Automatic Analysis of Facilitated Taste-liking

Yifan Chen[*]
University of Cambridge, UK
yc462@cantab.ac.uk

Zhuoni Jie[†]
University of Maryland, USA
jiezn96@gmail.com

Hatice Gunes
University of Cambridge, UK
Hatice.Gunes@cl.cam.ac.uk

ABSTRACT

This paper focuses on: (i) Automatic recognition of taste-liking from facial videos by comparatively training and evaluating models with engineered features and state-of-the-art deep learning architectures, and (ii) analysing the classification results along the aspects of facilitator type, and the gender, ethnicity, and personality of the participants. To this aim, a new beverage tasting dataset acquired under different conditions (human vs. robot facilitator and priming vs. non-priming facilitation) is utilised. The experimental results show that: (i) The deep spatiotemporal architectures provide better classification results than the engineered feature models; (ii) the classification results for all three classes of liking, neutral and disliking reach F1 scores in the range of 71%-91%; (iii) the personality-aware network that fuses participants' personality information with that of facial reaction features provides improved classification performance; and (iv) classification results vary across participant gender, but not across facilitator type and participant ethnicity.

CCS CONCEPTS

- **Human-centered computing → Interactive systems and tools**;
- **Computing methodologies → Computer vision**.

KEYWORDS

Taste-liking, facial reactions, affective computing, engineered features, deep spatiotemporal networks, personality

Reference this paper as:
Yifan Chen, Zhuoni Jie, and Hatice Gunes. 2020. Automatic Analysis of Facilitated Taste-liking. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3395035.3425645

1 INTRODUCTION

An objective, non-invasive, and instant way to measure and analyse people's liking of the taste of food products will have applications in creating robot connoisseurs and waiters for the hospitality industry, predicting consumer acceptance and satisfaction, as well as studying people's nutritional behaviours and food preference developments. The majority of the methods for measuring taste-liking have focused on explicit methods based on self-reported ratings (*e.g.*, [21]). These have limitations caused by time and ability to give feedback [8], difficulty in quantifying rating metrics [28], and inherent bias due to people's conscious and rational processing in answering the questions [23]. Automatic analysis and understanding of the nonverbal behavioural aspects of the customer taste-liking and response, including their facial and bodily expressions, and nonverbal aspects of their speech, can be utilised to mitigate this issue.

Facial expressions are the most widely used cues for measuring affective states [12], and for predicting behaviour and attitude changes [24]. Within interpersonal interaction, the communication of liking expressions has been identified as critical in the advancement of relational development [13]. Facial expressions are reliable indicators of spontaneous feelings [11, 26], and can reveal spontaneous appreciation or dislike while eating and drinking [42, 43].

This work is a continuation of our research on investigating taste-liking with a humanoid robot facilitator [19] where we conducted the first beverage tasting study with a human versus a humanoid social robot facilitator, with priming versus non-priming instruction styles. We reported that the facilitator type and facilitation style had no significant influence on cognitive taste-liking. However, in robot facilitator scenarios, people were more willing to follow the instruction and felt more comfortable when facilitated with priming. In this previous work we did not focus on developing methods for automatic recognition of facial taste-liking.

Therefore, in this paper, we specifically focus on automatic analysis of facial taste-liking and utilise the beverage tasting dataset we have introduced in [19]. The contribution of our work includes: (1) automatic recognition of facilitated taste-liking from facial videos through comparatively training and evaluating several models with engineered features and a number of state-of-the-art deep learning architectures, and (2) analyses of classification results along the aspects of facilitator type, and the gender, ethnicity, and personality of the participants. The former is an important component for creating efficient and engaging robot facilitators for the hospitality industry, and the latter is expected to provide insights on whether the developed models are impacted by user demographics which will enable informed adaption of the robot facilitators to different user groups.

2 RELATED WORK

2.1 Facial Reactions to Tastes

Several studies in infants have shown that differential facial reactions elicited by varying concentrations of taste and odour stimuli [34, 38] and with different taste concentrations [15] have an innate basis and genetic origin [42], and remain more or less stable into

[*] The first two authors have contributed equally to this work.
[†] The first two authors have contributed equally to this work.

adulthood [16, 39]. Studies have also consistently demonstrated newborns showing expressions indicating pleasure in response to sweet flavours, and indicating negative emotions in response to sour, bitter, and sometimes salty flavours [42]. People have innate behaviours to show a preference for sweetness and aversion to bitterness, which is found to be independent of culture [32]. People's liking and disliking of different flavours can evoke corresponding facial reactions. Sensory studies have demonstrated that negative facial reactions are more intense, quicker to appear, and easier to recognise than positive facial expressions [46]. Consistent facial responses of nose wrinkling, furrowing of the forehead, as well as brow lowering and mouth opening after swallowing are found to be caused by the bitter taste [40]. The sour taste elicits consistent facial reactions including elicited lip pursing and nose wrinkling [38]. The abovementioned findings related to taste-related facial reactions can be utilised for automatic taste-liking estimation.

2.2 Automatic Facial Reaction Analysis

Facial gestures and movements are mostly analysed in terms of the emotional information they communicate which has led to the development of automatic facial expression recognition (FER) systems and tools. The research field of FER has seen significant progress in recent years due to the availability of novel sensors, publicly available datasets, crowdsourced labels, and novel machine learning techniques [37]. FER approaches usually extract hierarchical feature representations using carefully hand-crafted features [37], [48] or, more recently, data-driven methodologies [25], to analyse and understand human facial expressions. The recent success of deep learning has further enhanced their performance by reducing the dependency on the choice of features used [22].

The Facial Action Coding System (FACS) [14] is another way of (manually) quantifying and detecting subtle changes in facial features, with a catalogue of 44 unique action units (AUs) that correspond to the movements of the face's individual muscles. Previous works have shown that facial AUs can communicate positive and negative emotional tones [4] and various automatic systems have been developed to analyse facial AUs (see [27] for a survey).

Previous research has demonstrated that there is a well established link between personality and the way people express their emotions. In particular, subjects who display higher levels of extroverson and neuroticism are reported to be more likely to express their emotions via their face [33]. There is also existing work reporting that it is possible to predict automatically the personality of a person from facial expressions in the context of online conversational video [3]. Therefore, in this paper we exploit the link between personality and the facial reactions for automatic analysis of taste-liking.

2.3 Automatic Analysis of Facial Taste-liking

Dibeklioglu and Gevers in [8] investigated algorithms for automatic taste liking prediction on a large-scale dataset to evaluate six different beers with the only independent variable being the type of beverage. They mapped the overall liking scores from 9-point scale to 3-point scale (1, 2, and 3 indicating dislike; 4, 5, and 6 indicating neutral state; 7, 8, and 9 indicating liking), and used these as class labels. Essentially, they evaluated several feature extraction and machine learning methodologies for automatic classification of taste-liking into three classes: dislike, neutral, and like. They achieved the best classification results using deep-learned expression dynamics encoded into Fisher Vectors (FV), which exploited expression dynamics such as the speed and acceleration of facial movements, achieving an accuracy of 70.37% for distinguishing between the three levels of taste liking. However, this dataset and its labels are not publicly available for research purposes.

Zhi et al. in [47] presented a direct mapping between hedonic rating and facial responses evoked by various taste stimuli using optical flow and genetic algorithms. Optical flow is employed to analyze facial characteristics of the subjects' facial responses evoked by taste stimuli, while support vector machine (SVM) was used for hedonic rating identification. They reported that the higher the number of categories used, the lower the recognition accuracy is. Accuracies along different modes were reported as follows: two-class mode "1–4/6–9", three-class mode "1–4/5/6–9", four-class mode "1–2/3–4/6–7/8–9", and five-class mode "1–2/3–4/5/6–7/8–9" achieved 64.9%, 45.7%, 36.3%, 26.3%, respectively. Genetic algorithm was also utilised to select facial regions that have high contribution to hedonic rating identification. They reported that the texture changes of eye area, wrinkles at the nasal root, and mouth area most significantly reflect the facial reaction corresponding to hedonic rating. However, this work did not investigate the usage of deep learning architectures that are known to provide state of the art recognition performance in many fields.

Jie and Gunes in [19] conducted a facial reaction analysis by focusing on AU features as the information source providing affective cues for self-reported taste-liking. They used the existing facial behaviour analysis toolkit OpenFace [2], and obtained AU intensity and presence for each tasting video clip. They also compared for each AU, their presence ratios and average intensities in videos by grouping them in terms of disliked and liked video labels. However, they did not report on automatic classification in terms of taste-liking.

In this paper, we utilise the beverage tasting dataset introduced in [19] and focus on (i) automatically recognising facial taste-liking from videos by comparatively training and evaluating both models with engineered features and state-of-the-art deep learning architectures, and (ii) analysing how the performance of the taste-liking classifiers evaluated vary across the aspects of facilitator type, and the gender, ethnicity and personality of the participants.

2.4 Deep Spatiotemporal Architectures

Deep learning-based methods, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and their variants, have already demonstrated their effectiveness in automatic expression recognition [10]. There has been a recent interest in deep spatiotemporal networks [17, 41], as they can encode temporal dependencies in successive frames and learn spatial features in conjunction with temporal features simultaneously, which boosts the performance in general. The works of [1, 30, 35] use 3D convolutional kernels with shared weights along the temporal axis instead of the traditional 2D kernels, which has been widely used for dynamics-based FER. The works of [20] and [18] combined the strength of multiple methods as they can cascade the outputs of CNNs with Long Short Term Memory networks (LSTMs) for various vision tasks involving time-varying inputs. In this paper, we employ the 3D-CNN architecture to exploit both spatial and temporal information to encode more subtle variations of facial behaviour.

3 AUTOMATIC RECOGNITION OF TASTE-LIKING

In this section, our goal is to achieve automatic recognition of taste-liking from facial videos by comparatively training and evaluating several models with engineered features and a number of state-of-the-art deep learning architectures.

3.1 Data and Labels

In this paper, we utilise the beverage tasting dataset introduced in [19]. The study was approved by the Ethics Committee of the Department and the participants volunteered to take part in the study without monetary compensation. The dataset contains data from twenty-seven volunteers (9 males and 18 females; mean age = 22.25 years, SD = 3.58 years; 14 Asian and 13 non-Asian as the most balanced ethnic split for meaningful analysis). Five beverages were used in the study: grape juice, lemonade, dark roast coffee, salty water, and non-alcoholic ginger beer with pepper sauce. After each tasting sample, participants were asked to provide liking scores on a 7-point Likert scale. As we have reported in [19], the five beverages were able to elicit significantly different taste-liking scores. The liking score difference was not influenced by facilitators and facilitation styles. Significant difference in Action Units was found for liking and disliked facial reactions, providing the motivation for automatic estimation of taste-liking.

Nonverbal reactions (face, upper body and audio) during the entire session were captured using a Logitech C920 high definition webcam positioned frontally to capture upper body and face. Videos were recorded with a resolution of 1280 × 720 pixels at a rate of 23 frames per second under controlled illumination conditions. Each subject had a recording of about 25 minutes. Each recording was segmented into short clips, with each clip containing a tasting sample. For some tasting samples, participants had more than one tasting attempt, resulting in several clips of one tasting sample. This yielded in 377 clips in a total of about 107K frames, with 197 clips of human facilitator, 180 clips of robot facilitator, 141 clips of priming facilitation style, and 236 clips of non-priming facilitation style. The full details about the study, data and labels, as well as the results of the various statistical analyses, can be found in [19].

In order to conduct facial reaction analysis to various tastes, we created ground-truth taste-liking labels similarly to [8]. We mapped the overall liking scores from 7-point scale to 3-point scale (1 and 2 indicating disliking; 3, 4, and 5 indicating neutral state; 6 and 7 indicating liking). We utilise these as class labels. This results in 172 disliking, 140 neutral, and 65 liking video clips. Fig. 1 shows representative frames from different participants, with human/robot facilitator type and self-reported taste-liking labels. These frames show the instantaneous expressions right after the participants took a sip from their beverage.

This dataset also contains the personality information of each participant as they were asked to fill in the Mini-IPIP Big-Five Personality questionnaire [9] before partaking in the study. Mini-IPIP is an simplified Inter-national Personality Item Pool–Five-Factor Model measure, which describes a person's personality along 5 dimensions: extroversion, agreeableness, conscientiousness, neuroticism and openness.

3.2 Engineered Features

In order to extract the more traditional engineered features from the face, we used OpenFace 2.0 [2] for facial landmark detection and tracking as well as head pose estimation. As a result, we obtained for each frame, 68 facial landmarks and head pose (translation and orientation) information. A sample frame with visualisations of facial landmarks and head pose information is shown in Fig. 2.

3.2.1 *Low-level Geometric Features.*

Geometric features are analysed at video-level. Frame-level geometric features are extracted first. The position X, velocity V, and acceleration A of facial landmarks with head pose were computed and concatenated into a vector $[X(t), V(t), A(t)]$ and L2 normalised, where t denotes the current frame. For video-level geometric features, the Fisher Vector (FV) encoding was implemented using a Gaussian Mixture Model (GMM). To reduce the dimensionality, Principle Component Analysis (PCA) was first applied to the geometric feature vectors reducing its dimensionality to $D = 70$ with an explained variance ratio of 96%, and then a GMM was fitted to the processed features. The number of Gaussians was set to $K = 64$ with a subset of 64000 features randomly sampled to fit the GMM, getting FV to represent each video clip with a dimension of $(2D + 1)K = 9024$. PCA was applied to transform the FV features to 300-dimensional vectors as inputs for classifiers, getting an explained variance ratio of more than 95%.

3.2.2 *Low-level Appearance Features.*

Appearance features are analysed at frame-level. For appearance features, we extracted HOGs features from the aligned 112×112 faces. We used blocks of 2×2 cells, 8×8 pixels, leading to 4464-dimensional vectors describing the face, as visualised in Fig 3. PCA was applied to reduce the dimensionality of HOG features to 300-dimensional vectors as classifier inputs, with an explained variance ratio of 86%. A sample frame with visualisations of the aligned face and the corresponding HOG features are shown in Fig. 3.

3.2.3 *High-level Features.*

In order to obtain frame-level and video-level high-level facial reaction features, we used OpenFace 2.0 [2] for Action Units (AUs) detection. OpenFace is able to recognise a subset of AUs. The output of AU detection module is 0/1 label for absence/presence, and intensity between 0 and 1 of each frame. For frame-level AU features, we used the detected intensity of 17 AUs to construct a 17-dimension feature vector, $[I_1, I_2, ..., I_{17}]$, where I_k denotes the intensity of $k - th$ AU. For video-level AU features, for each video clip, we computed three parameters: the proportion of the frames that each AU was present Pre, mean intensity of each AU MI, and the standard deviation of the intensity of each AU SI. By computing AU-present proportions rather than the number of frames, we ensured that variation in video duration did not

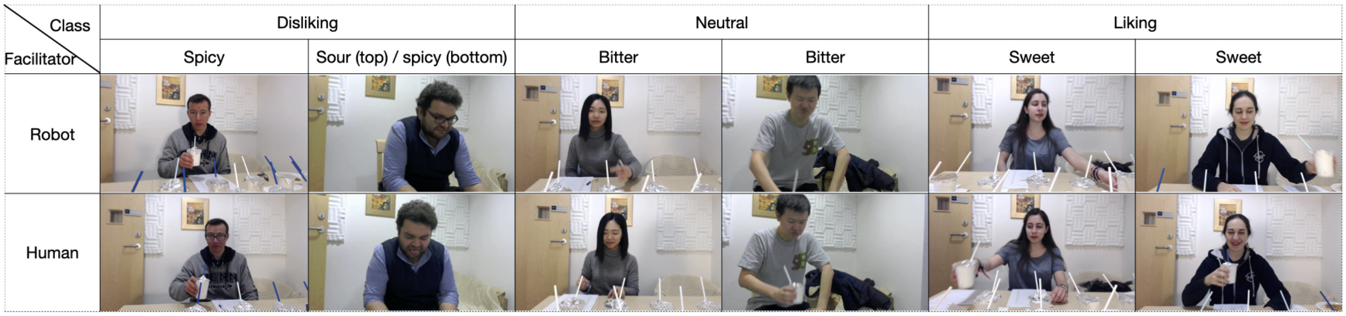

Figure 1: Representative examples from the dataset with human/robot facilitator type and self-reported taste-liking.

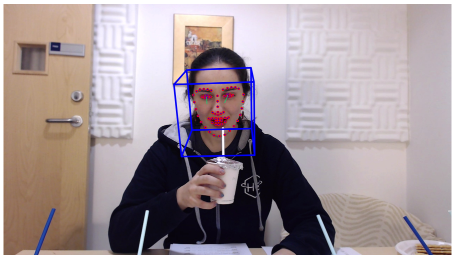

Figure 2: Facial landmark and head pose tracking sample frame.

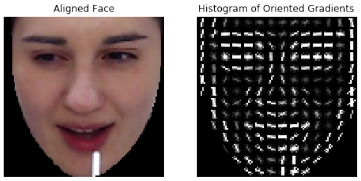

Figure 3: Visualisation of aligned face and HOG features.

influence parameter estimates. The video-level AU feature was then concatenated as $[Pre, MI, SI]$ with a dimensionality of 51.

3.3 Deep Spatiotemporal Networks

Inspired by prior work [35], we implemented three kinds of 3D CNN models and refer to them as FrameCNN, FregionCNN and FusionCNN, to aggregate both spatial and temporal information across a consecutive sequence of frames. FrameCNN uses the raw video frames, with the full upper-body region including the experiment background. FregionCNN uses the full facial region, while only the eye and mouth regions are taken as input in FusionCNN. The details of these models are provided in Table 1. Each model is composed of 3D convolutional layers, 3D max-pooling layers, ReLU activation functions, fully connected layers and dropouts. We use $w \times h \times d$ to represent the input width, height and depth, and the value of d is 18 dependent on the video frame sequence length. FrameCNN and FregionCNN have the same network structure, where w and h are fixed to 64. FusionCNN receives two parts of input data and then concatenates them after flattening, so its input width and height are halved.

Specifically, the input video frames are firstly stacked in sequence for the convolution operation by 3D kernels in 3D convolutional (Conv3D) layer. The 3D-CNN uses the filters in both temporal and geographical direction, while conventional 2D CNN kernel only focuses on the latter. After 3D max-pooling, the output dimension is reduced and important spatial-temporal features are retained. Multi-dimensional input are then flattened into a one-dimensional array for full connection. The fully-connected layers introduce more non-linearity using ReLU activation function [29] to extract feature hierarchically with the formula:

$$Y = \theta(W_d X + b_d) \tag{1}$$

where the θ is ReLU, W_d and b_d represent weights and bias of a dense (fully connected) layer. The dropout layers are added to prevent the network from overfitting, which can also regularise the network. Lastly, The softmax layer generates the final classification results.

3.3.1 *Personality-aware Network.*

The goal of creating a personality-aware network is to incorporate the self-reported personality information to obtain a more complete representation of the participant's behaviour when attempting to classify their facial response to tasting and to design a more holistic model of taste-liking prediction. To do this, we incorporate participants' responses to all 20 questions of the personality questionnaire as features into the 3D CNN model. The network is a dual-input end-to-end architecture, inspired by the attribute-aware network proposed in [44], and can utilise visual and textual information. Personality scores are encoded as one hot vector and upsampled to match the dimensionality of the features extracted from video frames. Both features are concatenated prior to being fed to the fully-connected layers.

4 EXPERIMENTAL SETUP

4.1 Experiments with Engineered Features

For comparative evaluation and analysis, four models were trained using the various engineered features explained in the previous section. The first model (Model 1) uses the frame-level AU intensity features as inputs to a linear Support Vector Machine (SVM) classifier. For the second model (Model 2), video-level AU intensity features were fed to the linear-SVM classifier. For the third model (Model 3), after performing PCA to reduce the dimensionality of

FusionCNN		FrameCNN/FregionCNN	
Input-1/2	32 × 32 × 18	Input	64 × 64 × 18
Layer	Output Dimension	Layer	Output Dimension
3D-Convolution-1/2	32×30×30×4	3D-Convolution	32×62×62×4
3D-Maxpooling-1/2	32×10×10×1	3D-Maxpooling	32×20×20×1
Dropout-1/2	32×10×10×1	Dropout	32×20×20×1
Flatten-1/2	3200	Flatten	12800
Concatenate	6400	Dense	1028
Dense	1024	Dropout	1028
Dropout	1024	Dense	128
Dense	512	Dropout	128
Dropout	512	Dense	3
Dense	128	Dropout	3
Dropout	128		
Dense	3		

Table 1: Network architecture of deep learning models in terms of the output dimension of different layers.

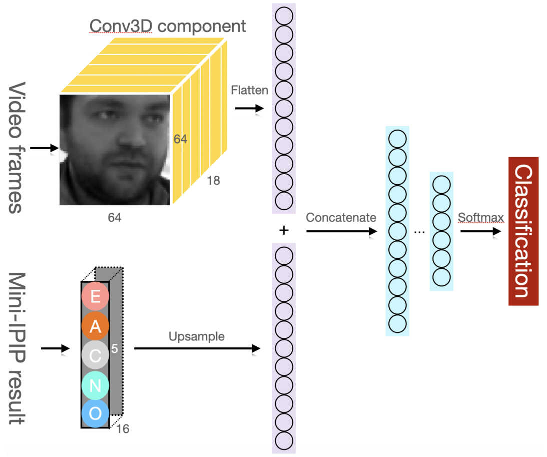

Figure 4: FregionCNN-P and FrameCNN-P structure. The only difference for them is the input for the Conv3D component. Numbers are the sizes of the feature vectors.

the features, frame-level HOG features were used as inputs to the linear-SVM classifier. For the fourth model (Model 4), PCA was applied on the video-level geometric features and the output was used as input to the linear-SVM classifier. A summary of these models is provided in Table 2 .

Scikit-learn library [31] was used to implement and evaluate the linear SVM models. For frame-level features, the linear-SVM used an overlapping processing window of 10 frames to smooth the features prior to classification. Before passing onto the SVM classifiers, all features were L2 normalised. All SVM classifiers used a one-vs-one decision strategy and class weights to balance the imbalanced aspect of the dataset, with regularisation parameter $C = 1$.

Method	Feature	Classifier
Model 1	Frame-level AU intensity features	linear-SVM
Model 2	Video-level AU intensity features	linear-SVM
Model 3	Frame-level HOG features + PCA	linear-SVM
Model 4	Video-level geometric features + PCA	linear-SVM

Table 2: A summary of the four automatic approaches implemented using engineered features.

To evaluate these models, we employed Leave-One-Subject-Out (LOSO) Cross-Validation technique (i.e., using data from 26 participants for model training and the remaining data from one participant for testing). Specifically, due to the relatively small test set size, leave-three-subject-out cross-validation was used for the geometric features and both frame- and video-level AU features. Leave-one-subject-out cross-validation was implemented for the other features. The final taste-liking classification results was obtained as the average of scores across all folds.

4.2 Experiments with 3D CNNs

The three 3D FrameCNN, FregionCNN and FusionCNN models were all implemented on Keras 2.2.4 with Tensorflow 1.13.1 at the backend. All three models were trained and tested on NVIDIA QUADRO RTX 8000 GPU. All the models use Categorical Cross Entropy loss function and the RMSprop optimiser, with default learning rate. Leave 6 or 9 subject-out cross validation is applied for these models.

To ensure subject-independent evaluation and testing, we employed Leave-N-Subject-Out Cross-Validation technique (i.e., using data from 6/8 participants for testing and the remaining data for model training). More specifically, data from 27 participants is divided into 4-subject-independent folds with data from 6 subjects in the folds 1-3, and remaining data from 8 subjects in fold 4. The aim of this process is to use approximately 20% of the overall data for testing and the remaining amount for training. Testing is done for each of these folds separately, and results are (weighted) averaged across these folds.

The optimal values for the batch size and epoch are 16 and 30 respectively chosen from the space of $\{8, 16, 32\}$ and $\{30, 50, 100\}$. Successive frame sequences are obtained by OpenCV-Python, and their length is fixed to 18. The input dimensions for FreignCNN and FrameCNN are $64 \times 64 \times 18$, and FusionCNN models are $[32 \times 32 \times 18, 32 \times 32 \times 18]$ respectively. The kernel size in 3D-CNN layers is

3 × 3 × 15 while the max pooling size is 3 × 3 × 3. The dropout rate is set to 0.2, tested from {0.1, 0.2, 0.5}. During training, we use 10% of the data as the hold-out evaluation. Early stopping is also introduced to avoid overfitting.

The facilitated taste-liking dataset of [19] is class-imbalanced with 172 disliking, 140 neutral, and 65 liking video clips. It is well-known that classification tends to be biased in favour of the majority class when using class-imbalanced data. Two principal strategies to address this problem are to oversample the minority class or to undersample the majority class. Because neural networks generally perform better with more training data, we chose to employ the Synthetic Minority Oversampling TEchnique (SMOTE) [5]. SMOTE works by selecting one example of a minority class and its *k* nearest neighbours that are close in the feature space. Then a line between the examples in the feature space are drawn, and a new synthetic sample at a point along that line is created as a convex combination of the two chosen instances segmented by that line. After oversampling, the training set contains equally distributed samples along the three classes of disliking, neutral, and liking.

For personality-aware classification, we utilise FregionCNN and the best performing network FrameCNN, and extend it to include personality information (FregionCNN-P and FrameCNN-P) shown in Figure 4. The hyperparameters remain the same as the original FregionCNN.

5 RESULTS AND ANALYSES

5.1 Results with Engineered Features

The classification results of the four models are shown in Table 3. The frame-level AU model (Model 1) performed best in all the hand-crafted feature models in terms of recall and f1-scores, with an overall accuracy of 41.14%, while video-level geometric model (Model 4) with with 42.63% performed best in terms of overall accuracy score. Similarly to what was observed in [8], although both of the AU-based methods (Model 1 and Model 2) seemed to provide the best performance overally, the accuracy of taste-liking classification might be heavily dependent on the accuracy of the estimated AU probabilities. Additionally, AU features alone may not be accurate in representing people's preferences - i.e., a spontaneous negative expression might not imply that the person dislikes this particular flavour. For example, when people taste pungent flavours such as lemon juice, this may lead to strong negative facial expressions even though in reality they might like this sour taste.

Among the four models compared, the frame-level HOG model (Model 3) obtained the worst classification performance, which probably indicates that frame-level HOG features alone without trajectory or temporal information are not suitable for taste-liking estimation. According to [8], fusing facial appearance (HOG descriptors in our case) in each frame of a video through FV encoding can generate a good performance. Coding HOG features with other temporal information may lead to better classification performance.

In terms of individual classes, the geometric model (Model 4) provided the highest recall score of 67.52% for classifying dislike, which indicates its sensitivity to people's dislike expressions. People may show significant changes in terms of facial dynamics when communicating nonverbal aspects of dislike. The video-level AU model (Model 2) provided the highest recall score of 44.56% for classifying the liking class, and frame-level AU model (Model 1) provided the highest f1-scores for both neutral and liking classes, which confirms the informative nature of the AU features.

Both models utilising low-level features, HOG model (Model 3) and geometric model (Model 4), had worse classification powers for classifying the liking class than the disliking and neutral classes. Considering the f1-scores, all models performed relatively worse for the liking class. This finding supports what was reported in [19], that dislike-related facial expressions may be more intense, more frequent, and more evoked, thus leading to better classification results than neutral and liking-related (positive) facial expressions.

5.2 Results for Deep Spatiotemporal Networks

Table 4 provides the classification results of the three deep learning models employed. FusionCNN achieves the best weighted F1-score without the need for oversampling. After oversampling, FrameCNN provides the best F1-score. Their performance for each class varies: FusionCNN gives the highest liking classification F1-score at 91.8%, FregionCNN has the best neutral F1-score at 87.77% and the highest disliking F1-score at 91.28% is provided by FrameCNN. Without oversampling, the F1-score values remain below 90%. In general, the results for disliking and neutral classes are marginally better than the liking class.

Oversampling helps improve the model classification performances at a small margin. Particularly, the overall f1-score for FrameCNN has increased from 82.57% to 89.12%, with a 3% increase for FusionCNN and 4% for FregionCNN. It benefits the neutral class most, but slightly impacts the scores for disliking class on FregionCNN and liking class on FrameCNN.

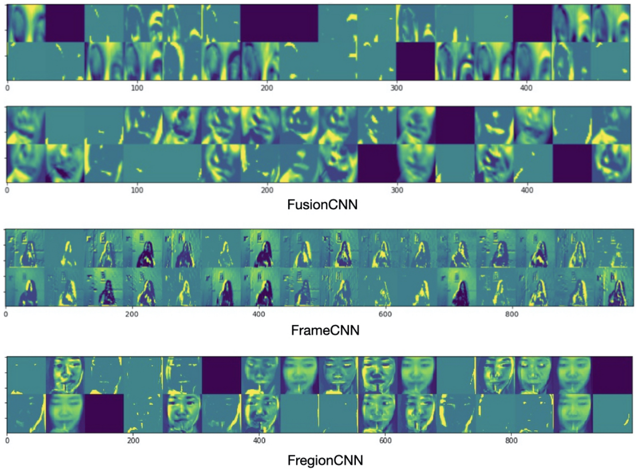

Figure 5: The intermediate activations of three 3D CNN models over their Conv3D layers.

5.2.1 *Visualisation of Network Layers.*

In order to understand how successive convolutional layers transform their input and to get a first idea of the meaning of individual filters [6], we visualise

model	dislike		neutral		liking		overall		
	Recall	F1	Recall	F1	Recall	F1	Precision	Recall	F1
Model 1	39.96%	43.48%	39.92%	**40.28%**	40.36%	**29.64%**	41.14%	**40.12%**	**37.78%**
Model 2	45.36%	**48.40%**	30.08%	31.72%	**44.56%**	26.48%	42.47%	40.01%	35.61%
Model 3	31.56%	27.15%	**47.44%**	31.67%	14.63%	12.81%	32.98%	31.19%	23.83%
Model 4	**67.52%**	48.04%	20.52%	13.48%	16.32%	7.32%	**42.63%**	34.79%	22.93%

Table 3: Classification results of the four models that employed engineered features.

model	dislike		neutral		liking		overall			
	Recall	F1	Recall	F1	Recall	F1	Precision	Recall	F1	Accuracy
FregionCNN(w/o)	78.49%	81.82%	87.86%	83.96%	72.31%	71.76%	81.11%	80.90%	80.88%	80.90%
FregionCNN(w/t)	86.05%	84.81%	87.14%	**87.77%**	76.92%	78.74%	84.88%	84.88%	84.86%	84.88%
FusionCNN(w/o)	87.79%	83.20%	78.42%	81.04%	78.46%	85.00%	83.19%	82.71%	82.71%	82.71%
FusionCNN(w/t)	90.12%	85.64%	79.86%	82.84%	86.15%	**91.80%**	86.11%	85.64%	85.67%	85.64%
FrameCNN(w/o)	90.70%	85.25%	80.00%	81.45%	67.69%	77.88%	83.30%	82.76%	82.57%	82.76%
FrameCNN(w/t)	91.28%	**91.28%**	88.57%	87.32%	84.62%	87.30%	89.17%	89.12%	**89.12%**	89.12%

Table 4: Classification results of FregionCNN (facial region only), FusionCNN (eye and mouth region), and FrameCNN (with background) . "W/o" refers to training without oversampling while "w/t" refers to with oversampling.

extracted feature representation in the intermediate layers of FusionCNN, FrameCNN and FregionCNN in Fig. 5. The 32-filter intermediate Conv3D component generates 62 × 62 feature maps in Conv3D layer in FrameCNN and FregionCNN's Conv3D layer. We observe that FrameCNN retains the full body information of the participant, while FregionCNN merely exploits the facial region information. By contrast, FusionCNN has 32 filters with a size of 3 × 3 used to extract key facial features, and thus 32-channel feature maps with a size of 30 × 30 are produced in Conv3D layer and activation layer. Features are less distinguishable and there are more filters that are not activated and left blank. When participants were drinking, they were more likely to lower their head down, potentially making the model's ability to identify eye features more challenging. However, this did not seem to effect the overall results. One possible explanation is that the lower face plays a more important role than the upper face in analysing and recognising facial taste-liking behaviours.

5.2.2 Results of Personality-aware Classification. Personality-aware classification results are presented in Table 5. The original FregionCNN and FrameCNN are improved by the fusion of participants' personality information with facial feature information. The best overall F1-score is obtained by FrameCNN-P without oversampling at 87.43%. The other three systems' F1-scores range from 84% to 86%, all higher than those of FregionCNN, FusionCNN and FrameCNN without oversampling. Relevant literature investigating links between taste preferences and personality concluded that generally personality is weakly or moderately related to self-reported taste preference [7], depending on specific flavours. This could potentially explain the higher classification results obtained in our experiments. Note however that, one limitation of the personality-aware network design is that it requires a personality questionnaire to have been filled in apriori, before beverage tasting behaviours can be evaluated. This limitation can potentially be addressed by utilising an automatic personality predictor (e.g., [36], [49]).

5.2.3 Facilitator Type, Gender and Ethnicity of the Participants. As the deep spatiotemporal architectures provide better classification results than the models that utilise engineered features, for further analysis, we focus on the 3D CNN models only. We are interested in analysing and understanding whether there are differences in terms of classification results when considering the following three aspects that pertain to this dataset: facilitator type (robot vs. human), gender (female vs. male) and ethnicity (Asian vs. non-Asian as this provides the most balanced split in terms of ethnicity). In order to do this, we split all the model predictions for the 27 participants along these three aspects, and compute the corresponding F1-scores as shown in Figures 6,7 and 8.

We observe that the model scores for robot facilitation increased slightly after oversampling. We also observe that, after oversampling, the classification results are more stable for the Non-Asian participants as compared to the Asian participants. Therefore, we conclude that the facilitator type and the race of participants do not affect automatic classification of taste-liking. Overall, the scores obtained for females are higher than those of males. This is in line with the results reported in [8]. The gender-related taste-liking analysis in [8] indicated that classification accuracy was higher for females as compared to males (73.08% vs. 69.58%). They also found that for disliking, mean expressiveness of eye and cheek regions were significantly higher for females, but for males eyebrow and forehead regions were significantly more expressive. Therefore, we conclude that, the gender of the participants does affect the results of automatic classification of taste-liking, and should be taken into account in design choices made for real-world applications.

5.2.4 Comparison of the Results to Related Work. In order to provide a comparison between the models designed and evaluated in this paper and related works, we focus on the works of [8] and [47]. Zhi et al. in [47] achieved 45.7% accuracy when they formulated the problem of hedonic rating identification as a three-class problem. Dibeklioglu and Gevers in [8] achieved an accuracy of 70.37% for

model	dislike		neutral		liking		overall			
	Recall	F1	Recall	F1	Recall	F1	Precision	Recall	F1	Accuracy
FregionCNN-P(w/o)	89.53%	88.70%	85.71%	86.76%	78.46%	**78.69%**	86.39%	86.21%	86.26%	86.21%
FregionCNN-P(w/t)	87.21%	86.50%	88.57%	87.92%	73.85%	73.85%	84.74%	85.41%	84.84%	85.41%
FrameCNN-P(w/o)	93.02%	**90.07%**	90.00%	**90.86%**	72.31%	73.05%	86.88%	88.33%	**87.43%**	88.32%
FrameCNN-P(w/t)	87.21%	86.87%	86.43%	86.85%	75.39%	76.28%	85.77%	84.88%	85.04%	84.88%

Table 5: Classification results of FregionCNN-P and FrameCNN-P.

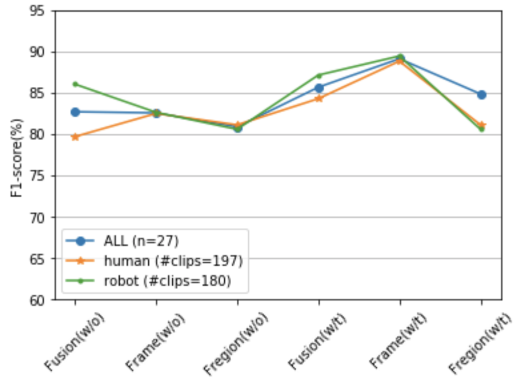

Figure 6: F1-scores reported separately for different facilitators across three 3D CNN models, with "ALL" referring to the original scores. The number of participants is represented by n. There are 197 video clips which participants facilitated by human and 180 clips by robot.

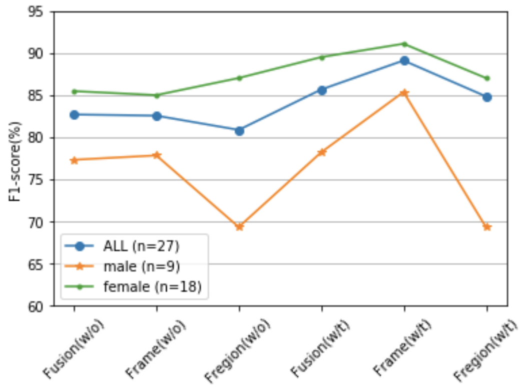

Figure 7: F1-scores reported separately for female and male groups. The number of participants in each group is represented by n.

distinguishing between the three levels of dislike, like, and neutral. FrameCNN, the best performing model in our work, provides a better classification performance by achieving 89.12% accuracy for classifying liking, neutral and disliking. Overall, it should be noted that due to the differences in the datasets used, these comparisons have only limited utility in exposing differences in performance between these approaches.

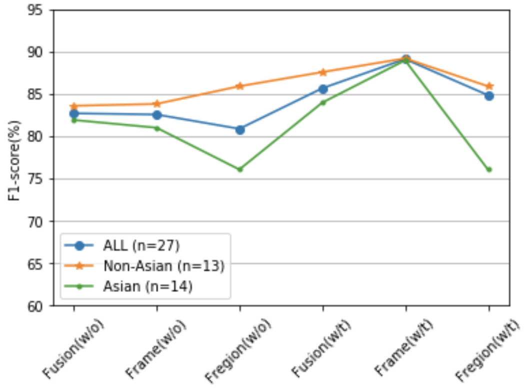

Figure 8: F1-scores reported separately for Asian and non-Asian participant groups.

6 CONCLUSIONS

This work applied two different strategies for automatic facial expression recognition on the taste-liking dataset introduced in [19]: One strategy utilises engineered geometric, appearance, and AU features while the other uses 3D Convolutional Neural Networks (3D-CNNs). The results of the extensive experiments we have conducted lead us to the following conclusions: (i) Developing robot facilitators (e.g., robot waiters) for the hospitality industries with automatic taste-liking capability is viable; (ii) deep spatiotemporal architectures are more promising for developing such capability; (iii) incorporating the personality information of the human users is promising in improving the automatic facial taste-liking recognition; and (iv) different automatic models and/or metrics might need to used for female vs. male user groups to achieve *fair* facial taste-liking recognition results (e.g., see [45] for an investigation of bias and fairness in facial expression analysis).

ACKNOWLEDGMENTS

The work of Hatice Gunes is supported by the EPSRC (grant ref. EP/R030782/1).

REFERENCES

[1] Iman Abbasnejad, Sridha Sridharan, Dung Nguyen, Simon Denman, Clinton Fookes, and Simon Lucey. 2017. Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1609–1618.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[3] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. FaceTube: Predicting Personality from Facial Expressions of Emotion in Online Conversational Video. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (Santa Monica, California, USA). Association for Computing Machinery, New York, NY, USA, 53–56. https://doi.org/10.1145/2388676.2388689

[4] Paul D Bolls, Annie Lang, and Robert F Potter. 2001. The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. *Communication Research* 28, 5 (2001), 627–651.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[6] Francois Chollet. 2017. *Deep Learning with Python* (1st ed.). Manning Publications Co., USA.

[7] Catherine J Day. 2009. *An exploration of the relationships between personality, eating behaviour and taste preference.* Ph.D. Dissertation. Sheffield Hallam University.

[8] Hamdi Dibeklioglu and Theo Gevers. 2018. Automatic Estimation of Taste Liking through Facial Expression Dynamics. *IEEE Transactions on Affective Computing* (2018).

[9] M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. 2006. The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological assessment* 18, 2 (2006), 192.

[10] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent Neural Networks for Emotion Recognition in Video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA). Association for Computing Machinery, New York, NY, USA, 467–474. https://doi.org/10.1145/2818346.2830596

[11] Paul Ekman and Dacher Keltner. 1997. Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture* (1997), 27–46.

[12] Paul Ekman, Robert W Levenson, and Wallace V Friesen. 1983. Autonomic nervous system activity distinguishes among emotions. *Science* 221, 4616 (1983), 1208–1210.

[13] Kory Floyd. 2000. Attributions for nonverbal expressions of liking and disliking: The extended self-serving bias. *Western Journal of Communication (includes Communication Reports)* 64, 4 (2000), 385–404.

[14] E Friesen and Paul Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3 (1978).

[15] Judith R Ganchrow, Jacob E Steiner, and Munif Daher. 1983. Neonatal facial expressions in response to different qualities and intensities of gustatory stimuli. *Infant Behavior and Development* 6, 4 (1983), 473–484.

[16] Ellen Greimel, Michael Macht, Eva Krumhuber, and Heiner Ellgring. 2006. Facial and affective reactions to tastes and their modulation by sadness and joy. *Physiology & Behavior* 89, 2 (2006), 261–269.

[17] Dami Jeong, Byung-Gyu Kim, and Suh-Yeon Dong. 2020. Deep Joint Spatiotemporal Network (DJSTN) for Efficient Facial Expression Recognition. *Sensors* 20, 7 (2020), 1936.

[18] Zirui Jiao, Fengchun Qiao, Naiming Yao, Zhihao Li, Hui Chen, and Hongan Wang. 2018. An Ensemble of VGG Networks for Video-Based Facial Expression Recognition. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).* IEEE, 1–6.

[19] Zhuoni Jie and Hatice Gunes. 2020. Investigating Taste-liking with a Humanoid Robot Facilitator. In *Proceedings of IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* 1–6.

[20] Sarasi Kankanamge, Clinton Fookes, and Sridha Sridharan. 2017. Facial analysis in the wild with LSTM networks. In *2017 IEEE International Conference on Image Processing (ICIP).* IEEE, 1052–1056.

[21] Silvia C King, Herbert L Meiselman, and B Thomas Carr. 2013. Measuring emotions associated with foods: Important elements of questionnaire and test design. *Food Quality and Preference* 28, 1 (2013), 8–16.

[22] Dimitrios Kollias and Stefanos Zafeiriou. 2018. Training Deep Neural Networks with Different Datasets In-the-wild: The Emotion Recognition Paradigm. In *Proceedings IJCNN.* 1–8. https://doi.org/10.1109/IJCNN.2018.8489340

[23] Egon Peter Köster. 2003. The psychology of food choice: some often encountered fallacies. *Food Quality and Preference* 14, 5-6 (2003), 359–373.

[24] Peter Lewinski, Marieke L Fransen, and Ed SH Tan. 2014. Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics* 7, 1 (2014), 1.

[25] Shan Li and Weihong Deng. 2018. Deep Facial Expression Recognition: A Survey. *CoRR* abs/1804.08348 (2018).

[26] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops.* IEEE, 94–101.

[27] Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. 2019. Automatic Analysis of Facial Actions: A Survey. *IEEE Transactions on Affective Computing* 10, 3 (2019), 325–347.

[28] Daniel McDuff, Rana El Kaliouby, et al. 2014. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing* 32, 10 (2014), 630–640.

[29] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML.*

[30] Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* 577–582.

[31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[32] Danielle R Reed, Toshiko Tanaka, and Amanda H McDaniel. 2006. Diverse tastes: Genetics of sweet and bitter perception. *Physiology & behavior* 88, 3 (2006), 215–226.

[33] Heidi R. Riggio and Ronald E. Riggio. 2002. Emotional Expressiveness, Extraversion, and Neuroticism: A Meta-Analysis. *Journal of Nonverbal Behavior* 26 (2002), 195–218.

[34] Diana Rosenstein and Harriet Oster. 1988. Differential facial responses to four basic tastes in newborns. *Child development* (1988), 1555–1568.

[35] Reddy Sai Prasanna Teja, Karri Surya Teja, Shiv Ram Dubey, and Snehasis Mukherjee. 2019. Spontaneous Facial Micro-Expression Recognition using 3D Spatiotemporal Convolutional Neural Networks. *International Joint Conference on Neural Networks* (2019).

[36] Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2016. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access* 5 (2016), 705–721.

[37] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2014. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence* 37, 6 (2014), 1113–1133.

[38] Jacob E Steiner. 1973. The gustofacial response: observation on normal and anencephalic newborn infants. (1973).

[39] Jacob E Steiner. 1979. Human facial expressions in response to taste and smell stimulation. In *Advances in child development and behavior.* Vol. 13. Elsevier, 257–295.

[40] Jacob E Steiner, Dieter Glaser, et al. 2001. Comparative expression of hedonic impact: affective reactions to taste by human infants and other primates. *Neuroscience & Biobehavioral Reviews* 25, 1 (2001), 53–74.

[41] Ning Sun, Qi Li, Ruizhi Huan, Jixin Liu, and Guang Han. 2019. Deep spatial-temporal feature fusion for facial expression recognition in static images. *Pattern Recognition Letters* 119 (2019), 49–61.

[42] Romy Weiland, Heiner Ellgring, and Michael Macht. 2010. Gustofacial and olfactofacial responses in human adults. *Chemical senses* 35, 9 (2010), 841–853.

[43] Karin Wendin, Bodil H Allesen-Holm, and Wender LP Bredie. 2011. Do facial reactions add new dimensions to measuring sensory responses to basic tastes? *Food quality and preference* 22, 4 (2011), 346–354.

[44] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating Bias and Fairness in Facial Expression Recognition. arXiv:2007.10075 [cs.CV]

[45] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating Bias and Fairness in Facial Expression Recognition. In *Proceedings of the 16th European Conference on Computer Vision Workshops.*

[46] Gertrude G Zeinstra, MA Koelen, D Colindres, FJ Kok, and C De Graaf. 2009. Facial expressions in school-aged children are a good indicator of 'dislikes', but not of 'likes'. *Food Quality and Preference* 20, 8 (2009), 620–624.

[47] Ruicong Zhi, Xin Hu, Chenyang Wang, and Shuai Liu. 2020. Development of a direct mapping model between hedonic rating and facial responses by dynamic facial expression representation. *Food Research International* 137 (2020), 109411. https://doi.org/10.1016/j.foodres.2020.109411

[48] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. 2012. Learning active facial patches for expression analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2562–2569.

[49] Oya Çeliktutan and Hatice Gunes. 2017. Automatic Prediction of Impressions in Time and across Varying Context: Personality, Attractiveness and Likeability. *IEEE Transactions on Affective Computing* 8, 1 (2017), 29–42.