# Feeling Positive? Predicting Emotional Image Similarity from Brain Signals

Tuukka Ruotsalo
University of Copenhagen
LUT University
tr@di.ku.dk

Kalle Mäkelä
University of Helsinki
kalle.o.makela@helsinki.fi

Michiel M. Spapé
University of Helsinki
michiel.spape@helsinki.fi

Luis A. Leiva
University of Luxembourg
name.surname@uni.lu

## ABSTRACT

The present notion of visual similarity is based on features derived from image contents. This ignores the users' emotional or affective experiences toward the content, and how users feel when they search for images. Here we consider valence, a positive or negative quantification of affective appraisal, as a novel dimension of image similarity. We report the largest neuroimaging experiment that quantifies and predicts the valence of visual content by using functional near-infrared spectroscopy from brain-computer interfacing. We show that affective similarity can be (1) decoded directly from brain signals in response to visual stimuli, (2) utilized for predicting affective image similarity with an average accuracy of 0.58 and an accuracy of 0.65 for high-arousal stimuli, and (3) effectively used to complement affective similarity estimates of content-based models; for example when fused fNIRS and image rankings the retrieval F-measure@20 is 0.70. Our work opens new research avenues for affective multimedia analysis, retrieval, and user modeling.

## CCS CONCEPTS

• **Information systems** → **Users and interactive retrieval**.

## KEYWORDS

Affective Computing; BCI; Ranking Relevance

**ACM Reference Format:**
Tuukka Ruotsalo, Kalle Mäkelä, Michiel M. Spapé, and Luis A. Leiva. 2023. Feeling Positive? Predicting Emotional Image Similarity from Brain Signals. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3613442

## 1 INTRODUCTION

Similarity comparisons are fundamental for many applications that manage and deliver multimedia content to Web users. For example, recommender systems use similarity in their loss functions, search engines use similarity to match content to queries or user profiles, and social media services use similarity to estimate user interests for content to be included in feeds. The present methods for assessing information similarity rely on a fairly simplistic assumption: items are similar when their content-based features or usage patterns are similar [25, 54, 55]. For example, researchers have used visual features extracted from images and videos or textual features associated with those (e.g. captions), ignoring the emotional experiences of users [61].

**(a) High valence (positive)**

**(b) Low valence (negative)**

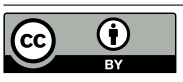
**Figure 1: An example of two images that are visually similar but dissimilar in an affect they are likely to evoke: fun and exciting (a), which evokes high valence emotion, vs. fear and avalanche (b), which evokes low valence emotion.**

Cognitive science has shown that human judgements about similarity do not rely merely on objective appraisals of semantic similarity between two items, but involve *affective* appraisals. For example, the time it takes to make a similarity judgement is strongly related to the experienced distance between two items [14, 45]. Likewise, decisions on emotional valence strongly depended on the congruence (i.e. affective similarity) between a briefly presented preceding prime and the target stimulus [3]. Therefore, if a model of content similarity is to approximate human judgement, then the affective similarity is of critical importance.

In terms of computational modeling, the lack of affective features in content similarity is often referred to as the 'affective gap' [63]. For example, Figure 1 shows two images with visually similar mountain scenery. These images, however, are affectively very different. Figure 1a shows an image of skiing associated with positive emotions, while Figure 1b shows an image of an avalanche associated with negative emotions. Such affective differences are not captured by current image similarity models that rely on content-based features as opposed to predictions on how users emotionally experience content. Even computational models of aesthetics [29] fall short in this regard.

In sum, current methodologies for capturing similarity do not consider affective features, as experienced by users. We also lack computational tools to predict, decode, and utilize affective information as part of models that can estimate visual content similarity.

In this paper, we focus on predicting *valence* (positivity or negativity) of visual stimuli and use the predicted valence to assess affective similarity. We propose a first-of-its-kind methodology using brain-computer interfacing (BCI) recorded using functional near-infrared spectroscopy (fNIRS). The advantage of using brain





*MM '23, October 29-November 3, 2023, Ottawa, ON, Canada*

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3613442

signals are that (1) recording such signals does not require any explicit user interaction to express their emotional experience, instead it can be directly recorded using wearable sensors; and (2) brain signals can be directly used to measure the affective processing of stimuli. Consequently, brain signals may be used in combination with (or even as an alternative for) conventional user input. Moreover, our approach incorporates subjective emotional experience as part of the similarity computation process as opposed to only relying on content-driven signals. Ultimately, in the future, BCI technology may strongly impact how users' affective states can be used to interact with computers. With this motivation, we formulate the following research questions:

(1) Can valence (positive and negative affective states) be detected, predicted, and used for affective similarity search?
(2) Does affective information complement similarity-based ranking models, as compared to their content-based alternatives?

To this end, we propose to use new technology to estimate affective information and incorporate it into information retrieval systems. We make the following contributions:

- We report the largest fNIRS in-lab experiment for studying affective image classification and ranking.
- We show how fNIRS brain data can be incorporated in predicting positive and negative affective states of visual content.
- We demonstrate fNIRS as a novel similarity index for ranking information according to the emotional experience they evoke in human users and compare them with models that use content-based features.

## 2 RELATED WORK

**Image similarity** methods have been largerly based on comparing either content representations learned from the visual information of images [55] or text associated to them, such as captions [10]. These can already effecively address the 'semantic gap' [47] by detecting images with semantically similar but visually varying content. However, these methods are not as successful in addressing the 'affective gap' [59, 63], which refers to classifying the affective response an image is likely to evoke in users, independently of its contents.

**The affective gap** has turned out to be more challenging to address. Researchers have proposed feature engineering methods that would better account for affective content [33, 57, 62], as well as deep learning to learn efficient representations [9, 18, 39]. Despite these advances, all content-based methods rely on the analysis of content (as opposed to user experience) to reach a consensus estimate on the perceived emotions that is then used as the dominant (average) emotion category. Recently, emotions have been studied in the context of web search and media content analysis [2, 23, 59]. However, the literature is very scarce, currently limited to clickstream data and explicitly annotated data that does not account for predicting individually experienced emotions.

**Physiological and neuroimaging methods**, as opposed to content-based methods, have been proposed to decode emotional reponses directly from human cognition [32]. These approaches allow to capture such emotional responses as they are experienced by each individual. Typical techniques that have been used include peripheral sensors [26], electrocorticography (ECoG) [44], electroencephalography (EEG) [21], and more recently fNIRS [19].

While EEG and peripheral wearable sensors have been extensively studied, their use in realistic applications is still limited [12, 13]. Peripheral sensors do not always provide reliable measurement data and EEG is prone to known artifacts, for example, due to motor activity and eye movements. fNIRS does not suffer from these shortcomings. It is a relatively recent technique of neuroimaging, which relies on the relationship between neural activity and blood oxygenation [4, 7]. Thus, by measuring absorption at different wavelengths of light, fNIRS may quantify cortical activity, particularly if such areas are near the surface and unimpeded by tissues interfering with light (e.g. hair). fNIRS has recently been used in measuring emotion-related activity in frontal areas, for both discrete emotions [20] and emotional dimensions [5]. For the latter, the emotional dimension of valence showed a particular effect on negative emotional pictures in the right prefrontal cortex on oxygenated hemoglobin (O2Hb).

**fNIRS technology for emotion recognition** has been far less studied. Recent studies suggest that emotional decoding from fNIRS data has clear potential [42], although high-accuracy reports have typically been found in the presence of certain methodological limitations. For example, studies using lengthy video clips of music videos [6], commercial ads [38], or video games [1] reported high accuracy, between 77% and 91% in detecting self-reported preferences and emotional valence. The use of limited selections (e.g. 3 per condition in [38] and 5 in [50]) from non-validated stimulus databases, means that confounds may have artificially boosted model accuracy. For example, if videos in one condition were rated as being more fun while simultaneously being more dynamic in content or presentation, classifiers may have profited from cognitive differences related to processing multiple scenes.

## 3 fNIRS DATA ACQUISITION

### 3.1 Participants

Thirty-one volunteers (18 male, 12 female, 1 non-binary with mean age 31.2 years) were recruited to participate in the study. Participants were fully informed about the nature of the study, and their rights as participants, including the right to withdraw at any time without fear of negative consequences. They signed informed consent. The study was approved by Ethical review board in the humanities and social and behavioral sciences of the University of Helsinki.

### 3.2 Stimuli

One hundred and twenty images from the IAPS database [27] were used in the study. The IAPS database was conceived as a catalog of pictures that represents the entire range of emotional reactions. It contains some images of violence, as well as images that are judged to be erotic, fear evoking, disgusting, and/or repellent by some viewers. Each image is associated with valence and arousal scores indicating their positivity and negativity. These scores have been validated across several studies on various populations. Of these, sixty were previously tested as having low, or negative, valence ($2.71 \pm 0.81$ on a scale from 1 to 9) and sixty as having high, or positive, valence ($6.94 \pm 0.53$). Each participant viewed a random

selection of forty images, half of them negative and half positive according to their valence scores. Images were scaled vertically to a maximum size of 1500x1024 px.

## 3.3 Procedure

For each participant, upon the signing of informed consent and after setting up the equipment, we recorded a 1-minute resting-state activity while a crosshair was shown on the screen. Then, experimental stimuli were displayed in two blocks of twenty trials each. Each trial began with an instruction to carefully consider the subsequent image and to freely associate any emotion. Upon pressing a key, a fixation cross was shown for 4 s, after which an image appeared onscreen. Each image was shown for 14 s, followed by an inter-trial interval showing a black screen for 0.1 s, after which the trial ended and the participant advanced to the next trial. The experiment took about 45 minutes to complete, excluding participant's preparation and device setup.

## 3.4 Apparatus

Stimulus display and timing used E-Prime 3 (Psychology Software Tools, Inc, Sharpsburg, PA) running on a PC under Windows 10. Synchronization of display with datastreams was achieved using the DCOM interface to send triggers from E-Prime to the fNIRS recording software, OxySoft (Artinis Medical Systems, Elst, The Netherlands).

Optical density data for fNIRS analysis were recorded using the Artinis Brite-24 device in a configuration with 10 diodes transmitting light at two wavelengths (760 and 850 nm) and 8 photodiodes detecting light. Diodes were positioned on an elastic cap and placed such that the distance between receivers and optodes approximated 30 mm and that, as illustrated in Figure 3, each receiver obtained light from three different transmitters, resulting in OD signals from 12 sources (dotted lines) per hemisphere. Optical densities from each of the resulting 24 channels were digitized at 50 Hz.

## 3.5 Data preprocessing

Raw optical density (OD) data were exported from OxySoft and processed using MNE.[1] To determine artefactual channels (e.g. due to poor placement, optode orientation, or hair blocking light), the scalp coupling index (SCI) [37] was calculated for each channel. SCI measures the negative correlation between HbO and HbR channels in the frequency range where the heartbeat is most apparent (0.7 - 1.5Hz). A strong negative correlation indicates that the optodes are adequately connected to the scalp. Channels suggesting poor contact with the scalp (SCI < 0.8) were marked as bad and interpolated from neighboring channels. Following, artefacts detected in the continuous signal were corrected using temporal derivative distribution repair [17]. The OD data were then converted to hemoglobin concentrations using the modified Beer-Lambert law [15] to derive oxy-hemoglobin (HbO) and deoxy-hemoglobin (HbR) levels. To these, a 0.1 Hz low pass filter was applied to remove physiological noise, while a 0.01 Hz high pass filter was applied to remove slow trends in signals that were unrelated to evoked activity. Finally, the continuous data were time-locked to the onset of stimuli and segmented into epochs of 17 s, including 5 s of pre-stimulus

[1] https://mne.tools/stable/auto_tutorials/preprocessing/70_fnirs_processing.html

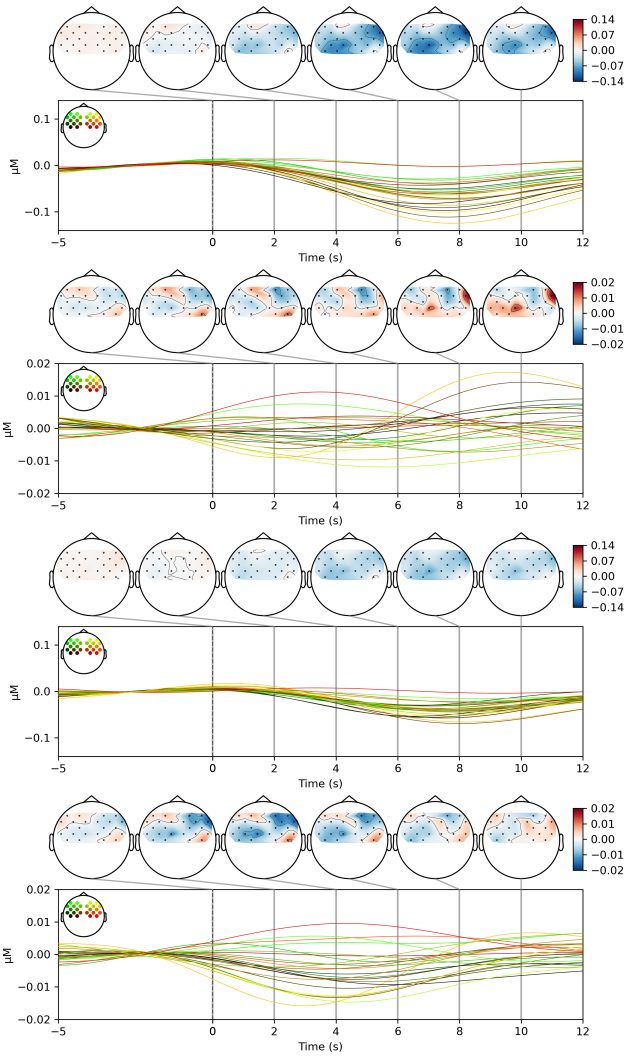



**Figure 2: Topographic and temporal plots of averaged HbO and HbR responses (in micro moles per liter) for low-valence (top) and high-valence (bottom) stimuli.**

baseline activity. The dataset is available for research purposes at https://osf.io/pd9rv/ [48].

## 3.6 Neural activity findings

To validate the test setup and estimate the mean effect of valence on stimulus-evoked HbO and HbR levels, the average baseline activity was subtracted from the first 12 s of post-stimulus activity, and averaged within low-arousal/positive, low-arousal/negative, high-arousal/positive, and high-arousal/negative conditions for each participant. To inspect spatial effects, we used a montage with only the transverse transmitter/diode pairs included. These were then grouped into a four-level *left-to-right* and a three-level anterior-to-posterior factor. These two location factors were combined with valence in a three-way repeated measures ANOVA. The analyses were conducted separately for HbO and HbR measures. Significance

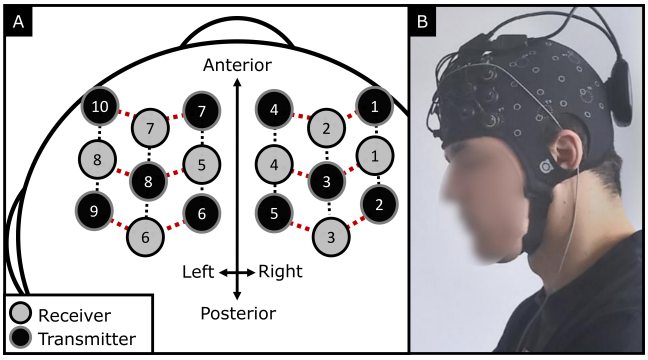



**Figure 3: Optode placement. A: schematic display of the transmitter and the receiver optodes across left and right hemispheres. Dotted lines: areas between diodes measured with fNIRS (in red: used in validation analysis). B: photograph of the fNIRS device (blurred for anonymity).**

is tested at alpha = .025 to account for inflated type-I error occurrence by performing multiple analyses. Non-significant location main effects are not reported to maintain brevity.

A repeated measures ANOVA on average HbO and HbR with left-to-right location, anterior-to-posterior location, and valence (positive vs negative) as factors, it was found that anterior-to-posterior significantly affected HbO, $F(2, 60) = 8.78$, $p = .0005$, $\eta_p^2 = .23$ (HbO: $p = .049$). This effect further interacted with left-to-right location in HbO, $F(6, 180) = 21.46$, $p < .0001$, $\eta_p^2 = .42$, and HbR, $p = .003$. Both the location main effect and interaction effect indicate regional differences in response to viewing an image, regardless of the emotional content. More interestingly, a main effect of valence was found for HbR, $F(1, 30) = 9.89$, $p = .004$, $\eta_p^2 = .25$, though not for HbO, $p = .31$. Negative valence images evoked stronger negativity ($-0.032 \pm 0.006$) than positive valence images ($-0.019 \pm 0.06$). The two location factors significantly mediated the interaction between left-to-right and anterior-to-posterior location in HbO $F(1, 180) = 12.32$, $p < .0001$, $\eta_p^2 = .29$ (HbR: $p = .035$), indicating the effect of valence was localized. As may be seen from Figure 2, the effect of valence on HbO could be characterized as a stronger negative effect for negative valence images in the prefrontal lateral areas of the right hemisphere as well as in the medial dorsal-frontal left hemisphere.

To determine whether the effects could also be observed once specifying the analysis towards the high-arousal stimuli, we repeated the same analysis but without the low arousal conditions. Thus, another repeated measures ANOVA on average HbO was conducted, with left-to-right location and anterior-to-posterior location, but now with valence (high-arousal/negative vs high-arousal/positive) as factors. Now, valence no longer had a significant main effect, $F(1, 30) = 1.83$, $p = .19$, $\eta_p^2 = .06$. The anterior-to-posterior location retained its main effect, $p = .02$, as did the interaction between the two location factors, $p < .0001$. Finally, a significant three-way interaction was observed, $F(6, 180) = 9.32$, $p < .0001$, $\eta_p^2 = .24$. Thus, valence continued to have a clear, lateralised effect.

# 4 AFFECT PREDICTION FROM fNIRS

The goal of this task is to classify presented stimuli as low-valence or high-valence based on the response in fNIRS signal. The same task is performed for three subsets of the IAPS database, one consisting of exclusively high-arousal stimuli, one of only low-arousal stimuli, and other containing both types of stimuli.

## 4.1 Feature extraction

Feature extraction was conducted for the preprocessed data (subsection 3.5). In fNIRS, a typical response to visual stimuli is some activity in a brain region approximately 4 to 12 seconds after the stimuli and then return to baseline [35, 60]. To capture this effect with a moderately sized feature space, we divided the 12-second post-stimulus period of each epoch into three 4-second windows and extracted the mean of each window for each channel. Finally, features from HbR channels were filtered out since HbO and HbR channel pairs are found strongly dependent, resulting in a 72-feature vector for each epoch.

## 4.2 Classification setup

We used a logistic regression classifier for this task. Individual models were trained for each participant using stratified sampling and 10-fold leave-one-out cross-validation. In each fold, fNIRS features were normalized to zero mean and unit variance based on the training set. The best combination of model hyperparameters where optimized with Optuna.[2] The classifier achieved the best results with $L_2$ regularization and regularization strength $C = 0.4$.

## 4.3 Classification results

Table 1 shows the fNIRS classification results and Figure 4 subject-specific ROC-AUC scores. The accuracy and AUC values show classification performance varying from AUC=0.58 to the best achieved for high-arousal images AUC=0.69. All results are statistically significant over a random classifier using permutation testing. As expected, the results suggest that low-arousal (uninteresting or non-exciting) affective states are more challenging to decode from brain signals, while high-arousal (interesting or exciting) affective states have a higher classification performance. However, according to previous experiments, arousal classification can also be conducted with comparable accuracy [43].

| Task | No. images | Accuracy | $p$-value | ROC AUC |
|---|---|---|---|---|
| High-arousal | 60 | $0.65 \pm 0.02$ | $< .01$ | $0.69 \pm 0.02$ |
| Low-arousal | 60 | $0.56 \pm 0.02$ | $< .01$ | $0.58 \pm 0.02$ |
| Combined | 120 | $0.58 \pm 0.02$ | $< .01$ | $0.60 \pm 0.02$ |

**Table 1: Valence classification results (Mean ± Std. Err.) for stimuli with different arousal classes.**

# 5 SIMILARITY AND RANKING EXPERIMENTS

We illustrate how our work can be used in a real-world scenario to retrieve affectively similar content, for which we conduct a series

[2] https://optuna.org/

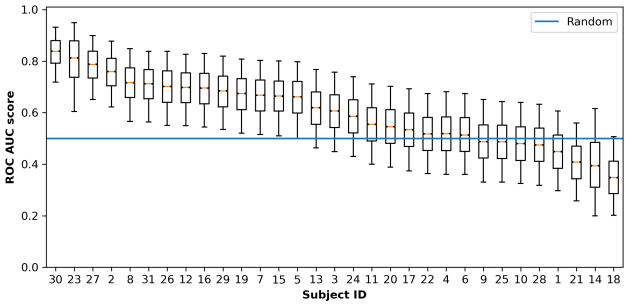



**Figure 4: Distribution of ROC AUC scores for each subject in low/high valence classification. Accuracy scores follow the same distribution pattern.**

of query-by-image information retrieval (IR) experiments: given a query image, all available images in a database are ranked by similarity and the top-$K$ are shown to the user.

We focus on the task of retrieving either high-valence (positive) or low-valence (negative) high-arousal images. This is motivated by the fact that currently, IR systems can only account for affective relevance by considering content-based analysis [59], but not retrieve relevant content based on whether it is positively or negatively *experienced* by an individual user. It is also understood that high-arousal stimuli correspond to generally interesting or relevant high-attention information [24], while low-arousal stimuli evokes less prominent responses.

The fNIRS classifier described before was used to assess the affective similarity of high-arousal IAPS images. The performance and rankings computed, based on the similarity assessments from brain responses, were compared to content-based retrieval models using textual (captions) and pixel-level image features.

## 5.1 Data preparation

The data from the participants who achieved a classification accuracy below 50% in our previous experiments were removed, motivated by the "BCI illiteracy" problem, which states that BCI does not work for a non-negligible portion of users, estimated to be around 15–30% of the population [53][3] This resulted in valid data for 23 participants (14% illiteracy rate).

## 5.2 Experimental setup

We focus on the cross-validated image similarity assessment setting. Each image, at a time, is used as a query $x$ to rank the rest of the images according to their similarity to that query. There are 60 images in our dataset, out of which 30 are positive (high-valence) and the remaining 30 are negative (low-valence) images.

We should note that ranking images instead of finding the closest matches according to similarity scores allows for a fair comparison between different retrieval models, as each model may generate similarity scores in completely unrelated scales. It also allows for combining different retrieval models easily, as shown later.

[3]BCI illiteracy is observed more often in active rather than passive BCI.

Given that we are interested in determining the similarity of images that evoke similar emotions as to a given query image, an image will be considered a "match" if it belongs to the same class (high-valence/positive or low-valence/negative) as the query image, and vice versa for non-matching images. This is important for computing Recall, since we know that the maximum number of matching images we can assess for any query image is 30 (half of the images in the dataset belong to either the positive or negative class). For each query image, we evaluate the performance of each model at different rank positions $K \in \{1, 5, 10, 20\}$.

## 5.3 Evaluation measures

The experiments used two distinct evaluation aspects: ranking similarity and similarity assessment performance. On the one hand, ranking similarity indicates whether the different models provide similar or distinct and complementary performance when compared to others. It is measured using the following metrics (higher is more similar): Overlap (Agreement rate between rankings),Rank Biased Overlap (RBO) (Agreement rate between rankings, also considering their order [56]), Intersection over Union (IoU) (Jaccard index between rankings). On the other hand, similarity assessment performance indicates the ability of a model in detecting similar images when compared to the query image. It is measured using the following metrics (higher is better): Precision (Fraction of items among the ranked top-$K$ items), Recall (Fraction of matching items in the ranked top-$K$ among all matching items in the database), and F-measure (The harmonic mean of Precision and Recall).

## 5.4 Ranking models

Aiming at testing different ranking models, together with our fNIRS classifier and a control IAPS (ground-truth) model, we also considered content-based and text-based models to quantify how the fNIRS model compares to similarity assessed by models that use conventional features from images or their associated text descriptions. Table 2 provides an overview of these models.

**IAPS:** Our reference model uses the arousal and valence scores from the original IAPS database [27] as image features. The ground-truth ranking is computed in a 2-dimensional vector space (arousal and valence scores) by using the Euclidean distance as the dissimilarity metric. For each query image we will compare this ground-truth ranking against the other ranking models.

**fNIRS:** This model uses the softmax vector of the classifier we trained in our previous experiments and estimates valence scores for each image assessed by each user. Since we focus on retrieving high-valence or low-valence high-arousal images, we set the arousal score to be the average of the centroid that represents the predicted class and use the following classification rule to estimate the valence score:

$$v(x,u) = \begin{cases} \min \mathcal{V}/p(x,u) & \text{if } x \text{ is predicted as negative} \\ \max \mathcal{V} \cdot p(x,u) & \text{if } x \text{ is predicted as positive} \end{cases} \quad (1)$$

where $\mathcal{V}$ is the set of all valence scores in our dataset and $p(x,u)$ is the classification probability of image $x$ belonging to either the positive or the negative class, according to our classifier's softmax vector for user $u$. The idea is to deviate from minimal and maximal values of valence based on the confidence of the predictions delivered by

| Model | Feature representation | Dimensionality | Distance | Notes |
|---|---|---|---|---|
| IAPS | Arousal & Valence scores | 2 | Euclidean | Ground-truth data |
| fNIRS | Positive/Negative softmax | 2 | Euclidean | Logistic regression |
| IMAGE | RGB Convolution maps | 512 | $\chi^2$ | Transfer learning w/ ResNet50 |
| TEXT | Sentence embeddings | 384 | Cosine | SBERT Lite uncased, image captions |

**Table 2: Overview of the similarity assessment models.**

our classifier so that each fNIRS ranking is tailored to each participant and query image. For example, if the softmax vector for a given user and a given query image $x$ is $[p(x,u|-), p(x,u|+)] = [0.1, 0.9]$, it means our classifier is highly confident about $x$ being a positive image for that user, so the estimated valence score would be $v(x,u) = 0.9 \max \mathcal{V}$ and the estimated arousal score would be the centroid $a(x,u) = 1/|\mathcal{X}^+| \sum_{i \in \mathcal{X}^+} a(x_i)$, where $x_i$ is the $i$th image in the set of positive $\mathcal{X}^+$ IAPS images. Then, the final ranking will comprise the closest images to these valence values according to the Euclidean distance to the query image $x$.

**IMAGE:** This is a visual content-based model based on the popular deep residual convolutional network ResNet50 [18] as a feature extractor. We follow the usual transfer learning procedure for feature extraction: load a pre-trained model, remove the last fully-connected output layers, and add a global average pooling layer in order to get a flattened convolutional map that summarizes the input image. In this case, we use the $\chi^2$ distance as a dissimilarity metric, since it has been shown to work better than the Euclidean distance with high-dimensional items such as images [34]. We experimented with other models such as a VGG16 [46] and Inception [49], however ResNet50 provided higher performance and so it was selected as the IMAGE model.

**TEXT:** This is a textual content-based model based on SBERT [41], a modification of the well-known pre-trained BERT architecture [16]. SBERT is a Transformer model that was trained using siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using the cosine similarity measure. Since the IAPS database does not provide textual descriptions, we use the Microsoft Cognitive Services[4] to extract image captions and use these as the textual descriptions for the images. We also extracted image tags and experimented with other text ranking models, such as the classic BM25 [22] and Doc2Vec [28], but SBERT using image captions provided higher performance and so it was selected as the TEXT model.

**Fused rankings:** We also compare rankings using the Reciprocal Rank Fusion (RRF) technique [11]. RRF ranks items based on their (inverted) position, resulting in a fused ranking where the top-ranked items have high agreement across the combined rankings and vice versa for the least ranked items [58]. Fused rankings were computed for fNIRS+IMAGE, and fNIRS+TEXT to study the performance of the models when brain signals were combined with other features. For each query, we first rank the whole dataset (60 images) with each model, then perform RRF, and finally keep the top-$K$ as the final ranking.

[4]https://azure.microsoft.com/en-us/services/cognitive-services/

## 5.5 Ranking similarity results

Table 3 shows the ranking similarity according to the selected evaluation measures. We can observe that similarity varies considerably across models and shows rather low overlaps with the ground-truth IAPS ranking. Interestingly, both fNIRS and IMAGE rankings are more similar to IAPS rankings. The likely reason is that, while fNIRS directly captures affective information, IAPS images have been suggested to be confounded with visual features [40]. We also report the results of comparing against random rakings, providing thus an empirical lower bound.

| Model | Overlap | RBO | IoU |
|---|---|---|---|
| fNIRS | 0.18 ± 0.13 | 0.11 ± 0.10 | 0.13 ± 0.11 |
| IMAGE | 0.21 ± 0.15 | 0.12 ± 0.12 | 0.13 ± 0.11 |
| TEXT | 0.11 ± 0.04 | 0.07 ± 0.03 | 0.06 ± 0.02 |
| fNIRS+IMAGE | 0.20 ± 0.14 | 0.12 ± 0.11 | 0.12 ± 0.10 |
| fNIRS+TEXT | 0.13 ± 0.09 | 0.09 ± 0.07 | 0.08 ± 0.06 |
| Random | 0.13 ± 0.11 | 0.06 ± 0.07 | 0.07 ± 0.07 |

**Table 3: Ranking similarity evaluation (± std) against the IAPS rankings, using the top-10 ranked results.**

We can see that TEXT rankings overlap less with the ground-truth IAPS rankings. This can be attributed to the quality of the automatic image captions, which were often inaccurate for many of the negative images. Recall that the IAPS database contains many images that can be considered out-ot-distribution for state-of-the-art image captioning systems. For example, there are many explicit images (e.g. pornographic and violent content) that can hardly be captioned correctly in an automated way, given that the vast majority of image captioning systems are trained on large datasets such as COCO [30] and Flickr [36], which mainly contain natural images and general-purpose non-explicit content.

As can be observed, the fused fNIRS+IMAGE rankings do not agree much more than the individual rankings alone, whereas the fused fNIRS+TEXT ranking showed a marginal improvement for the TEXT ranking. A possible explanation for this is that the three different ranking models were providing orthogonal information. To study this, we conducted an additional experiment where we compare the fNIRS rankings against IMAGE and TEXT rankings. Table 4 shows the results of this comparison and reveals that, indee, the three ranking models yield different, orthogonal rankings. This observation, however, is encouraging since it suggests that the

fNIRS model may provide additional diversity to the rankings provided by the other models, i.e., information that is emotionally related to the query images.

| $K$ | Model | Overlap | RBO | IoU |
|---|---|---|---|---|
| 10 | fNIRS vs IMAGE | 0.16 ± 0.10 | 0.09 ± 0.07 | 0.10 ± 0.07 |
| 10 | fNIRS vs TEXT | 0.11 ± 0.08 | 0.06 ± 0.07 | 0.06 ± 0.05 |
| 20 | fNIRS vs IMAGE | 0.35 ± 0.14 | 0.21 ± 0.12 | 0.22 ± 0.11 |
| 20 | fNIRS vs TEXT | 0.21 ± 0.09 | 0.12 ± 0.06 | 0.12 ± 0.06 |

**Table 4: Ranking similarity evaluation (± std) of fNIRS against the other models using top-$K$ results.**

## 5.6 Similarity assessment results

**Individual models:** The three leftmost plots in Figure 5 show the results of the similarity assessment experiments for the individual models. Statistical analysis were run to compare the performance of the individual models at $K = 20$ for Precision and Recall. A $\chi^2$ test of proportions as omnibus test revealed significant differences between models in terms of Precision ($\chi^2(2, N = 422) = 35.672, p < .001$). Pairwise tests of proportions (Bonferroni-Holm corrected) as post-hoc test revealed significant differences between fNIRS and TEXT ($p < .01$), and IMAGE and TEXT ($p < .001$). No differences were found between IMAGE and fNIRS. This suggests that image features and fNIRS features best capture the emotional information and yield to comparable precision, around 65% in both cases. In terms of Recall, the omnibus test was significant ($\chi^2(2, N = 422) = 15.188, p < .001$) and post-hoc tests revealed significant differences between IMAGE and fNIRS ($p < .001$) and TEXT ($p = .021$). In summary, we can conclude that the affective states measured from the brain are equally precise as the image representations, but image representations yield higher recall.

**Fused models:** The two rightmost plots in Figure 5 show the results of the similarity assessment experiments for the fused rankings. Again, statistical analysis were run to compare the performance of the fused models and their components at $K = 20$ for Precision and Recall. When comparing fNIRS against the fused models, no significant differences were found in terms of Precision, but there were statistically significant differences in terms of Recall ($\chi^2(2, N = 422) = 87.670, p < .001$). The post-hoc test of pairwise comparisons (Bonferroni corrected) revealed significant differences between both fused models and fNIRS ($p < .001$). This suggests that, to achieve the best similarity assessment performance, one should combine fNIRS with visual features to benefit from both image and brain representations. When comparing the IMAGE model against the fused models, no significant differences were found. When comparing the TEXT model against the fused models, significant differences were found in terms of Precision ($\chi^2(2, N = 422) = 35.672, p < .001$) but not for Recall. The post-hoc test revealed significant differences for Precision between TEXT and both fused models ($p < .001$). In summary, fNIRS and IMAGE models both assess precise similarity and overall the combination of fNIRS+IMAGE delivers the most accurate similarity assessments.

## 5.7 Visual analysis

To provide more insights into our results, we show examples of top-10 rankings in Figure 6 for both positive (high-valence) and negative (low-valence) query images. Overall, we can see that the type of similarity assessed by each model is rather different. This is in part explained by the tendency of the content-based models to exploit high-similarity content and "lock" the user by assessing similarity based on visual and textual similarities that do not count for affective diversity, whereas fNIRS tends to assess more diverse similarity features that match the affective content of the image, independently of the content features.

Notably, the IAPS database contains very diverse images that can be considered emotionally similar for a given query image, but that are not topically or visually similar. This is particularly detrimental for the IMAGE model, since in IAPS there might be only a handful of images that are visually similar to a given query. For example, an image about gun violence is considered a good affective match when searching for negative images, but there are quite many more images representing negative emotional content that have nothing to do with gun violence. While this may hurt the performance of content-based models, it also reflects real-world scenarios where users may prefer diverse content matching their information needs rather than results with highly similar or almost identical content.

# 6 DISCUSSION

Current similarity assessment strategies rely on objective/semantic measures rather than subjective/affective perception measures. Such strategies insufficiently account for the emotional experiences of users. We have shown that affective information can be directly decoded from brain activity and incorporated into similarity assessment models. In the following, we answer the research questions posed at the beginning of this paper.

**Can valence (positive and negative affective states) be detected, predicted, and used for affective ranking?** Our results show that valence has significant effects in fNIRS responses. Particularly strong effects were found in valence for high-arousal content. That is, content that is *a priori* shown to be attention-grabbing shows the most robust responses between positive (high-valence) and negative (low-valence) emotional experiences. Classification results from single-trial decoding experiments show that valence can be decoded with reasonable accuracy, however the decoding of low-arousal content is less effective. This is somewhat expected, as low-arousal content is known to evoke diminutive emotional responses [8]. High-arousal content is also of higher interest for many applications, as it is associated with more vigorous user attention [27].

**Does affective information complement similarity-based ranking models, as compared to their content-based alternatives?** While the performance of affective information decoded directly from fNIRS is generally less effective than the IMAGE model in determining affective similarity, the IMAGE model performance is higher only for high similarity ranks. This suggests, as expected, that highly similar content features (look-alike images) are good indicators of affective similarity. However, the performance of the IMAGE model decays rapidly, indicating that it fails to capture similarity that has an affective dimension, but different content features

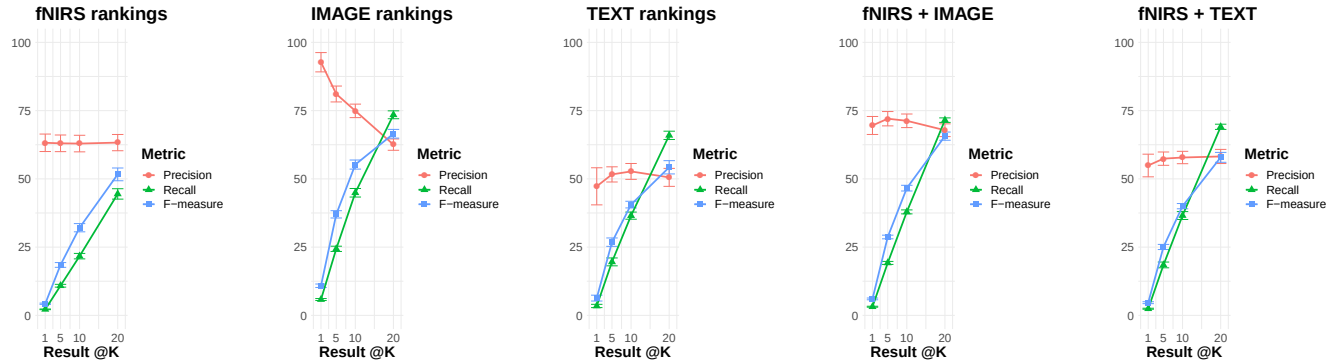
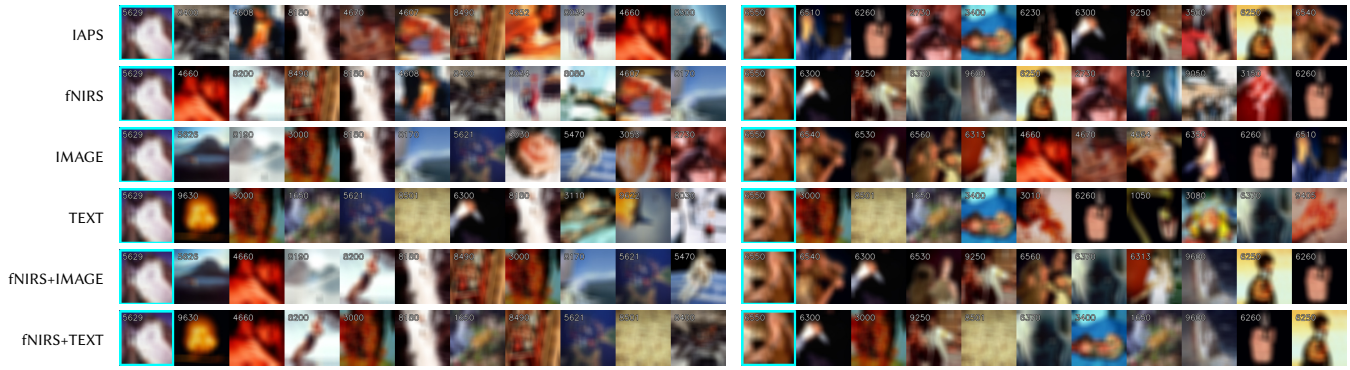



**Figure 5: Similarity assessment performance evaluation results, both for individual models (three leftmost plots) and fused rankings (two rightmost plots). Error bars denote the standard error of the mean.**




**(a) High-valence retrieval examples** **(b) Low-valence retrieval examples**

**Figure 6: Sample rankings of *positive* (high valence) and *negative* (low valence) images for two randomly-chosen users. The first image in each ranking, highlighted in cyan, is always the query image. Images are blurred because IAPS images cannot be published in any format, to retain their integrity for use in experimental studies. We have indicated the IAPS image IDs to assist in replication.**

(i.e., affectively similar but visually dissimilar images). This suggests that fNIRS and other physiological and brain-computer interfaces can provide crucial additional information about affective similarity that goes beyond content-based features.

**Limitations and future work.** The visual stimuli we used were selected from a standard database widely used for studying emotional reactions [27]. Yet, there there might be a risk of fNIRS signals interfering with other affective dimensions influencing similar brain regions such as approach/avoidance that can typically be triggered by some of the IAPS images. Nevertheless, we should note that the achieved decoding performance from fNIRS is in line with previous studies [31, 35, 51, 52] using subject-specific models. While the ability to transfer features between subjects is desirable, cross-subject performance remains an area for future research.

The content-based models we used represent state-of-the-art approaches, and they can also be considered fair control conditions to study the effects of emotional and affective dimensions of similarity. Nevertheless, we cannot exclude the possibility that finetuning the content-based models might lead to improved performance results. In particular, the full 72-D fNIRS feature vector could be directly used to learn improved models instead of 2-D logits.

## 7 CONCLUSION

This work goes beyond the conventional goal of simply optimizing for content similarity. The affective signature obtained from fNIRS brain signals is utilized to determine affective similarity. Brain signals are not just based on the similarity of content, but on the similarity of the user-evoked emotional responses. This technique may have fundamental implications for search, recommendation, and multimedia content personalization when affective similarity can be deployed in real-world application scenarios.

## ACKNOWLEDGMENTS

This work is supported by the Academy of Finland (grants 352915, 350323, 336085, 322653), the Horizon 2020 FET program of the European Union through the ERA-NET Cofund funding grant CHIST-ERA-20-BCI-001, and the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147).

# REFERENCES


[1] A. R. Andreu-Perez, M. Kiani, J. Andreu-Perez, P. Reddy, J. Andreu-Abela, M. Pinto, and K. Izzetoglu. 2021. Single-Trial Recognition of Video Gamer's Expertise from Brain Haemodynamic and Facial Emotion Responses. *Brain Sci.* 11, 1 (2021).

[2] I. Arapakis, J. M. Jose, and P. D. Gray. 2008. Affective Feedback: An Investigation into the Role of Emotions in the Information Seeking Process. In *Proc. SIGIR.*

[3] P. Avero and M. G. Calvo. 2006. Affective priming with pictures of emotional scenes: The role of perceptual similarity and category relatedness. *Span. J. Psychol.* 9, 1 (2006).

[4] H. Ayaz, M. Izzetoglu, K. Izzetoglu, and B. Onaral. 2019. The use of functional near-infrared spectroscopy in neuroergonomics. In *Neuroergonomics.*

[5] M. Balconi, E. Grippa, and M. E. Vanutelli. 2015. What hemodynamic (fNIRS), electrophysiological (EEG) and autonomic integrated measures can tell us about emotional processing. *Brain Cogn.* 95 (2015).

[6] D. Bandara, L. Hirshfield, and S. Velipasalar. 2019. Classification of affect using deep learning on brain blood flow data. *J. Near Infrared Spectrosc.* 27, 3 (2019).

[7] S. C. Bunce, M. Izzetoglu, K. Izzetoglu, B. Onaral, and K. Pourrezaei. 2006. Functional near-infrared spectroscopy. *IEEE Eng. Med. Biol. Mag.* 25, 4 (2006).

[8] L. Carretié. 2014. Exogenous (automatic) attention to emotional stimuli: a review. *Cogn. Affect. Behav. Neurosci.* 14, 4 (2014).

[9] M. Chen, L. Zhang, and J. P. Allebach. 2015. Learning deep features for image emotion classification. In *Proc. ICIP.*

[10] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. 2020. Uniter: Universal image-text representation learning. In *Proc. ECCV.*

[11] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proc. SIGIR.*

[12] K. M. Davis, M. Spapé, and T. Ruotsalo. 2022. Contradicted by the Brain: Predicting Individual and Group Preferences via Brain-Computer Interfacing. *IEEE Trans. Affect. Comput.* (2022).

[13] K. M. Davis III, M. Spapé, and T. Ruotsalo. 2021. Collaborative Filtering with Preferences Inferred from Brain Signals. In *Proc. WWW.*

[14] C. de la Torre-Ortiz, M. Spapé, and T. Ruotsalo. 2023. The P3 indexes the distance between perceived and target image. *Psychophysiology* 60, 5 (2023), e14225.

[15] D. T. Delpy, M. Cope, P. van der Zee, S. Arridge, S. Wray, and J. Wyatt. 1988. Estimation of optical pathlength through tissue from direct time of flight measurement. *Phys. Med. Biol.* 33, 12 (1988).

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL.*

[17] F. A. Fishburn, R. S. Ludlum, C. J. Vaidya, and A. V. Medvedev. 2019. Temporal derivative distribution repair (TDDR): a motion correction method for fNIRS. *Neuroimage* 184 (2019).

[18] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. CVPR.*

[19] D. Heger, C. Herff, F. Putze, R. Mutter, and T. Schultz. 2014. Continuous affective states recognition using functional near infrared spectroscopy. *Brain Comput. Interfaces* 1, 2 (2014).

[20] X. Hu, C. Zhuang, F. Wang, Y.-J. Liu, C.-H. Im, and D. Zhang. 2019. fNIRS evidence for recognizably different positive emotions. *Front. Hum. Nneurosci.* 13 (2019).

[21] R. Jenke, A. Peer, and M. Buss. 2014. Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 5, 3 (2014).

[22] K. S. Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments. *Inf. Proces. Manag.* 36, 6 (2000).

[23] G. Kazai, P. Thomas, and N. Craswell. 2019. The Emotion Profile of Web Search. In *Proc. SIGIR.*

[24] J. Kenemans, M. Verbaten, W. Sjouw, and J. Slangen. 1988. Effects of task relevance on habituation of visual single-trial ERPs and the skin conductance orienting response. *Int. J. Psychophysiol.* 6, 1 (1988).

[25] A. Khosla, A. Das Sarma, and R. Hamid. 2014. What makes an image popular?. In *Proc. WWW.*

[26] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2011. DEAP: A database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* 3, 1 (2011).

[27] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. 2008. *International affective picture system (IAPS): Technical manual and affective ratings.* Technical Report A-8. NIMH Center for the Study of Emotion and Attention.

[28] Q. Le and T. Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. ICML.*

[29] L. A. Leiva, S. Morteza, and A. Oulasvirta. 2022. Modeling How Different User Groups Perceive Webpage Aesthetics. *Univers. Access Inf. Soc.* (2022).

[30] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. 2014. Microsoft COCO: Common Objects in Context. In *Proc. ECCV.*

[31] L. Liu and Z. Chen. 2021. The Application of Closed-loop Brain Training: Near-infrared Spectroscopy (NIRS) Neurofeedback. In *Proc. ISAIMS.*

[32] Z. Liu, J. Shore, M. Wang, F. Yuan, A. Buss, and X. Zhao. 2021. A systematic review on hybrid EEG/fNIRS in brain-computer interface. *Biomed. Signal Process. Control* 68 (2021).

[33] J. Machajdik and A. Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proc. MM.*

[34] S. Patil and S. Talbar. 2010. Content Based Image Retrieval Using Various Distance Metrics. In *Proc. ICDEM.* LNCS 6411.

[35] P. Pinti, I. Tachtsidis, A. Hamilton, J. Hirsch, C. Aichelburg, S. Gilbert, and P. W. Burgess. 2020. The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Ann. N. Y. Acad. Sci.* 1464, 1 (2020).

[36] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. 2017. Flickr30K Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *Int. J. Comput. Vis.* 123, 1 (2017).

[37] L. Pollonini, C. Olds, H. Abaya, H. Bortfeld, M. S. Beauchamp, and J. S. Oghalai. 2014. Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. *Hear. Res.* 309 (2014).

[38] K. Qing, R. Huang, and K.-S. Hong. 2021. Decoding Three Different Preference Levels of Consumers Using Convolutional Neural Network: A Functional Near-Infrared Spectroscopy Study. *Front. Hum. Neurosci.* (2021).

[39] T. Rao, X. Li, and M. Xu. 2020. Learning multi-level deep representations for image emotion classification. *Neural Process. Lett.* 51, 3 (2020).

[40] C. Redies, M. Grebenkina, M. Mohseni, A. Kaduhm, and C. Dobel. 2020. Global image properties predict ratings of affective pictures. *Front. Psychol.* 11 (2020).

[41] N. Reimers and I. Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. EMNLP-IJCNLP.*

[42] T. Ruotsalo, K. Mäkelä, and M. Spape. 2023. Crowdsourcing Affective Annotations via fNIRS-BCI. *IEEE Trans. Affect. Comput.* (2023).

[43] T. Ruotsalo, K. Mäkelä, M. M. Spapé, and L. A. Leiva. 2023. Affective Relevance: Inferring Emotional Responses via fNIRS Neuroimaging. In *Proc. SIGIR.*

[44] G. Schalk and E. C. Leuthardt. 2011. Brain-computer interfaces using electrocorticographic signals. *IEEE Rev. Biomed. Eng.* 4 (2011).

[45] W. C. Shipley, J. I. Coffin, and K. C. Hadsell. 1945. Affective distance and other factors determining reaction time in judgments of color preference. *J. Exp. Psychol.* 35, 3 (1945).

[46] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. ICLR.*

[47] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intellig.* 22, 12 (2000).

[48] M. Spape, K. Mäkelä, and T. Ruotsalo. 2023. NEMO: A Database for Emotion Analysis Using Functional Near-infrared Spectroscopy. *Preprint. To appear.* (2023).

[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going Deeper with Convolutions. In *Proc. CVPR.*

[50] K. Tai and T. Chau. 2009. Single-trial classification of NIRS signals during emotional induction tasks: towards a corporeal machine interface. *J. Neuroeng. Rehabil.* 6, 1 (2009).

[51] L. R. Trambaiolli, A. Tiwari, and T. H. Falk. 2021. Affective neurofeedback under naturalistic conditions: a mini-review of current achievements and open challenges. *Front. Neuroergonomics* 2 (2021).

[52] L. R. Trambaiolli, J. Tossato, A. M. Cravo, C. E. Biazoli Jr, and J. R. Sato. 2021. Subject-independent decoding of affective states using functional near-infrared spectroscopy. *PLOS One* 16, 1 (2021).

[53] C. Vidaurre and B. Blankertz. 2010. Towards a Cure for BCI Illiteracy. *Brain Topogr.* 23, 2 (2010).

[54] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *Proc. MM.*

[55] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proc. CVPR.*

[56] W. Webber, A. Moffat, and J. Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4 (2010).

[57] V. Yanulevskaya, J. C. van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek. 2008. Emotional valence categorization using holistic image features. In *Proc. ICIP.*

[58] B. A. Yilma and L. A. Leiva. 2023. The Elements of Visual Art Recommendation: Learning Latent Semantic Representations of Paintings. In *Proc. CHI.*

[59] Q. You, J. Luo, H. Jin, and J. Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proc. AAAI.*

[60] M. A. Yücel, A. v. Lühmann, F. Scholkmann, J. Gervain, I. Dan, H. Ayaz, D. Boas, R. J. Cooper, J. Culver, C. E. Elwell, A. Eggebrecht, M. A. Franceschini, C. Grova, F. Homae, F. Lesage, H. Obrig, I. Tachtsidis, S. Tak, Y. Tong, A. Torricelli, H. Wabnitz, and M. Wolfc. 2021. Best practices for fNIRS publications. *Neurophotonics* 8, 1 (2021).

[61] S. Zhao, G. Ding, Q. Huang, T.-S. Chua, B. W. Schuller, and K. Keutzer. 2018. Affective Image Content Analysis: A Comprehensive Survey. In *Proc. IJCAI.*

[62] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *Proc. MM.*

[63] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B. W. Schuller, and K. Keutzer. 2021. Affective image content analysis: Two decades review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intellig.* (2021).