Image Reconstruction of Tablet Front Camera Recordings in Educational Settings

Rafael Wampfler
Dept. of Computer Science
ETH Zurich, Switzerland
wrafael@inf.ethz.ch

Andreas Emch
Fision AG, Switzerland
ae@fision-technologies.ch

Barbara Solenthaler
Dept. of Computer Science
ETH Zurich, Switzerland
solenthaler@inf.ethz.ch

Markus Gross
Dept. of Computer Science
ETH Zurich, Switzerland
grossm@inf.ethz.ch

ABSTRACT

Front camera data from tablets used in educational settings offer valuable clues to student behavior, attention, and affective state. Due to the camera's angle of view, the face of the student is partially occluded and skewed. This hinders the ability of experts to adequately capture the learning process and student states. In this paper, we present a pipeline and techniques for image reconstruction of front camera recordings. Our setting consists of a cheap and unobtrusive mirror construction to improve the visibility of the face. We then process the image and use neural inpainting to reconstruct missing data in the recordings. We demonstrate the applicability of our setting and processing pipeline on affective state prediction based on front camera recordings (i.e., action units, eye gaze, eye blinks, and movement) during math-solving tasks (active) and emotional stimuli from pictures (passive) shown on a tablet. We show that our setup provides comparable performance for affective state prediction to recordings taken with an external and more obtrusive GoPro camera.

Keywords

Front Camera Setup, Inpainting, Affective Computing, Classification, Deep Learning

1. INTRODUCTION

Tablet computers have found quick application in education [14] as the technology offers new opportunities to students and teachers. It has been shown that tablets can influence learning pathways [19] and improve digital skills [47]. Moreover, tablets typically have built-in cameras, which can be used to unobtrusively record the student during the learning. Such data offers valuable clues to experts about the student's learning behavior and attention. Student observation has been implemented in studies with external camera setups [56]. Such frontal-view camera data can also be used for predictions of the affective states of a student based on facial feature extraction [46], which works robustly even with low-resolution recordings [43]. Affective states are psychophysiological constructs describing emotions (short-term) and moods (long-term) elicited by a stimulus [36, 51], and their impact on learning has attracted considerable attention in research on intelligent tutoring systems and education [3, 13, 41]. For example, Craig et al. [12] have found a positive correlation between learning and flow and a negative correlation between learning and boredom.

Using external cameras for frontal view recordings of students provides an optimal viewing angle for robust facial feature extraction and affective state prediction. However, such setups require externally positioned cameras, which can be obtrusive and further depend on timestamp synchronization with the digital learning environment. Using tablet computers for learning circumvents these problems, as the built-in camera can be leveraged and timestamps are inherently in sync. Built-in cameras have, however, a sub-optimal viewing angle, leading to partially occluded and skewed faces in the recordings that makes it difficult to robustly extract facial features for affect prediction.

In this paper, we therefore propose a camera setup for tablet computers and a deep learning-based image processing pipeline to reconstruct high-quality facial recordings of students. The setup requires a small mirror to be attached to the camera to improve the visibility of the face. Then, the image is reconstructed using a neural inpainting approach. We demonstrate the advantage of this setup and our reconstruction by an application for predicting affective states. The high quality of the reconstructed image enables facial feature extraction, such as head pose, eye gaze, and facial landmarks. We compare our method with an external camera setup (GoPro camera) and show that we can achieve a similar performance for predicting two levels (high and low) of valence and arousal for students performing active tasks, i.e., solving math tasks (up to 0.73 AUC) and students performing passive tasks, i.e., exposed to emotional stimuli from pictures (up to 0.80 AUC).

2. RELATED WORK

Inpainting. Image inpainting is an image processing method to reconstruct missing or corrupted regions of an image. Common application areas include image restoration (e.g.,

Rafael Wampfler, Andreas Emch, Barbara Solenthaler and Markus Gross "Image Reconstruction of Tablet Front Camera Recordings in Educational Settings" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 245 - 256

removing scratches and text) [34], photo-editing (e.g., object removal) [50], and image coding and transmission (e.g., recovering the missing blocks) [54]. In this work, we focus on the specific task of face completion. Popular non-learning based approaches applied to faces consist of patch-based methods, where image patches are copied to missing areas. Similar patches can be identified by using a face image dataset [58]. We refer to Guillemot and Le Meur [21] for a complete overview of non-learning based models.

While non-learning based methods can have difficulties to ensure consistent image structures [24, 45, 55], learning-based approaches typically generate smoother results. A popular line of learning-based methods uses generative adversarial networks (GAN) to inpaint missing regions of an image. GANs consist of a generative network to create a new image and a discriminator network to distinguish the new image from actual ground truth images. Using such a GAN approach, Malesevic et al. [37] reported a peak signal-to-noise ratio (PSNR) of up to 20.57 for inpainting missing regions in faces. A similar performance of up to 20.2 PSNR and 0.84 structural similarity (SSIM) was achieved by Li et al. [31] using an encoder-decoder network as the generator, a local and global loss function and a semantic regularization term. On the other hand, Liao et al. [32] used a collaborative model by training a GAN simultaneously on multiple tasks (i.e., face completion, landmark detection, and semantic segmentation). Using this knowledge-sharing approach, they reported a PSNR of up to 31.5 and an SSIM of 0.97 on face inpainting.

Convolutional neural networks (CNN) have been used for image inpainting as well. The encoder compresses the image with convolutional operations into a latent space, and the decoder reconstructs the image from the compressed representation. Guo et al. [22] proposed an encoder-decoder network using full-resolution residual blocks. For face inpainting, they reported a PSNR of 29 and an SSIM of 0.95. On the other hand, Liu et al. [35] achieved a PSNR of 34.69 and an SSIM of 0.99 by adding a coherent semantic attention layer to the encoder. One disadvantage of this method is its long runtime of 0.82 seconds per image of size 256×256 rendering this method inapplicable for real-time video processing with more than one frame per second. Another problem with existing CNN-based methods is that the convolution operations are applied both to the valid and missing pixels at the same time, which can lead to visual artifacts (e.g., color discrepancy and blurriness). To overcome this issue, Liu et al. [34] proposed partial convolutions, where the convolution operations are only applied to valid pixels by masking regions that need to be inpainted. The mask is updated during training of the network, including newly inpainted values. The authors demonstrated that the approach could produce semantically meaningful predictions also for inpainting regions with different shapes and sizes, achieving a PSNR of up to 34.34 and an SSIM of up to 0.95. We use this partial convolution approach to inpaint missing regions in images from front camera recordings. The dataset used for training the network is tailored to our use case.

Affective State Prediction. In our work, we focus on the prediction of affective states in the educational domain, such as in classroom settings and online courses. It was shown that affective states have an impact on learning gain in general, and during math learning in particular [29, 44]. For example, Csikszentmihalyi [13] showed that engaged concentration has a positive effect on learning, while boredom negatively influences learning. Affective states are often grouped into basic emotions identified by Ekman [16] (i.e., anger, disgust, fear, happiness, sadness, and surprise) or described by the valence and arousal dimensions [40]. Valence indicates if an emotion is perceived as positive or negative, and arousal represents the intensity of an emotion.

Different modalities have been used to predict affective states using the valence-arousal space in educational settings. Acoustic features from student voices during interaction with tutors have been used to predict three levels of valence [33]. On the other hand, bio-sensors (i.e., skin conductance, heart rate, and skin temperature) and handwriting data have been successfully used to predict affective states in the valence-arousal space during math solving tasks [53]. Another line of research predicted valence and arousal using mouse and keyboard interaction data collected during text writing [49]. Multi-modal approaches fusing different modalities have also been introduced for the prediction of affective states. We refer to D'Mello et al. [15] that provides a concise overview of such methods.

Prediction of affective states from video recordings is one of the most popular approaches nowadays as it allows different features to be exploited, such as body language and posture, head movement, eye gaze and facial expressions [57]. Bosch et al. [6] calculated statistics (i.e., maximum, median and standard deviation) of the frame-level likelihood values of 19 different action units (AU) (i.e., facial muscle movements), the head position and gross body movement from webcam video recordings of students playing an educational physics game. They predicted two levels of boredom (0.61 AUC), confusion (0.65 AUC), delight (0.87 AUC), engagement (0.68 AUC) and frustration (0.63 AUC). Based on this work, Kai et al. [26] found that an interaction-based model using timing and counting-based features performs worse than the video-based model. Similarly, using a math tutor, Arroyo et al. [2] found facial expressions to be more predictive for confidence, frustration, excitement, and interest than conductance bracelets, pressure mice, and a posture analysis seat. Also in other domains facial expressions have found to be a good predictor for affective states. In text comprehension tasks, confusion (0.64 AUC), engagement (0.55 AUC), and frustration (0.61 AUC) have been successfully predicted using 20 different AUs [11]. On the other hand, Grafsgaard et al. [20] found upper face movements predictive for engagement, frustration, and learning in a setting consisting of a programming tutor and a webcam. Finally, based on eye gaze features (e.g., fixation and view angle) extracted from a specialized eye capturing device, boredom (69 %) and curiosity (73 %) have been successfully predicted on two levels each [25]. A survey of different video-based approaches for predicting affective states is provided by Zeng et al. [57].

A majority of the existing vision-based approaches use external devices, such as webcams, and rely on posed facial expressions to predict basic emotions [57]. In contrast, we present a novel setup for reliably recording the face of users based on the front camera of tablet computers only, and hence without the need for expensive devices or synchroniza-

tion between the devices. We demonstrate the usefulness of our setting by predicting affective states in terms of valence and arousal using data from an experiment containing spontaneous (non-posed) facial expressions. Finally, for our vision-based model, we fuse different existing approaches with novel features.

3. CAMERA SETUP

We present a low-cost hardware setup for recordings from the integrated front camera of a tablet computer, maximizing the visibility of the face of the users. Videos and images captured by the front camera are preprocessed, and missing parts are inpainted using a deep learning model to reconstruct the face of the users. Our approach is image-based and processes captured videos frame by frame.

3.1 Hardware Setup

While working on a tablet (e.g., writing with a stylus) it is convenient to have the device lying on the table (see Figure 1a). Due to the field of view of the front camera, only part of a users' face is visible. To adjust the field of view of the front camera, we attached a circular mirror (3 cm radius) to the tablet using a hinge (see Figure 1b). The hinge was fixed with glue so that the mirror would remain in a stable position. The mirror was mounted with an angle of 75 degrees relative to the tablet. This angle was chosen so that the visibility of the face was maximized. Due to the mirror setup, the upper part of the recordings is mirror-inverted (see Figure 1c). Depending on the conditions of the illumination of the recording environment, the exposure time of the camera of the recording device (e.g., tablet) needs to be adapted accordingly so that the camera focuses on the face instead of the background. This adjustment of the exposure time can lead to an overexposed background (see Figure 1c).

3.2 Image Processing Pipeline

A raw image captured by the front camera is split by the mirror into two parts with the upper part of the image being mirror-inverted (see Figure 2A). To reconstruct the image, we propose a series of processing steps applied to the image (i.e., flattening the splitting boundary, face composition, image rotation, and extracting the face area). Image rotation and extraction of the face area are conducted as a preprocessing step for inpainting. Further, to train our inpainting model at a later stage, we assume that we have access to a dataset Ψ of square-shaped face images.

Splitting boundary. We apply a transformation to flatten the splitting boundary of the image (green line in Figure 2A), which simplifies image processing in the later stages and improves the final results qualitatively. We divide the image into 16 rectangles with equal width. An example of such a rectangle is shown in purple in Figure 2A. For each such rectangle, we transform the region defined by the vertices p_1, p_2, p_3, and p_4 into the region defined by the vertices p_1, p_2, p_5, and p_6 using a perspective transformation with linear interpolation. The location of these points can be calculated beforehand (or read from the image) because the mirror remains in a fixed position. The result of the transformation is shown in Figure 2B, where the splitting boundary (green) is a straight line.

Face composition. We rearrange the image by moving the part below the splitting boundary to the top and the flipped upper part to the bottom (see Figure 2C). The cut line defined by the mirror is shown in black. In addition, we adapt the height of this cut line because depending on the distance of the face, the missing part is increasing (increasing distance) or decreasing (decreasing distance). As a next step, we push the bottom corner of the upper face towards the middle by applying a second perspective transformation to the image so that the upper and lower part of the face are matching (see Figure 2D).

Image rotation. We then rotate the front camera image so that the eyes are horizontally aligned (see Figure 2E). Using dlib [28], we extract the coordinates of the facial landmarks belonging to the left and right eye. From these landmarks, we calculate the position of the center of each eye and rotate the image around the midpoint between the eye centers so that the line connecting the center of the eyes is horizontally aligned.

Face area. We extract the face area by computing a square bounding box encompassing the face (see the orange box in Figure 2E). This bounding box is defined by the vertices $p_7 = (x_7, y_7)$ and $p_8 = (x_8, y_8)$ and is given by

$$x_7 = c_{x,I} - \frac{w_{I_\Psi}}{2} * \frac{\delta_I}{\delta_{I_\Psi}} \tag{1}$$

$$x_8 = c_{x,I} + \frac{w_{I_\Psi}}{2} * \frac{\delta_I}{\delta_{I_\Psi}} \tag{2}$$

$$y_7 = c_{y,I} - \frac{c_{y,I_\Psi}}{h_{I_\Psi}} * (x_8 - x_7) \tag{3}$$

$$y_8 = c_{y,I} + \frac{h_I - c_{y,I_\Psi}}{h_{I_\Psi}} * (x_8 - x_7), \tag{4}$$

where I and I_Ψ denote an image of the front camera and an image in the dataset Ψ, respectively. The width and height in pixels of an image are given by w and h. The x- and y-coordinate of the midpoint between the left and right eye are denoted by c_x and c_y, respectively, and δ is the distance between the eyes. Here, we assume that the origin is located at the top left of the image.

The part of the front camera image I outlined by the orange bounding box is then resized to the resolution $w_{I_\Psi} \times h_{I_\Psi}$ using bilinear interpolation. If the head of the user is close to the mirror, the face covers the full height of the image, and the bounding box might go over the upper and/or lower image borders. In such a case, we fill the parts overlapping the image with black pixels to get consistently sized bounding boxes (note that for visualization purpose only, the orange box in Figure 2E does not reflect this but instead is cut at the image border). We use the face detector of dlib [28] to test if a face and hence the landmarks of the eyes are identified in the image. In cases where the face cannot be detected, we use the landmarks of the eyes of the last image where the face could be identified (assuming that we have a video recording available, i.e., a series of images).

Inpainting missing area. As the last step in our image preprocessing pipeline, we inpaint the missing parts in the bounding box of the image (black region of the orange box in Figure 2E) with the neural inpainting approach of Liu

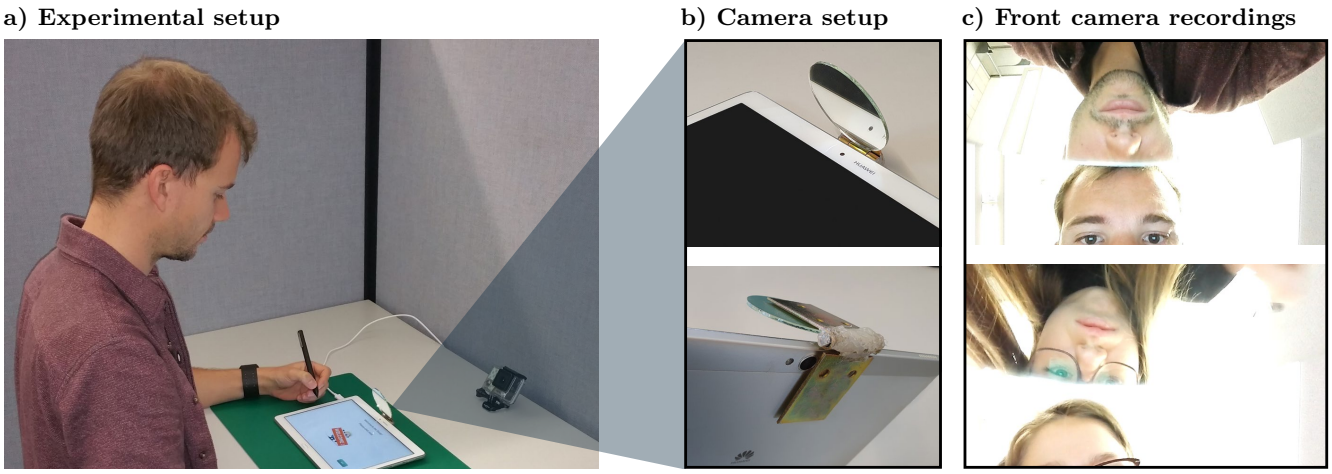

Figure 1: The hardware setup. A user is working on the tablet (a). A mirror is attached to the tablet using a hinge (b). Due to the mirror reflections, the field of view of the front camera is changed so that the face of the participant is visible (c).

et al. [34] described in Section 3.3. We apply the neural inpainting only to the bounding box because it contains the important parts of the face (i.e., eyebrows, eyes, and mouth). We inpaint other parts of the image outside the bounding box using a simple Navier-Stokes based inpainting method provided by OpenCV [8] which is based on a circular neighborhood of three pixels for each inpainted pixel. Finally, we rotate the image back to its original orientation. This then leads to the final reconstructed image shown in Figure 2F.

3.3 Neural Inpainting

For the neural inpainting approach, we use the dataset Ψ of square-shaped face images with customized missing regions tailored to our application of tablet front camera recordings and then train the network on this dataset.

Training dataset. The model is trained on a large corpus of images from the dataset Ψ together with a mask for each image indicating the missing parts (a mask is a matrix with the same size as the image having a $'1'$ entry for missing pixels and a $'0'$ entry otherwise). We create the corresponding mask randomly and similar in shape (rectangle) to the expected mask in our front camera recordings (see Figure 3 for an example of two such masks applied to two images from the CelebA-HQ dataset [27]). Note that the mask (missing image region) is not necessarily horizontal but rotates if a user is rotating the tablet or the head (vertical in the extreme).

Inpainting method. Liu et al. [34] use a neural network that consists of an encoder E and a decoder D. The encoder network transforms the input image $\mathbf{I} \in \mathbb{R}^{M \times N}$ into a low-dimensional (latent) space $\mathbf{z} = E(\mathbf{I})$. The decoder then reconstructs the original image based on this low-dimensional representation $\hat{\mathbf{I}} = D(\mathbf{z})$. The encoder and decoder networks consist of $n = 8$ partial convolutional layers denoted as $E_1, \ldots, E_n$ and $D_1, \ldots, D_n$ for the encoder and decoder networks, respectively. Before each convolution operation, the image is constrained by the mask to condition the operation on only valid pixels. The mask is updated for the next layer removing masking for pixels where the convolutional operation operated on unmasked values. In addition, each layer in the encoder network E_i is connected to the corresponding layer in the decoder network $D_i, \forall i \in \{1 \ldots, n\}$ using skip links. These skip links allow for copying unmasked pixels directly from the encoder to the decoder without passing the bottleneck (latent space). To direct the training of the network towards semantically meaningful inpaintings, a combination of four loss functions is used (i.e., per-pixel loss, perceptual loss, style loss, and total variation loss). Using these loss functions smooth transitions of the predicted masked values into their neighboring pixels is also taken into account. As activation functions Rectified Linear Unit (encoder) and a leaky version of a Rectified Linear Unit (decoder) are used.

4. AFFECTIVE STATE PREDICTION

We present the prediction of affective states as an example application of our mirror setup and image processing pipeline. Our classification pipeline can be generally applied to any recordings captured with a tablet front camera or an external camera (such as a GoPro). Our method assumes that we have access to reports of affective states of users based on the circumplex model of affect [48]. The circumplex model defines affective states in a two-dimensional space spanned by valence and arousal. The classification task then amounts to preprocessing the camera recordings to adjust the brightness and the frame rate and predicting valence and arousal based on features extracted from the adjusted camera recordings. Affectiva [39] provides out of the box predictions of the basic emotions and valence based on images and video recordings. However, initial tests revealed that these predictions are not of sufficient quality when applied to our use case. Thus, we developed our own set of features incorporating some additional features not taken into account by Affectiva, such as movement and fidgeting. Moreover, by using our own extracted features, we can predict arousal in addition to valence.

4.1 Preprocessing

First, we resample the camera recordings using FFmpeg [5] to a constant frame rate close to the mean frame rate. Depending on the recording device, the frame rate can vary (e.g., the frame rate can drop due to the higher load of the

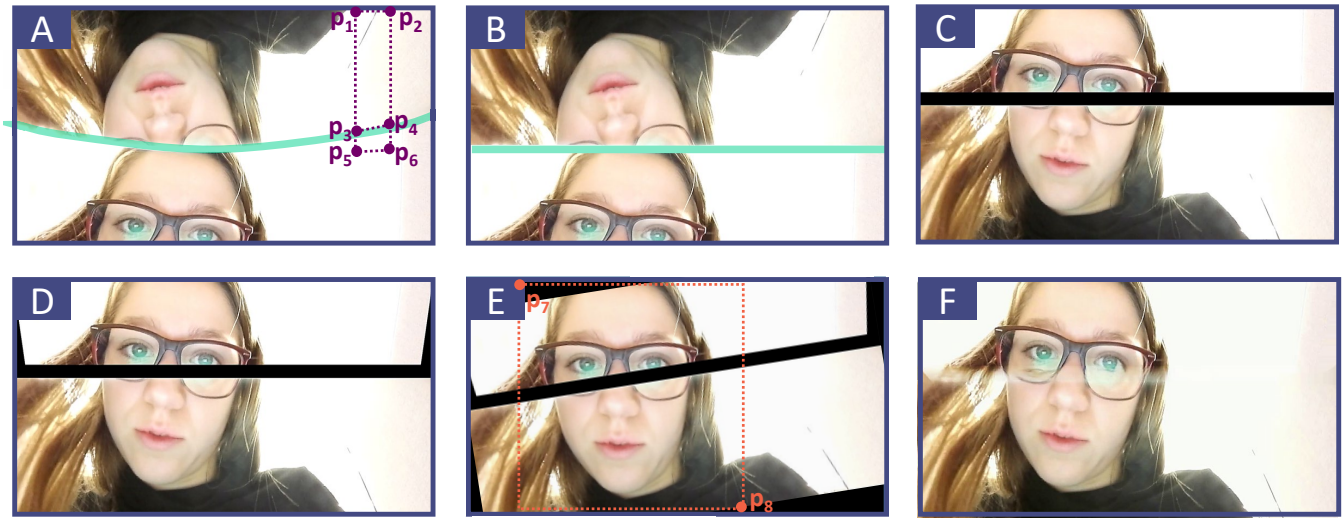

Figure 2: The main inpainting steps. The splitting boundary of front camera recordings (A) is flattened using a perspective transformation (B). The face is reconstructed from the upper and lower parts (C) and warped so that the upper and lower part match (D). Finally, after horizontally aligning the eyes (E), the missing regions (black) are inpainted (F).

Figure 3: Two example masks applied to images of the CelebA-HQ dataset [27].

device). A constant frame rate facilitates the extraction of the features and the processing of the recordings in later stages. In addition, we adjust the brightness of the recordings based on the brightness estimation of Affectiva [39] to improve the lighting of the face for the analysis. Depending on the conditions of illumination at recording time the face can be underexposed (too dark) or overexposed (too bright, e.g., when the camera is directed towards a lamp). This can hinder the accurate detection and extraction of facial features such as landmarks.

4.2 Feature Extraction

From the camera recordings, we extract several different feature types. We design all features such that they are independent of the frame rate (e.g., using percentages instead of absolute positions) to support cameras with different frame rates. To extract facial landmarks, eye gaze, and head position from the camera recordings, we rely on OpenFace [4] using static extraction (i.e., per frame without calibrating to a person). OpenFace also provides a confidence value $c(i) \in [0, 1]$ for each frame i indicating the confidence in the landmark detection estimate. If $c(i) < 0.82$, we discard the frames $i-5, \ldots, i+5$ (i.e., 11 frames). The number of frames to discard (11) and the threshold (0.82) were heuristically determined. All features are computed over a window containing N frames. If, after considering the confidence value, less than 80 % of the frames are remaining, we discard the window and the corresponding data point. Again, this threshold was determined heuristically. Where appropriate, we calculate for the different feature types basic statistics over the window (i.e., maximum, minimum, relative position of minimum and maximum, mean, standard deviation, and the slope of a fitted linear regression line), providing 282 features in total. In addition, to correct for differences between individuals related to facial expressions and posture, we normalize each feature according to a baseline by subtracting the feature calculated over a baseline period (e.g., watching a nature video putting the individuals in a relaxed state).

Action units. Facial action units (AUs) are based on the Facial Action Coding System (FACS) and identify independent motions of the face [17]. We extract basic statistics of the intensity (from 0 to 5) of 17 AUs covering motions in the eye, cheek, nose, mouth, and chin region. In addition, for each AU, we calculate the percentage of the presence (absent versus present) in the window. Moreover, the AUs can be directly mapped to the six basic emotions identified by Ekman [16]. Thus, for each basic emotion, we also calculate the basic statistics of the corresponding added up AUs.

Eye blinks. Researchers have found a correlation between eye blink frequency and stressful situations in a car driving simulation [23]. Similarly, a correlation between eye blinks and affective states in learning environments was found [38]. Here, we base the eye blink detection on the signal from the AU that represents eye closure as a continuous signal (from 0 to 5) with peaks indicating potential eye blinks. We detect peaks belonging to an eye blink by thresholding the signal according to the ratio between the prominence (how much a peak stands out measured as the vertical distance between the peak and its lowest contour line) and width of a

a) Eye gaze regions b) Mouth aspect ratio

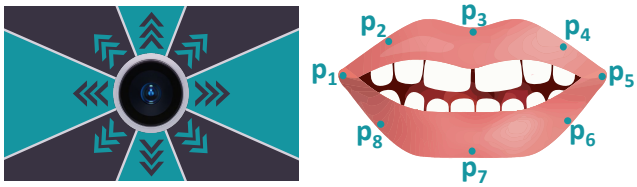

Figure 4: Eye gaze regions and mouth aspect ratio (MAR). The gaze angle is discretized into nine different gaze regions, including the center (gazing towards the camera lens) (a). MAR is calculated based on the height and width of the mouth (b).

peak. Heuristically, we found a threshold of 0.026 to provide the best results. We found that taking into account the width of the peaks is necessary to accurately detect peaks belonging to eye blinks because the prominence of the peaks differs among users and head pose. We extract the number of blinks and the basic statistics of the duration between blinks, the prominence, and the width of each blink. In addition, inspired by interbeat intervals (time interval between individual heartbeats) and the calculation of heartbeats thereof, we linearly interpolate the duration between two consecutive peaks surviving the threshold (i.e., eye blinks) to infer a continuous signal. We then calculate the number of eye blinks for every frame by taking the inverse of this interpolated signal. Subsequently, we again calculate the basic statistics over the number of eye blinks.

Eye gaze. The intention behind features related to eye gaze is that individuals might look away when thinking while solving math tasks or when looking at emotionally disturbing pictures. Thus, we compute the basic statistics on the angle in the x-direction (looking left-right) and y-direction (looking up-down) of the eye gaze averaged for both eyes and measured in radians in world coordinates. In addition, we discretize the eye gaze angle by defining nine different gaze regions (see Figure 4a). The center corresponds to a line of gaze directed towards the camera lens. For each of the nine regions, we count the number of occurrences and normalize it over $s * \text{fps}$, where s is the window size and fps is the frame rate per second (so that it is independent of the used camera, i.e., the frame rate).

Mouth aspect ratio. Previously, the mouth aspect ratio (MAR) was used to detect driver drowsiness [52]. It is defined by the ratio between the height and the width of the mouth, which is increased when opening the mouth (see Figure 4b):

$$\text{MAR} = \frac{\|p_2 - p_8\| + \|p_3 - p_7\| + \|p_4 - p_6\|}{3 * \|p_5 - p_1\|}. \quad (5)$$

Each point $p_i, \forall i \in \{1, \ldots, 8\}$, is defined as the average of the inner and outer mouth landmarks. From the MAR, we calculate the basic statistics.

Head Movement. From the longest head moving sequence of an individual in the window, we extract the position of the first frame of the sequence in relation to the beginning of the window, the duration of the movement, and the total distance of the movement. The position of the first frame and the duration are normalized by $s * \text{fps}$. We also sum up the total

a) Original b) Fidgeting

Figure 5: Fidgeting of a user. From the original image (a), the fidgeting image (b) is calculated by pixel-wise thresholding the difference of the current (a) to the past grayscale images.

distance moved over the entire window to capture individuals continually moving back and forth. In addition, we calculate the basic statistics of the velocity and acceleration of the head movements in the window. All these features are extracted for the x-axis, y-axis, and z-axis separately. Finally, we also extract the basic statistics of the distance of the head to the camera in the three-dimensional space.

Fidgeting. Navarathna et al. [42] introduced a fidgeting index for predicting movie ratings from audience behavior by calculating the total energy individuals are using for the movement. In contrast to features related to the head movement, fidgeting captures all the movement in the video (i.e., also body and face). First, we define the grayscale adaptive background b_{gray}, which is a weighted average of past frames. To calculate the energy E for a new frame f_{gray} (converted into grayscale), we subtract the adaptive background b_{gray} from f_{gray}, binarize the image by thresholding it, and then calculating the percentage of surviving pixels with respect to the camera resolution (see Figure 5b). We have chosen the threshold such that noise from the background is minimized, and the visibility of movements is maximized. Finally, the adaptive background is updated using

$$b_{\text{gray}} = (1 - a) * b_{\text{gray}} + a * f_{\text{gray}}, \quad (6)$$

where a is a weight term (we found $a = 0.2$ to provide the qualitatively best results). From the energy E of each frame in the window, we calculate basic statistics, sum up the energies over all frames and use the position of the frame with minimum and maximum energy normalized by $s * \text{fps}$.

4.3 Classification

We build the ground truth for our classifiers by splitting valence and arousal into two levels (high and low). We then use classifiers to predict these levels based on the features extracted from the camera recordings. In addition, we remove features having a correlation greater than a threshold, select features based on the ANOVA F-value between the class labels and the features, and standardize the features to have zero mean and unit variance. We use four different classifiers (i.e., Random Forest, Support Vector Machine, k-Nearest Neighbors and Gaussian Naive Bayes) because these classifiers have been most promising in initial tests and they have shown to provide good results for predicting affective states from video data in other works [25, 10, 6]. We use leave-one-user-out cross-validation to evaluate our models, which ensures that data of a participant is not used for training and testing at the same time. Finally, we optimize the hyperparameters (i.e., number of selected features, the threshold for

removing correlated features, and parameters of the model) using random search with nested cross-validation.

5. RESULTS

We conducted a qualitative and quantitative evaluation of our mirror setup and image processing pipeline with neural inpainting and investigated the applicability of our setup to predict affective states during math-solving tasks (active) and during exposure to emotional stimuli from images (passive). For training the neural inpainting model, we have used the celebA-HQ dataset [27] consisting of 30000 face aligned colored images from celebrities with a resolution of 1024×1024 pixels (we downsampled the images to 512×512 pixels). We split the dataset into a training set of 25000 images, a test set of 2500 images and a validation set of 2500 images. We set the parameters for the network in the same way as proposed by Liu et al. [34]. The results of the affective state prediction are based on a Random Forest classifier since this was the best performing model. Hyperparameters were optimized using random search with 50 iterations. Finally, for measuring the performance of our model, we used the area under curve (AUC) of the receiver operating characteristic curve and accuracy (chance level = 0.5).

5.1 Experiment

We reused a dataset that we collected in a controlled lab experiment [53]. The dataset consists of data from 88 participants (45 female) from age 18 to 29 (mean = 22.1, standard deviation SD = 2.0) of university students in the bachelor program. The participants used a Huawei MediaPad M2 10.0 tablet running Android 5.1 during the experiment. They were recorded by the front camera (resolution of 1280×720 pixels) using our proposed mirror construction setup and a GoPro HERO3 camera (frame rate per second FPS of 59.94 and a resolution of 1920×1080 pixels) (see the setup in Figure 1a). Due to the varying load of the tablet during the experiment, the fps was variable (mean = 20.02, SD = 1.92). We resampled the recordings from the tablet and the GoPro to an fps of 25 and 60, respectively. To synchronize the timestamps between the GoPro and the tablet, a beep signal was played on the tablet before the start of each session.

The study procedure consisted of three main steps conducted on the tablet to collect baseline data and trigger different affective states. First, each participant was watching a seven minutes nature video, which served as a baseline. Second, the participants were presented 40 pictures in random order from the International Affective Picture System (IAPS) [30] for around 20 minutes. The IAPS is a collection of 1182 pictures standardized in terms of valence and arousal and is widely used in psychological research for the study of emotions. Each image was shown for ten seconds and was followed by a ten seconds fixation cross. The 40 images have been selected from the IAPS dataset such that a wide range of the valence-arousal space was covered.

Finally, each participant solved multiple-choice math tasks for approximately 30 minutes. The math tasks were selected from a collection of math tasks provided by ACT [1] and divided into three different conditions varying in difficulty level, available completion time, and monetary reward (participants were rewarded and penalized depending on the correctness of the solution and started with a credit of CHF 40). In the repetitive condition, easy and repetitive (i.e., similar) tasks were presented with more than enough available time to solve the tasks and a minor reward (+CHF 0.2) and penalty (−CHF 0.2). In the challenge condition, tasks with medium difficulty levels were shown with sufficient time to solve the tasks, and a large monetary reward (+CHF 2) but an only minor penalty (−CHF 0.2). The overchallenge condition consisted of tasks with a high difficulty level, insufficient time to solve the tasks and a small monetary reward (+CHF 0.2) but a large penalty (−CHF 2). The tasks were presented in six blocks. Each block contained tasks from a specific condition, and each condition was assigned randomly to two blocks.

After each image and math task, participants were asked to fill in the self-assessment manikin (SAM) [7] to judge their current valence and arousal level on a 9-point Likert scale. To build our affective prediction model, we split the valence and arousal ratings of the participants into two classes (low $\in \{1, \ldots, 3\}$ and high $\in \{7, \ldots, 9\}$). For IAPS, the number of data points amounted to 843 (1206) and 1218 (982) for low and high valence (arousal), respectively. On the other hand, for math tasks, the number of low and high valence (arousal) ratings amounted to 724 (1380) and 1422 (726), respectively.

5.2 Face Recognition

We provide qualitative and quantitative results of our setup using neural inpainting. In particular, we compare our results to recordings taken by the GoPro camera.

Qualitative evaluation. Figure 6 shows the facial landmarks detected by OpenFace for three participants from the front camera without inpainting, using neural inpainting, and from the GoPro. The positions of the detected landmarks without inpainting are inferior compared to neural inpainting. For participant 3, the landmarks at the upper face (eyebrows, eyes, and nose) are misaligned without inpainting. Often no facial landmarks could be detected (see Figure 6 participants 1a and 2a). With our neural inpainting approach, we achieved a qualitatively good recovered image independent of the position of the missing region (e.g., eyes and mouth). It is noteworthy that the inpainting and facial landmark detection also worked for participants wearing glasses. The detected landmarks after neural inpainting are similar to the landmarks detected from the GoPro recordings (see Figure 6c). Depending on the position of the head, the landmarks of the eyes and the mouth can become locally condensed in the GoPro recordings, and it might be hard to distinguish slight facial movements. On the other hand, from the front camera, the recordings are frontal, and the variations of facial parts (e.g., eye and mouth) are better visible.

Quantitative evaluation. Table 1 presents the average confidence in landmark detection of OpenFace over all frames for the IAPS and math-solving tasks and the full recordings (including also parts not belonging to the IAPS and math tasks). Reported confidence values by OpenFace are between 0 (not confident) and 1 (fully confident). Without inpainting, the confidence values are low, and standard deviations are high due to the imperfect recognition of landmarks. Without inpainting landmarks were often only detected correctly

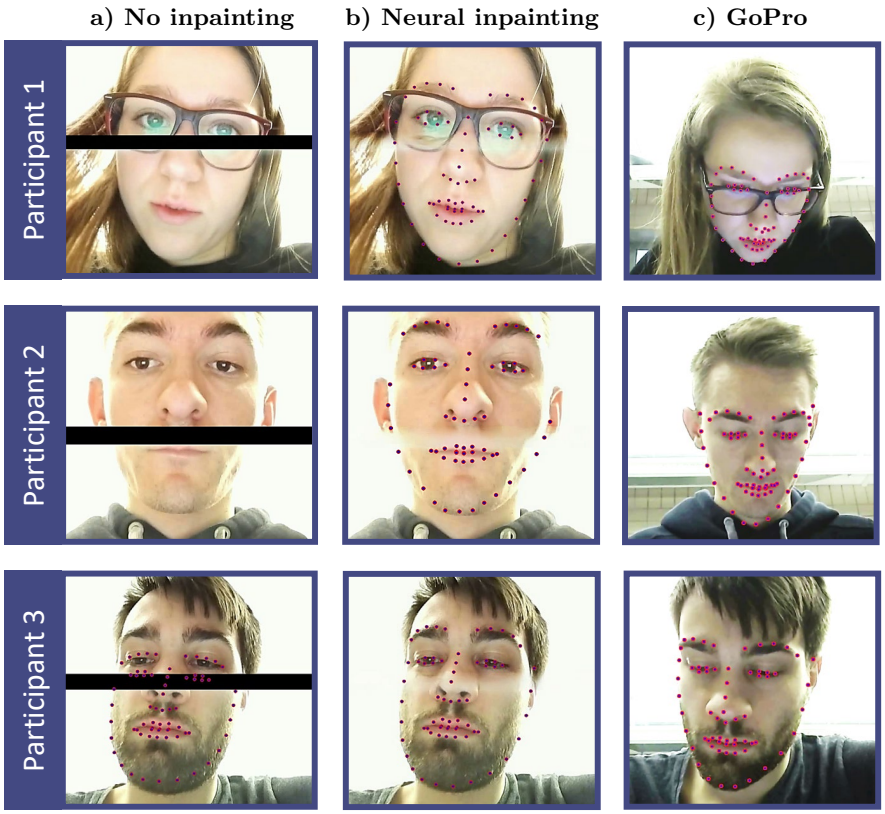

Figure 6: Recordings of three participants. The facial landmarks were detected from the front camera recordings without inpainting (a) and with neural inpainting (b) and from the external GoPro camera (c). If no landmarks are visible, no landmarks were detected by OpenFace.

Table 1: Means of framewise confidence in landmark detection for different camera sources, tasks (math and IAPS) and the full recordings. Confidence values range from 0 (not confident) to 1 (fully confident). Standard deviations are given in brackets.

Source	IAPS	Math	Complete
Front (no inpainting)	0.79 (0.36)	0.48 (0.45)	0.68 (0.42)
Front (inpainting)	0.94 (0.14)	0.90 (0.22)	0.93 (0.18)
GoPro	0.97 (0.08))	0.93 (0.17)	0.95 (0.12)

when the missing regions were situated above the eyebrows (i.e., no landmarks have been affected). After applying neural inpainting, the confidence values increased by 19 % and 88 % during IAPS and math sequences, respectively. When considering the full video recordings, the increase amounts to 37 %. In addition, the standard deviation decreased substantially. This increase of the confidence leads to an increase in the number of samples (if a window used during feature extraction contained less than 80 % frames with a confidence value above 0.82 we discarded the corresponding data point). For IAPS, this lead to 348 and 383 additional samples for valence and arousal, respectively. For the math tasks, this amounted to 1233 and 1179 additional samples for valence and arousal, respectively. Finally, the confidence in landmark detection of the GoPro recordings is comparable to the front camera recordings with neural inpainting. In general, for recordings taken during exposure to a stimulus set of images

Table 2: Performance of Random Forest on the math and IAPS data from two levels (low and high) of valence and arousal based on the front camera recordings with neural inpainting and the GoPro recordings. The chance level for accuracy and AUC is 0.5.

Source	Data	AUC	Accuracy
Front camera	Math (valence)	0.73	68 %
	Math (arousal)	0.54	57 %
	IAPS (valence)	0.80	73 %
	IAPS (arousal)	0.70	66 %
GoPro	Math (valence)	0.76	72 %
	Math (arousal)	0.58	62 %
	IAPS (valence)	0.78	72 %
	IAPS (arousal)	0.73	67 %

the mean confidence is higher than during math tasks. This can be attributed to the fact that while solving math tasks, participants were moving more, which leads more often to suboptimal head positions for landmark detection. This finding is also reflected in the higher standard deviations of the confidence values for math tasks.

5.3 Classification Performance

Before predicting the affective states, the reconstructed front camera recordings and the GoPro recordings were preprocessed (see Section 4.1). Features were extracted using a ten seconds window encompassing the on-screen time of each pic-

Table 3: Number of occurrences of each feature type in the ten most predictive features. The numbers are provided for each of the four models (MV = math valence, MA = math arousal, IV = IAPS valence, IA = IAPS arousal).

Feature Type	MV	MA	IV	IA
Action units	0	2	2	3
Eye blinks	1	4	0	1
Eye gaze	1	2	2	0
Mouth aspect ratio	0	0	0	0
Head Movement	5	2	5	6
Fidgeting	3	0	1	0

ture and the last ten seconds of each math task because each picture was presented for ten seconds and the minimum task duration was ten seconds (see Section 4.2). Table 2 presents the performance of our model for predicting two levels (low and high) of valence and arousal. Based on the findings that the confidence in landmark detection increased up to 88 % with neural inpainting, we used only the front camera recordings with neural inpainting. Using these recordings, our model achieved a performance of 0.73 AUC and 0.80 AUC for predicting valence on math tasks and IAPS, respectively. For predicting arousal, the performance drops and is only at random level for math tasks (0.54 AUC), while for IAPS it is above random (0.70 AUC). A similar pattern is visible for the GoPro recordings. While for predicting arousal based on the math tasks, the performance is close to random (0.58 AUC), all other predictions are above random. In summary, the predictions using the front camera are comparable to using the GoPro recordings with a maximum difference of 0.04 AUC. For predicting valence based on IAPS, the performance from the front camera recordings (0.80 AUC) exceeds the performance achieved by using the GoPro (0.78 AUC).

Feature importance. Table 3 presents the number of occurrences of each feature type in the ten most important features for each of the four models. We analyzed the feature importance using the Gini importance measure provided by the Random Forest classifier. Features related to head movement contributed the most for predicting valence based on math tasks (five features) and valence and arousal based on IAPS (five and six features). For predicting arousal based on math tasks, eye blinks provided four out of the ten most important features. There were no MAR features among the top ten features for any model. However, all feature types appeared in the top 30 ranked features of each model. For the model based on the math tasks, the maximum moved distance in the x-direction and the number of eye blinks were the highest scoring features for predicting valence and arousal, respectively. For the model based on IAPS, the mean acceleration in the x-direction and mean velocity in the x-direction were most important for predicting valence and arousal, respectively. Interestingly, head movement along the x-axis (left and right) was more informative than along the z-axis (forward and backward).

5.4 Runtime

We conducted a runtime analysis of the different parts of our inpainting pipeline and affective state prediction model. Our computing environment consisted of an Intel® Core™ CPU i9-9900K @ 3.60GHz and an NVIDIA GeForce® RTX 2080 Ti. Processing one frame consisted of flattening the splitting boundary, face composition, image rotation and extracting the face area (mean = 17.07 ms, SD = 4.74 ms), detecting the position of the eyes using dlib (mean = 74.66 ms, SD = 6.43 ms), using the deep learning model to inpaint missing regions in the face (mean = 76.25 ms, SD = 13.81 ms) and inpainting the background of the image (mean = 47.01 ms, SD = 11.87 ms). Summing up these values leads to a processing time for one frame of 214.99 ms. Prediction of a new data point consisted of feature extraction (mean = 16.37 ms, SD = 2.18 ms) and using the Random Forest classifier for predicting valence and arousal (mean = 6.43 ms, SD = 10.52 ms), leading to a total prediction time of 22.8 ms.

6. DISCUSSION

Our findings show that it is possible to use our tablet-based front camera setup and processing pipeline to accurately capture users for extracting features such as facial landmarks and movement of the head and body. Our neural inpainting pipeline provides a qualitatively accurate restoration of missing regions caused by our mirror construction setup and increases the confidence in landmark detection by up to 88 %. Compared to recordings from a GoPro camera, our setup provides better results in terms of face visibility (frontal view). Thus, it potentially facilitates the recognition of minor facial movements (e.g., mouth and eyes). In particular, for solving math tasks we found the recording conditions of the GoPro more challenging due to the viewing angle (participants were bending over the tablet). This resulted in lower confidence in landmark detection (0.93 for math tasks versus 0.97 for IAPS). Similarly, the front camera recordings with neural inpainting showed higher confidence in landmark detection during IAPS (0.94) compared to solving math tasks (0.90). During the exposure to a stimulus set of images from the IAPS dataset, participants were sitting straight, implicating that the splitting boundary was located at the forehead, which made inpainting easier. In contrast, during solving math tasks, the splitting boundary was often located in the middle (eye) or lower part of the face (mouth), creating a more challenging situation for our neural inpainting model.

We showed the applicability of our setup for predicting affective states during active (math-solving) and passive (exposure to pictures) tasks based on the recordings from the front camera. Our model achieved better performance on IAPS (up to 0.80 AUC) than on the math tasks (up to 0.73 AUC). Due to the active involvement of the participants while solving math tasks, participants were moving more, which made accurate tracking of facial landmarks, AUs, and eye gaze more demanding. In addition, our model performed better for predicting valence (0.73 AUC and 0.80 AUC) than arousal (0.54 AUC and 0.70 AUC). One-third of the participants rated arousal constantly as low or high without showing much variation. This finding can affect the generalization of our model to other participants for predicting arousal. In addition, although affective states are universal, they also have components that are individual to a person [18]. This makes it harder to predict an affective state of a person without having training data available of that person. Comparing the performance of our affective prediction pipeline to other research is difficult because most existing work [10, 57] predicted basic emotions and used other settings.

Our analysis of the feature importance showed that head movement is a predictive feature in contrast to MAR. Some AUs capture movements of the mouth. Thus, we analyzed the correlation between MAR and AUs specific to the mouth region. The correlations between the MAR feature and the AUs specifying lip corner puller (-0.15, p-value $= 0.15$), opening the mouth (0.25, p-value $= 0.13$) and jaw drop (0.045, p-value $= 0.26$) have all been low and not significant.

In comparison to recordings from the GoPro, our model based on front camera recordings performed equally well and even better for predicting valence on IAPS (0.80 AUC versus 0.78 AUC). This renders our setup a viable alternative to more expensive equipment such as a GoPro. Our setup comes at low costs (CHF 5), is unobtrusive, can easily be mounted, is flexible in the application (e.g., in classrooms or at home), and eliminates the need for synchronizing different devices. In contrast to external cameras, the camera (i.e., the lens) in our setup is small and unobtrusive. Some participants reported after the experiment that they got slightly distracted by the GoPro but not by our mirror setup. Similarly, in the video recordings, we recognized that participants were sometimes glancing at the GoPro. Finally, with a processing time of 214.99 ms per frame, our pipeline can handle four frames per second. Our affective prediction pipeline is capable of making 43 new predictions every second.

Limitations. We acknowledge potential limitations to our approach presented in this paper. Our setup is constrained by the lighting conditions, head pose, and occlusions from hand movement. We believe that other camera setups suffer from the same constraints. Further, our mirror construction is a prototype and not yet ready for production. Although during the experiment the construction proved to be stable, it can be improved in terms of stability and flexibility. Neural inpainting provided qualitatively satisfactory results for most facial parts. However, if the splitting boundary is covering the eyes (i.e., both eyes are occluded), it is hard for the inpainting model to reconstruct the eyes at a qualitatively high level. Consequently, the landmark detection cannot recover eye gaze and eye blinks, but still detects other facial features. In addition, although the CelebA-HQ dataset consists of facial images from celebrities with diverse ethnicity, age and facial characteristics (e.g., glasses and facial hair), our inpaining method might be less appropriate for students who are underrepresented in the CelebA-HQ dataset. We further acknowledge that our experiment is restricted to math tasks and exposure to emotional stimuli from pictures in a lab environment with bachelor students. We are optimistic that our approach generalizes to a broader population and to other tasks given that we used active (math-solving) and passive (exposure to pictures) tasks and assuming a proper baseline normalization of the features. Finally, we have predicted valence and arousal on two levels omitting data points in the medium range (4 to 6). Our main contribution is the novel mirror construction and the processing pipeline. We have mainly built our affective prediction model for demonstrating the applicability of our setup. Nevertheless, we believe that our features and pipeline can be interesting for other researchers predicting affective states based on video data.

Future work. Future research comprises refining and extending our hardware setup and inpainting pipeline, as well as evaluating our affective prediction model in other domains. In particular, realtime performance would be desirable for on-the-spot assessment of a student's affective state. The CelebA-HQ dataset, which we used to train our inpainting model, contains only images with a frontal view of faces. In our recordings, individuals are captured at different angles. Thus, rotation of the recordings or using a dataset providing faces at different angels can improve the neural inpainting model. In addition, a deep learning model could be trained on our features for affective prediction, and the feature set could be extended by gesture-based features. Such features have shown to be promising for predicting affective states [9].

7. CONCLUSION

In this paper, we presented a hardware setup consisting of a cheap and unobtrusive mirror construction to improve the visibility of the face in tablet-based front camera recordings. Recordings were processed using an inpainting pipeline consisting of a neural network for reconstructing missing data in the recordings. We showed that the mirror construction improved the visibility of the face in situations where external cameras (e.g., GoPro) struggle. With a qualitative and quantitative evaluation, we demonstrated that we could achieve results comparable to a GoPro camera. In particular, neural inpainting improved confidence in facial landmark detection by up to 88 %. We showed the applicability of our setup and processing pipeline on affective state prediction based on front camera recordings. Our model consisted of features capturing information from movement, eyes, and face. We evaluated our affective prediction model on data from a lab experiment with 88 participants using leave-one-user-out cross-validation. Participants were solving math tasks (active) and were exposed to emotional stimuli from pictures (passive). Our model accurately predicted two levels (low and high) of valence (up to 0.80 AUC) and arousal (up to 0.73 AUC) using data from the front camera. These results were comparable to results obtained using recordings from a GoPro camera (up to 0.78 AUC for valence and up to 0.73 AUC for arousal). The novelty of our contribution consists of the hardware setup and processing pipeline. In addition, we proposed features for affective state prediction, which can be useful for other researchers. Our setup is cheap (CHF 5), easy to mount, and can be used in classrooms or at home. Besides affective state prediction, it can be used to monitor students or analyzing attention. Most existing approaches use external cameras such as GoPros or webcams, which are more expensive, more difficult to handle, and are exposed to time synchronization problems. In our setup, the camera data is recorded on the same device as the task is conducted, and thus we circumvent such time synchronization issues in an elegant way. The findings of this work are important because they support the emerging trend of using tablet computers in the classroom and for learning at home by simplifying student recording and assessment.

Acknowledgments. We thank Katja Wolff and Fraser Rothnie for their assistance in creating the figures.

8. REFERENCES

[1] ACT. The act technical manual, 2017.

[2] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson. Emotion sensors go to school. In *AIED*, volume 200, pages 17–24. Citeseer, 2009.

[3] R. S. J. d. Baker, S. M. Gowda, M. Wixon, J. Kalka, A. Z. Wagner, A. Salvi, V. Aleven, G. W. Kusbit, J. Ocumpaugh, and L. Rossi. Towards sensor-free affect detection in cognitive tutor algebra. *International Educational Data Mining Society*, 2012.

[4] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[5] F. Bellard. Ffmpeg. `https://ffmpeg.org/`.

[6] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 379–388, 2015.

[7] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[9] D. M. Bustos, G. L. Chua, R. T. Cruz, J. M. Santos, and M. T. Suarez. Gesture-based affect modeling for intelligent tutoring systems. In *International Conference on Artificial Intelligence in Education*, pages 426–428. Springer, 2011.

[10] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.

[11] Y. Chen, N. Bosch, and S. D'Mello. Video-based affect detection in noninteractive learning environments. *International Educational Data Mining Society*, 2015.

[12] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of educational media*, 29(3):241–250, 2004.

[13] M. Czikszentmihalyi. *Flow: The psychology of optimal experience*. New York: Harper & Row, 1990.

[14] C. Ditzler, E. Hong, and N. Strudler. How tablets are utilized in the classroom. *Journal of Research on Technology in Education*, 48(3):181–193, 2016.

[15] S. K. D'Mello, N. Bosch, and H. Chen. *Multimodal-Multisensor Affect Detection*, page 167–202. Association for Computing Machinery and Morgan & Claypool, 2018.

[16] P. Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[17] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement: Investigator's Guide 2 Part*. Consulting Psychologists Press, 1978.

[18] H. A. Elfenbein and N. Ambady. Universals and cultural differences in recognizing emotions. *Current directions in psychological science*, 12(5):159–164, 2003.

[19] G. Falloon. Young students using iPads: App design and content influences on their learning pathways. *Computers & Education*, 68:505–521, 2013.

[20] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*, 2013.

[21] C. Guillemot and O. Le Meur. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144, 2013.

[22] Z. Guo, Z. Chen, T. Yu, J. Chen, and S. Liu. Progressive image inpainting with full-resolution residual network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2496–2504, 2019.

[23] M. Haak, S. Bos, S. Panic, and L. J. M. Rothkrantz. Detecting stress using eye blinks and brain activity from EEG signals. *Proceeding of the 1st driver car interaction and interface (DCII 2008)*, pages 35–60, 2009.

[24] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[25] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo. Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems*, pages 29–38. Springer, 2014.

[26] S. Kai, L. Paquette, R. S. Baker, N. Bosch, S. D'Mello, J. Ocumpaugh, V. Shute, and M. Ventura. A comparison of video-based and interaction-based affect detectors in physics playground. *International Educational Data Mining Society*, 2015.

[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR*, 2018.

[28] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[29] V. Kostyuk, M. V. Almeda, and R. S. Baker. Correlating affect and behavior in reasoning mind with state test achievement. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 26–30. ACM, 2018.

[30] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, 2008.

[31] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919, 2017.

[32] H. Liao, G. Funka-Lea, Y. Zheng, J. Luo, and K. S. Zhou. Face completion with semantic knowledge and collaborative adversarial learning. In *Asian Conference on Computer Vision*, pages 382–397. Springer, 2018.

[33] D. J. Litman and K. Forbes-Riley. Recognizing student

emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech communication*, 48(5):559–590, 2006.

[34] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[35] H. Liu, B. Jiang, Y. Xiao, and C. Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4170–4179, 2019.

[36] K. Lochner and M. Eid. *Successful emotions: how emotions drive cognitive performance*. Springer, 2016.

[37] D. Malesevic, C. Mayer, S. Gu, and R. Timofte. Photo-realistic and robust inpainting of faces using refinement gans. In *Inpainting and Denoising Challenges*, pages 129–144. Springer, 2019.

[38] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser. Facial features for affective state detection in learning environments. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pages 467–472, 2007.

[39] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. Kaliouby. Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 3723–3726, 2016.

[40] A. Mehrabian and J. A. Russell. *An approach to environmental psychology.* MIT Press, 1974.

[41] M. Miserandino. Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of educational psychology*, 88(2):203, 1996.

[42] R. Navarathna, P. Lucey, P. Carr, E. Carter, S. Sridharan, and I. Matthews. Predicting movie ratings from audience behaviors. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1058–1065. IEEE, 2014.

[43] B. T. Nguyen, M. H. Trinh, T. V. Phan, and H. D. Nguyen. An efficient real-time emotion detection using camera and facial landmarks. In *2017 Seventh International Conference on Information Science and Technology (ICIST)*, pages 251–255. IEEE, 2017.

[44] Z. A. Pardos, R. S. J. D. Baker, M. O. C. Z. San Pedro, S. M. Gowda, and S. M. Gowda. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proc. LAK*, pages 117–124. ACM, 2013.

[45] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[46] P. Pham and J. Wang. Predicting learners' emotions in mobile mooc learning via a multimodal intelligent tutor. In *International Conference on Intelligent Tutoring Systems*, pages 150–159. Springer, 2018.

[47] D. Reid and N. Ostashewski. iPads in the classroom–new technologies, old issues: Are they worth the effort? In *EdMedia+ Innovate Learning*, pages 1689–1694. Association for the Advancement of Computing in Education (AACE), 2011.

[48] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[49] S. Salmeron-Majadas, R. S. Baker, O. C. Santos, and J. G. Boticario. A machine learning approach to leverage individual keyboard and mouse interaction behavior from multiple users in real-world learning scenarios. *IEEE Access*, 6:39154–39179, 2018.

[50] G. K. Sarpate and S. K. Guru. Image inpainting on satellite image using texture synthesis & region filling algorithm. In *2014 International Conference on Advances in Communication and Computing Technologies (ICACACT 2014)*, pages 1–5. IEEE, 2014.

[51] K. R. Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[52] A. Singh, C. Chandewar, and P. Pattarkine. Driver drowsiness alert system with effective feature extraction. *International Journal for Research in Emerging Science and Technology*, 5(4):26–31, 2018.

[53] R. Wampfler, S. Klingler, B. Solenthaler, V. Schinazi, and M. Gross. Affective state prediction in a mobile setting using wearable biometric sensors and stylus. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, pages 198–207, 2019.

[54] C. Wang, X. Sun, F. Wu, and H. Xiong. Image compression with structure-aware inpainting. In *2006 IEEE International Symposium on Circuits and Systems*, pages 4–pp. IEEE, 2006.

[55] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017.

[56] J. Zaletelj and A. Košir. Predicting students' attention in the classroom from kinect facial and body features. *EURASIP journal on image and video processing*, 2017(1):80, 2017.

[57] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008.

[58] Y. Zhuang, Y. Wang, T. K. Shih, and N. C. Tang. Patch-guided facial image inpainting by shape propagation. *Journal of Zhejiang University-SCIENCE A*, 10(2):232–238, 2009.