Two participants had a prior knowledge about the head pointing technique as they participated in Experiment 1. However, they did not have any experience in interacting with computers by visual gestures and were considered as novices for a given task. All participants had normal or corrected to normal vision: two participants had eye glasses. All participants regularly used physical QWERTY keyboard.

3.1.2 *Design*

In this experiment, the participants were typing text by pointing at the keys of a virtual keyboard by head movement and selecting the keys using *mouth open* and *brows up* facial gestures. Head pointing was used in this experiment because it allowed for testing visual gesture detector in a presence of noticeable head movements and rotations produced by the participants while pointing at different keys of the keyboard. Thus, it supported ecologically valid evaluation of both selection techniques as compared to, for example, gaze or mouse (hand) pointing which assumes or requires that the head stays relatively still.

The experiment was a within-subjects single factor repeated measures design. The experiment consisted of two sessions (mouth interaction and brow interaction) with 5 trials in each. The experiment was counterbalanced regarding a selection technique. The total number of target phrases typed was *13 participants × 2 selection techniques × 5 trials = 130.*

3.1.3 Apparatus

The experiment was performed in the same laboratory conditions and with the same equipment as explained in Experiment 1, with exceptions that a virtual keyboard of the middle size was utilized, head tracking [3] was used for pointing and facial gesture detection [3,23] was utilized as a selection technique. The target phrases were taken from Finnish [8] and English [18] corpuses. The system produced a 'click' sound once a facial gesture was detected. The output of the face processing software was visible to the participants.

As it is well known, there is a tradeoff between the event detection rate of the system and the amount of false alarms it produces. More sensitive systems detect nearly all events but possibly produce a large number of false alarms. Less sensitive systems give a fewer number of false alarms at the expense of having some missed events. Different errors of the facial gesture detector will have different effect on the text entry performance. Thus, false alarms would result in unintentional entry of characters, increasing text entry error rate. A miss of a gesture would cause a user to repeat the gesture, resulting in slow typing speed and, possibly, frustration of the user. Based on the earlier findings [3] reported for this technology in detecting *mouth open* and *brows up* gestures, the expected misdetections due to the system are as follows: false alarm rate of 7% and missed event rate of 11.5%.

3.1.4 Procedure

The procedure was similar to the one in Experiment 1. Each session started with training of a classifier for categorization between facial gesture and non-gesture events. For more detailed description of the training procedure as well as technical characteristics of the classifier refer to [3]. Once a classifier has been trained, the participants had a short practice in using the technique by typing their names. In the following experiment, Finnish participants typed Finnish phrases and non-Finnish participants typed English phrases. The system outputted a log file that contained a history of the participants' actions, as explained in Experiment 1. The experiment lasted about 60 minutes.

3.2 Results

A one-way ANOVA was used for data analysis with selection technique as the only independent variable which had two levels (mouth interaction and brow interaction). The trials were treated as repetitions. The same dependent variables were analyzed as in Experiment 1: text entry speed, relative error rate and self-reported subjective ratings of the user experiences. The data for one participant was completely removed from the analysis due to technical reasons. For three participants the brow interaction session was not performed due to the fact that they had difficulties in producing *brows up* gesture long enough (1-2 minutes) to complete the training of a facial classifier. One participant did not have time to perform the brow interaction session of the experiment.

3.2.1 Text entry speed

The ground mean of the text entry speed was *M=3.07* wpm (*SD=0.30*) in case of *mouth open* selection technique and *M=2.85* wpm (*SD=0.43*) in case of *brows up* selection technique. The effect of the selection technique on the text entry speed is shown in Figure 5. A one-way ANOVA did not reveal any statistically significant differences for these results.

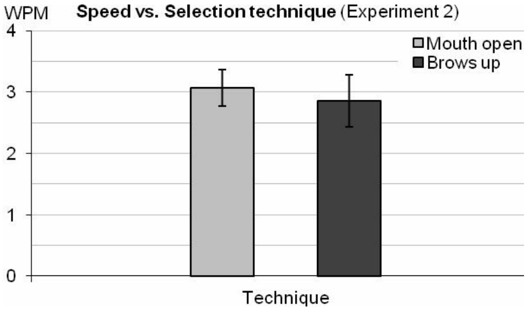

Figure 5: Mean text entry speed in words per minute.

3.2.2 Error rate

The error analysis revealed the ground mean of the relative error rate of *M=6%* (*SD=3*) for mouth interaction and *M=21%* (*SD=15*) for brow interaction. Figure 6 shows the average relative error rate for both selection techniques. A one-way ANOVA showed that there was a significant difference in the mean values of the relative error rate among the selection techniques (*$F_{1,7}$=9.538*, *p<0.05*).

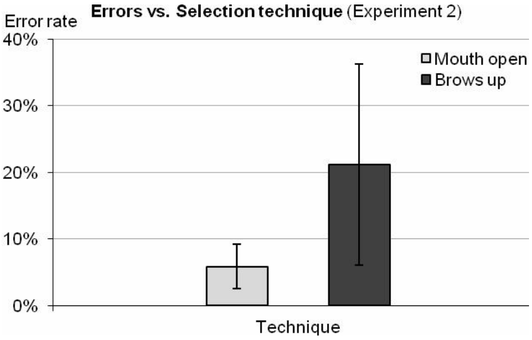

Figure 6: Mean relative error rate.

3.2.3 Self-reported subjective ratings

Figure 7 indicates that there was no agreement among the participants about their subjective experiences regarding the selection techniques. A Wilcoxon signed ranks test was applied for a pairwise comparison of the subjective ratings for different categories (Table 2). There were no significant differences in the evaluation results.

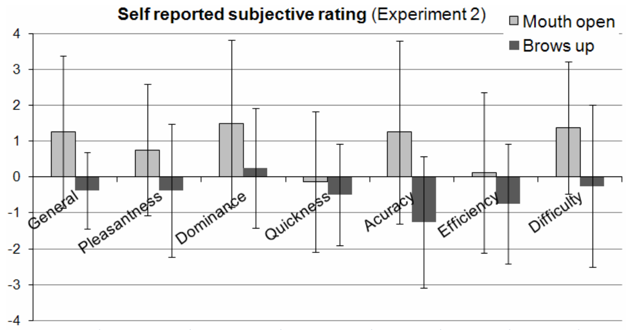

Figure 7: Subjective evaluations.

Table 2. Wilcoxon signed ranks test results with mean *M*, standard deviation *SD*, minimum *Min* and maximum *Max* scores, *Z* and *p* values.

Rating	Selection	M	SD	Min	Max	Test
General	mouth	1.25	2. 12	-2	4	*Z* = -1.763,
	brows	-0.38	1.06	-2	1	*p* = 0.078
Pleasantness	mouth	0.75	1.83	-2	3	*Z* = -0.938,
	brows	-0.38	1.85	-2	3	*p* = 0.348
Dominance	mouth	1.5	2.33	-2	4	*Z* = -0.954,
	brows	0.25	1.67	-2	3	*p* = 0.340
Quickness	mouth	-0.13	1.96	-3	3	*Z* = -0.791,
	brows	-0.50	1.41	-2	1	*p* = 0.429
Accuracy	mouth	1.25	2.55	-3	4	*Z* = -1.706,
	brows	-1.25	1.83	-3	2	*p* = 0.088
Efficiency	mouth	0.13	2.23	-3	3	*Z* = -0.987,
	brows	-0.75	1.67	-3	2	*p* = 0.323
Difficulty	mouth	1.38	1.85	-2	3	*Z* = -1.411,
	brows	-0.25	2.25	-3	4	*p* = 0.158

4. DISCUSSION

4.1 Head vs. Gaze Pointing

The gaze pointing technique resulted in significantly faster text entry speed than the head pointing technique. This was an expected result due to the fact that eye movements are typically much faster than head movement. However, being faster in pointing, gaze input caused a great variation in typing speed between the participants: the fastest typists were almost 3 times faster than the slowest, while head pointing caused more stable typing speed among the participants. Further, gaze pointing resulted in about doubled error rate comparing to head pointing. This fact confirmed the earlier results [6] which found gaze as a very fast but erroneous means of pointing than head or hand (mouse) input in text entry applications.

The dependency of the text entry speed and relative error rate on the keyboard size in case of gaze pointing seems to be higher than in case of head pointing. Although it was not recognized as statistically significant result in our experiment, this tendency may become stronger if size of the keys will be further decreased. For example, previous studies [12] demonstrated that if a dimension of the object becomes smaller than 50 pixels in diameter, interaction by gaze becomes virtually impossible. Imperfect calibration of an eye tracker or calibration drift may significantly increase this limit.

In gaze pointing sessions with small keyboard of Experiment 1, in some trials the typed phrases were almost impossible to understand, for example, one participant typed "*mi,di kysyg hsdsjjs kysymy,diä*" instead of "*miksi kysyt hassuja kysymyksiä?*" and another typed " *junst lbst sins myöhäsdä?*" instead of "*junat ovat aina myöhässä.*". This was not the case for typing with head pointing. We assume that with further decrease in the keyboard size, gaze pointing will become difficult or even impossible while head pointing will not undergo significant changes. However, this is to be confirmed in the future experiments.

The obtained results allowed us to conclude that head pointing, causing noticeably slower text typing, has advantages over gaze pointing in accuracy, stability and lesser dependency on the size of the keys. The latter is especially beneficial for developing small-sized typing applications which occupy only a small part of the screen real estate and allow the user a better possibility in observing the typed text and interacting with interface widgets and other applications. This opposes the current requirement in designing gaze-based typing interfaces, in which key layouts usually take a large part of the screen real estate to compensate for inaccuracy in eye/gaze input. Still, the participants rated text entry with gaze pointing higher than head pointing in many respects, emphasizing a fast interaction speed that can be achieved with gaze input.

4.2 *Mouth open* vs. *Brows up* Selection

The results obtained from Experiment 2 revealed that both selection techniques were approximately equally fast with no significant differences observed. The error analysis further allows making a clear conclusion: *mouth open* selection gesture caused much less errors in text typing than *brows up* selection gesture. In addition, mouth interaction caused more consistent results with small deviation from the mean than brow interaction. Subjective ratings did not reveal any significant differences between the techniques.

A comparison of the results from head pointing sessions of Experiments 1 and 2 showed a decrease in the text entry performance when facial gestures substituted hand input (button press) for key selection. Taking an assumption that button press takes only 0.005 s on average [1] and is not a source of typing errors, it is possible to directly compare the results. Thus, *mouth open* gesture itself decreased typing speed by about 1 wpm and increased the error rate by about 2%. This is considered as a quite low decrease in the text entry performance. Otherwise, the achieved text entry speed is comparable or superior to the results reported for similar computer vision-based systems which use dwell time [6,1] or facial gestures [2,5] as a key activation command.

Anatomically, facial expressions result from contractions and relaxations of different facial muscles. Due to differences in facial muscle control, a range of visual appearances for different facial gestures varies significantly among the participants. Figure 8 shows examples of *brows up* gestures (the pictures

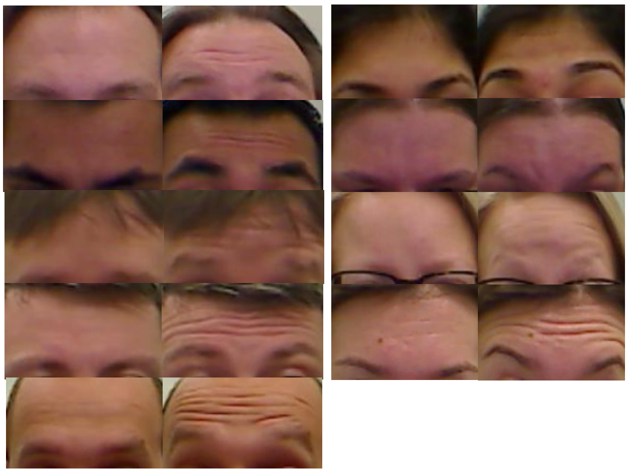

Figure 8: Examples of neutral state and *brow up* gesture in the upper face for male (left) and female (right) individuals.

come from our pilot study [3]). Note that the eye region was analyzed by the classifiers although cut from the pictures in order to hide identities of the participants. It is visible from the figure that in some cases *brows up* gestures may result in slight skin displacements with no wrinkles visible in the forehead region. This might affect the detection performance of facial classifiers for *brows up* gesture in Experiment 2 and explain a large amount of errors for this technique. Whereas *mouth open* gesture resulted in a strong visual pattern that was in average reliably detected by the system for all participants. Additionally, the participants reported that *mouth open* gesture felt more natural and was easier to control than *brows up* gesture.

4.3 General Discussion

The results from Experiment 1 allowed for a direct comparison between effects of gaze and head pointing on the text entry performance of the novice users. The maximum average text entry speed was achieved in gaze typing session: the novices were able to type with the average speed of 11.78 wpm when gaze pointing was combined with the biggest keyboard and hand input (button press) for key selection. The maximum average text entry speed of 4.48 wpm was achieved by the novices when head pointing was combined with the smallest keyboard and hand input for key selection. In Experiment 2 when head pointing was combined with facial gesture selection method, a speed of the text entry decreased. The maximum average speed of 3.07 wpm was achieved by the novices in the interface that combined head pointing with *mouth open* key selection (4-6 wpm are attainable by advanced users).

In average, gaze-based interface caused more errors then computer vision-based interfaces. In the current study no error correction was allowed; however, it would be interesting to consider techniques for correcting errors and their impact on the total text entry speed for gaze-based and computer vision-based interfaces. We consider that in video-based text entry interfaces, it is a desirable feature to switch between input techniques depending on available hardware or task/application at hand. As argued in [6], there are many applications which do not require high precision of text entry due to natural language redundancy. Therefore, gaze pointing can be used for fast text entry, for example, in chatting applications; whereas head pointing (as less erroneous and more accurate technique) can be used in tasks which require high correctness in the typed text. In the future, both techniques could be merged together in a similar manner as hand and gaze pointing were merged in [22]: gaze could be used for quick and rough pointer relocation followed by further adjustment by head.

In overall, the participants reported that it was easy to understand and operate with video-based pointing and selection interfaces. The novices were able to adjust to the system by developing different strategies of key pointing and selecting. For example, some participants preferred to point at the keys by using head movement and rotation, while others kept their faces frontal to the camera and moved the torso instead. The later usually resulted in a better performance of the gesture detectors. Head pointing was generally recognized as more tiring technique than gaze pointing as it required more physical effort from the participants to move a computer pointer to a desired position. A choice of intensity and duration of facial gestures used for key selection was also developed during the experiment. Participants quickly figured out that short but pronounced facial gestures are recognized better by the system.

In video-based interfaces it is important to give sufficient feedback to the user about functioning of the computer vision methods. As observed during our experiments, there were situations when the face detector continuously failed to detect a face (due to strong face rotations or when a participant went outside of the camera's field of view) and still the participants were trying to point at the keys of the keyboard. It would be beneficial for the participants to receive, for example, a sound alarm when the face detector loses a face.

5. CONCLUSIONS AND FUTURE WORK

Because visual interaction is natural for humans and many movements and gestures can be made on a voluntary basis [17], video-based interaction is one promising technology to support hands-free human computer interaction. We presented the first results from our ongoing experimental verification that aimed to compare different video-based pointing and selection techniques to find the best ones for the purpose of hands-free text entry.

Encouraged by the obtained results, we plan to extend this work in several ways. Firstly, the robustness and speed of computer vision methods will be improved. Secondly, the results of the experiments suggested that individual differences and abilities in visual interaction may vary widely between people. Therefore, a range of visual gestures that may satisfy different user preferences will be explored and tested in a series of future studies. Thirdly, we will continue searching for the most appropriate design of a typing application to be utilized with a video-based interface. Finally, a longitudinal study of the performance of a video-based interface in text entry is an evident continuation of this work.

6. ACKNOWLEDGMENTS

The authors thank the Academy of Finland (project 129354) and the University of Tampere.

7. REFERENCES

[1] Betke, M., Gips, J., and Fleming, P. 2002. The Camera Mouse: visual tracking of body features to provide computer access for people with severe disabilities. *IEEE Trans. Neural Systems and Rehabilitation Engineering*. 10, 1 (Mar. 2002), 1-10.

[2] De Silva, G.C., Lyons, M.J., Kawato, S., and Tetsutani, N. 2003. Human factors evaluation of a vision-based facial gesture interface. In *Proceedings of IEEE Computer Vision and Pattern Recognition Workshop* (Madison, Wisconsin, USA, June 16 - 22, 2003). CVPRW'03. IEEE Computer Society, NY, USA, 52-52.

[3] Gizatdinova, Y., Špakov, O., Surakka, V. 2012. Face typing: vision-based perceptual interface for hands-free text entry with a scrollable virtual keyboard. In *Proceedings of the Workshop on the Applications of Computer Vision* (Breckenridge, CO, USA, January 9 - 11, 2012). WACV'12. IEEE, 81-87, in press.

[4] Gorodnichy, D. and Roth, G. 2004. Nouse 'use your nose as a mouse' perceptual vision technology for hands-free games and interfaces. *Image and Vision Computing*. 22, 12 (Oct. 2004), 931-942.

[5] Grauman, K., Betke, M., Lombardi, J., Gips, J., and Bradski, G.R. 2003. Communication via eye blinks and eyebrow raises: video-based human-computer interfaces. *Universal Access in the Information Society*. 2, 4 (Nov. 2003), 2-4.

[6] Hansen, J.P., Tørning, K., Johansen, A.S., Itoh, K., and Aoki, H. 2004. Gaze typing compared with input by head and hand. In *Proceedings of the Symposium on Eye tracking research & applications* (San Antonio, TX, USA, March 22 - 24, 2004). ETRA'04. ACM, NY, USA, 131-138.

[7] Levenshtein, V.I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*. 10, 8 (Feb. 1966), 707-10.

[8] MacKenzie, I.S. and Soukoreff, R. W. 2003. Phrase sets for evaluating text entry techniques. In *Proceedings Extended Abstracts of the Conference on Human Factors in Computing Systems* (Ft. Lauderdale, FL, USA, April 5 - 10, 2003). CHI'03. ACM, NY, 754-755.

[9] Majaranta, P. and Räihä, K-J. 2007. Text entry by gaze: utilizing eye-tracking. In *Text entry systems: Mobility, accessibility, universality*, I.S. MacKenzie and K. Tanaka-Ishii, Eds. San Francisco: Morgan Kaufmann, 175-187.

[10] Majaranta, P., Ahola, U.-K., and Špakov, O. 2009. Fast gaze typing with an adjustable dwell time. In *Proceedings of the Conference on Human Factors in Computing Systems*, (Boston, MA, USA, April 4 - 9, 2009) CHI'09, ACM, NY, USA, 357-360.

[11] Manresa-Yee, C., Ponsa, P., Varona, J., and Perales, F.J. 2010. User experience to improve the usability of a vision-based interface. *Interacting with Computers*. 22, 6 (Nov. 2010), Elsevier, NY, USA, 594-605.

[12] Miniotas, D., Špakov, O., and MacKenzie, I. S. 2004. Eye gaze interaction with expanding targets. In *Proceedings Extended Abstracts of the Conference on Human Factors in Computing Systems* (Vienna, Austria, April 24 - 29, 2004). CHI'04, ACM, NY, 1255-1258.

[13] Morris, T. and Chauhan, V. 2006. Facial feature tracking for cursor control. *J. Network and Computer Applications*. 29, 1 (Jan. 2006), 62-80.

[14] Palleja, T., Rubion, W., Teixido, M., Tresanchez, M., del Viso, A.F., Rebate, C., and Palacin, J. 2009. Using the optical flow to implement a relative virtual mouse controlled by head movements. *J. Universal Computer Science*. 14, 19 (Nov. 2008), 3127-3141.

[15] Porta, M. 2002. Vision-based user interfaces: Methods and applications. *Int. J. Human–Computer Studies*, Elsevier, 57, 1 (Jul. 2002), 27-73.

[16] Špakov, O. and Miniotas, D. 2004. On-line adjustment of dwell time for target selection by gaze. In *Proceedings of Nordic Conference on Human-Computer Interaction* (Tampere, Finland, October 23 - 27, 2004). NordiCHI'04. ACM, NY, USA, 203-206.

[17] Surakka, V., Illi, M., and Isokoski, P. 2004. Gazing and frowning as a new human-computer interaction technique. *ACM Trans. Applied Perception*. 1, 1 (Jul. 2004), ACM, NY, USA, 40-56.

[18] Tuisku, O., Majaranta, P., Isokoski, P., and Räihä, K.-J. 2008. Now Dasher! Dash away!: longitudinal study of fast text entry by Eye Gaze. In *Proceedings of the Symposium on Eye tracking research & applications*, (Savannah, GA, USA, March 26 - 28, 2008). ETRA'08. ACM, NY, 19-26.

[19] Turk, M. and Kölsch, M. 2005. Perceptual interfaces. *Emerging Topics in Computer Vision*. G. Medioni and S. Kang. Eds. Upper Saddle River, NJ: Prentice Hall, 456-520.

[20] Urbina, M.H. and Huckauf, A. 2007. Dwell time free eye typing approaches. In *Proceedings of the Conference on Communication by Gaze Interaction*, (Leicester, UK, September 3 - 4, 2007). COGAIN'07. 65-70.

[21] Wobbrock, J.O., Rubinstein, J., Sawyer, M.W., and Duchowski, A.T. 2008. Longitudinal evaluation of discrete consecutive gaze gestures for text entry. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Savannah, GA, USA, March 26-28, 2008). ETRA'08. ACM, NY, 26 - 28 March 2008. Savannah, Georgia: ACM Press, 11-18.

[22] Zhai, S., Morimoto, C., and Ihde, S. 1999. Manual and gaze input cascaded (MAGIC) pointing. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit (CHI '99)*. ACM, NY, USA, 246-253.

[23] Zhao, G., Huang, X., Gizatdinova, Y., and Pietikäinen, M. 2010. Combining dynamic texture and structural features for speaker identification. In *Multimedia Workshop on Multimedia in Forensics, Security and Intelligence*, (Scottsdale, AZ, USA, Nov. 28 - Dec. 1, 2011). MiFOR'11. ACM, NY, 93-98.