This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2832044, IEEE Transactions on Affective Computing

Automatic Estimation of Taste Liking through Facial Expression Dynamics

Hamdi Dibeklioğlu, *Member, IEEE*, and Theo Gevers, *Member, IEEE*

Abstract—The level of taste liking is an important measure for a number of applications such as the prediction of long-term consumer acceptance for different food and beverage products. Based on the fact that facial expressions are spontaneous, instant and heterogeneous sources of information, this paper aims to automatically estimate the level of taste liking through facial expression videos. Instead of using handcrafted features, the proposed approach deep learns the regional expression dynamics, and encodes them to a Fisher vector for video representation. Regional Fisher vectors are then concatenated, and classified by linear SVM classifiers. The aim is to reveal the hidden patterns of taste-elicited responses by exploiting expression dynamics such as the speed and acceleration of facial movements. To this end, we have collected the first large-scale beverage tasting database in the literature. The database has 2970 videos of taste-induced facial expressions collected from 495 subjects. Our large-scale experiments on this database show that the proposed approach achieves an accuracy of 70.37% for distinguishing between three levels of taste-liking. Furthermore, we assess the human performance recruiting 45 participants, and show that humans are significantly less reliable for estimating taste appreciation from facial expressions in comparison to the proposed method.

Index Terms—Taste liking, taste appreciation, facial expression dynamics, spontaneous expression, taste-induced expression.

1 Introduction

Food is one of the primary necessities of life. Nowadays, the quality of food (e.g. lower fat and sugar) is important to prevent obesities and promote healthier ingredients. To obtain different food composition (e.g. lower fat, sugar and salt) with similar taste liking, the challenge is to measure the appreciation of food in an objective, spontaneous and instant way. In general, the human face can be used as a cue to determine if someone likes a particular taste or not as it offers rich and spontaneous data in terms of facial expressions. Previous studies show that the face reveals appreciation or dislike while eating and drinking [1], [2]. Such spontaneous facial expressions can be used to measure quality and intensity of the taste.

In contrast to above studies based on human observations, in this paper, the aim is to automatically recognize taste-induced facial expressions for taste liking. Many studies in human facial analysis categorize facial expressions and connect them to emotional states [3], [4]. In tasting, however, facial expressions are not directly indicative of these inner emotional states, but rather a spontaneous motor response to flavor. Therefore, facial analysis for emotion classification (e.g. Action Units) is not directly applicable to taste liking. For instance, a person may display facial Action Units (AU) that correspond to disgust expression (e.g. AU 15: lip corner depressor, AU 9: nose wrinkler) when tasting lemon juice, yet, this does not necessarily mean that he/she dislikes the taste. Similarly, we cannot expect to observe a joy expression (e.g. AU 12: lip corner puller) in response to every positive taste. Beside the appearance of taste-induced facial expressions, subtle dynamic information hidden in such expressions is important. The aim is to discover this hidden information by analyzing the expression dynamics such as the acceleration and speed of the facial movements. In this paper, the focus is on facial expressions for the estimation of taste liking as they provide spontaneous, instant and heterogeneous human data.

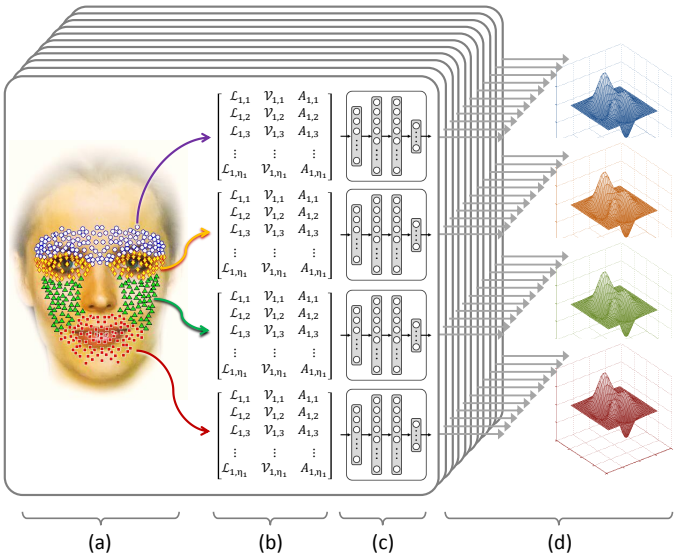


Fig. 1. Overview of the proposed approach: (a) Facial landmark tracking, (b) extraction of facial dynamics (location, speed, and acceleration), (c) deep learning of regional representations through stacked denoising autoencoders, and (d) computation of regional Fisher vectors to represent videos.

We aim to automatically measure taste liking by means of a holistic interpretation of facial expressions. Facial analysis is

Manuscript received May XX, 2016; revised XXXX XX, XXXX.. This study was supported by the Dutch national program COMMIT.
H. Dibeklioğlu is with the Computer Engineering Department, Bilkent University, 06800 Ankara, Turkey, and also with the Pattern Recognition and Bioinformatics Group, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: dibeklioglu@cs.bilkent.edu.tr).
T. Gevers is with the Informatics Institute, University of Amsterdam, 1098 XH Amsterdam, The Netherlands (e-mail: th.gevers@uva.nl).

done by considering the expression dynamics such as acceleration and speed of regional facial movements. To this end, inner facial movements are tracked in videos. Facial regions are proposed corresponding to important parts of faces to incorporate locality as pixels within these regions move together. Per frame, deep-learned regional dynamics are obtained using stacked denoising autoencoders, which are then coded into a Fisher vector for video representation. The regional Fisher vectors are concatenated and used as input of an SVM classifier to classify taste liking. Overview of the proposed approach is visualized in Figure 1.

Unfortunately, no large-scale taste datasets of facial expressions are available today. Therefore, to test and compare the proposed method, a new large-scale taste database has been collected containing spontaneous facial expressions while drinking different types of beer. Such a database is a milestone to automatically interpret taste-induced facial behavior in real-life scenarios. To differentiate from using hand-crafted features, deep learning is applied to obtain efficient feature representations. This is made possible by the enormous amount of newly collected spontaneous facial expression data. Although automatic face analysis has been used in different computer vision applications, this is the first paper on automatic taste liking tested on a large-scale database.

People are quite reliable and accurate in distinguishing between emotional facial expressions, however, human ability to predict taste appreciation from facial expressions has not been investigated yet. To this end, we have recruited 45 participants, and assess human performance for this task in comparison to the proposed method.

Our key contributions can be listed as follows: (1) First time in the literature, we propose to learn a deep representation for per-frame dynamics of facial responses by jointly encoding location, speed, and acceleration parameters of densely sampled facial landmarks. (2) We introduce an efficient temporal representation for taste-induced facial expressions by combining the deep-learned per-frame dynamics of each frame in a given video through Fisher vector encoding. (3) We propose the first automatic approach for estimating the level of taste liking using facial expression dynamics. (4) We have collected the first large-scale beverage tasting database in the literature. (5) We provide new empirical findings on taste-elicited facial dynamics. (6) We assess the human performance for estimating taste appreciation from facial expressions. Finally, (7) we show that the proposed method can also be used for other face analysis tasks such as smile classification.

2 Related Work

Taste of food and beverage products are extensively evaluated to predict consumer acceptance, before market introduction. Current methods for evaluating taste liking depend almost entirely on self-report ratings that may bias the participants to respond in a non-spontaneous, rational way. Consequently, instant, objective and spontaneous rating of the respondents about the product cannot be measured by these methods. Facial expressions, however, can reveal such likings.

Facial expressions are strong indicators of spontaneous feelings and emotions, as well as displaying innate responses to basic tastes. Many studies report that positive facial expressions are elicited by liked (sweet) tastes, while disliked (bitter) tastes induce negative expressions in neonates [5], [6], [7]. Moreover, taste-elicited facial responses of adults are shown to be similar to those found in newborns [1], [8]. Recent findings indicate that liking is associated with more subtle and neutral facial expressions, while unpleasant tastes evoke more facial responses with higher intensities [9], [10].

Although the scientific interest on taste-elicited facial expression analysis is rapidly increasing, most studies use manual coding of facial action units [11] to analyze the relations of facial responses with liking level and with basic tastes such as bitter, salty, sour, sweet, and umami [1], [2], [10]. A few recent works use automatically recognized facial expressions for these tasks. Due to the limited accuracy of automatic facial action unit estimation, detectors of basic emotional expressions (e.g. anger, happiness, disgust, sadness, fear, and surprise) are employed in such studies [9], [12], [13]. Whilst the use of automatic analysis is promising, [14] indicates that a large number of emotions are needed to be measured to fully characterize the emotional response to foods. Therefore, the use of a few carefully selected measures, such as basic emotional expressions, can be argued to miss potentially valuable information.

Our approach is different from previous work because our goal is to automatically measure taste liking using a holistic, efficient representation based on deep learning for the full interpretation of dynamic facial expressions. Furthermore, none of the methods for taste-elicited facial expression analysis exploit subtle dynamic patterns of expressions such as speed and acceleration. In contrast to all published material, in this paper, we use facial expression dynamics for estimating the level of taste liking, as well as proposing the very first automatic approach for this task.

Temporal information is shown to be discriminative for several face analysis tasks including facial AU detection [15], emotional expression recognition [16], [17], spontaneity detection [18], [19], facial age estimation [20], and kinship verification [21], [22]. To this end, while some studies focus on engineering descriptors to capture temporal dynamics such as amplitude, speed, and acceleration of fiducial point displacements [18], [19], [20], or to represent temporal change in appearance [23], [24], others aim to learn changes in facial shape and appearance during expressions using temporal models such as hidden Markov models [16], [25].

Following the recent dramatic improvements in the field of deep learning, newer approaches [15], [17], [26] have shifted the focus to the deep architectures for temporal analysis of facial expressions. For instance, Jung *et al.* [26] models temporal appearance and shape of basic expressions using a deep Convolutional Neural Network (CNN), and a two-hidden-layer neural network, respectively. Yet, since such networks require a fixed input dimensionality, the duration of facial videos is downscaled to a fixed length. Obtained frames are then fed to a CNN so as to use each frame as a different input channel. Normalized coordinates of the fiducial points in these frames are combined into a single vector and fed to a feedforward network to model facial shape. In [17], 3-dimensional (3D) CNNs are used for learning regional changes in facial appearance during emotional expressions. However, the size of spatio-temporal blocks needs to be equal for 3D convolutions. Thus, the method is applied to videos using a sliding window approach. Once the whole video is processed, estimations for all windows are fused to obtain the final prediction.

Jaiswal *et al.* [15] propose to jointly model temporal change in appearance and shape through a combined architecture of CNN and Bi-directional Long Short-Term Memory Neural Network (BLSTM) [27] for facial AU recognition. To include shape information in the analysis, regional binary masks are used together with texture images. In order to capture dynamics of shape and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2832044, IEEE Transactions on Affective Computing

appearance changes, each frame of binary mask and texture sequences is described not only by itself but also by its difference from adjacent frames in a neighborhood of two frames, resulting a 5-frames temporal window representation. These frame-based shape and texture features are fed to two parallel convolution blocks. Their responses are fused, and followed by two additional convolution layers and a fully connected layer. Once per-frame representations are learned by the CNN, temporal dependencies in the sequence of obtained features are modeled by a BLSTM.

All the aforementioned temporal models except BLSTMs tend to learn characteristics of the temporal flow instead of capturing dynamics. However, facial behavior is complex, and temporally ordered facial responses cannot be expected during tasting. Furthermore, downscaling the duration of expressions to obtain a fixed-length representation causes the loss of temporal dynamics information (e.g. speed and acceleration). Relying on a fixed-length temporal window to learn the dynamic characteristics, on the other hand, limits the use of available temporal information. Although recurrent architectures such as long short-term memory neural networks [27], [28], allow efficiently learning from varying-length sequences as described above for facial AU recognition, they require substantial amount of data when the given sequences include long lags and heavy noise between informative intervals. Unlike well-defined intervals of AUs and emotional expressions, taste-elicited expressions are combinations of facial responses which have not been fully discovered/defined yet. Thus, indicative durations/frames of such facial responses cannot be explicitly labeled to train temporal models. In order to overcome such issues, this paper presents the first attempt to deep learn per-frame dynamics of facial responses in an unsupervised manner so as to reveal dependencies between location, speed, and acceleration of dense facial landmarks. Furthermore, we propose to encode per-frame dynamics of a given tasting video to a Fisher vector to model their pattern of co-occurrence for different appreciation levels. Since this paper aims to reveal the importance/informativeness of inner-facial movement dynamics in the analysis of taste-elicited expressions, appearance (facial texture) features are not employed in the proposed method.

Fisher vector representation and stacked denoising autoencoders have been successfully employed for face analysis in recent studies [29], [30], however, the use of these approaches conceptually differ in the current study. For instance, while [29] combines spatio-temporal information obtained from different facial regions through Fisher vector encoding, the current study uses Fisher vector representation for temporal pooling of per-frame dynamics. Similarly, while stacked denoising autoencoders are employed to visually transform expressive face images to neutral ones in [30] for more accurate face recognition, we use them to learn an efficient spatio-temporal representation for each frame in a facial video by modeling the non-linear relations of location, speed, and acceleration parameters of facial landmarks.

3 Method

Our approach aims to automatically estimate taste liking through facial expressions. In this section, details of the proposed method are given. The flow of the system can be given as follows. Initially, facial landmark points and head pose are tracked in videos. The tracked points are pose and scale normalized. After normalization, speed, and acceleration of the displacement of each facial landmark are computed. Landmarks are grouped into four different facial regions, namely: eyebrows/forehead, eyes, cheeks, and mouth. For each region, location, speed, and acceleration of the points at each frame of the videos are fed to a Stacked Denoising Autoencoder (SDAE) in order to learn efficient regional representations. The learned regional representations are computed for each frame of the test video, and coded into a Fisher vector. Concatenated regional Fisher vectors are used to train Support Vector Machine (SVM) classifiers to distinguish between three levels of taste liking, i.e. disliking, neutral, and liking.

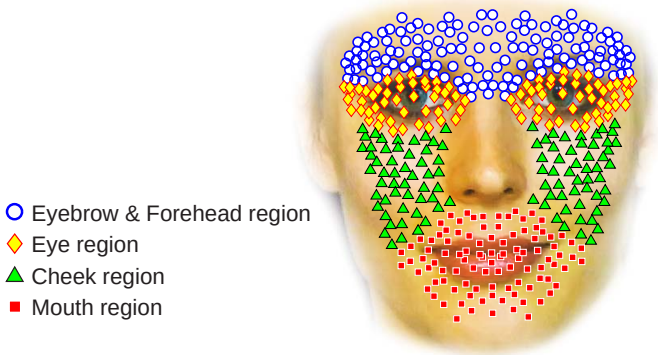


Fig. 2. Tracked facial landmarks, and the defined regions.

3.1 Facial Landmark Tracking

For a detailed analysis of the inner facial dynamics, we track 3D locations of 429 facial landmark points using a state-of-the-art tracker recently proposed by Jeni *et al.* [31]. The tracked 429 facial fiducial points on the eyebrows, forehead, eyes, cheeks, and mouth are shown in Figure 2. The tracker employs a combined 3D supervised descent method [32], where the shape model is defined by a 3D mesh and the 3D vertex locations of the mesh [31]. A dense parameterized shape model is registered to an image such that its landmarks correspond to consistent locations on the face. The accuracy and robustness of the method for 3D registration and reconstruction from 2D video was validated in a series of experiments in [31].

Once the facial landmarks are tracked, similarity normalized 3D shape representation is used for further analysis. Similarity normalized representation is the set of vertex locations after removing the global rigid transformations such as translation, rotation and scale. Since the normalized representation is frontal with respect to the camera, we ignore the depth (Z) values of the facial points. To leverage regional properties, tracked landmarks are grouped into four facial regions, namely: eyebrows & forehead, eyes, cheeks, and mouth as shown in Figure 2. The time series of the location of the points in region j (where, $j \in \{1, 2, 3, 4\}$) for a video is denoted as $\mathcal{L}_j$, and denoised by using the 4253H-twice smoothing method [33]. Facial movement dynamics are discriminative for facial expression recognition as shown in previous research [18], [19]. Therefore, speed $\mathcal{V}$ and acceleration $\mathcal{A}$ sequences are computed as

$$\mathcal{V}(t) = \frac{d\mathcal{L}}{dt}, \tag{1}$$

$$\mathcal{A}(t) = \frac{d^2\mathcal{L}}{dt^2} = \frac{d\mathcal{V}}{dt}, \tag{2}$$

and used together with the location sequence $\mathcal{L}$ of landmarks for facial representation. Including speed and acceleration measures

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2832044, IEEE Transactions on Affective Computing

in the per-frame analysis does not only allow capturing dynamic patterns of facial responses but also provides temporal phase information. For instance, while a lip corner puller (AU 12) displayed in a video frame can be identified based on the landmark locations, without using temporal information (displacement/speed) we cannot determine whether the action unit just starts (onset) or it almost ends (offset).

3.2 Learning Face Representation

The computation of location, speed, and acceleration measures for the facial representation may be complex and redundant due to tracking noise or correlated movements of the facial points. Deep architectures can learn efficient feature representations and are able to cope with high dimensionality and redundancy of data. Since we do not have additional information (i.e. class label) to learn per-frame facial representation, an unsupervised approach is required. Deep learners can progressively reveal low-dimensional, nonlinear structures in an unsupervised manner [34]. To this end, we employ the Stacked Denoising Autoencoders (SDAE) [35] to learn a transformation of raw features to an effective representation that is able to capture discriminative facial cues for classifying different levels of taste liking.

A deep autoencoder can be described as a neural network with multiple hidden layers. Such a network is trained to reconstruct its inputs, where hidden layers learn efficient representations of the inputs. In SDAE, each hidden layer is learned using a denoising autoencoder [36], which maps a corrupted version $\tilde{\mathbf{x}}$ of input $\mathbf{x} \in \mathbb{R}^p$ to a latent representation $\mathbf{y} \in \mathbb{R}^q$, and then maps it back to the original space to obtain the reconstructed input $\mathbf{z} \in \mathbb{R}^p$ as follows:

$$\mathbf{z} = \mathrm{g}(\mathbf{y}) = \mathrm{g}\big(\mathrm{f}(\tilde{\mathbf{x}}_i)\big), \tag{3}$$

where f and g denote the encoding and decoding functions, respectively. Then, the parameters Q of the denoising autoencoder is optimized by minimizing the average reconstruction error:

$$\begin{aligned} Q^* &= \arg\min_{Q} \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{x}_i, \mathbf{z}_i) \\ &= \arg\min_{Q} \frac{1}{N}\sum_{i=1}^{N} \ell\Big(\mathbf{x}_i, \mathrm{g}\big(\mathrm{f}(\tilde{\mathbf{x}}_i)\big)\Big), \end{aligned} \tag{4}$$

where ℓ is a loss function, and in this study it is defined as the squared error:

$$\ell(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|^2. \tag{5}$$

$\mathbf{x}_i$ is the i^{th} training sample, and $\tilde{\mathbf{x}}_i$ shows its corrupted version. N indicates the total number of training samples. In this way, the first hidden layer is trained to reconstruct the input data. Then each of the hidden layers are trained to reconstruct the states of the layer below, respectively. $\tilde{\mathbf{x}}$ is obtained by randomly setting a fraction w of input vector $\mathbf{x}$ to 0. Transformation weights are initialized at random and then optimized by stochastic gradient descent. Once the pre-training is completed, the entire deep autoencoder is trained to fine-tune all the parameters together to obtain optimal reconstruction, using backpropagation of error derivatives.

For each facial region, a separate 4-layer architecture is designed. To ensure a compact final representation, the number of units q at the 4^{th} hidden layer of each network is set to $\lceil \frac{d}{6} \rceil$, where d denotes the feature dimensionality of the input data. Let η_j be the number of landmarks of the facial region j, then $\lceil \frac{d}{6} \rceil = \eta_j$, because 2D measures of the location, speed, and acceleration are used as raw representation of the face at each frame. The number of units for the first three hidden layers, and other hyperparameters of SDAE are determined by minimizing the validation error (during the training of classification model; see Section 3.3). The list of the hyperparameters, and other considered values are given in Table 1.

TABLE 1
List of the hyperparameters, and considered values.

Hyperparameter	Considered values
Number of hidden layers	$\{2, 3, 4\}$
Number of units for the final hidden layer	$\{\lceil \frac{d}{6} \rceil\}$
Number of units for other hidden layers	$\{\lceil \frac{d}{4} \rceil, \lceil \frac{d}{2} \rceil, d, \lceil \frac{3d}{2} \rceil, 2d\}$
Fixed learning rate	$\{0.001, 0.01\}$
Number of epochs	$\{30, 50\}$
Corruption noise level (w)	$\{0.1, 0.2, 0.4\}$

To employ facial dynamics in the analysis, we use derivatives of location coordinates (speed and acceleration) as features. However, they are sensitive to noise in location measures. Therefore, the sequence of location coordinates are smoothed using the 4253H-twice method [33] before extracting speed and acceleration features (as described in Section 3.1). For further noise removal, each regional SDAE is trained to reconstruct these smoothed measures (location, speed, acceleration), using their raw version. Note that such a smoothing step is not applied to the hidden layers.

3.3 Video Representation and Classification

When the SDAEs are trained, regional feature vectors for each frame are encoded to the learned $\lceil \frac{d}{6} \rceil$ dimensional representation. By combining these frame based features, the tasting videos can be described. However, since the duration of the videos differ, a fixed-length feature vector is required for the representing videos. Although time series can be classified by temporal models without having a fixed-length representation, such models tend to learn characteristics of the temporal flow. Yet, taste-induced facial expressions display a complex behavior, and temporally ordered facial responses cannot be expected during tasting. To this end, an improved Fisher vector (IFV) coding is employed to describe the videos [37]. The use of such a representation aims to reveal pattern of co-occurrence of facial responses instead of capturing their temporal order.

Using a Gaussian mixture model (GMM) with 64 Gaussian distributions[1], a $128\lceil \frac{d}{6} \rceil$ dimensional IFV is computed for each facial region per video. These Fisher vectors are normalized by power normalization and l^2-norm as described in [37]. Computed regional (eyebrows & forehead, eyes, cheeks, and mouth) Fisher vectors are then concatenated and modeled by linear Support Vector Machines (SVMs). We opt to use linear kernel for SVM since Fisher vectors can be effectively modeled by linear models [37], [38]. Regularization parameter of SVM is optimized by minimizing the validation error.

1. The number of distributions in the Gaussian mixture model is empirically set to 64 so as to keep the representation as compact as possible while providing a decent validation accuracy.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2832044, IEEE Transactions on Affective Computing

4 Beverage Tasting Database

In order to assess informativeness of facial responses for estimating person-independent taste liking, a large-scale video database of beer tasting (i.e. drinking) has been collected. We opt for beers as stimuli due to the complexity of the tastes incorporated (e.g. some combination of wet, bitter, sweet, sour, carbonated, and malty) in a single product [39].

502 respondents were recruited to evaluate six different beers. Recruitment was according to the following criteria: all respondents consumed beer or lager at least once a week. 78% of the respondents were male, and 22% were female; 89% worked full-time and 11% part-time; 28% were in socio-economic group[2] AB, 68% in C1 and 4% in C2DE. Age of the respondents range from 21 to 60 (14% aged 21-24, 31% aged 25-34 and 55% aged 35-60).

Products were served at $5 \pm 2°$C according to a randomized design to minimize first sample bias. Samples were served one at a time, 120ml per product, and respondents did not have to consume all of the sample. For each sample, respondents were first asked to follow the procedure summarized below:

- Before tasting a beer, ensure you have thoroughly cleansed your pallet with a piece of cracker and some water. Be sure you do not have any of the cracker left in your mouth.
- Sit up straight, facing forward and tuck your chair into the booth.
- When you are ready to taste your sample hold up the green piece of card in front of your face, have a neutral expression on your face, and put the card back down.
- Pick up your glass, making sure the 3-digit code is facing the camera and not obscured by your hand.
- Take a sip of your drink but do not swallow right away.
- Put the glass down, look into the camera and swallow the beer.
- Remain looking at the camera for a few seconds.
- Take a further couple of sips before proceeding with the questionnaire.
- Complete the questionnaire by selecting a response for each question as instructed.

Facial expressions during the entire session were captured using a Logitech C920 high definition webcams frontally positioned to the face. Videos were recorded with a resolution of 1280×1024 pixels at a rate of 15 frames per second under controlled illumination conditions. Data were collected in many booths in parallel, thus several computers (PCs) were required. PCs provided by a third party were unable to acquire videos with more than 15 frames per second. Each subject has a recording of about one hour. The respondents were requested to show a green card to the camera just before starting each of the beer tasting. Green cards were automatically detected, and one video segment for each beer was identified from the detection of green card until click bursts, indicating the start of questionnaire responding after the beer tasting. Afterwards, the frame just before the initial sip in each video segment was manually annotated. Frames before the initial sip were removed. Each beer was evaluated by each participant by completing a questionnaire. The questionnaire comprised of an overall liking score on a 9-point scale ("dislike extremely" to "like extremely"), as well as other evaluations such as 5-point scale purchase intent ("would definitely not buy" to "would definitely buy").

2. In terms socio-economic classification [40], group AB represents higher and intermediate managerial, administrative, professional occupations; group C1 indicates supervisory, clerical and junior managerial, administrative, professional occupations; group C2DE consists of skilled, semi-skilled and unskilled manual occupations, unemployed and lowest grade occupations.

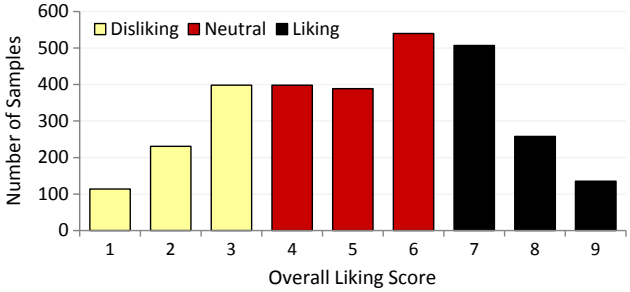


Fig. 3. Distribution of the overall liking scores and the corresponding liking classes in the beverage tasting database.

Visual data for 7 subjects were lost due to recording problems. Data from 495 subjects are used in our experiments, yielding 2970 videos of beer tasting (495 subjects $\times$ 6 beers) composed of about 700K frames. Overall liking scores were linearly mapped from 9-point scale to 3-point scale (indicating disliking, neutral state, and liking), and used as class labels for distinguishing between different levels of taste liking. Resulting database is composed of 743 disliking, 1327 neutral, and 900 liking videos. Distribution of the overall liking scores and the corresponding liking classes are shown in Figure 3. Since we don't have the consent of the participants, we are unable to share the database.

5 Experiments & Results

To evaluate the proposed approach and assess the discriminative power of facial expression dynamics for taste liking, we use our newly collected beer tasting database of 495 subjects. The regularization parameter of SVMs, and hyperparameters of stacked denoising autoencoders (see Table 1) are optimized on a seperate validation set. To this end, a two level cross-validation scheme is used. While 10-fold cross-validation is used for testing, remaining nine folds (at each iteration) are three-fold cross-validated to optimize the parameters. There is no subject overlap between folds in the database. In the experiments, performance of three-class classification (disliking, neutral, liking) is evaluated for the estimation of taste-liking level. At each fold of cross-validation, the number of training samples for each class is equalized by randomly choosing n samples per class, where n is the number of training samples of the class with minimum sample-size. Folds and the randomly selected training samples are kept same for all experiments.

5.1 Facial Regions

In this paper, we propose an approach that distinguishes between different taste liking levels based on the dynamic movement characteristics of facial regions. To assess the informativeness of the facial regions, we evaluate the accuracy of the proposed approach using eyebrow & forehead, eye, cheek, and mouth regions, individually. We then compare these results with the combined use of regional representations, and with the use of a global (holistic) face representation. The global face representation is learned by the same approach by encoding all facial features together.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2832044, IEEE Transactions on Affective Computing

TABLE 2
Correct classification rates for individual and combined use of regional descriptors, and using global face representation. Best results are boldfaced.

Region	Disliking	Neutral	Liking	Total
Eyebrow & Forehead	0.7227	0.5373	0.6422	0.6155
Eye	0.6635	0.5569	0.6178	0.6020
Cheek	0.6366	0.5102	0.5833	0.5640
Mouth	0.6393	0.5026	0.5733	0.5582
Combined	**0.7927**	**0.6413**	**0.7222**	**0.7037**
Global Face	0.7658	0.6240	0.6733	0.6744

As shown in Table 2, eyebrow & forehead region performs best with a classification accuracy of 61.55%, and followed by the eye region (60.20%). This finding shows the importance of upper facial movements for detecting taste preferences. The lowest correct classification rates are provided by cheek, and mouth regions, respectively. The use of mouth region performs with an accuracy of 55.82% which is 5.73% (absolute) less than that of eyebrow & forehead region. These results may be explained by the fact that movements of mouth and cheek regions are highly affected by drinking and swallowing.

When we fuse regional descriptors, classification accuracy is increased to 70.37%, which is significantly (8.82% higher, $t = -7.16$, $df = 5938$, $p < 0.001$) higher than that of best performing facial region. Global face representation achieves an accuracy of 67.44%, outperforming individual use of regional descriptors. Yet, the accuracy of the global approach is significantly ($t = 2.44$, $df = 5938$, $p = 0.015$) lower than that of the combined use of regional descriptors. This finding can be explained by two reasons. First, a four-layer SDAE is employed to learn both global and regional face representations using the same number of data samples. While the global approach may learn relations between regional dynamics, it has to deal with much higher complexity in comparison to the regional approach. Second, both regional and global descriptors are modeled using 64 Gaussian distributions during Fisher vector computation. As a result, based on the higher complexity of global features, regional video descriptors can better represent dynamic characteristics.

5.2 Facial Dynamics

One of the main contributions of this paper is the exploitation of per-frame dynamics of facial responses to capture subtle temporal cues. In order to evaluate whether facial dynamics provide useful information for classifying taste-liking levels, we implement a modified version of the proposed method by discarding speed and acceleration measures. Resulting method solely uses the facial landmark locations to deep learn face representation. We compare the accuracy of the modified method to that of the proposed approach. Both individual facial regions and their combined use are evaluated for comparison.

As shown in Figure 4, discarding per-frame dynamics (i.e. speed and acceleration) significantly decreases ($t = 3.29$, $df = 5938$, $p < 0.001$) the accuracy for all facial regions and for their combined use. While the proposed method achieves a classification accuracy of 70.37%, only 63% of the samples are correctly classified by the sole use of landmark locations. In other words, combining speed and acceleration information with landmark locations provides a relative accuracy improvement of 10%. For the individual regions, using facial dynamics achieves a 9% (relative) improvement on average.

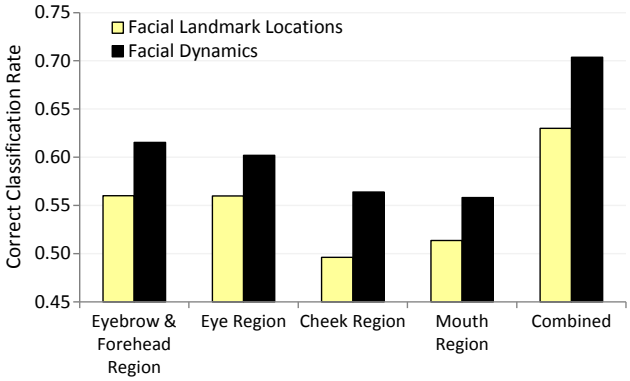


Fig. 4. Influence of facial dynamics on correct classification rates.

TABLE 3
Correct classification rates for location- and dynamics-based analysis with the use of 2D and 3D face normalization. Best results are boldfaced.

Method	Disliking	Neutral	Liking	Total
2D + Landmark Locations	0.6934	0.5702	0.6036	0.6111
3D + Landmark Locations	**0.7121**	**0.5832**	**0.6313**	**0.6300**
2D + Facial Dynamics	0.7532	0.6108	0.6710	0.6646
3D + Facial Dynamics	**0.7927**	**0.6413**	**0.7222**	**0.7037**

One of the reasons behind these results is the fact that Fisher vector encoding cannot capture temporal information based on the displacement of points while modeling the distribution of landmark locations (in a video) unless speed and acceleration measures are used as additional inputs. Furthermore, the significant accuracy decrease due to discarding speed and acceleration measures validates the importance of including dynamic/temporal information in frame-based representations.

Since the proposed method solely relies on displacement measurements of the fiducial points, accurate shape normalization of faces is vital for a reliable analysis of facial dynamics. Note that even small/moderate variations in 3D head pose may cause significant amount of noise in the displacement measurements because the computation of change in point locations requires exact alignment of consecutive face images. Therefore, as described in Section 3.1, 3D facial tracking is employed for more accurate alignment through 3D shape normalization. To assess the effectiveness of using 3D tracking/normalization, we compare the accuracy of our method with that of its modified version, where 2D coordinates (of the same landmarks) are used for facial alignment, discarding the normalization of yaw and pitch rotations.

As shown in Table 3, when we solely employ facial landmark locations for the analysis, accuracy with the use of 3D shape normalization is only 1.89% higher than that of using 2D normalization. Note that this is not a statistically significant ($t = -1.50$, $df = 5938$, $p > 0.05$) improvement. Yet, once we employ facial dynamics (including speed and acceleration) for the classification, the use of 3D shape normalization significantly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2832044, IEEE Transactions on Affective Computing

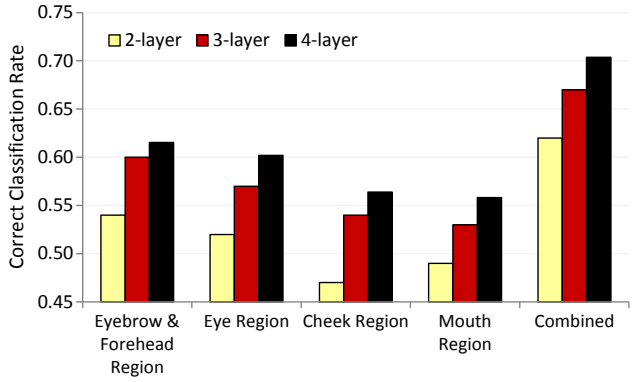


Fig. 5. Correct classification rates for different number of hidden layers.

outperforms ($t = -3.24$, $df = 5938$, $p = 0.001$) its 2D competitor with an accuracy improvement of 3.91%. This finding can be explained by the fact that displacement measures and their derivatives are more sensitive to noise compared to the point locations. As a result, even subtle misalignments between faces in consecutive frames significantly affects the reliability of the speed and acceleration measures, causing a accuracy decrease when 2D normalization is used.

5.3 Hidden Layers

The proposed method uses four-layer stacked denoising autoencoders to learn regional descriptors. However, in order to evaluate the effect of number of hidden layers, we evaluate the use of two, three, and four hidden layers in the network architecture, and compare the resulting correct classification rates. The same hyperparameters are considered (see Table 1) for the optimization of each network. Figure 5 shows the obtained correct classification rates using different number of hidden layers.

Classification accuracy is improved at each additional layer for all regions as well as for their combined use. Increasing the number of hidden layers from two to three improves the classification accuracy of regional representations by 10% on average. The fourth layer can only provide an improvement of 4% over the third layer. For the combined use of regional descriptors, the accuracy increase by the third and fourth layers, are 7%, and 5%, respectively. Obtained accuracy improvements confirm that SDAE can gradually reveal nonlinear structure in the facial data, and consequently learn a better representation for regional dynamics.

Next, we analyze the number of units for each hidden layer. Although the number of units are determined for each of the 10 folds, separately, in most cases the same set of values are chosen in our experiments. These configurations of hidden units for different regions, and layers are given in Table 4. Note that the number of units in the highest hidden layer is set to one sixth of the input dimensionality (see Section 3.2) for a compact representation.

5.4 Comparison to Other Methods

To the best of our knowledge, this is the first study proposing an algorithm to automatically classify the level of taste liking from facial videos. Therefore, we implement seven different baselines using related methods to compare with the proposed approach.

The first baseline is a modified version of the proposed approach, where the stacked denoising autoencoders are replaced by principle component analysis (PCA). For each region, one sixth of the original dimensionality is obtained using PCA. In the second method, a Fisher kernel is derived by modeling location, speed, and acceleration features using a hidden Markov model (HMM) as described in [41]. The number of hidden units in HMM is determined by minimizing the validation error. 2, 5, and 10 hidden units are used as candidate configurations.

As the third baseline, we use the improved trajectories method [42] that achieves state-of-the-art results for action recognition. Trajectories are computed for every 15 frames. Extracted trajectory features, histograms of oriented gradient (HOG), histograms of optical flow (HOF), and motion boundary histograms (MBH) are combined, and fed to PCA to reduce the feature dimensionality to half. Obtained principle components are then encoded to improved Fisher vector representation (IFV) using 128 Gaussian distributions.

In the fourth baseline, facial appearance in each frame is described by HOG features. Initially, faces are normalized with respect to roll rotation, resized, and cropped as to obtain a 128×128 pixels resolution. HOG features are extracted from 2×2 cells in 8×8 equally sized blocks, and 9 bins are used to compute histograms. For a fair comparison, dimensionality of the HOG features is reduced to 210, yielding a similar feature dimensionality with the proposed approach. Videos are then represented by IFV using 128 Gaussian distributions.

The fifth baseline method analyzes each frame of the videos to detect 11 facial action units (AU) that are shown to signal taste-related cues [1], [8]: AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU6 (cheek raiser), AU9 (nose wrinkler), AU12 (lip corner puller), AU15 (lip corner depressor), AU18 (lip puckerer), AU20 (lip stretcher), AU23 (lip tightener), and AU26 (jaw drop). Then, the estimated AU probabilities are encoded to improved Fisher vector representation (IFV) using 128 Gaussian distributions. To detect the AUs, we use the method proposed in [43]. In this method, facial surface is divided into 27 regions using facial landmarks. Then, local binary patterns (LBP) and local phase quantization (LPQ) features are extracted from each region, and used together to train SVMs as regional AU detectors. The computed posterior probabilities for each region are fused using weighted SUM rule. To estimate these posterior probabilities, sigmoids of SVM output distances are used. Weights are determined by the validation performance of the classifiers. To train detectors for the indicated 11 AUs, we combine subsets of the DISFA [44], Bosphorus [45], extended Cohn-Kanade [4], and Affectiva-MIT [46] databases.

For the sixth and seventh baselines, we modify the CNN-LSTM architecture proposed in [47]. In the sixth baseline, CNN layers of [47] are removed and the sequences of facial dynamics are fed to the 3 stacks of LSTMs. Since we do not have per-frame annotations for taste liking in our database, a single softmax classifier is placed on top of the last time step of the stacked LSTMs. In the seventh baseline, while the CNN and LSTM architectures of [47] are kept intact, fusion of the outputs of CNNs and LSTMs are disabled and one softmax classifier is connected to the last time step of the stacked LSTMs. Normalized face images are used as inputs to CNNs. The outputs of the CNNs are connected to LSTMs so as to allow an end-to-end learning. For a fair comparison sixth and seventh baselines are trained from scratch using our database. The same hyperparameters/settings with [47] are used in these baselines.

All baseline methods are trained to distinguish between three

TABLE 4
Determined number of units at each hidden layer of 2-, 3-, 4-layer architectures. Values from left to right show the number of hidden units from initial layer to higher layers. Note that d denotes the dimensionality of input features.

Region	2-layer	3-layer	4-layer
Eyebrow & Forehead	$\lceil \frac{3d}{2} \rceil \rightarrow \lceil \frac{d}{6} \rceil$	$d \rightarrow 2d \rightarrow \lceil \frac{d}{6} \rceil$	$\lceil \frac{3d}{2} \rceil \rightarrow \lceil \frac{3d}{2} \rceil \rightarrow 2d \rightarrow \lceil \frac{d}{6} \rceil$
Eye	$\lceil \frac{d}{2} \rceil \rightarrow \lceil \frac{d}{6} \rceil$	$d \rightarrow \lceil \frac{3d}{2} \rceil \rightarrow \lceil \frac{d}{6} \rceil$	$d \rightarrow 2d \rightarrow 2d \rightarrow \lceil \frac{d}{6} \rceil$
Cheek	$d \rightarrow \lceil \frac{d}{6} \rceil$	$\lceil \frac{d}{2} \rceil \rightarrow \lceil \frac{d}{4} \rceil \rightarrow \lceil \frac{d}{6} \rceil$	$d \rightarrow d \rightarrow \lceil \frac{3d}{2} \rceil \rightarrow \lceil \frac{d}{6} \rceil$
Mouth	$d \rightarrow \lceil \frac{d}{6} \rceil$	$\lceil \frac{d}{2} \rceil \rightarrow \lceil \frac{d}{2} \rceil \rightarrow \lceil \frac{d}{6} \rceil$	$\lceil \frac{d}{2} \rceil \rightarrow d \rightarrow d \rightarrow \lceil \frac{d}{6} \rceil$

TABLE 5
Description of the compared methods, and achieved correct classification rates. Note that the "Facial Dynamics" denotes the use of speed, and acceleration measures together with landmark locations. Best results are boldfaced.

Method	Description	Disliking	Neutral	Liking	Total
Proposed	Facial Dynamics + SDAE + IFV (64-GMM)	**0.7927**	0.6413	**0.7222**	**0.7037**
Baseline 1	Facial Dynamics + PCA + IFV (64-GMM)	0.6501	0.4748	0.5867	0.5525
Baseline 2	Facial Dynamics + Fisher Kernel	0.5464	0.3753	0.3589	0.4131
Baseline 3	Improved Trajectories + PCA + IFV (128-GMM)	0.6258	0.5667	0.5956	0.5902
Baseline 4	HOG Features + PCA + IFV (128-GMM)	0.6514	0.6127	0.5922	0.6162
Baseline 5	Facial AU Levels + IFV (128-GMM)	0.5303	**0.7038**	0.5944	0.6273
Baseline 6	Facial Dynamics + LSTM	0.6322	0.5616	0.5846	0.5862
Baseline 7	Face Images + CNN-LSTM	0.6301	0.6282	0.5760	0.6129
	Number of samples	743	1327	900	2970

levels of taste liking: disliking, neutral, and liking. While sixth and seventh baselines employ softmax classifier, other baselines use linear SVM. As shown in Table 5, the proposed approach significantly c outperforms all the baseline methods with a correct classification rate of 70.37%. AU-based method (*baseline 5*) follows the proposed approach with an accuracy of 62.73%. Although the AU-based method provides the second best performance, the accuracy of taste liking classification may drastically drop in case of inaccurate estimation of AU probabilities.

Fusing facial appearance (HOG descriptors) in each frame of a video through IFV encoding (*baseline 4*) provides an accuracy of 61.62%. While appearance features can capture subtle changes in a better way compared to shape features (e.g. facial landmark locations), both our proposed approach and its modified version (see Section 5.2) that discards speed and acceleration measures, outperform *baseline 4*. Based on this finding, we can confirm the informativeness of deep-learned representations.

End-to-end modeling of facial image sequences using a CNN-LSTM architecture (*baseline 7*) can correctly classify only 61.29% of the videos. LSTM modeling of facial dynamics (*baseline 6*) performs even worse with an accuracy of 58.62%. These results may suggest that taste-liking levels are correlated with pattern of co-occurrence of specific facial responses rather than temporal flow of the responses. Another reason may be the large data requirement of recurrent architectures for modeling sequences with long lags and heavy noise between informative intervals.

The correct classification rate achieved by the improved dense trajectories method (*baseline 3*) is only 59.02%. This result can be explained by the fact that improved trajectories do not leverage the knowledge of facial morphology in comparison to facial tracking methods. Consequently, it performs 11.35% (absolute) worse than the proposed method.

When the stacked denoising autoencoders are replaced with PCA (*baseline 1*) in the proposed approach, a 15.12% accuracy decrease is observed. This finding shows that SDAEs can learn very efficient and informative descriptors for this task by revealing non-linear relations between facial movements and their dynamics.

Fisher kernel representation (*baseline 2*) computed from location, speed, and acceleration measures of landmarks, performs worst in our experiment with an accuracy of 46.84%. This is an expected result since the dynamic characteristics of taste-induced facial behavior is complex, and taste-induced responses do not follow a specific temporal pattern.

When we analyze per-class accuracies, it is seen that all methods except AU-based baseline provide higher accuracy for disliking condition. This result is in line with the findings of [9] and [10] indicating that unpleasant tastes evoke more facial responses with higher intensities, which can be better differentiated than pleasant tastes since liking is associated with more subtle and neutral facial expressions.

5.5 Influence of Gender

In order to explore the gender-based differences in taste-elicited facial dynamics, the correct classification rates are obtained for each gender. While the accuracy is 73.08% for females, a correct classification rate of 69.58% is obtained for male participants. Consequently, there is no significant ($t = 1.7181$, $df = 2968$, $p = 0.0859$) accuracy difference between male and female subjects. Next, for each taste-liking level, we compute the amount of features (in the final video representation) that significantly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2832044, IEEE Transactions on Affective Computing

TABLE 6
Confusion matrices for human prediction and the proposed method. Correct classification rates for each class are boldfaced.

	Human Prediction			Proposed Method		
Actual \ Predicted	Disliking	Neutral	Liking	Disliking	Neutral	Liking
Disliking	**0.6400**	0.2178	0.1422	**0.7333**	0.1733	0.0933
Neutral	0.2844	**0.3689**	0.3467	0.1600	**0.5600**	0.2800
Liking	0.1200	0.3333	**0.5467**	0.0800	0.2533	**0.6667**
	Total Accuracy: 0.5185			Total Accuracy: 0.6533		
	Weighted Cohen's κ: 0.366			Weighted Cohen's κ: 0.546		

($p < 0.05$) differ between male an female subjects. Our results show that gender significantly affect 34.48%, 28.88%, and 22.67% of the features during disliking, neutral, and liking conditions, respectively.

Since the proposed video-level representation is not directly interpretable in terms of facial dynamics, we extract the point displacements for each region, and compute their first principle components (through PCA). Using the first principle component sequences of displacement, regional mean displacements over each video are calculated. Notice that mean displacement is equivalent to mean speed since each video has been sampled at the same frame rate. Our analysis of gender effects on these regional measures show that mean expressiveness of eye and cheek regions are significantly ($p < 0.01$) higher for females during disliking. Yet, eyebrow & forehead region of males is significantly ($p < 0.01$) more expressive than that of females during disliking.

5.6 Influence of Age

To assess the influence of age on taste-elicited facial expression dynamics, we parse our results and analyze the correct classification rates for different age groups. To this end, we split the subjects into two groups based on their age as *young* (21-34 years) and *mid-aged* (35-60 years), representing 45% and 55% of the participants, respectively. Obtained results show that taste-liking levels of 69.13% of the young group is correctly classified, while the accuracy for mid-aged group is 71.38%. Yet, the accuracy difference between the age groups is not statistically significant ($t = 1.3371$, $df = 2968$, $p > 0.1813$). Furthermore, we calculate the amount of features (in the final video representation) that significantly ($p < 0.05$) differ between young and mid-aged subjects, for each taste-liking level. As a result, we find that age significantly affect 75.63%, 87.06%, and 92.51% of the features during disliking, neutral, and liking conditions, respectively.

To explore the effects of age on taste-elicited facial expressions, we analyze the regional mean displacement measures as in Section 5.5. Our results indicate that during liking, mouth ($p < 0.005$) and eyebrow & forehead ($p < 0.02$) regions of young subjects are significantly more expressive than that of mid-aged group, while eye region of young subjects displays significantly ($p < 0.01$) lower expressiveness.

5.7 Comparison to Human Accuracy

To comprehend general human knowledge and ability to judge and classify taste appreciation of other individuals from their facial expressions, we gathered human predictions of taste-liking levels for a subset of beer tasting videos in the collected database. To this end, we randomly selected 75 videos for each of disliking, neutral, and liking classes in a way that each level of overall liking (9-point scale) had 25 samples. In total, 225 videos were used from 146 male and 42 female (gender distribution is similar that of whole database), mainly Caucasian.

For the experiment, we recruited forty-five participants, 23 male and 22 female. Each participant was shown 15 videos, and asked to rate the perceived taste-liking level for each video as liking, neutral, or disliking. None of the participants were experts on face analysis or took a special training in facial expressions. The participants ranged in age from 23 to 56 years (mean: 30.2) and were of 14 different nationalities, i.e. British, Chinese, Costa Rican, Dutch, German, Greek, Hungarian, Indian, Indonesian, Iranian, Portuguese, Romanian, Serbian, and Turkish. Taste-liking level for each video was predicted by three different participants. A different set of videos were shown to each participant. In order to compare the reliability of human prediction to that of the proposed method, we have tested our method on the same subset of 225 videos. We assess and compare the performance of humans and our method based on confusion matrices, total accuracy, and linear weighted Cohen's κ that is the proportion of ordinal agreement above what would be expected to occur by chance [48].

As shown in Table 6, our method performs better than humans for each of the three liking classes. Total accuracy of the proposed method reaches to 65.33% which is significantly higher than human accuracy (13.48% higher, $t = -3.50$, $df = 898$, $p < 0.001$). Yet, confusion patterns of human and computer predictions are similar. For instance, most confusion occurs between neutral and liking classes, which is followed by the confusion of neutral and disliking classes. While 34% of human predictions for neutral and liking samples are confused with each other, this rate is 26.67% for the proposed method. As expected, the lowest confusion rate is observed between liking and disliking classes (i.e. 13.11% and 8.67% for human and computer predictions, respectively). Consequently, the most accurate classification is achieved for disliking class by both of human and computer predictions.

Based on weighted Cohen's κ, the predictions by different participants have been found to be fairly consistent ($0.20 < \kappa \leq 0.40$). Weighted κ for automatic predictions (by our method), on the other hand, is 49.18% higher (relative) and represents a moderate level of agreement ($0.40 < \kappa \leq 0.60$). These findings suggest that humans are less accurate and less reliable for estimating taste appreciation from facial expressions in comparison

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2832044, IEEE Transactions on Affective Computing

Fig. 6. Automatically extracted frames which correspond to the highest scores computed by SVMs for disliking (top row), neutral (middle row), and liking classes (bottom row).

to the proposed method. It is important to note that these results are based on knowledge and ability of a non-expert population. In other words, participants had no expertise in facial expressions at the time of the experiment.

5.8 Visual Analysis

Our proposed approach uses deep learning to compute frame-based descriptors, and Fisher vector encoding for video representation. Since these methods perform in an unsupervised manner, no labels are required for the per-frame liking levels. If such labels are provided, a better understanding of taste-elicited facial expressions could be obtained. And these labels could allow the implementation of single image based approaches. To this end, we explore the most discriminative frames (single images) for distinguishing between different liking levels.

Fisher vector encoding provides a fixed-length representation for varying-duration videos. As an extreme case, we can even compute a Fisher vector for a single frame, and evaluate it using the models learned on the videos. In this way, we can detect the frames that correspond to the highest scores computed by SVMs for disliking, neutral, and liking classes.

Since the respondents in the collected database did not allow us to publish their images, we have collected an additional small-scale database for visualization purposes. To this end, we recorded six respondents' (three female, three male) facial responses during the tasting of four different beers. Tasting durations for each beer are segmented in the same way as described in Section 4. Consequently, we have obtained four videos for each of the six subjects. For each frame in this database, a Fisher vector is computed, and fed to the three-class (disliking, neutral, and liking) SVM classifier. Note that the stacked denoising autoencoders, Fisher vector encoders, and SVM models are all trained on the database of 495-subjects. Then, for each subject, three frames are extracted which correspond to the highest scores for disliking, neutral, and liking classes, as show in Figure 6.

Extracted frames show interesting facial expression patterns. Disliking-related frames mostly display lowered eyebrows, lowered eyelids, and stretched lips, as well as having raised upper lips. Almost all liking-related frames show lip sucking, and some of them have raised eyebrows. Finally, frames corresponding to the highest score for the neutral class, show perfect neutral faces. These facial expression responses are similar to the taste-related facial actions reported by previous studies [1], [8]. Please note that, the deep-learned descriptors also include speed and acceleration information, but they are not visualized here.

5.9 Application to Smile Classification

To assess the generalization of the proposed method to other face analysis tasks, we evaluate the method for spontaneous versus posed smile classification and compare its accuracy to that of the state-of-the-art smile classification systems proposed in the literature [19], [23], [25], [49], [50], [51]. Task of spontaneous versus posed smile classification is chosen for this experiment since effective modeling of dynamics and/or spatio-temporal characteristics of smiles are crucial in order to provide a reliable and accurate spontaneity analysis. In our experiment, we employ the UvA-NEMO smile database [50] that has 1240 smile videos (597 spontaneous, 643 posed) from 400 subjects (185 female, 215 male). Videos were recorded with a resolution of 1920×1080 pixels at a rate of 50 frames per second.

TABLE 7
Classification accuracy of different methods for spontaneous versus posed smile classification on the UvA-NEMO smile database, and the features employed by these methods. Highest correct classification rate is boldfaced.

Method	Feature	Accuracy
Proposed Method	Deep-learned Facial Dynamics	**0.9177**
Dibeklioğlu *et al.* (2015) [19]	Facial Dynamics + Age	0.9056
Wu *et al.* (2014) [49]	Spatio-temporal Appearance	0.9140
Dibeklioğlu *et al.* (2012) [50]	Facial Dynamics	0.8702
Pfister *et al.* (2011) [23]	Spatio-temporal Appearance	0.7306
Dibeklioğlu *et al.* (2010) [25]	Eyelid Dynamics	0.7105
Cohn and Schmidt (2004) [51]	Lip Corner Dynamics	0.7726

As shown in Table 7, the proposed method provides an accuracy of 91.77% and improves the state of the art. Although the accuracy improvement over the work of Wu *et al.* [49] (91.40%) is marginal, it is important to note that the methods proposed by Wu *et al.* [49] and Pfister *et al.* [23] exploit spatio-temporal appearance of face (by extracting completed Local Binary Patterns from three orthogonal planes descriptor [23] and its discriminative variant [49] from a given smile video) instead of the sole use of displacement dynamics of facial landmarks. Yet, such spatio-temporal approaches could not perform better than the proposed method. This finding suggests the importance of deep-learned displacement dynamics for face analysis tasks.

6 Conclusions

In this paper, we have proposed the first approach for automatic estimation of taste liking from facial expression videos. Instead of using handcrafted features, the proposed approach deep learns regional facial dynamics per frame, and encodes them to a Fisher vector per region to describe videos. Regional Fisher vectors are then concatenated and classified by linear SVM classifiers.

We have presented the first large-scale beverage tasting database (2970 videos of 495 subjects) in the literature for detailed and precise analysis of taste-elicited spontaneous facial expressions. On the collected database, the proposed approach has achieved an accuracy of 70.37% for distinguishing between three levels of taste-liking (liking, being neutral, and disliking), outperforming all other methods by more than 8.65% (absolute). The results have indicated that the combined use of regional dynamics are more discriminative than the global face representation for this task. Relying on SVM scores, the most discriminative facial responses of six young adults for taste-liking estimation have been obtained, and shown to be similar to those reported in previous studies.

Our experiments for distinguishing between spontaneous and posed enjoyment smiles have confirmed the generalization power of the proposed method, suggesting that deep learning can indeed provide efficient representations of regional facial dynamics. Recruiting 45 participants, we have evaluated the ability and reliability of humans for estimating taste appreciation of others' from their facial expressions. Our findings have shown that humans are significantly less reliable for this task in comparison to the proposed method.

References

[1] R. Weiland, H. Ellgring, and M. Macht, "Gustofacial and olfactofacial responses in human adults," *Chemical senses*, vol. 35, no. 9, pp. 841–853, 2010.

[2] K. Wendin, B. H. Allesen-Holm, and W. L. Bredie, "Do facial reactions add new dimensions to measuring sensory responses to basic tastes?" *Food quality and preference*, vol. 22, no. 4, pp. 346–354, 2011.

[3] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *California Mental Health Research Digest*, vol. 8, no. 4, pp. 151–158, 1970.

[4] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 94–101.

[5] J. R. Ganchrow, J. E. Steiner, and M. Daher, "Neonatal facial expressions in response to different qualities and intensities of gustatory stimuli," *Infant Behavior and Development*, vol. 6, no. 2, pp. 189–200, 1983.

[6] N. A. Fox and R. J. Davidson, "Taste-elicited changes in facial signs of emotion and the asymmetry of brain electrical activity in human newborns," *Neuropsychologia*, vol. 24, no. 3, pp. 417–422, 1986.

[7] D. Rosenstein and H. Oster, "Differential facial responses to four basic tastes in newborns," *Child development*, vol. 59, no. 6, pp. 1555–1568, 1988.

[8] E. Greimel, M. Macht, E. Krumhuber, and H. Ellgring, "Facial and affective reactions to tastes and their modulation by sadness and joy," *Physiology & Behavior*, vol. 89, no. 2, pp. 261–269, 2006.

[9] R. A. de Wijk, W. He, M. G. Mensink, R. H. Verhoeven, and C. de Graaf, "Ans responses and facial expressions differentiate between the taste of commercial breakfast drinks," *PLoS ONE*, vol. 9, no. 4, p. e93823, 2014.

[10] G. G. Zeinstra, M. Koelen, D. Colindres, F. Kok, and C. De Graaf, "Facial expressions in school-aged children are a good indicator of 'dislikes', but not of 'likes'," *Food Quality and Preference*, vol. 20, no. 8, pp. 620–624, 2009.

[11] P. Ekman and W. V. Friesen, *Facial action coding system*. Palo Alto, CA: Consulting Psychologists Press, 1977.

[12] L. Danner, L. Sidorkina, M. Joechl, and K. Duerrschmid, "Make a face! implicit and explicit measurement of facial expressions elicited by orange juices using face reading technology," *Food Quality and Preference*, vol. 32, pp. 167–172, 2014.

[13] R. A. de Wijk, V. Kooijman, R. H. Verhoeven, N. T. Holthuysen, and C. de Graaf, "Autonomic nervous system responses on and facial expressions to the sight, smell, and taste of liked and disliked foods," *Food Quality and Preference*, vol. 26, no. 2, pp. 196–203, 2012.

[14] S. C. King and H. L. Meiselman, "Development of a method to measure consumer emotions associated with foods," *Food Quality and Preference*, vol. 21, no. 2, pp. 168–177, 2010.

[15] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–8.

[16] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Computer Vision and image understanding*, vol. 91, no. 1, pp. 160–187, 2003.

[17] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian Conference on Computer Vision*, 2014, pp. 143–157.

[18] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 02, pp. 121–132, 2004.

[19] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Recognition of genuine smiles," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 279–294, 2015.

[20] H. Dibeklioğlu, F. Alnajar, A. Ali Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1928–1943, 2015.

[21] H. Dibeklioğlu, A. Ali Salah, and T. Gevers, "Like father, like son: Facial expression dynamics for kinship verification," in *IEEE International Conference on Computer Vision*, 2013, pp. 1497–1504.

[22] E. Boutellaa, M. B. López, S. Ait-Aoudia, X. Feng, and A. Hadid, "Kinship verification from videos using spatio-temporal texture features and deep learning," in *International Conference on Biometrics*, 2016, pp. 1–7.

[23] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *International Conference on Computer Vision Workshops*, 2011, pp. 868–875.

[24] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[25] H. Dibeklioğlu, R. Valenti, A. A. Salah, and T. Gevers, "Eyes do not lie: Spontaneous versus posed smiles," in *ACM International Conference on Multimedia*, 2010, pp. 703–706.

[26] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, "Deep temporal appearance-geometry network for facial expression recognition," *arXiv preprint arXiv:1503.01532*, 2015.

[27] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 255–259.

[30] C. S. N. Pathirage, L. Li, W. Liu, and M. Zhang, "Stacked face de-noising auto encoders for expression-robust face recognition," in *International Conference on Digital Image Computing: Techniques and Applications*, 2015, pp. 1–8.

[31] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[32] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[33] P. F. Velleman, "Definition and comparison of robust nonlinear data smoothing algorithms," *Journal of the American Statistical Association*, vol. 75, no. 371, pp. 609–615, 1980.

[34] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[36] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine learning*, 2008, pp. 1096–1103.

[37] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, 2010, pp. 143–156.

[38] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *British Machine Vision Conference*, 2011, pp. 76.1–76.12.

[39] L. Lee, S. Frederick, and D. Ariely, "Try it, you'll like it the influence of expectation, consumption, and revelation on preferences for beer," *Psychological Science*, vol. 17, no. 12, pp. 1054–1058, 2006.

[40] Market Research Society, *Occupation Groupings: A Job Dictionary*, 2006.

[41] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*, no. 11, 1998, pp. 487–493.

[42] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *International Conference on Computer Vision*, 2013, pp. 3551–3558.

[43] B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic, "Decision level fusion of domain specific regions for facial action recognition," in *International Conference on Pattern Recognition*, 2014, pp. 1776–1781.

[44] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[45] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *Biometrics and Identity Management*, 2008, pp. 47–56.

[46] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected In-the-Wild," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 881–888.

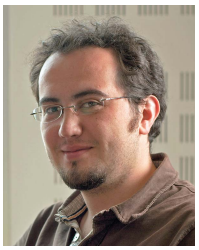
[47] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 25–32.

[48] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[49] P. Wu, H. Liu, and X. Zhang, "Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1240–1244.

[50] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in *European Conference on Computer Vision*, 2012, pp. 526–539.

[51] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 2, pp. 121–132, 2004.



Hamdi Dibeklioğlu (S'08–M'15) received the the M.Sc. degree from Boğaziçi University, Istanbul, Turkey, in 2008, and the Ph.D. degree from the University of Amsterdam, Amsterdam, The Netherlands, in 2014. He is currently an Assistant Professor in the Computer Engineering Department of Bilkent University, Ankara, Turkey. He is also a Research Affiliate with the Pattern Recognition & Bioinformatics Group at Delft University of Technology, Delft, The Netherlands. Earlier, he was a Visiting Researcher at Carnegie Mellon University, University of Pittsburgh, and Massachusetts Institute of Technology. His research interests include Affective Computing, Intelligent Human-Computer Interaction, Pattern Recognition, and Computer Vision.

Dr. Dibeklioğlu was a Co-chair for the Netherlands Conference on Computer Vision 2015, and a Local Arrangements Co-chair for the European Conference on Computer Vision 2016. He served on the Local Organization Committee of the eNTERFACE Workshop on Multimodal Interfaces, in 2007 and 2010.

Theo Gevers (M'01) is a Full Professor of Computer Vision with the University of Amsterdam, Amsterdam, The Netherlands. He is a co-founder of Sightcorp and 3DUniversum, spinoffs of the University of Amsterdam. His main research interests are in the fundamentals of image understanding, 3-D object recognition, human-behavior analysis and color in computer vision.

Prof. dr. Gevers is a Co-chair for various conferences including the European Conference on Computer Vision 2016. He is a Program Committee Member for a number of conferences and an Invited Speaker at major conferences. He has given lectures at various major conferences.