Facial Composite Generation with Iterative Human Feedback

Florian Strohm, Ekta Sood, Dominike Thomas, Mihai Bâce, Andreas Bulling
University of Stuttgart
{name.surname}@vis.uni-stuttgart.de

Abstract

We propose the first method in which human and AI collaborate to iteratively reconstruct the human's mental image of another person's face only from their eye gaze. Current tools for generating digital human faces involve a tedious and time-consuming manual design process. While gaze-based mental image reconstruction represents a promising alternative, previous methods still assumed prior knowledge about the target face, thereby severely limiting their practical usefulness. The key novelty of our method is a collaborative, iterative query engine: Based on the user's gaze behaviour in each iteration, our method predicts which images to show to the user in the next iteration. Results from two human studies (N=12 and N=22) show that our method can visually reconstruct digital faces that are more similar to the mental image, and is more usable compared to other methods. As such, our findings point at the significant potential of human-AI collaboration for reconstructing mental images, potentially also beyond faces, and of human gaze as a rich source of information and a powerful mediator in said collaboration.

1 Introduction

Methods and tools to generate human faces have broad applicability in computer graphics, visual design, human-computer interaction (HCI), and beyond. For example, creating facial composites or photofits – visual reconstructions of a human face from someone's mind – is widely used in criminal investigations to reconstruct the appearance of a wanted person [15, 14, 16]. Another popular application domain is gaming in which users frequently want to generate digital avatars that are visually similar to them as quickly and effortlessly as possible. Previous work has explored two main approaches to achieve this goal: One approach involves methods to automatically convert a real face image into an avatar using image processing techniques [22, 20]. However, these methods require the target face to be visually available in advance, which is not always possible in particular for facial composites or photofits. The second and more common approach is to rely on software tools that, while allowing for fine-grained control of the generated images [28], require a lot of tedious and time-consuming manual work, and are usually geared to a particular application domain.

A promising alternative are computational methods that directly reconstruct mental images from brain activity recorded while users look at carefully crafted visual stimuli [4, 17, 30, 33, 8, 29, 24]. However, measuring brain activity is impractical for most real-world systems given that it requires expensive and special-purpose equipment as well as extensive operator training. In contrast, human gaze can be measured using off-the-shelf hardware that has recently become both significantly more affordable and usable also by non-experts [21]. Given these advances, a recent line of work has started to investigate gaze-based mental image retrieval [35] as well as visual reconstruction [25, 26]. While the former uses eye gaze features to retrieve a mental image from an existing database, gaze-based mental image reconstruction is profoundly more challenging given that the target image exists only in a user's mind.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Strohm et al. recently demonstrated the first method to reconstruct a mental image solely from eye gaze fixations [32]. While their face reconstructions looked promising visually, their method still required the target face to be known in advance, which renders the method impractical for real applications. With the goal to address this fundamental limitation, we propose a novel method for gaze-based mental image reconstruction that does not require any prior knowledge about the target image. To achieve this goal, our iterative method implements the principle of human-AI collaboration [2] (see Figure 1): On the AI side, we propose a query engine that predicts the most relevant set of face image features for the next iteration while taking prior iterations into account. These features are then decoded into images using a decoder network and presented to the human. The human user looks at these images, searching for similar features to their mental image while their gaze behaviour is being recorded using an eye tracker. Subsequently, new face image features are extracted using the gaze-guided extractor proposed by [32]. These steps are repeated multiple times with gaze-guided image features being extracted in each iteration. Finally, all extracted image features are combined into a single feature vector that is decoded into the final facial composite.

The contributions of our work are two-fold: First, we propose a novel gaze-based collaborative method for mental image reconstruction (GBC-MIR). In stark contrast to prior work [32], our method does not require prior knowledge about the target face but only a pre-trained gaze-guided image feature extractor. Furthermore, our method is domain independent – while we demonstrate an application for face image reconstruction, our method can be applied for any mental image reconstruction task. Second, we evaluate our method in a 12-participant user study and show that it outperforms the current state of the art in reconstruction quality while, at the same time, achieving higher usability.

2 Related work

Mental image retrieval and reconstruction. In mental image retrieval, a query engine typically presents selected, existing images to a user who then manually chooses the closest fit [11, 9, 13, 12, 10]. Based on this choice, the system then presents a new set of images and the user input process continues iteratively until the target image has been found. Instead of selecting existing images from a database, target images can also be created synthetically. In mental image reconstruction, synthetic images are refined in every round of user feedback, until the user is satisfied with the result. Bontrager et al. integrated user feedback in such a fashion and used deep interactive evolution to reconstruct mental images [5], while Xu et al. improved their method by allowing control over specific facial features through relevance feedback [36]. By choosing only pictures with appropriate features, the user allows these features to contribute to the next iteration of pictures. Zaltron et al. proposed Composite Generating Generative Adversarial Networks (CG-GAN) allowing the user to traverse the latent space of a pre-trained GAN [37]. Users were able to select proposed faces to combine their information through a process called mutation. However, these methods rely on explicit user feedback rather than implicit behavioural cues.

Gaze-based mental image retrieval and reconstruction. Human gaze has recently attracted increasing research interests as a promising implicit feedback modality for mental image retrieval and reconstruction. Leading these efforts was work by Sattar et al. who used bags of visual words to address the task in an open-world setting [27]. Stauden et al. further improved their feature extraction component by adding a pre-trained CNN [31], while Barz et al. improved the encoding of fixation sequences with a pre-trained SegNet and used a SVM for the final prediction task [3]. In later work, Sattar et al. went beyond predicting only the target instance by also predicting the target class and attributes [26], while Wang et al. explored a setting in which gaze behaviour used for the target prediction was collected after showing a stimulus [35]. Finally, Strohm et al. were the first to demonstrate the feasibility of reconstructing mental images using gaze behaviour only [32]. Their method consisted of an encoder to extract image features and activation maps as well as a scoring network to compare human gaze maps with neural activation maps to predict a relevancy score for each extracted feature. These scores were used to combine features from multiple images into a single feature vector that was finally decoded into the mental image. While achieving promising results, their method requires prior knowledge severely limiting the usability of their method.

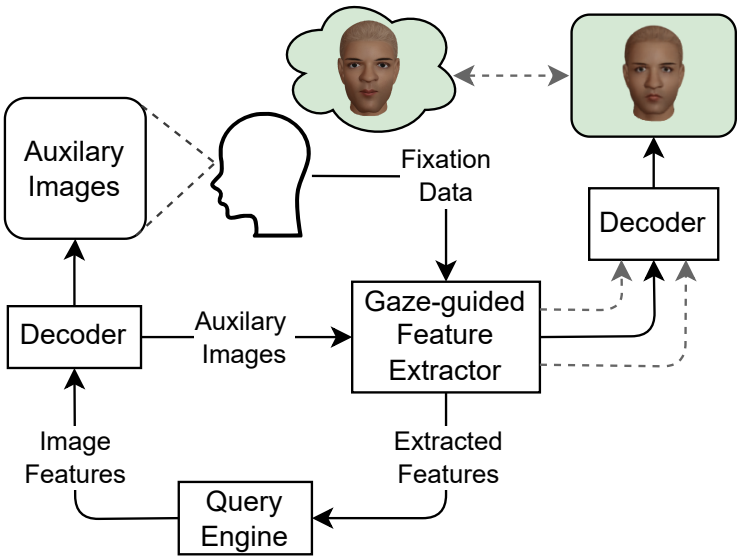

Figure 1: Overview of our method for gaze-based collaborative mental image reconstruction (GBC-MIR). With a specific mental image in mind, users observe auxiliary images proposed by a *query engine* and visually reconstructed by a *decoder*. A *gaze-guided feature extractor* extracts image features that are relevant to the mental image. Based on these features, the query engine proposes image features for the next iteration. This loop continues for several iterations until all extracted feature vectors are combined to reconstruct the final mental image.

3 Gaze-based collaborative mental image reconstruction (GBC-MIR)

The computational task that we study is that of visually reconstructing mental images from eye gaze fixations without prior knowledge of the target. We approach this by showing multiple, generated auxiliary images to a user and recording their gaze while they look at these faces. The auxiliary images are encoded and relevant image features are extracted based on the user's gaze behaviour. The extracted features can subsequently be decoded to reconstruct the mental image. Strohm et al. formulated this task as a mapping $\{(I_i, G_i)|i = 1...n\} \mapsto I_M$, where I_i are the auxiliary images shown to the user, G_i are eye gaze fixations of that user on these images, and I_M is the to-be-reconstructed mental (target) image that the user has in mind [32]. The fundamental difficulty of this task is to extract sufficient information about the mental image from the fixations on the given set of auxiliary images. They used a small set of six auxiliary images, generated using prior knowledge. This ensured that enough information about the target could be extracted from the gaze data. However, given that information about the target image is not available for most real world use cases (e.g., facial composites in law enforcement), this renders their method impractical for actual use.

To address this fundamental limitation, we present a collaborative human-AI system for mental image reconstruction that does not require any prior knowledge of the target image. Instead, our method iteratively reconstructs the target image by presenting sets of images to the user, thus performing a mapping $\{(I_{i,j}, G_{i,j})|i = 1...n\} \mapsto F_{M,j}$, $j = 1...m$, where $F_{M,j}$ represents extracted gaze-guided image features from iteration $j \in [1, 2, ..., m]$. After m iterations, in which our query engine dynamically selects image features based on the user's gaze, our system combines the extracted features of each round to generate the target image. The overall architecture of our system is shown in Figure 1 and its core components are described in the following.

Gaze-guided feature extractor (GFE). A central component of our collaborative system is the gaze-guided feature extractor (GFE). It is represented by a pre-trained model that implements the function $\{(I_i, G_i)|i = 1...n\} \mapsto F$, i.e., given a set of auxiliary images I, it extracts a single feature vector F based on the user's gaze fixations G. The extractor decides, for each feature dimension $f \in F$, from which auxiliary image I_i to select a feature. That is, F is composed of features from the set of auxiliary images I.

Algorithm 1 Algorithm implemented by the selection layer in our end-to-end neural implementation. This layer first calculates the distance of each auxiliary feature value to the target feature. Afterwards, it selects the k-th best feature by sampling k from the probability distribution P_{GFE}. By multiplying the auxiliary image feature vector with a one-hot encoding of the k-th index, this layer selects the k-th best value while allowing to propagate gradients through the auxiliary features $f_{I,j}$.

Input: Auxiliary and target image features $f_{I,j}$ and f_M. Probability distribution P_{GFE}
Output: Selected features $f_{M,j}$

1: differences $\leftarrow$ abs($f_{I,j} - f_M$)
2: args $\leftarrow$ arg_sort(differences)
3: $k \sim P_{\text{GFE}}$
4: selected_arg $\leftarrow$ one_hot(args[k])
5: $f_{M,j} \leftarrow$ (selected_arg $* f_{I,j})^T$
6: **return** $f_{M,j}$

Decoder. The decoder maps feature vectors F into the image domain I using a pre-trained model. This allows us to visualise features for the user (in the form of faces) and to collect gaze data, which in turn can be used to select relevant features through the GFE. It is crucial that both the decoder and the GFE operate in the same feature space $\mathbb{F}$ to allow for an iterative loop of visually decoding features for users to look at, and encoding the images again using the recorded gaze information.

Query engine. For our system to maximise the information about the user's mental image, it needs to decide which images to show in each iteration. For this, we propose a novel query engine which predicts the image features $F_{i,j+1}$ for the next iteration, given the previously shown and extracted features, $F_{i,j}$ and $F_{M,j}$:

$$P(F_{i,j+1}|F_{i,j}, F_{M,j}), \text{ with } F_{i,0} \text{ and } F_{M,0} = \text{constant} \tag{1}$$

This allows our system to dynamically show images to the user based on their previous gaze behaviour. Thus, in each iteration, it can dynamically decide for which features to collect more information.

Iterative collaboration. Together, the GFE, decoder, and query engine form an interactive feedback loop, enabling the human and the system to collaboratively reconstruct the mental image as shown in Figure 1: The query engine proposes a set of feature vectors about which our system wants to gain knowledge, which are visually decoded into the image space by the decoder. Our system then shows these images to the user while their gaze fixations are being recorded. Using the joint gaze and image information, the GFE then extracts the feature vector that is most relevant in this iteration for eventually reconstructing the mental image. This cycle continues for m iterations after which all the feature vectors extracted by the GFE are combined and visually decoded into the final mental image.

3.1 Simulating the gaze-guided feature extractor

Training this system end-to-end is impractical given that human gaze data for the GFE would have to be collected dynamically during training for each set of images predicted by the query engine. A key innovation of our method is that we instead simulate the GFE in the form of a selection layer during training, and only use a pre-trained GFE model, which requires gaze data, at test time.

The GFE is simulated as a probability distribution over the possible features that it can select from the auxiliary images. Given the user's fixation data, the GFE predicts, for each feature dimension $f \in F$, from which of the shown auxiliary images I to select the value of f. Using a small amount of labelled data $\{(I_i, G_i, I_M)|i = 1...n\}$ we can estimate how often the pre-trained GFE selects the feature value closest to the target, the second closest, and so on. Specifically, we estimate $P_{\text{GFE}}(f_M = f_k)$, where k represents the similarity rank of f_k to the mental image feature f_M (see Appendix A for details). Using these estimated probabilities we define a selection layer in Algorithm 1.

Using the target feature value f_M, we calculate a distance ranking of the proposed feature values $f_{I,j} = (f_{1,j}, f_{2,j}, ..., f_{n,j})$, $f_{i,j} \in F_{i,j}$ for iteration j. The layer selects the k-th best feature value f_I by sampling k from $P_{\text{GFE}}(f_M = f_k)$. Instead of directly indexing the selected value from the auxiliary image, we construct a one-hot selection vector and calculate the dot product with the features of the auxiliary images $f_{I,j}$. This way we model the selection as an external random node, allowing us to calculate gradients for each parameter in the neural network.

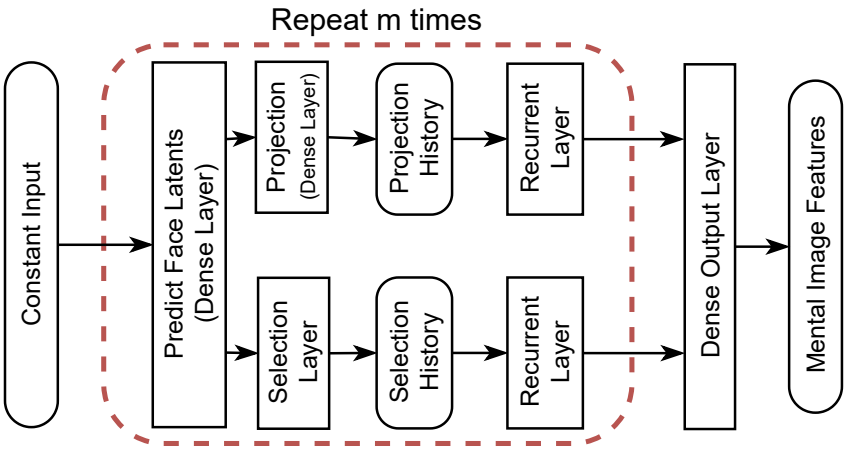

Figure 2: Overview of our end-to-end neural implementation of the collaborative reconstruction system. Dense layers predict the auxiliary image features for the next iteration. For this, they use the input of two recurrent layers that combine information about the shown and selected features of previous iterations. A final dense output layer combines the information of all selected feature vectors to predict the mental image features.

The GFE performs the mapping $\{(I_i, G_i)|i = 1...n\} \mapsto F$, while the selection layer implements the mapping $\{F_i|i = 1...n, F_M\} \mapsto F$. This approach allows us to predict the extracted features F of each iteration without gaze data but requires the target image features F_M as an input. Since the target face and the corresponding features are known at training time, this proxy mapping allows us to train GBC-MIR end-to-end.

3.2 End-to-end Training

The only component that has to be learned is the query engine since pre-trained models are used for the GFE and decoder. To train the query engine end-to-end we combine it with the selection layer defined in Algorithm 1 to create the GBC-MIR feedback loop. Figure 2 shows the neural network structure used to learn a set of recurrent and dense layers forming the query engine. The network is composed of m stacked iteration modules as indicated by the dotted red line in Figure 2. Each module consists of a dense layer predicting n auxiliary image feature vectors for that iteration based on the output of the previous module, which are fed into two separate network paths: One path starts with a selection layer that simulates the pre-trained GFE and receives $n * |F|$ features as input. It outputs the selected features $F_{j,M}$ of iteration j based on Algorithm 1. These features are input to a recurrent layer together with the history of features from all previous iterations $\{(F_{M,k})|k = j-1, j-2, ...1\}$, combining the vector sequence into a single feature vector. The second path starts with a dense projection layer which also receives $n * |F|$ features as an input and outputs a lower-dimensional representation of these features. Together with the history of projected features from previous iterations, these are input into another recurrent layer combining the features. The projection path enables the network to keep track of which features it already proposed in any prior iteration. The outputs of each recurrent layer from the two paths, selection and projection, are concatenated and input to the next module of iteration $j + 1$. After m iteration modules, the output of the last module is fed into a dense output layer which predicts the target image features.

The network is trained by minimising the mean squared error between predicted and target image features. While the network's input is only a constant value, it receives information about the target through the selection layer in each iteration module. At test time the selection layer is replaced with a pre-trained GFE, allowing the system to make user-specific predictions based on their gaze behaviour. In this case a pre-trained decoder generates the auxiliary images for each iteration given the predicted image features of the query engine. These images can then be shown to the user while recording their gaze behaviour. The generated images of the first iteration are constant, as our system has no knowledge about the mental image yet. However, subsequent predictions are made based on

the previously extracted features of the GFE, which are user-specific and encode increasingly more information about their mental image.

4 Reconstruction experiments

For evaluation, we used the pre-trained GFE and decoder provided by Strohm et al. [32] and thus evaluate GBC-MIR in the FaceMaker face image domain [28]. FaceMaker is an interactive tool to manually generate human-like faces by manipulating a set of 28 face appearance sliders, giving fine grained control over different facial features. We compare GBC-MIR with a baseline method and a FaceMaker control.

4.1 Methods

Ours – Gaze-based collaborative mental image reconstruction (GBC-MIR). To train our system we estimated the values P_{GFE} of the pre-trained GFE model for the selection layer (see appendix). As we have to reconstruct 28 features and the used GFE requires six input images, the dense layers of our system depicted in Figure 2 consist of $28 * 6 = 168$ neurons with a Sigmoid activation. The projection layers consist of 50 neurons with a ReLU activation, projecting the 168 image features into a 50 dimensional space. The recurrent layers are implemented by gated recurrent unit (GRU) [7] layers, each extracting 50 features from the history of selected and projected features, respectively. We set the number of iterations to $m = 10$, resulting in ten stacked iteration modules. The output dense layer consist of 28 neurons with Sigmoid activation.

The model was trained for 50 epochs on a V100 GPU with a batch size of 32, where each batch consists of randomly generated target feature vectors of size 28. We used the Adam optimiser [23] with a learning rate of 0.0004 and default parameters otherwise. For the user study, we replaced the selection layer with the pre-trained GFE.

As in Strohm et al. [32], the task of the participants in each iteration was to rank the six auxiliary images according to the similarity with their mental image. They had 30 seconds to complete this task after which the gaze data was used to extract features with the GFE. Since the used GFE requires 30 seconds of gaze recordings to extract image features, and we used ten iterations, the resulting total reconstruction time amounted to five minutes. This keeps the interaction time with the system short, reducing possible eye strain and ensuring fast reconstruction times.

Baseline – Interactive evolution. As the baseline we compared to an adapted version of the state-of-the-art mental image reconstruction method proposed by Zaltron et al. [37]. This system initially proposes nine randomly generated faces. Users then have the option to either generate new random faces or mutate faces: Any number of the proposed faces can be locked so that they are not replaced by new random faces. Similarly, any number of faces can be selected for mutation. For this, the average of the selected faces is calculated and new faces are generated by adding random noise to this average face. Participants can control the amount of this noise, i.e., how large the changes should be. Additionally, participants can choose if they want all features to be changed while mutating or only a single random feature. This system can be used to iteratively traverse the latent space until participants decide that one of the proposed faces adequately resembles their mental image. The original method by Zaltron et al. also allowed for explicit manipulation of certain face attributes but given that this requires a dataset with labelled attributes, it is not possible to consider this functionality.

Control – FaceMaker. FaceMaker [28] is a tool to manually generate face images using a set of sliders and is used in this work as a control. More specifically, starting with a mean face, participants can manipulate 28 facial features by moving an associated slider. FaceMaker updates the face in real time. Users can adjust these sliders until they are satisfied with the created face resembling their mental image sufficiently well. Since FaceMaker was used to train the GFE and feature decoder [32], and allows for full control over the generation of the faces, we use it as a control condition.

Table 1: Mean absolute feature distance (MAFD) for different facial regions of our gaze-based method compared to the baseline by Zaltron et al. [37] and the FaceMaker control condition.

	MAFD				
Method	Eyes	Nose	Mouth	Jaw	Overall
Ours	$\mathbf{46.3 \pm 8}$	$\mathbf{44.5 \pm 16}$	$\mathbf{41.8 \pm 12}$	$\mathbf{45.7 \pm 14}$	$\mathbf{40.9 \pm 6}$
Zaltron et al. [37]	55.8 ± 13	50.6 ± 17	54.9 ± 15	53.0 ± 22	48.0 ± 7
FaceMaker	30.0 ± 9	33.5 ± 11	28.9 ± 10	33.6 ± 11	29.6 ± 6

4.2 User study

We compared these three methods in a user study with 12 participants (four female) aged between 18 and 29 years (M=23.5, SD=3.6). All participants had normal or corrected-to-normal vision and were recruited through university mailing lists. For our GBC-MIR method, binocular gaze was recorded at 2,000 Hz using a stationary EyeLink 1000 Plus eye tracker. To increase gaze tracking accuracy, we used a chin rest to stabilise participants' heads. Each stimuli covered about $8.9°$ degrees of visual angle. We counterbalanced the conditions using a within-group Latin square study design.

For each method, participants had to complete three trials in succession. Each of these trials started with a 30 second memorisation step of the target face. After memorisation, participants were instructed to reconstruct the face using the respective method. To prevent memorisation effects of the target faces between methods, each target face was used for each method between groups, not within. After completing three trials of one method they were asked to complete a System Usability Scale (SUS) [6] and NASA-TLX questionnaire [19]. Once participants finished all nine trials, they were briefly interviewed about the different systems and finally compensated for their participation.

4.3 Reconstruction results

Metrics. We report the mean absolute feature distance (MAFD) defined as

$$\text{MAFD} = \frac{1}{28} \sum_{i=1}^{28} |f_i^p - f_i^M|, \tag{2}$$

where f_i^p is the predicted value and f_i^M is the target value for feature f. The feature values for f are normalised between 1 and 182 as this is the range given by FaceMaker. Since we calculate the distance between prediction and target feature value a lower MAFD is better. In addition, we report performance for the facial areas *Eyes*, *Nose*, *Mouth* and *Jaw* by grouping features as defined by Strohm et al. [32].

Results. Table 1 shows the performance of our system compared to the baselines. As can be seen, our gaze-based method achieves a MAFD of 40.9 compared to 48.0 by the method of Zaltron et al. [37], representing an improvement of about 15%. Using the FaceMaker control condition participants achieved an overall MAFD of 29.56. This MAFD is quite considerable given that participants had full and explicit control of the face generation process using FaceMaker, underlining the difficulty of the face reconstruction task.

Figure 3 shows four sample target faces, reconstructions produced by our gaze-based method, as well as baseline reconstructions produced with the method by Zaltron et al. and FaceMaker. Colour-coded feature groups indicate the reconstruction quality of the different facial regions. The colour codes represent three equidistant bins between the minimum and maximum MAFD of the test set reconstructions. We can observe that our method can produce visually plausible mental image reconstructions for most facial regions. Furthermore, as indicated by the MAFD, participants struggle to reconstruct some facial features even when having full control with FaceMaker.

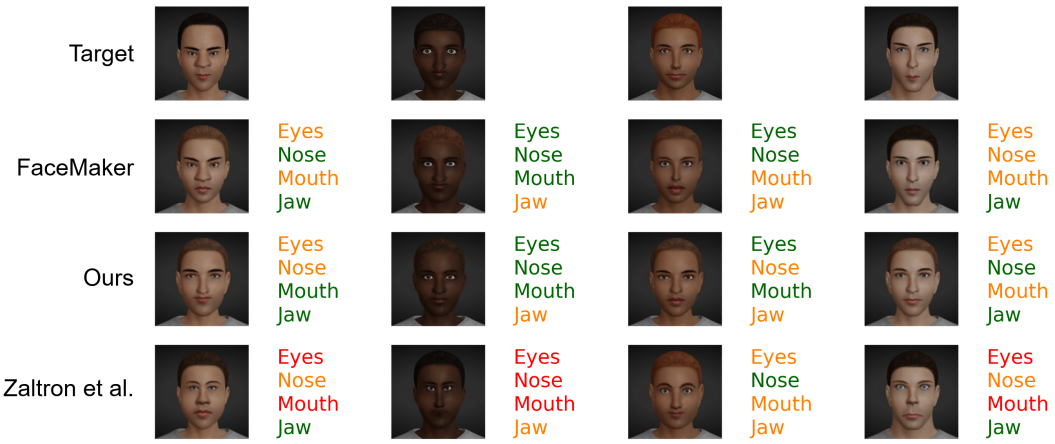

Figure 3: Example reconstructions created during the user study. It shows the reconstructions with our gaze-based collaborative system GBC-MIR (Ours) compared to the two baselines. Colour-coded labels (high, medium, or low) indicate the reconstruction quality of different facial regions. The colour codes represent three equidistant bins between the minimum and maximum MAFD of the test set reconstructions.

4.4 Qualitative evaluation

We conducted an additional 22-participant user study to assess the subjective quality of our reconstructions. Participants were shown sets of four faces, each containing a target face from our test set as well as the reconstructions from the three methods in random order. Participants were asked to rank the three reconstructions from most similar (rank one) to least similar (rank three) according to their similarity with the target. We also asked them to provide text explanations for the reasoning behind their ranking. The average rank given to our reconstructions was 2.28 ± 0.7 while the method by Zaltron et al. [37] ranked 2.40 ± 0.7 and the FaceMaker [28] control 1.33 ± 0.5.

5 Analysis and discussion

Reconstruction quality. Our results show that gaze-based mental image reconstruction without prior knowledge about the target is possible and outperforms the previous best manual reconstruction method by $\sim$15% in MAFD (see Table 1). An additional user study to evaluate the reconstructions underlines this as users tend to rate our reconstructions as more similar to the target image than the reconstructions based on the method by Zaltron et al. [37]. However, analysing the user feedback and comparing qualitative results to the FaceMaker [28] control condition indicate that there is still room for improvement. As faces are perceived holistically [15], a few mismatched features might make a face unrecognisable or, in our case, appear vastly different than the target image. Furthermore, we observed a large variance in the predictions of our system with up to 16 in MAFD for nose related features. This variance can be partly explained by humans' differing ability to memorise images, and in particular faces [34], as the variance for the FaceMaker control condition is similarly high. This is likely also the reason for the relatively high MAFD of the FaceMaker control method, since participants recreated their mental image of the target, not the target face directly. Finally, we evaluated the SUS questionnaires for each condition. The resulting SUS scores are 52 for the method by Zaltron et al., 67 for our proposed system and 76 for FaceMaker (see appendix C for details). Overall these results tend towards a better usability of GBC-MIR compared to Zaltron et al..

Feature extractor importance. Since we simulate the GFE during training of GBC-MIR by replacing it with a selection layer (see Section 3.2), we are able to analyse the influence of different probability distributions P_{GFE} on the performance. Figure 4 shows the simulated MAFD for different probability distributions P of the feature extractor over number of iterations. At the extremes, P_{ideal} assumes a feature extractor always selecting the best possible feature, while P_{equal} has an equal chance of selecting any feature. The MAFD for P_{equal} remains at a constant value of 52.56 independent of the number of iterations. Since the system does not receive any information about the target due to the

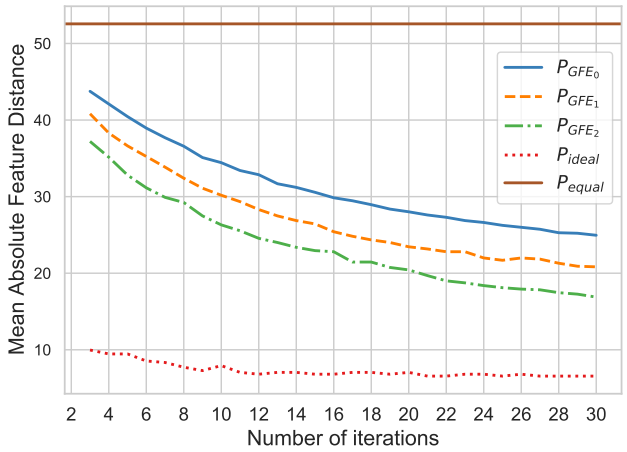

Figure 4: Simulated mean absolute feature distance (MAFD) over number of iterations for different probability distributions of the feature extractor. P_{GFE_0} are the probabilities estimated for our gaze-guided feature extractors while P_{GFE_1} and P_{GFE_2} are better, theoretical versions. P_{equal} represents a random feature extractor while P_{ideal} always selects the best feature value possible.

random feature extractor, our network learns to reconstruct the mean face. P_{GFE_0} shows the simulated MAFD for the probabilities we estimated for the pre-trained GFE used in our system. To analyse how GBC-MIR scales with the quality of the GFE, we simulated results with improved versions $P_{\text{GFE}_{1,2}}$, as shown in Figure 4. For P_{GFE_1}, we subtracted 10% from the probabilities of selecting sub-optimal features (ranks 2-6) and added it to the probability of selecting the best feature (rank 1). In the case of P_{GFE_2}, 20% was redistributed instead. As can be observed in Figure 4, the MAFD of GBC-MIR scales well with the performance of the feature extractor. The MAFD of P_{GFE_1} is consistently lower over the number of iterations compared to P_{GFE_0} and analogously between P_{GFE_2} and P_{GFE_1}. We found that the query engine learns different strategies depending on the GFE performance. For unreliable GFEs it repeatedly samples six equidistant areas in the feature space to determine a value range that is most likely. More reliable GFEs allow to explore the feature space more thoroughly as less repeated measurements to increase confidence are required (see appendix B for details).

6 Conclusion

We have proposed GBC-MIR, the first collaborative system for reconstructing a specific instance of a mental face image from users' eye gaze. In contrast to previous work, our method requires neither sophisticated sensing setup with limited practical relevance as with fMRI, nor any time-consuming and laborious manual image generation process. Through evaluations with a 12-participant eye tracking user study as well as a 22-participant qualitative study, we have shown that our gaze-based method can generate faces that are more similar to the mental target image, and with higher usability than the previous state-of-the-art method. Since our proposed end-to-end collaborative system is independent of any specific image domain, it can be adapted for other image reconstructions tasks. Furthermore, it is even possible to combine our proposed method with pre-trained image feature extractors using human feedback other than gaze information. As such, our results mark an important advance towards collaborative mental image reconstruction and also underline the potential of human gaze for this challenging task.

7 Ethical statement

Data collection. The data collection procedure was approved by the university's ethics committee and is in line with the General Data Protection Regulation (GDPR) of the European Union. Participants were informed about the task they had to complete and the type of data that was collected. All collected data was fully anonymised and, as such, cannot be traced back to a specific individual, thus ensuring their privacy. The data collection only started after participants agreed to these conditions

and gave their informed consent. They were able to interrupt the study at any time and leave the laboratory. Participants were financially compensated, even if they dropped out early for any reason.

Potential risks. Future improvements on this challenging task will likely allow to reconstruct mental images more quickly and with even higher visual fidelity. While they thus have many beneficial applications, such as in police investigations or by helping users with limited expressive skills, involuntary reconstruction poses a non-negligible risk of misuse. However, given that our method still requires users to explicitly search for similar facial features, our current method has little potential for misuse. Nevertheless, it is important to be aware of this potential and to avoid misuse of future methodological developments in this active area of research.

Acknowledgments and Disclosure of Funding

Florian Strohm and Andreas Bulling were funded by the European Research Council (ERC) under the grant agreement 801708. Ekta Sood was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. Mihai Bâce was funded by a Swiss National Science Foundation (SNSF) Early Postdoc.Mobility Fellowship.

References

[1] Hervé Abdi and Lynne J Williams. Tukey's honestly significant difference (hsd) test. *Encyclopedia of research design*, 3(1):1–5, 2010.

[2] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(08):18–28, 2020.

[3] Michael Barz, Sven Stauden, and Daniel Sonntag. Visual search target inference in natural interaction settings with machine learning. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–8, 2020.

[4] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. In *Advances in Neural Information Processing Systems*, pages 6517–6527, 2019.

[5] Philip Bontrager, Wending Lin, Julian Togelius, and Sebastian Risi. Deep interactive evolution. In *International Conference on Computational Intelligence in Music, Sound, Art and Design*, pages 267–282. Springer, 2018.

[6] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[8] Hiroto Date, Keisuke Kawasaki, Isao Hasegawa, and Takayuki Okatani. Deep learning for natural image reconstruction from electrocorticography signals. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2331–2336. IEEE, 2019.

[9] Yuchun Fang and Donald Geman. Experiments in mental face retrieval. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 637–646. Springer, 2005.

[10] Yuchun Fang, Donald Geman, and Nozha Boujemaa. An interactive system for mental face retrieval. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 193–200, 2005.

[11] Yuchun Fang, Zhengye Xiao, Wei Zhang, Yan Huang, Liang Wang, Nozha Boujemaa, and Donald Geman. Attribute prototype learning for interactive face retrieval. *IEEE Transactions on Information Forensics and Security*, 16:2593–2607, 2021.

[12] Yuchun Fang and Qiulong Yuan. Attribute-enhanced metric learning for face retrieval. *EURASIP Journal on Image and Video Processing*, 2018(1):1–9, 2018.

[13] Yuchun Fang, Wei Zhang, and Ningjie Liu. On the perception analysis of user feedback for interactive face retrieval. *ACM Transactions on Applied Perception (TAP)*, 17(3):1–20, 2020.

[14] Charlie D Frowd, Derek Carson, Hayley Ness, Jan Richardson, Lisa Morrison, Sarah Mclanaghan, and Peter Hancock. A forensically valid comparison of facial composite systems. *Psychology, Crime & Law*, 11(1):33–52, 2005.

[15] Charlie D Frowd, Peter JB Hancock, and Derek Carson. Evofit: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on applied perception (TAP)*, 1(1):19–39, 2004.

[16] Ben George, Stuart J Gibson, Matthew IS Maylin, and Christopher J Solomon. Efit-v- interactive evolutionary strategy for the construction of photo-realistic facial composites. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 1485–1490, 2008.

[17] Yağmur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, and Marcel A. J. van Gerven. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in Neural Information Processing Systems*, pages 4246–4257. Curran Associates, Inc., 2017.

[18] Ralitza Gueorguieva and John H Krystal. Move over anova: progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. *Archives of general psychiatry*, 61(3):310–317, 2004.

[19] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[20] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017.

[21] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, page 1151–1160, New York, NY, USA, 2014. Association for Computing Machinery.

[22] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Yunfeng Lin, Jiangbei Li, and Hanjing Wang. Dcnn-gan: Reconstructing realistic image from fmri. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019.

[25] Hosnieh Sattar, Andreas Bulling, and Mario Fritz. Predicting the category and attributes of visual search targets using deep gaze pooling. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2740–2748, 2017.

[26] Hosnieh Sattar, Mario Fritz, and Andreas Bulling. Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. *Neurocomputing*, 387:369–382, 2020.

[27] Hosnieh Sattar, Sabine Müller, Mario Fritz, and Andreas Bulling. Prediction of search targets from fixations in open-world settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pages 981–990, 2015.

[28] Valentin Schwind, Katrin Wolf, and Niels Henze. *FaceMaker - A Procedural Face Generator to Foster Character Design Research*, volume Game Dynamics: Best Practices in Procedural and Dynamic Game Content Generation, pages 95–113. Springer International Publishing, Cham, 2017.

[29] Sophia M Shatek, Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. Decoding images in the mind's eye: The temporal dynamics of visual imagery. *Vision*, 3(4):53, 2019.

[30] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):1–23, 2019.

[31] Sven Stauden, Michael Barz, and Daniel Sonntag. Visual search target inference using bag of deep visual words. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 297–304. Springer, 2018.

[32] Florian Strohm, Ekta Sood, Sven Mayer, Philipp Müller, Mihai Bâce, and Andreas Bulling. Neural photofit: Gaze-based mental image reconstruction. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1–10, 2021.

[33] Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, 2(1):1–10, 2019.

[34] Roeland J. Verhallen, Jenny M. Bosten, Patrick T. Goodbourn, Adam J. Lawrance-Owen, Gary Bargary, and J. D. Mollon. General and specific factors in the processing of faces. *Vision Research*, 141:217–227, 2017.

[35] Xi Wang, Andreas Ley, Sebastian Koch, David Lindlbauer, James Hays, Kenneth Holmqvist, and Marc Alexa. The mental image revealed by gaze tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[36] Caie Xu, Ying Tang, Masahiro Toyoura, Jiayi Xu, and Xiaoyang Mao. Generating users' desired face image using the conditional generative adversarial network and relevance feedback. *IEEE Access*, 7:181458–181468, 2019.

[37] Nicola Zaltron, Luisa Zurlo, and Sebastian Risi. Cg-gan: An interactive evolutionary gan-based approach for facial composite generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2544–2551, 2020.

A Evaluating the gaze-guided feature extractor (GFE)

We fine-tuned and evaluated the GFE proposed by Strohm et al. [32] for mental human-like face reconstruction from FaceMaker [28]. In their work they utilised knowledge about the target face to generate sets of six auxiliary faces, ensuring that the target facial features were reflected in the faces shown to the user. In stark contrast, our work does not use prior knowledge, therefore target features are not necessarily reflected in the shown faces. As this might influence the gaze behaviour of users, we collected additional training data to fine-tune their GFE.

A.1 Data Collection

We recorded gaze data of 7 female and 3 male participants aged between 19 and 35 years (M=27.1, SD=4.7). They were recruited through university mailing lists and compensated for their participation with 15€ per hour. The eyesight of all participants was normal or corrected-to-normal. We recorded their binocular gaze at 2,000Hz using a stationary EyeLink 1000 Plus eye tracker. To increase gaze tracking accuracy, we used a chin rest stabilising participants' head. A 24.4-inch screen with a resolution of 1920×1080 pixels was placed 90cm in front of the participants to show the face stimuli. Each stimuli covered $8.9°$ degrees of visual angle.

Once participants agreed to our consent form they completed three trials. Each trial started with a calibration-validation procedure to ensure accurate eye tracking. Following this, a target face was shown for 30 seconds that participants had to memorise. After the memorisation phase we iteratively displayed ten sets of six auxiliary faces for 30 seconds each and gave participants the task to rank the faces from one to six according to similarity with their mental image. All face stimuli were generated randomly and independently of each other by sampling the FaceMaker features from a uniform distribution.

A.2 Fine-tuning and evaluation

We fine-tuned the GFE using 28 out of 30 trials we collected for each participant. Strohm et al. [32] were able to train their GFE to perform binary classification, since a feature either was the target feature or not. However, since our work does not use prior knowledge, we cannot create binary labels and as such fine-tune their GFE using continuous labels by minimising the mean squared error[1]. The labels are created by calculating the similarity $1 - \text{abs}(f_M - f_i)$ between each feature $f_i \in F_i$ and the target features $f_M \in F_m$.

[1]The architecture and all hyper-parameters are equivalent to Strohm et al. [32].

Using the remaining two trials of each participant we estimated P_{GFE} by counting how often the model selects the best feature, the second best and so on. Table 2 shows the estimated probability distribution P_{GFE} for the original model [32] and our model fine-tuned with additional training data and continuous labels. We observe that by adding more task-specific training data, the GFE improved in performance, where the probability of extracting a top-3 feature increases from 61.1% to 68.2%.

	Similarity Rank					
Model	1	2	3	4	5	6
Strohm et al.	18.2%	20.2%	22.7%	16.9%	12.9%	9.1%
Ours	22.9%	21.8%	23.5%	16.7%	10.0%	5.1%

Table 2: Comparison of P_{GFE} (i.e., the probability distribution for selecting the k-th best feature value) for the pre-trained model by Strohm et al. [32] and our fine-tuned model. For our model we observe that the probability of extracting a top-3 feature increases from 61.1% to 68.2%.

B Query Engine Performance

To gain a better understanding of how GBC-MIR works, we analyse the behaviour of our query engine. Figure 5 shows the predicted feature value distribution for two trained ten-iteration query engines trained using P_{GFE_0} (Figure 5a) and P_{GFE_2} (Figure 5b). We observe that our query engine in Figure 5a learned to split the feature space into six roughly equidistant areas, due to the six faces shown.

When our query engine is trained with P_{GFE_0}, it learns the strategy to take multiple samples from the same area in order to increase its confidence in a feature's rank, rather than searching the entire feature space for the exact values of the mental image. This can be seen (Figure 5a) in the higher probability peaks and the surrounding gaps of zero probability.

Figure 5b shows the predicted feature value distribution for our query engine trained with the better simulated GFE P_{GFE_2}. We observe a more uniform distribution of the predictions across the feature space, as there are no longer any surrounding gaps of zero probabilities. With higher confidence in the selected feature's rank, the query engine does not require redundant samples and can use a more fine-grained search strategy.

C Usability and Workload

System Usability Scale We asked participants to fill in a System Usability Scale (SUS) questionnaire [6] to assess the usability of our three conditions. The SUS allows us to calculate a single

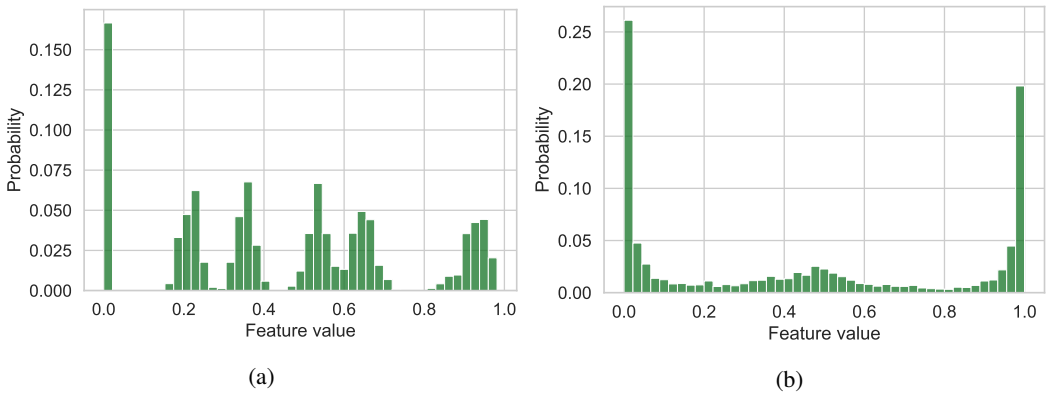

Figure 5: The plots show the probability of query engines to propose a certain auxiliary feature value. (a) shows the feature value distribution for the query engine trained using our estimated values P_{GFE_0}, while (b) shows the feature value distribution for a better estimate P_{GFE_2}.

score on a scale from 0-100 (higher is better), by answering ten standardised questions on a five point Likert scale. The resulting SUS scores are 52 for the method by Zaltron et al. [37], 67 for our proposed system and 76 for FaceMaker. A repeated measures one-way ANOVA test [18] indicated a significant difference (p=0.007) between the SUS scores of the three conditions. Pairwise Tukey's HSD post-hoc tests [1] indicate a significant difference only between the method by Zaltron et al. and FaceMaker (p=0.026). Overall these results tend towards a better usability of GBC-MIR compared to Zaltron et al..

NASA-TLX Workload Additionally to the SUS, we asked participants to complete the Raw-NASA-TLX questionnaire [19], allowing us to assess the perceived workload of the participants for each condition on a scale from 1-100 (lower is better). The resulting values are 46 for the method by Zaltron et al. [37], 47 for our proposed system and 30 for FaceMaker. A repeated measures one-way ANOVA test [18] indicate a significant difference (p=0.002) between the TLX values of the three conditions. Pairwise Tukey's HSD post-hoc tests [1] indicate a significant difference between the method by Zaltron et al. and FaceMaker (p=0.022) as well as between our method and FaceMaker (p=0.015). Overall, these results suggest that the usage of FaceMaker was the least demanding task, while there is no significant difference between GBC-MIR and the method of Zaltron et al..