Pattern Analysis and Applications (2019) 22:841–855
https://doi.org/10.1007/s10044-018-0741-0

ORIGINAL ARTICLE

Touchless typing with head movements captured in thermal spectrum

Adam Nowosielski[1] · **Paweł Forczmański**[1]

Received: 20 November 2017 / Accepted: 20 July 2018 / Published online: 30 July 2018

Abstract
Many physically challenged people are unable to operate standard electronic equipment or computer input devices. They need special assistive technologies and one of the options is the head operated interface. Face-oriented algorithms often assume a particular level of lighting with adequate intensity and spatial configuration. In the paper, we propose a thermal-imaging-based algorithm of head operating typing. It does not assume the visible light illumination. We investigated, in context of thermal imagery, several contemporary general purpose object detectors known to be accurate in case of images captured by the visible light camera. Then, a selected face detector is employed in the head operated interface analysing head movements in the thermal spectrum. The attention has been focused on the problem of touchless typing which is performed in the existing solutions either through the camera mouse or through traverse procedure with an addition mechanism (like eye blink or mouth open) needed for clicking events in both cases. Our novel solution for touchless typing with head movements combines the thermal imaging for capturing user action with the hierarchical letter selection procedure. The solution employed allows to reach any alphabet character in just three steps, i.e. with directional head movements, without the need of any additional mechanisms for clicking events.

Keywords Thermovision · Face detection · Human–computer interaction · Touchless typing · Typing with head movements · Virtual keyboard

1 Introduction

Many physically challenged people are unable to operate standard electronic equipment or typical computer input devices. At the same time, new solutions for human–computer interfaces introduce touchless interaction and offer hands-free control using gestures. Unfortunately, not all of these natural user interfaces are applicable for users with motor impairments. People with physical handicap need special assistive technologies designed particularly for them. The head operated interface constitutes one of the options. It primarily provides its users with means for operation in the electronic world, information access, computer-mediated communication with others, etc. Finally, it improves their independence in everyday life and increases their participation in social activities. Several alternative techniques for touchless interfaces exist and the most popular ones include [19]: hand gesture recognition, brain computer interfaces, eye tracking, speech recognition and silent speech recognition (lip movement analysis). As it was noted, not all alternatives can be applicable for everyone and the choice of the appropriate solution is dependent on the specific form of disability.

Head operated interface offers touchless interaction through the analysis of the movements of user's head. Thanks to the detection and tracking of the user's face or facial features, it is possible to execute certain actions in no-contact interface. From the historical perspective, the first solutions employed markers attached to distinctive parts of the user's face. With a marker attached e.g. to the middle of the forehead, the process of the detection and tracking was simplified. Thanks to the progress in computer vision and pattern recognition, current solutions widely operate without the need for additional facilities.

✉ Paweł Forczmański
pforczmanski@wi.zut.edu.pl

Adam Nowosielski
anowosielski@wi.zut.edu.pl

[1] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Zolnierska Str 52, 71-210 Szczecin, Poland

Existing head operated approaches, in general, focus on conventional mouse replacement offering the *camera mouse* interface (e.g. [24, 36]). The interaction is based on the pointer manipulation through head movements in the Graphical User Interface (GUI). Rotation and translation of the head are denoted as rigid motions [36]. For confirmation purposes (i.e. clicking events), different non-rigid motions are employed. Some of the examples include: eye blinks [24, 30, 37], mouth shape changes (opening, closing or stretching) [5, 14, 33, 36], brows movements [14] or cheeks twitch.

With the help of rigid and non-rigid motion modelling a successful mouse alternative can be achieved, which is sufficient in most cases. However, human–computer interaction, besides the pointer manipulation, often requires a text input. Text entry in the camera mouse approach is usually performed on the standard on-screen QWERTY keyboard through pointer manipulation. The process is not comfortable, requires substantial precision, is tedious and time-consuming. Another possibility is the traverse procedure where keys or groups are accessed in sequence according to the direction of the head movement.

As it was observed, current research on non-haptic interfaces based on computer vision methods predominantly focus on visible spectrum omitting the thermal imaging. It should be remembered, that face-oriented algorithms often assume a certain level of illumination (in terms of intensity and spatial configuration). On the other hand, there are many situations when environmental conditions are not fully controlled [10]. Figure 1 presents two sets of face images taken at the same moment using visible spectrum (left column) and thermal cameras (right column) placed side by side. In the first case (first row), there is a strong directional sunlight coming from a window. The second case (second row) presents a scene of a poorly lit room. Both cases present a challenge for visible light imagery. In such case, thermal imaging seems to be a good choice. Images registered by infrared or thermal sensors can be used to perform face detection and recognition without the necessity to properly illuminate the subject [7, 21]. As can be seen from examples presented in Fig. 1, thermal images are definitely more stable in context of diversified lighting, particularly in severe lighting conditions. Moreover, in a broader biometrical context, such data are resistant to spoofing attempts (e.g. using previously captured photo or video stream [34]).

Fig. 1 Comparison of face images taken at the same moment in unfavourable lighting conditions using visible (left) and thermal (right) camera

In this paper, a novel solution for touchless typing with head movements is introduced. It combines the thermal imaging for capturing user actions with the hierarchical letter selection procedure we proposed earlier in [27]. Thermal imaging makes the algorithm independent on the lighting parameters, hence making it possible to work in complete darkness. The hierarchical character selection procedure offers substantial acceleration of the typing process. It requires three directional head movements only to reach base character without any additional mechanism for confirmation (eye blink, mouth open or other).

The rest of this paper is organized as follows: Sect. 2 presents some related works. Research on head operated interfaces is referred here together with face detection and tracking. The section includes keyboard layouts discussion. The main concepts of the interface are presented in Sect. 3 where the appropriate on-screen layout and adequate interaction techniques are introduced. In Sect. 4, several contemporary general purpose object detectors are investigated, addressed in the literature as effective in case of visible light illumination. They are applied to the face detection and tracking in thermal spectrum images. Finally, the proposed interface is evaluated and the results are analysed in Sect. 5. The article ends with a summary including conclusions and the discussion of the results.

2 Related works

According to the authors' best knowledge, there is no similar solution available on both commercial market, but also in the scientific literature. Hence, the review of related works was done on a basis of general head operated interfaces and elementary approaches which are combined in the presented system, i.e. face detection, face tracking, and also keyboard layouts.

2.1 Head operated interfaces

Based on the concept of camera mouse, some interfaces have been reported in the scientific literature. Different approaches are utilized to handle pointer manipulation in

GUI using head movements. In [5, 23] and [37] the position of user's nostrils related to the face region is used. Interestingly, in [5], a depth imaging technique is adapted. From a depth image, the nose position and the mouth status are detected and used for steering. In [30] and [33], the mouse cursor navigation is based on the estimation of eyes in the image plane. The mouse control in [36] is obtained by 3D head pose evaluation. The 3D pose estimation for camera mouse is also employed in [24]. As can be observed, some of the proposed solutions focus on tracking user's face only, while other proposals address facial features.

The conventional mouse replacement in non-contact head operated interfaces is an important research topic. Text typing in such environment, as it was mentioned before, is usually offered through on-screen keyboard operated by pointing mechanism or traverse procedure. A few examples can be found in literature: [14, 26] and [33]. Some solutions have also been shared publicly as open source projects [1] and [28]. The Assistive Context-Aware Toolkit (ACAT) [1] was originally created for Professor Stephen Hawking. Beside typing it offers tools for wide range of applications like documents management or Internet navigation. It can operate with eye blinks, eyebrows movements, cheeks twitch or mouth opening. The QVirtboard offers mouse replacement through face movement detection and a dedicated keyboard operating with the traverse procedure. It also supports other modes of control (eye tracking or hand movements) [25, 28].

Text typing by head movement analysis can be time-consuming. In the traverse procedure it requires many steps to reach the intended letter. On the other hand, in the pointing approach, it needs a significant precision. Nevertheless, since head operated interface may be the only solution for some people making an access to information or communication possible, improvements in existing or creating new ideas are necessary.

2.2 Face detection and tracking

The foundations of head operated interfaces are detection and tracking of face or facial features. The problem of human face detection in static images is quite well researched, and there are many complete solutions available [7, 12, 38]. From the practical point of view, it is equivalent to the determination of the scene area containing the searched face. The false positive rate, in such case, should be as low as 10^{-6} [38]; however, it is true for objects captured under similar imaging conditions (at the learning and testing stages). In such case, it is important to select proper discriminative features used to build a face model.

In case of thermal imaging, a visual representation of faces depends mostly on camera calibration parameters, which may not be reproducible. Exemplary facial portraits in terms of different camera calibration parameters are presented in Fig. 2. As it can be seen, many parts of faces are represented in different manner; thus, it makes the discussed task particularly difficult, and highly dependant on training samples availability.

The other important aspects of this problem are the mechanisms for feature matching and scanning of the source image. Hence, since we have no information about probable face position and its size, it is required to perform search procedure in all possible locations, taking into consideration all probable window (or image) scales, which increases the overall computational overhead. In the task discussed in this paper, the searching area and the scale pyramid can be significantly reduced, assuming constant and predictable position of user's head.

Later, in the paper, we focus on feature extractors and classifiers that enable proper facial portrait detection in thermal images. The algorithms were selected taking into consideration the computational complexity, the simplicity of implementation and the accuracy. Hence, we have investigated several well-known and recently proposed approaches.

In the next stage, the detected face is tracked in order to capture its movement. Recently, different methods of general object tracking have been proposed, yet it still remains challenging due to factors like abrupt appearance changes and severe object occlusions.

Face trackers use different approaches, which primarily focus on designing sophisticated appearance models and/or motion models to deal with challenging factors such as scale variations, three-dimensional rotations and illumination changes. There are three general classes of object trackers [40]: point trackers, kernel trackers, and silhouette trackers.

In most cases, each tracked face is associated with an information about its bounding box, last and predicted position and a serial number (or identification number). One of the most popular tools is the Kalman filter. It predicts further positions of objects [6], based on the idea of frame-to-frame analysis. It works in a stepwise manner using predicted

Fig. 2 Exemplary images taken with different temperature calibration parameters: 25.4–35.7 °C, 25.4–36.6 °C, and 26–36.9 °C, respectively (images from [16])

position or correspondence between their last position and foreground blobs (faces).

Another very popular method is Mean-Shift algorithm [40] since it is independent on the object appearance; hence, it makes it possible to track objects that are partially occluded or change their silhouettes over time.

As it was noted in [40], selecting proper features plays a crucial role in tracking. In the presented approach, where characteristic points on the face are selected, the optical flow seems to be the proper method. It is based on a dense field of displacement vectors which defines the translation of certain pixel in an image. It is often based on the about brightness constancy of corresponding pixels in consecutive frames. One of the most popular applications of optical flow is a tracker described in [20]. Thanks to its features, it was incorporated to our algorithm.

2.3 Layout of the keyboard

The QWERTY keyboard has been designed in the 19th century and still remains the standard, although more efficient and ergonomic layouts have been introduced. Many argue that the QWERTY key arrangement is not suited for current needs and interfaces. It is an issue in case of non-contact interfaces or mobile devices. Changes in the form of interaction are introduced and swipe typing (employed also in the QWERTY keyboard) or swipe gestures are the examples. Another trend is the reduction of the number of keys. Different characters are often placed on individual keys or in the sectors similar to the older phone keyboards. The 8pen keyboard (http://www.8pen.com/) and the 5-Tiles (http://fivetiles.com/) are good examples of such trend. The keyboard in the 8pen is divided into four directional sectors (top, bottom, left or right) with specifically arranged letters on the borders. The characters are accessible by circular swipe movements through directional sectors and the central sector. In the 5-Tiles keyboard, five separate areas (tiles) are sufficient to type. The typing proceeds with keystrokes and swipe gestures simultaneously. Although such new alternatives to the QWERTY key arrangement can be a source of inspiration for head typing interface, there are many difficulties in the adaptation process. Interactions like swipe gestures, circular movements or, sometimes, key-pressing present a substantial challenge.

Users are reluctant to learn new key arrangements and it seems that the only widely acceptable alternative to the QWERTY key arrangement is the alphabetical order. The alphabetical arrangement of characters has been successfully used in 12-keys mobile phone or 5-keys pagers. Nowadays, it is still frequently employed and the 5-Tiles or smart TVs are the examples. The alphabetical arrangement of characters can occur in a single-row or a multi-row layouts. The first solution is particularly interesting from the perspective of head operated interfaces. In our previous work [26], we demonstrated that such keyboard can be operated with only directional head movements without the need for additional gestures like eye blink or mouth opening for confirmation. To achieve this, in the most primary version, left and right directional movements are mapped to shifts of the active letter on the on-screen keyboard. The nodding gesture (tilt of the head down, denoted as downward direction) is interpreted as pressing while the opposite upward gesture is used for backspace.

The frequent need of moving through many characters in the single row of the alphabetical keyboard is a fundamental problem. Suggestions or dictionary support which may deactivate letters, similarly as in GPS navigation devices when typing the name of desired town, can alleviate the problem. Completely different approach has been proposed in the 3-Steps Keyboard [27] where any keyboard letter can be accessed with only three directional head movements. Since this concept is employed and extended in the current paper, it will be thoroughly presented hereafter.

3 Interface concept

To achieve the goal of imperceptible and user-friendly head operated interface for text entry, the appropriate on-screen layout and adequate interaction techniques are required.

3.1 3-Steps Keyboard

The 3-Steps Keyboard offers a new concept to touchless typing with head movements. It utilizes directional head movements and any keyboard letter can be accessed with just three steps (directional head movements). The keyboard has the form of a single row with alphabetically arranged characters as presented in Fig. 3. As it can be seen, keys are vertically translated in relation to each other and form distinctive groups. This presentation form has the aim to guide the user according to the appropriate interaction model. The displacement indicates the direction which has to be taken in order to reach certain group and individual characters.

The access to all the keys in the 3-Steps Keyboard is hierarchical and requires three consecutive directional head

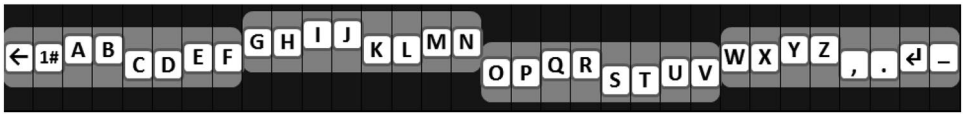

Fig. 3 The layout of the reduced interaction 3-Steps Keyboard [27]

movements. In the first step, the user selects the main group. There are four groups, each consisting of 8 keys. The first and the fourth groups are located on the same level, while the two middle ones are shifted, one upward and the other downward, respectively. The movement of the head in one of four main directions selects the appropriate group in the following way: the first group is selected with the left-direction movement, the second with the up movement, the third with the down movement, and the fourth with the movement to the right [27]. When the main group is selected, the others are deactivated (darkened) as depicted in Fig. 4.

After selecting the main group in the first step, a subgroup of two letters is marked in the second step. The procedure is similar, since subgroups are translated in relation to each other in the same manner as the main groups. A pair that corresponds to the head movement direction remains active, while the others are deactivated. The user proceeds to the third step where the choice of an appropriate letter continues with left or right directional movement. The selected letter is transcribed and the keyboard returns to the initial appearance with all the keys being active. The example of the interaction of '*k*' character entry is presented in Fig. 5. As it was assumed, the character is reached with just three movements UDL (up, down, left).

Besides alphabetic characters, the 3-Steps Keyboard also provides an access to the other symbols. The extreme buttons represent backspace and space. Dot, comma and enter are also available for a direct selection. Numbers and other symbols are accessed by switching the keyboard state (accomplished by the second key and LLR combination of movements) [27].

3.2 Interaction through head movements

The procedure of hierarchical letter selection offers considerable acceleration in accessing the keyboard keys. The typing requires three consecutive directional head movements to reach base character without any additional mechanism for confirmation (eye blink, mouth opening, etc.). The most important is a recognition and an interpretation of head actions. From the user's perspective, head movements should be simple and straightforward, easy to understand, learn and operate. The simplest situation concerns vertical directions. The up and down directions are achieved with the rotary upward or downward nod gesture (pitch). For the user, they are natural and very easy. More alternatives exist with horizontal directions. Left and right directions can be accomplished with head tilt (roll), rotation (yaw) or shift. Each individual has his/her own predispositions and the interface should not restrict any form of interaction.

For steering purposes we assume that the face is detected and specified by some surrounding frame (bounding box). This frame's central point is denoted with a pair of coordinates and represents a central position of the head. While the head moves, the centre coordinates change. Treating the initial coordinates as a base, any change of position allows to calculate the offset value in horizontal and vertical directions. On a basis of the solution presented in [27], we have adopted the following assumptions:

Fig. 4 Beginning of the interaction with the 3-Steps Keyboard

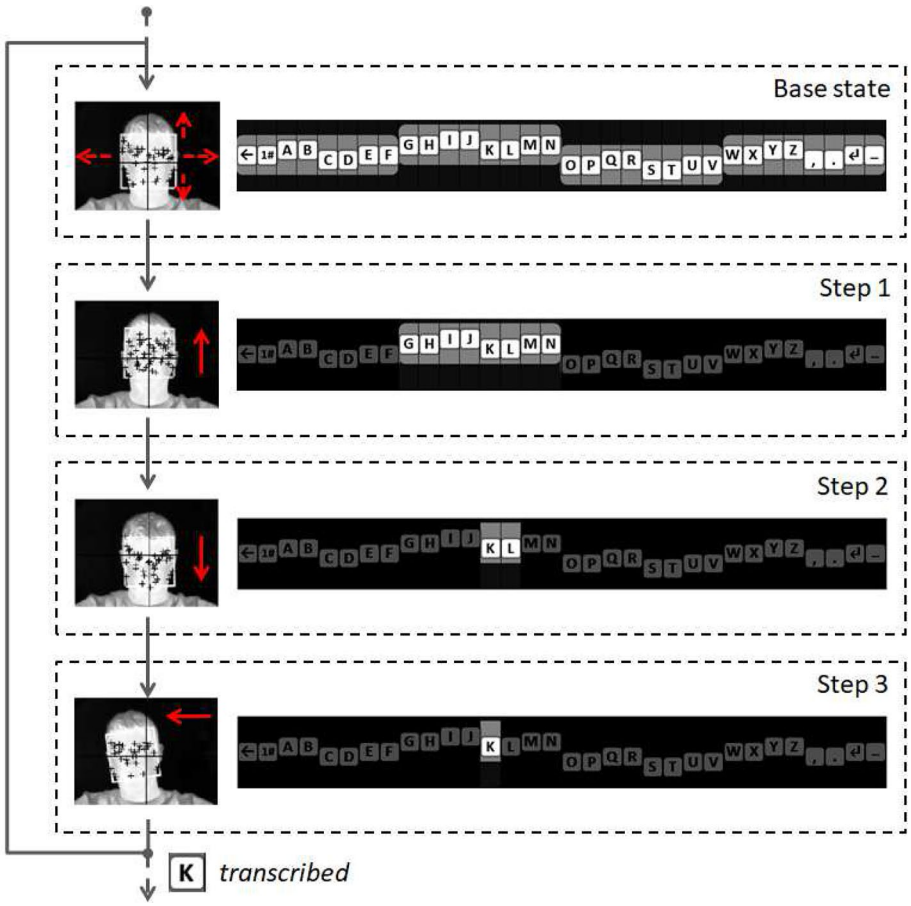

Fig. 5 The layout of the reduced interaction 3-Steps Keyboard. Exemplary interaction: letter 'a' transcription with the left, up and left combination of movements

- reference head centre is surrounded by a neutral area defined by thresholds determined at the calibration stage;
- coordinates of the reference head centre are updated with small, natural user movements (adaptation);
- movements exceeding the threshold in any direction are used for steering;
- when no other action follows after directional movement, it is assumed that no steering was intended, the user has taken a more comfortable position, and a new reference centre is calculated;
- the direction with a higher value of shift wins (e.g. during horizontal head movements with the tilt approach the abscissa coordinate changes jointly with the ordinate coordinate);
- diagonal movements are allowed to shorten the path (e.g. after performing the left gesture, user can directly move diagonally to the up position, instead of performing the left gesture with the return to the centre followed by the up gesture with the return to the centre);
- after a new character is typed and the face returns to the central position the new reference centre is calculated (continuous adaptation of the user position);

The interaction with the keyboard interface is based on a recognition of the user actions performed with the head. In the following chapter, we provide the details of our investigations regarding face detection and tracking in the thermal spectrum.

4 Method description

The algorithm consists of several steps, which are depicted in the Fig. 6. It works in the loop, in which a final sequence of characters is entered. The timeout blocks are introduced to protect the algorithm from infinite loop formation.

4.1 Face detection

The first step is to detect the user's face in the image. We do not assume anything about the face, except that it is sufficiently large (occupies more than approximately 25% of image area) and it is in the frontal orientation. The performance of a face detector influences the whole process in terms of computational efficiency and accuracy. According to the previous research [11, 12], a detector based on Local

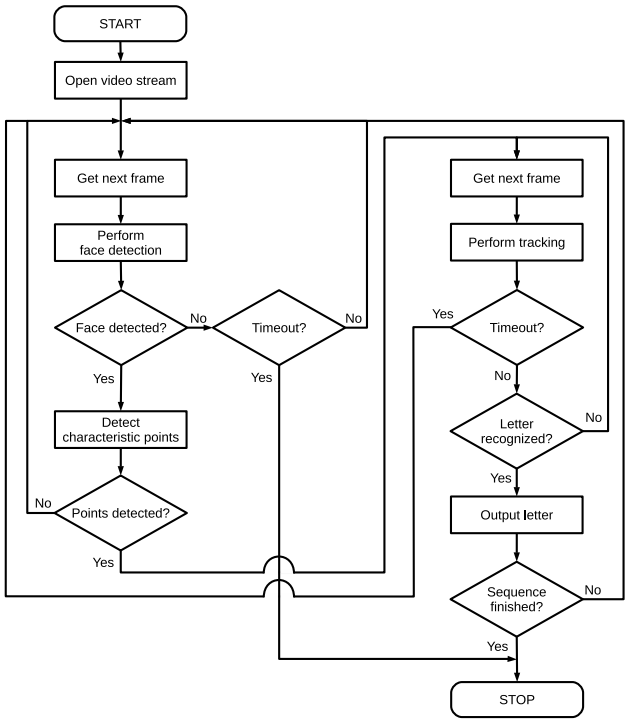

Fig. 6 An algorithm of face tracking-based touchless text input

Binary Patterns is a natural candidate. It was chosen based on the highest recall rate, yet taking into consideration the lowest possible computational overhead [31].

In order to verify the above mentioned observations, we tested several approaches, that have been successfully applied to many object detection tasks [3, 38, 39], namely cascading classifiers based on Haar-like features (Haar), Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP), together with HOG learned by Max-Margin Object Detection Algorithm (HOG + MMOD) and deep-learning Convolutional Neural Networks (DNN). The first three detectors were implemented using Open Computer Vision library (OpenCV), while the remaining two were implemented using Deep-Learning Library – dlib [18]. The OpenCV implementation used an Intel i7 second-generation processor, while dlib programs used the NVidia GTX780 GPU. The details of the experiments together with the discussion of the results are presented in [12].

The benchmark set used in the experiments consists of WIZUT database [17], Caltech (Courtesy NASA/JPL-Caltech) [16], OTCBVS Terravic Facial IR Database [22] and images taken from the Internet (UC). Collected faces include not only fully frontal portraits: some of them include glasses, head cover, or both. Ground-truth bounding boxes representing faces were marked manually, covering most informative facial part. The exemplary cropped faces used for training the detectors are presented in Fig. 7. The details of particular databases are given in Table 1.

We created two experimental setups, related to two possible application scenarios (called later experimental setups no. 1 and no. 2). The setup no. 1 recalls the

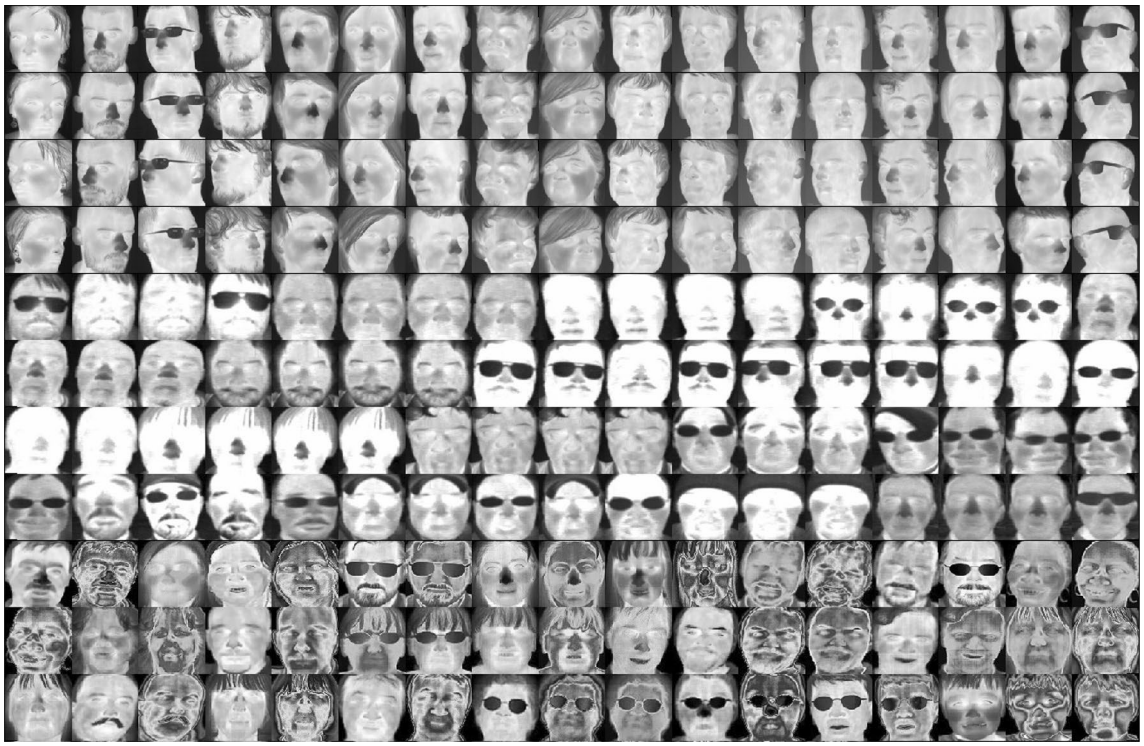

Fig. 7 Selected images used for training in the experimental setup no. 2

Table 1 Face detector training/testing-datasets characteristics

Dataset	WIZUT	Caltech	OTCBVS	UC
No. images/faces	505	64	2000	63/123
No. subjects	101	28	20	123
Image width	320	285	320	167–1920
Image height	240	210	240	129–1215
Min. face size	119 × 124	69 × 73	96 × 120	22 × 26
Max. face size	182 × 178	121 × 135	144 × 212	354 × 317
Rotation angle [°]	± 45	± 5	± 20	n/a

situation, when a user uses the touchless interface in an indoor conditions, while the setup no. 2 resembles fully uncontrolled conditions. The training dataset in the setup no. 1 consists of WIZUT (b) images which contain frontal-only faces captured in fully controlled laboratory conditions, using single camera, while the training dataset in the setup no. 2 consists of all WIZUT (a)-(e) groups, Caltech and OTCBVS images. The latter images contain faces in various orientations, captured with various cameras and with different sensor calibration parameters.

The testing dataset in the setup no. 1 consists of the remaining images gathered in WIZUT (a), (c), (d), and (e), Caltech and OTCBVS. In the setup no. 2, the testing images come from UC.

The negative samples for cascading classifiers were provided in an automatic manner, depending on the classifier training method. They were extracted from various images captured in the thermal spectrum, collected from the Internet, containing no faces.

Table 2 Training parameters for all detectors

Detector	Cascade			HOG+ MMOD	DNN
	Haar	HOG	LBP		
Pos. smpl.	800–1000	800–1500	500–1000	5138	2569
Neg. smpl.	2000–5000	4000	2000–4000	–	–
No. stages	11–12	15–17	11–13	54	3500

At the learning stage, in case of cascading classifiers based on Haar-like features, LBP and HOG, we used Gentle AdaBoost, with a varying numbers of positive/negative samples and learning stages. As it can be seen (compare Tables 1 and 2), the actual number of positive samples is often larger than the maximum number of the images in the base dataset. It is caused by the fact that the training samples are created from altered images (by rotation, changes in brightness, with noise added, or cropped) in order to increase their variability. In case of HOG+MMOD and DNN, the number of positive samples is the only parameter.

The number of positive/negative samples and the number of training stages presented in Table 2 varies, since it presents several variants of training procedure. It should be noted that cascading classifiers use a limited number of samples at each training stage, while DNN uses all the samples. HOG+MMOD uses a doubled number of positive samples, extended by symmetrical copies of the original images. The number of iterations depends on the learning rate and is automatically calculated.

The quality of the detector is calculated by means of Intersection over Union–*IoU*, often employed in object detection challenges [8]. The evaluation procedure takes into consideration bounding boxes associated with the object(s) in the image: ground-truth bounding box(es) of an area A_{gt} representing known object(s) and the predicted bounding box(es) from the model of an area A_{det} representing detected object(s). Its value is equal to the area of overlap between the bounding boxes divided by the area of their union:

$$IoU = \frac{A_{\text{gt}} \cap A_{\text{det}}}{A_{\text{gt}} \cup A_{\text{det}}}. \tag{1}$$

An *IoU* score higher than 0.5 is often considered a *good* prediction. However, in order to increase the recall indicator, we set the threshold to 0.3. Exemplary values of *IoU* for various detected faces are presented in Fig. 8.

In practice, in order to find a face in the image, we select the largest detected object. On the other hand, in the experimental evaluation, for each image, a matrix of

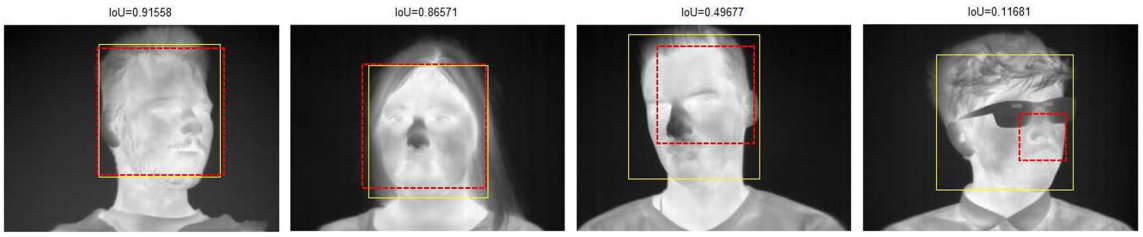

Fig. 8 A comparison of various *IoU* scores: successful detection (the two first cases), border line detection (the third case) and unsuccessful detection (the last case)

Table 3 Main parameters of evaluated detectors

Detector	Haar/HOG/LBP	HOG+MMOD	DNN
Min window size	16×16	40×40	40×40
Max window size	360×360	80×80	200×200
Scaling factor	1.02	1.2	1.2
No. neighbours	2	1	1

IoUs is calculated for all combinations of ground-truth and detected bounding boxes, and only the highest score is taken into consideration. As the performance metric, we selected precision and recall measures, taking into consideration *IoU* values.

Several experiments have been performed, which involved trained detectors with various parameters, presented in Table 3, namely the size of search window, scaling step in the image pyramid and candidates rejection rule based on the number of adjacent detections (number of neighbouring detection indicating true detection, in case of cascading classifiers).

As it was mentioned above, the experiments were aimed at comparing the selected LBP-based face detector with the other ones taking into consideration the detection rate (precision and recall measures) as well as computational overhead. While the detection rate for rather easy images (see Table 4, setup no. 1) is very high (almost perfect), for images taken in more complex conditions (see Table 4, setup no. 2) drops.

It is clearly visible that cascading detectors process a single frame in the shortest time (see Table 5). t_1 is a processing time without using scale pyramid, while t_2 is a processing time with scale pyramid (scaling factor in Table 3). The first case mimics the application when a user sits close to the camera and the screen. On the other hand, DNN-based solution, using more sophisticated computations, is the slowest one (regardless on the scaling method). Taking it into consideration, one can come to the conclusion that its use is not justified in the task discussed in this paper.

The results support the conclusion that when we take recall as the main indicator, having in mind the processing time, the LBP-based detector should be used. Another interesting observation is that the detectors learnt with presented samples are able to detect also some faces in images taken in the visible spectrum (see Fig. 9). It leads to the conclusion that probably it is possible to construct a detector that works independently on the imaging technology. However, a confirmation would need some extra investigations.

Table 5 A comparison of single frame processing times [msec] (t_1—without scale pyramid, t_2—with scale pyramid)

Detect.	Haar	HOG	LBP	HOG + MMOD	DNN
t_1	0.41	0.42	0.33	14	35
t_2	19	21	19	60	170

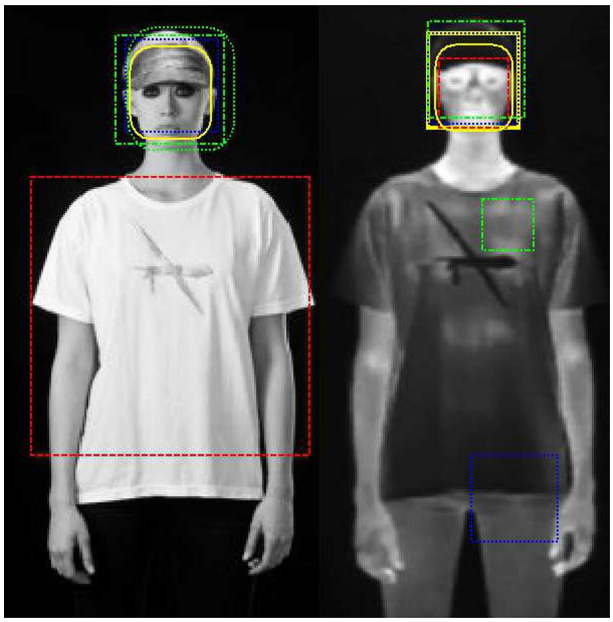

Fig. 9 Exemplary face detection results. The yellow solid line squares represent ground-truth, blue dotted line squares–HOG results, red dashed line–Haar results, green dash-dot line–LBP results, yellow solid line, rounded rectangles–HOG + MMOD results, and green rounded rectangles with dotted line–DNN results

4.2 Face tracking

To capture head movements and convert them into the interaction with the keyboard, a feature-tracking algorithm is employed. It is initialized with a face detected at the previous stage and detects some distinctive initial points inside the face region. It should be noted that typical facial features can not be used, since they appear blurred and not clearly

Table 4 Precision/Recall rates for different detectors/setups together with single frame processing time (t_1–without scale pyramid, t_2–with scale pyramid)

Detect.	Haar		HOG		LBP		HOG + MMOD		DNN	
Setup	1	2	1	2	1	2	1	2	1	2
Precision	0.74	0.46	0.89	0.63	0.75	0.47	0.91	0.87	0.90	0.88
Recall	0.92	0.46	0.90	0.49	0.95	0.51	0.90	0.37	0.98	0.30

visible when using the thermal spectrum image, in opposition to visible light images [13]. Many other distinctive points (not related to particular face parts), however, can be easily detected.

We applied selected popular characteristic points detectors on the thermal spectrum face images. Exemplary results for images from the WIZUT database [17] are presented in Fig. 10. We considered corners detected using FAST algorithm (green squares) [29], Harris–Stephens features (blue x) [15] and SURF features (red circle) [4]. Unfortunately, they occurred to be less precise than Shi and Tomasi features (see Fig. 11). Most of the characteristic points are located on the outline of the face and in the area of facial features.

Hence, the procedure of distinctive point selection developed by Shi and Tomasi [32] has been chosen. Shi and Tomasi modified the original Harris–Stephens corner detector by replacing the scoring function which now takes into consideration the smaller eigenvalue from the pair (minimum eigenvalue algorithm) for a given region consisting of an examined pixel and its neighbourhood. It is accepted as a corner when the smaller eigenvalue exceeds the predefined threshold.

Initially detected points are used for tracking. The Kanade—Lucas–Tomasi (KLT) feature-tracking procedure is employed. The procedure has been introduced first in [20] by Kanade and Lucas as an image registration method, and it was modified later. For each point, the tracker attempts to find a counterpart point in the new frame. The procedure is iterative and the initial guess of the point locations is refined with each step. The region centred on an interest point is evaluated by computing the affine transformation between the corresponding patches in consecutive frames [35]. If the sum of square differences between the current patch and the projected patch exceeds the threshold, the feature is eliminated. Otherwise, the feature is still being tracked [35]. Finally, all matched pairs are used for calculation of the geometric transformation. It is afterwards applied to the bounding box of the previous face location. The centre of that bounding box serves for steering purposes.

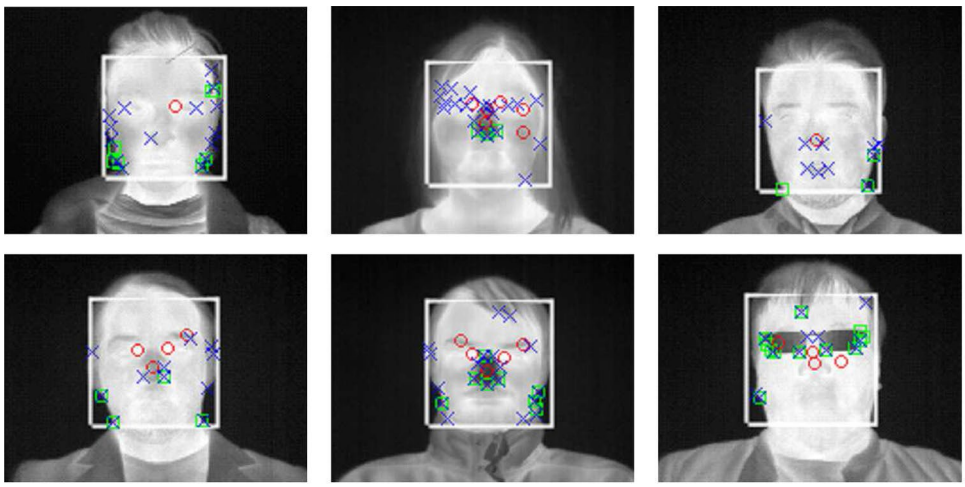

Fig. 10 Comparison of points detected on thermal face images from the WIZUT database [17] (first six persons, frontal images): FAST features (green square), Harris–Stephens features (blue x) and SURF features (red circle)

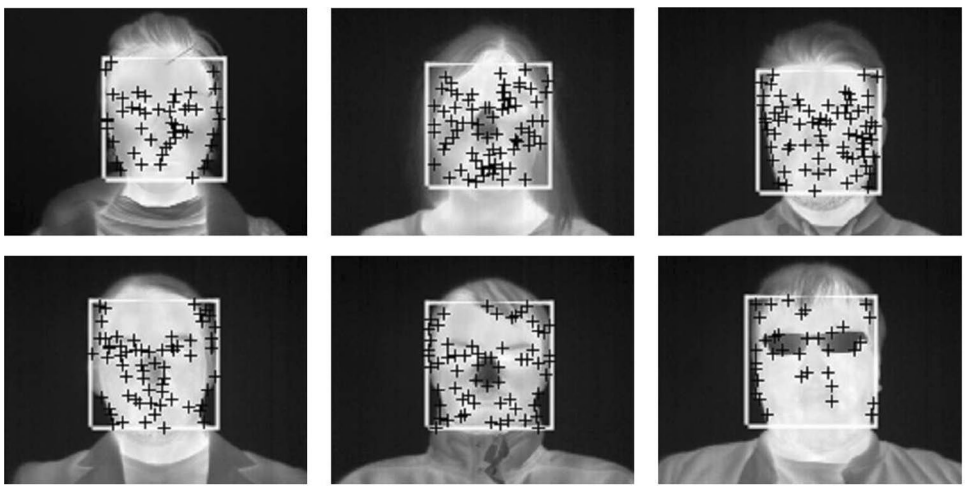

Fig. 11 Points detected using minimum eigenvalue algorithm on thermal face images from the WIZUT database [17] (first six persons, frontal images)

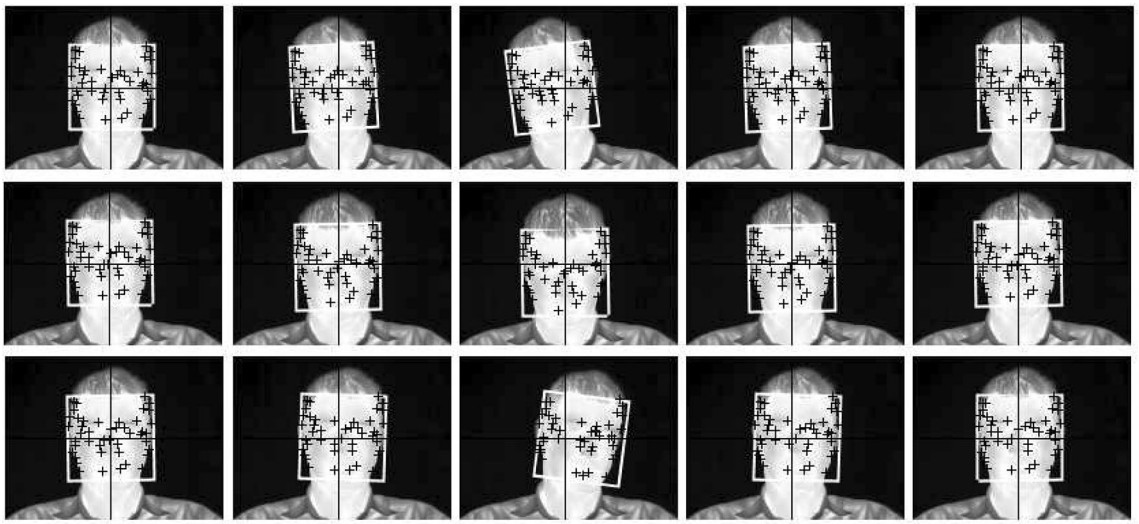

Fig. 12 Results of point tracker: left-direction movement (top row), downward movement (middle row), right direction movements (bottom row)

In Fig. 12, the results of point tracker applied to thermal image sequence is presented. In this case, a user performs a sequence of actions representing an input of 'd' letter by the following moves: left, downward and right. The black lines indicate the position of the reference centre (initial head position). Each row in the figure corresponds to the consecutive directional movements with subsequent images denoting: initial position, beginning of the gesture, maximal shift, on return and return to the starting position.

The algorithm of point tracking is known to be stable for images captured in the visible spectrum and rigid objects that do not change shape [35]. When tracking a face in the thermal spectrum for several seconds under extensive head movements many points can be lost. The assumptions formulated in Sect. 3.2 allow to avoid that problem. After a new character is typed and the face returns to the central position, a new reference centre is calculated which forces new face detection and a selection of new points for tracking. The procedure is fast and imperceptible for the user. The initialization of new points is also performed when too many points are lost and, periodically, when the head is out of the neutral area. Without the described recalibration procedure, the tracking is unstable in longer terms and may result in an improper rotation and scaling of the bounding box of a tracked face as presented in Fig. 13.

5 Interface evaluation

Evaluation of text entry methods is predominantly user-based. Usually, the performance measures are collected while the participant performs the task of typing a given text phrase quickly with respect to accuracy. Unfortunately, while typing, users make errors. Substitutions, insertions, or deletions of letters appear. There are three approaches to error correction [2]: none, recommended, and forced. In the first case, it is prohibited to correct errors and each mistake is taken into account when measuring the error rate. A quite different approach is offered with the forced error correction where the participant has to correct all the errors giving as the result a text that is identical to the one presented. Such procedure lengthens the typing time but offers good basis for different interfaces comparisons. For an uncomfortable or error-prone interface, the measured time of the text input is significantly longer. Since the final text does not contain errors, beside the typing speed, the number or the ratio of

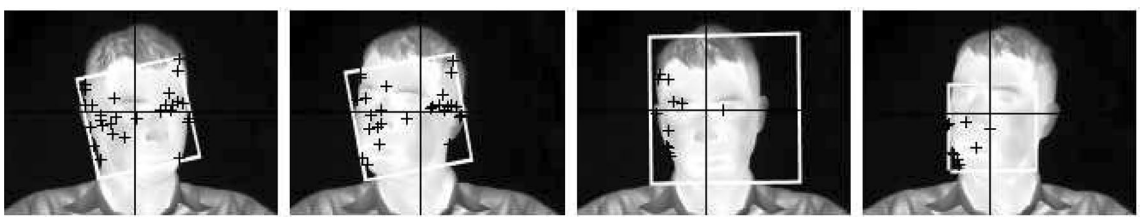

Fig. 13 Errors that may occur during long tracking

erroneous keystrokes are also considered as performance measures. In the recommended error correction condition, the participant is allowed to correct errors if he/she identifies them. Since the resultant text contains corrected and uncorrected errors, more elaborated error measures are required. They are based on counting the number of elementary operations required to correct wrongly transcribed text to the ideal one.

The evaluations of the proposed interface were user-based and have been performed with the third year computer science students of the West Pomeranian University of Technology, Szczecin, aged 22–23 years. The experiments have been organized during classes in Human–Computer Interaction Course and have lasted for a few months in a cyclic routine. Each new version of a prototype was evaluated by a group containing usually 6–12 students and after amendments and enhancements a new group continued testing. These experiments, conducted at first using visible spectrum imaging, allowed to develop the interaction routine with the 3-Steps Keyboard through head movements adopted hereby to thermal imagery. All the experiments have been conducted in the forced error correction condition.

The experimental setup is presented in Fig. 14. The experimental stand is situated perpendicularly to the bright window which in the normal daylight condition significantly impedes the work in the visible spectrum and hinders detection and tracking of faces. The hardware layer is a photographic stand, which consists of an infrared (IR) camera mounted on a 131 cm tall tripod with tubular and bull's eye spirit levels, denoted as '1' in Fig. 14. The IR camera is FLIR SC325, with 16-bit sensor of 320 × 240 pixels working at 60 Hz, having 25° × 18.8° FOV, interfaced by Ethernet [9]. When setting the camera, the most important factor is a correct alignment of the lens that should capture the whole head of the subject.

The software part of the interface has been launched on a notebook equipped with the first generation Intel Core i7-740QM processor and 8 GB of RAM ('2' in Fig. 14). All experiments have been performed in front of the 24" Full HD screen ('3' in Fig. 14) attached to the notebook. The participants ('4' in Fig. 14) did not have access to other standard input devices. They have to type text quickly and accurately (the forced error correction condition) using head movements.

The text input by head movements using the 3-Steps Keyboard approach has been well received by volunteers participating in the conducted experiments. The results for a group of 14 participants are presented in Fig. 15. There is a cpm (chars per minute) measure acquired for each individual on the left-hand side and the number of errors committed on the right-hand side. No bar in the second case means that a certain participant made no mistakes. Each participant has to type the same sentence consisting of 28 characters. For the 3-Steps Keyboard operated in the thermal spectrum, the mean value of 10.23 cpm has been obtained with 3.18 of standard deviation. The best participant achieved 16.31 cpm and the worst – 5.25 cpm. Nine users did not make any mistake, four made one mistake, and one participant has to correct three letters.

Previously, we have reported the typing speed of 12–14 cpm for the 3-Steps Keyboard operated in the visible spectrum in case of new users after a brief acquaintance [27]. In the thermal version of the interface, a mean value of 10.23 cpm for a whole group has been obtained. The most probable reason is the reduction of the time used to familiarize with the interface. When using the thermal version of the interface, the participants, after the explanation of the operating principles, proceeded to test typing of their name and surname, and immediately after, typed the test sequence. The previous experiments have been performed under classroom conditions, and participants were able to observe others and gather some experience.

To check whether the fact that the participants observed the others during the experiments influenced the final result and to provide the appropriate basis for comparisons, additional experiments using the visible spectrum have been conducted. Another group of 14 participants with no previous contact with the 3-Steps Keyboard have been gathered.

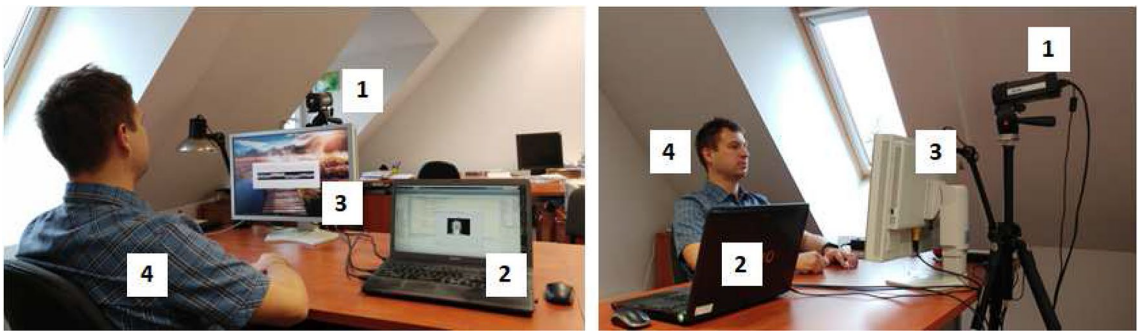

Fig. 14 A prototype of the test bench

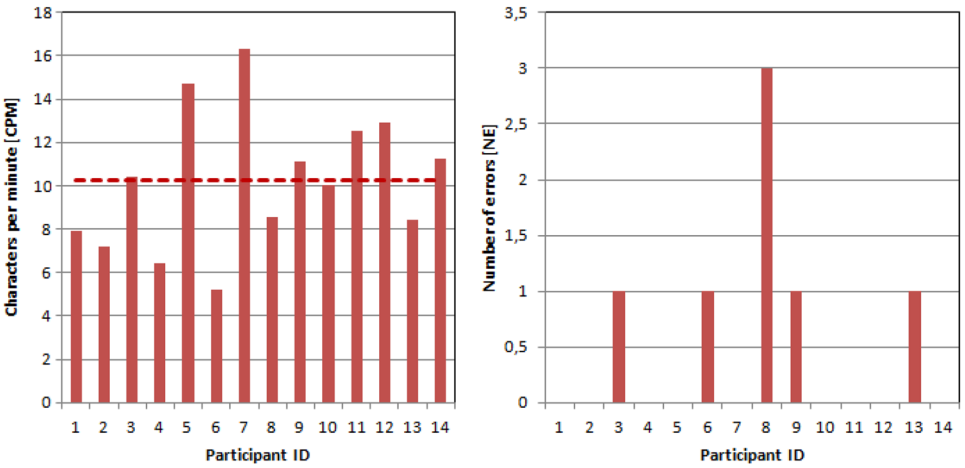

Fig. 15 The results of the experiments in the thermal spectrum

Experiments have been performed in the similar manner as in the thermal version of the interface: participants after the explanation of the operating principles, proceeded individually to the typing of their name, surname, and finally the test sequence. There was no audience during that experiments. The experiments have been performed using the same setup as in the thermal stand presented in Fig. 14. The thermal camera has been replaced with the icon7 Cyris T620 webcam. To provide good lighting conditions for the visible spectrum, the outside light was suppressed by means of a window blind and artificial light was switched on. The results are presented in Fig. 16.

When the users did not have a possibility to observe the others during the experiments, hence no experience have been gathered so, as expected, lower cpm values have been obtained. These results are very similar to those observed in thermal version of the interface. The average value of cpm equals to 10.15 with 2.13 of the standard deviation. The fastest participant achieved speed of 14.61 and the slowest – 7.18. Six of the participants committed errors during experiments, wherein four users did two mistakes and two users – one mistake. It is interesting that in both cases (the thermal and the visible spectra) the participants who committed errors, despite the fact of forced error correction condition, did not achieved the longest times of typing and they were able to accomplish the task in a reasonably short time. Returning to the higher cpm reported in [27] (12–14 cpm), it can be concluded that even a short time devoted to the observation of other participants during text entry can increase the performance of individuals.

Finally, in the context of target users, it must be noticed that the measuring of the typing speed for interfaces similar to the proposed one can be arguable. Depending on the type and the degree of disability, the usage of various solutions may be problematic or even impossible. Some users will cope better with one type of the interaction, while the other users will prefer other methods. For that reasons, the quality measures are not always reported by the researchers. What is more, the performance observed for disabled users, despite the fact that such interfaces are adopted for them, demonstrate worse performance (e.g. average typing speed of the test phrase in the 'Spelling board' [33] by able-bodied users

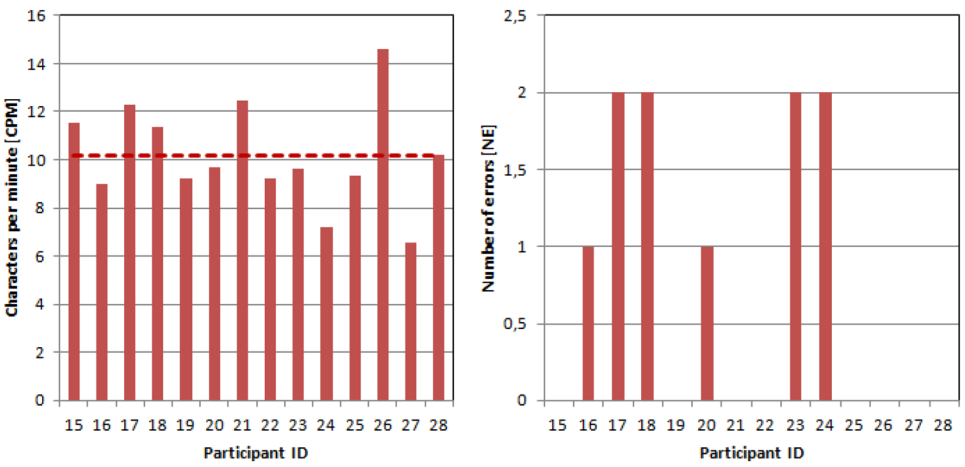

Fig. 16 The results of the experiments in the visible spectrum

reaches 16.95 while disabled users managed to complete the task with the mean value of 37.12 seconds [33] which is 119% longer). Some typing speeds of vision-based interfaces are compared in [14] where the authors reference works with 25, 31 and 44 cpm for the camera mouse approach and pointing procedure. As it was mentioned before, the pointing procedure requires substantial precision and may be not adequate for some users. A letter-scanning (traverse) procedure [14] used in the interface proposed by Grauman et al. allows users to achieve 5.7 cpm.

6 Summary

In this paper, a novel solution for touchless typing with head movements has been introduced. Its main concept is a fusion of the thermal imaging for capturing user action with the hierarchical letter selection procedure. Existing head operated interfaces require substantial precision when working in a camera mouse routine or suffer from a time-consuming process when operating with the traverse procedure. They also need a supplementary action (eye blinks, eyebrows movements, cheeks twitch, mouth opening or some others) to simulate the key stroke. The solution presented in this paper allows to reach an alphabet character in only three steps, i.e. directional head movements, excluding additional mechanisms for key-pressing. Despite not using the faster pointing procedure, the performance of our algorithm is decent. Its hierarchical nature of accessing characters, despite its uncommon nature, occurred to be easy for new users. In future, it is possible to further improve the interface by introducing a sentence-level suggestions.

The main target of the presented interface are people with disabilities. They need special assistive technologies which contribute to the improvement of their independence in everyday life and increase their participation in social activities. Considering the uncontrolled environment with dynamic lighting conditions, the usage of the thermal spectrum for capturing user action is justified. An evaluation of the selected thermal-imaging-based face detectors has been performed and presented in the paper. The results support the conclusion that when we take Recall as the main indicator, having in mind the processing time, the LBP-based face detector should be used. On the other hand, the tracking is particularly efficient when we apply KLT routine.

In future, we plan to perform evaluations with target users (in the performed experiments there were no handicapped participants). We also plan to further investigate the possibility of constructing a universal face/head detector that works independently on the imaging technology. We observed that in many cases the detector prepared for the thermal spectrum was able to detect faces also in images taken in visible light.

References

1. Assistive Context-Aware Toolkit (ACAT) (2017). Project page. https://01.org/acat. Accessed 20 Nov 2017
2. Arif AS, Stuerzlinger W (2009) Analysis of text entry performance metrics. In: IEEE Toronto international conference science and technology for humanity (TIC-STH). pp 100–105
3. Azzopardi G, Greco A, Vento M (2016) Gender recognition from face images using a fusion of SVM classifiers. In: Campilho A, Karray F (eds) Image Analysis and Recognition. ICIAR 2016. Lecture Notes in Computer Science, vol 9730. Springer, Cham
4. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) SURF: speeded up robust features. Comput Vis Image Underst 110(3):346–359
5. Bian Z-P, Hou J, Chau L-P, Magnenat-Thalmann N (2016) Facial position and expression-based human-computer interface for persons with tetraplegia. IEEE J Biomed Health Inform 20(3):915–924
6. Cannons K (1991) A review of visual tracking. Dept. Comput. Sci. Eng., York Univ., Tech. Rep. CSE-2008-07
7. Chang H, Koschan A, Abidi M, Kong SG, Won C-H (2008) Multispectral visible and infrared imaging for face recognition. In: 2008 IEEE Computer society conference on computer vision and pattern recognition workshops, pp 1–6
8. Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A (2015) The PASCAL visual object classes challenge: a retrospective. Int J Comput Vis 111(1):98–136
9. FLIR Instruments. Thermovision SDK User's manual, 2.6 sp2 edition (2010)
10. Forczmański P, Kukharev G, Shchegoleva N (2013) Simple and robust facial portraits recognition under variable lighting conditions based on two-dimensional orthogonal transformations. In: 7th International conference on image analysis and processing (ICIAP), LNCS 8156, pp 602–611
11. Forczmański P (2017) Human face detection in thermal images using an ensemble of cascading classifiers. In: Kobayashi S, Piegat A, Pejas J, El Fray I, Kacprzyk J (eds) Hard and soft computing for artificial intelligence, multimedia and security. ACS 2016. Advances in Intelligent Systems and Computing, vol 534. pp 205–215
12. Forczmański P (2018) Performance Evaluation of Selected Thermal Imaging-Based Human Face Detectors. In: Kurzynski M, Wozniak M, Burduk R (eds) Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017. CORES 2017. Advances in Intelligent Systems and Computing 578, pp 170–181
13. Fornalczyk K, Wojciechowski A (2017) Robust face model based approach to head pose estimation. In: Proceedings of the 2017 Federated conference on computer science and information systems, FedCSIS 2017, pp 1291–1295
14. Gizatdinova Y, Spakov O, Surakka V (2012) Face typing: vision-based perceptual interface for hands-free text entry with a scrollable virtual keyboard. In: IEEE workshop on applications of computer vision 2012, Breckenridge, CO. USA pp 81–87
15. Harris C, Stephens M (1988) A Combined Corner and Edge Detector. In: Proceedings of the 4th Alvey Vision Conference, pp 147–151

16. Hermans-Killam L Cool Cosmos/IPAC website, Infrared Processing and Analysis Center, http://coolcosmos.ipac.caltech.edu/image_galleries/ir_portraits.html. Accessed 10 May 2016
17. Jasiński P, Forczmański P (2016) Combined imaging system for taking facial portraits in visible and thermal spectra, image processing and communications challenges 7. AISC 389:63–71
18. King D (2015) Dlib 18.6 released: Make your own object detector! http://blog.dlib.net/2014/02/dlib-186-released-make-your-own-object.html. Accessed 27 Jan 2017
19. Kumar DK, Arjunan SP (2015) Human–Computer Interface Technologies for the Motor Impaired, CRC Press, p 214
20. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on Artificial intelligence, vol 2, pp 674–679
21. Małecki K, Nowosielski A, Forczmański P (2017) Multispectral Data Acquisition in the Assessment of Driver's Fatigue. In: Mikulski J (eds) Smart Solutions in Today's Transport. TST 2017. Communications in Computer and Information Science, vol 715 pp 320–332
22. Miezianko R IEEE OTCBVS WS Series Bench–Terravic Research Infrared Database. http://vcipl-okstate.org/pbvs/bench/. Accessed 20 May 2016
23. Morris T, Chauhan V (2006) Facial feature tracking for cursor control. J Netw Comput Appl 29(2006):62–80
24. Nabati M, Behrad A (2015) 3D Head pose estimation and camera mouse implementation using a monocular video camera. SIViP 9(1):39–4
25. Nowosielski A, Chodyła Ł (2013) Touchless input interface for disabled. In: Burduk R et al. (eds) Proceedings of the 8th international conference on computer recognition systems CORES 2013. Advances in Intelligent Systems and Computing, vol 226. pp 701–709
26. Nowosielski A (2016) Minimal interaction touchless text input with head movements and stereo vision. In: Chmielewski LJ, Datta A, Kozera R, Wojciechowski K. (eds) Computer vision and graphics. LNCS, vol 9972. pp 233–243
27. Nowosielski A (2018) 3-Steps keyboard: reduced interaction interface for touchless typing with head movements. In: Kurzynski M, Wozniak M, Burduk R (eds) Proceedings of the 10th international conference on computer recognition systems CORES 2017. CORES 2017. Advances in Intelligent Systems and Computing, vol 578. pp 229–237
28. QVirtboard. Project page (2017). http://qvirtboard.sourceforge.net. Accessed 20 Nov 2017
29. Rosten E, Drummond T (2005) Fusing points and lines for high performance tracking. Proc IEEE Int Conf Comput Vis 2:1508–1511
30. Santis A, Iacoviello D (2009) Robust real time eye tracking for computer interface for disabled people. Comput Methods Programs Biomed 96(1):1–11
31. Schaefer G, Krawczyk B, Doshi NP (2013) Improved LBP texture classification using ensemble learning. In: IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, pp 1–6
32. Shi J, Tomasi C (1994) Good features to track. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 593–600
33. Shin Y, Ju JS, Kim EY (2008) Welfare interface implementation using multiple facial features tracking for the disabled people. Pattern Recogn Lett 29(2008):1784–1796
34. Smiatacz M (2012) Liveness measurements using optical flow for biometric person authentication. Metrol Meas Syst 19(2):257–268
35. Yilmaz A, Javed O, Shah M (2006) Object tracking: a survey. ACM Computing Surveys (CSUR) 38(4), Article No. 13, ACM New York, NY, USA, pp 1–45
36. Tu J, Tao H, Huang T (2007) Face as mouse through visual face tracking. Comput Vis Image Underst 108(2007):35–40
37. Varona J, Manresa-Yee C, Perales FJ (2008) Hands-free vision-based interface for computer accessibility. J Netw Comput Appl 31(4):357–374
38. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vision 57(2):137–154
39. Wang X, Han TX, Yan, S (2009) An HOG-LBP Human Detector with Partial Occlusion Handling. In: IEEE 12th international conference on computer vision, Kyoto, pp 32–39
40. Yilmaz A, Javed O, Shah M (2006) Object tracking: a survey. ACM Comput Surv 38(4):13