Group Happiness Assessment Using Geometric Features and Dataset Balancing

Vassilios Vonikakis[1], Yasin Yazici[1,2], Viet Dung Nguyen[1], Stefan Winkler[1]
[1]Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore
[2]School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore
{bbonik, yasin.y, vietdung.n, stefan.winkler}@adsc.com.sg

ABSTRACT

This paper presents the techniques employed in our team's submissions to the 2016 Emotion Recognition in the Wild contest, for the sub-challenge of group-level emotion recognition. The objective of this sub-challenge is to estimate the happiness intensity of groups of people in consumer photos. We follow a predominately bottom-up approach, in which the individual happiness level of each face is estimated separately. The proposed technique is based on geometric features derived from 49 facial points. These features are used to train a model on a subset of the HAPPEI dataset, balanced across expression and headpose, using Partial Least Squares regression. The trained model exhibits competitive performance for a range of non-frontal poses, while at the same time offering a semantic interpretation of the facial distances that may contribute positively or negatively to group-level happiness. Various techniques are explored in combining these estimations in order to perform group-level prediction, including the distribution of expressions, significance of a face relative to the whole group, and mean estimation. Our best submission achieves an RMSE of 0.8316 on the competition test set, which compares favorably to the RMSE of 1.30 of the baseline.

CCS Concepts

•**Computing methodologies** → *Activity recognition and understanding;*

Keywords

Emotion Estimation, Group Happiness, Facial Expression Analysis, Partial Least Squares, Geometric Features, Dataset Balancing

1. INTRODUCTION

Facial expression analysis (also known as emotion estimation, or analysis of facial affect) has attracted significant attention in the computer vision community during the past decade, since it lies at the intersection of many important applications, such as human computer interaction, surveillance, crowd analytics etc.

The majority of existing approaches focus on estimating emotions for an individual face. As such, they usually attempt to classify 7 prototypical expressions, which have been found to be universal across cultures and subgroups. A very detailed and recent review of these techniques can be found in [18].

Recently, there is interest in estimating the affective state of a group of people as a whole. This may have direct applications in personal photo collections [24], crowd affective analytics, or even security. The Emotion Recognition in the Wild (EmotiW) contest and its group-level emotion recognition sub-challenge focus on this problem. Images are selected from Flickr and are annotated for their overall happiness intensity level by human annotators. These annotations serve as ground truth, which contesting models attempt to predict.

One of the main challenges of this task is that images are captured "in the wild". As a result, images may include occlusions, non-frontal poses, non-uniform illumination, various face sizes, as well as a range of ages and ethnicities. The dataset comprises 2638 images for the combined Training (TR) and Validation (VA) datasets. This size makes it difficult to train large-scale models, which may be prone to overfitting.

It has been shown that the perception of the overall happiness intensity of a group is affected by both top-down and bottom-up components [2]. Top-down attributes include factors external to the individual members of the group, such as the scene or arrangement of people. Bottom-up attributes are relevant to each member of the group such as their facial expressions or facial attributes.

In this paper we approach group-level happiness predominately from a bottom-up perspective. We introduce a regression model for face-level happiness intensity with promising performance across various head poses. The model is based on geometric features estimated from 49 facial points and trained on a balanced subset of all the images of the HAPPEI dataset (TR+VA). The trained regression model offers at the same time a semantic interpretation of facial distances which may contribute positively and negatively to a happy expression. In addition to the face-level estimations, we also include some top-down features, such as sizes of faces and the distances between them. Various training methods are explored for learning group-level happiness, including simple feed-forward neural networks and linear models based on Partial Least Squares (PLS). Our best submission exhibits a RMSE of 0.8316 on the competition test set, which compares favorably to the RMSE of 1.30 of the baseline [3].

The rest of the paper is organized as follows. Section 2 describes previous works which are related to our study. Section 3 describes in detail the proposed approach for face-level happiness estimation. Section 4 discusses our approach for making group-level predictions based on face-level happiness estimates. Section 5 evaluates

the results of our submissions and discusses some interesting findings. Finally, concluding remarks are provided in Section 6.

2. RELATED WORK

The literature regarding face-level emotion estimation is extensive. A comprehensive review can be found in [18]. Most existing methods are based on *appearance* features, such as Local Binary Patterns [11], Histogram of Gradients (HoG) [13], Gabor features [19], or even raw pixel values in combination with deep learning methods [12]. Very few employ geometric features [21]. This is due to the fact that landmark detection adds additional overhead and – at least until recently – did not achieve robust performance. In recent years however, facial landmark detection and tracking has improved considerably. This has led to an increase in the number of methods preferring geometric over appearance-based features. For example, the baselines of the 2015 and 2016 AVEC challenges [16,20] employ geometric features for video-based emotion recognition. Moreover handcrafted geometric features have also been used on the EmotiW 2015 dataset [10].

Geometric features are derived from facial registration points. They may include the actual coordinates of these points [17] or pairwise distances between them [14]. A large part of the literature focuses on detecting and tracking facial registration points, usually referred to as "face alignment". An extensive review of this topic can be found in [15]. Our approach employs distance-based geometric features similar to [14].

Using face-level emotion estimations in order to infer the overall emotion of a group of people is a more complicated task. So far, the literature regarding group-level estimation of emotions is rather limited. One of the most important works is [2], whose authors performed a crowdsourcing study in order to estimate the main factors that contribute to the impression of happiness in a group photo. Based on these findings, they developed a model using both bottom-up and top-down characteristics in order to predict the happiness intensity of unseen images. Another work in this direction is [4], which addresses the problem of classifying positive/neutral/negative group-level emotions in images using both local and contextual features.

Contextual features may play an important role in predicting the characteristics of a group [2], as well as affecting face-level estimations. One of the first studies of this issue was [7]. Although this study did not focus on emotion, it demonstrated the importance of such approaches, since they reported a significant increase in performance for age and gender estimations.

3. FACE-LEVEL ESTIMATION

Estimating the happiness level of individual faces is the first step in assessing the overall happiness level of an image. The main challenge however is the limited size of training data. The HAPPEI dataset consists of 2638 images, which contain approximately 9400 faces in total. Although this might seem like an adequate training set, it is not enough to cover the broad range of variability that faces exhibit, such as differences in pose, occlusions, identity, gender, age, or illumination. As a result, training directly with face pixel values or texture features in such a limited-size dataset may result in overfitting, limiting the generalization ability of the system.

Our main objective was to create a system that is as *invariant* as possible to these factors of face variability. For this reason, we decided to base our system on *geometric features* rather than pixel values or texture. If facial points are estimated reliably, geometric features can capture the structure of the face, while being affected minimally by identity, gender, age, illumination, and minor occlusions like spectacles [15,18]. Instead of heuristically selecting specific geometric features, we follow a data driven approach, in which the most relevant geometric facial distances are discovered by means of learning from the HAPPEI dataset. This is discussed in the following.

3.1 Geometric Features

In all our submissions, we took the following approach to extract geometric features. Faces were detected using OpenCV's Viola-Jones frontal face detector [22], with a minimum face size of 25×25 pixels, and the Intraface library [26] in order to detect 49 facial points. Intraface was selected because it most accurate and robust facial point detection methods [15]. The score of the alignment model provided by Intraface was used to discard false positive faces; any detection with a score lower than 0.3 was considered a non-face. Additionally, the head-pose estimations of yaw and pitch provided by Intraface were also stored for later use (Section 3.2).

After detecting all the faces from the whole HAPPEI dataset (TR and VA), only the ones with annotations were kept. This resulted in approximately 9400 faces. Since faces are roughly symmetric along the horizontal axis, a mirror image of a face should have the same emotion annotation. By exploiting the facial symmetry and mirroring the detected faces, the dataset size can be nearly doubled (the final number of faces is not exactly double because Intraface did not converge in some of the mirrored faces). The final training set consisted of 18767 annotated faces.

Let $\mathbf{p} \in \mathbb{R}^{2N}$ be a vector containing the estimated x and y coordinates of N facial points (in our case $N = 49$). In order to account for different sizes of faces and to introduce scale invariance, the coordinates of the facial points should be normalized according to their scale. A straightforward approach is to normalize according to the eye-to-eye distance. However, this could be heavily affected by yaw; if a face is not frontal, the eye-to-eye distance becomes smaller, thus affecting the scale. For this reason we use the root mean squared distance from the average of the points of a face. We consider a subset of only 13 *less-deformable* facial points (points that change minimally with expression). These include the eyes, the nose, and the middle of the upper lip. Eyebrows and the majority of the mouth points are excluded, since their coordinates may change significantly with different expressions, thus, affecting the scale. Let $\mathbf{p}_s \in \mathbb{R}^{2n}$ be a vector containing the coordinates of the 13 less-deformable points ($n = 13$) of a face. Then the scale S is estimated as follows:

$$S = \frac{\|\mathbf{p}_s - \bar{\mathbf{p}}_s\|_2}{2\sqrt{n}}, \tag{1}$$

where $\bar{\mathbf{p}}_s$ is the vector containing the mean coordinates among the set of the less-deformable points of the face. $\hat{\mathbf{p}}$ with the scale-invariant coordinates can now be estimated as follows:

$$\hat{\mathbf{p}} = \frac{1}{S}\mathbf{p} \tag{2}$$

The scale-invariant points $\hat{\mathbf{p}}$ are used to extract the facial geometric features. We define these features as "all possible combinations of Euclidean distances among $\hat{\mathbf{p}}$". As such, there can be $\binom{N}{2}$ distances (in our case 1176) for each face. Although the dimensionality of these features is high, they offer the advantage of being invariant to roll, as well as allowing for a direct *semantic interpretation* of facial deformations (section 5.2). Additionally, their high dimensionality can be addressed by using a training approach which incorporates dimensionality reduction (section 3.3). All our submissions utilized this type of geometric features for the face-level estimation of happiness intensity. Fig. 1 depicts the flowchart of all our 8 submissions, as well as their performance for different dataset splits.

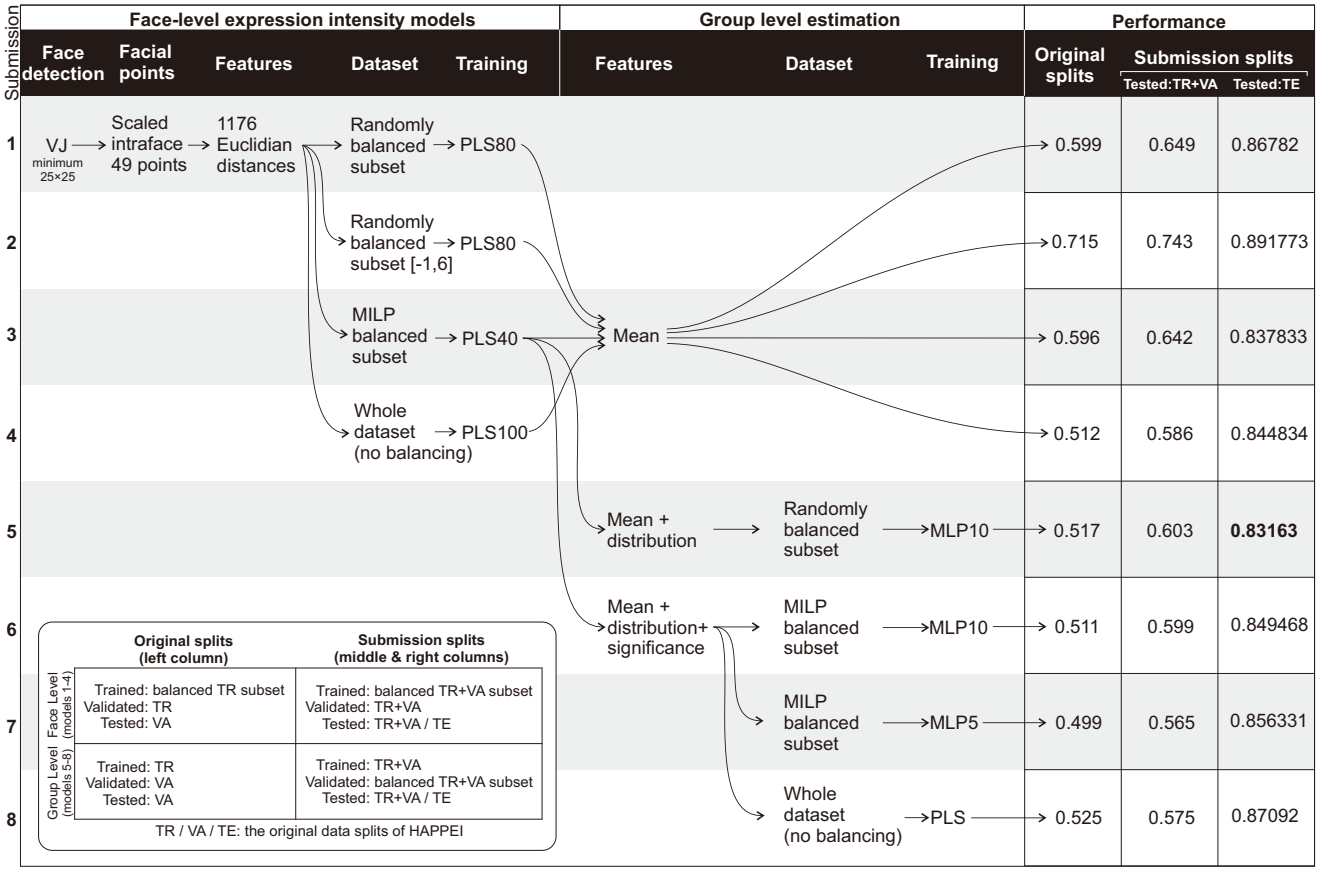

Figure 1: Flowchart of the different approaches used in our submissions.

3.2 Balancing the Data

The left column of Fig. 2 depicts the distribution of the full HAPPEI dataset (18767 faces from TR+VA) for 3 different attributes: intensity of happiness, yaw, and pitch. It is evident that the dataset is significantly skewed towards frontal faces of intensity 3. Most of the other happiness intensities or head poses are under-represented. Training directly with such skewed data may compromise the generalization ability of the system.

For this reason, we leveraged on the existing redundancies of the dataset in order to create a more *balanced* subset, which we will subsequently use as a training set for face-level happiness estimation. More specifically, we aimed for a uniform training distribution, which would include all the training examples of the most under-represented intensity quantization bin: those annotated as 0 (neutral) and those annotated as 5 (thrilled). We opted to have 1432 training examples per intensity bin, which is the number of faces annotated as 5. This results in a training subset with a total of $1432 \times 6 = 8592$ training faces.

Two different techniques for creating a balanced subset were explored. The first was to "trim" the intensity quantization bins which were over-represented by randomly selecting only 1432 training examples. Submissions 1 and 2 were based on this technique, with the latter including a change in the interval of labels from [0,5] to [-1,6]. Although simple, this balancing approach takes into consideration only happiness intensity and does not consider headpose. As such, the resulting randomly balanced subset may be significantly skewed in terms of headpose.

Creating a balanced subset which enforces a uniform distribution in happiness intensity as well as yaw and pitch is a challenging combinatorial problem. For this reason we employed a method we presented in [23].[1] This approach utilizes Mixed Integer Linear Programming (MILP) in order to enforce a specific distribution across different dimensions when selecting a subset from a larger dataset. Yaw and pitch were normalized to standard scores, truncated to the interval $[-3\sigma, 3\sigma]$ and quantized into 6 bins in order to match the quantization of happiness intensity. The quantized data were used in the MILP optimization, which found the optimal combination of training examples that would result to a distribution as close to uniform as possible, across happiness, yaw and pitch.

The right column of Fig. 2 depicts the distributions of the resulting MILP subset, all of which are closer to uniform, as much as the redundancy of the original dataset allows. The yaw and pitch dimensions are significantly under-represented for angles greater than $\pm 20^\circ$; there are simply not enough training examples to form a perfectly uniform distribution. Notice however that *all* the under-represented training examples both for intensity (0 and 5), as well as for yaw and pitch ($\leq -20^\circ$ and $\geq 20^\circ$) are *simultaneously* included in the balanced subset; the 'before' and 'after' distributions are identical in their outer regions. This would be very difficult to achieve with a simple random sub-sampling, and ensures that the resulting subset includes the maximum available information across all 3 dimensions. Submission 3 and submissions 5-8 are

[1] Code is available at https://sites.google.com/site/vonikakis/software-code/dataset_shaping.

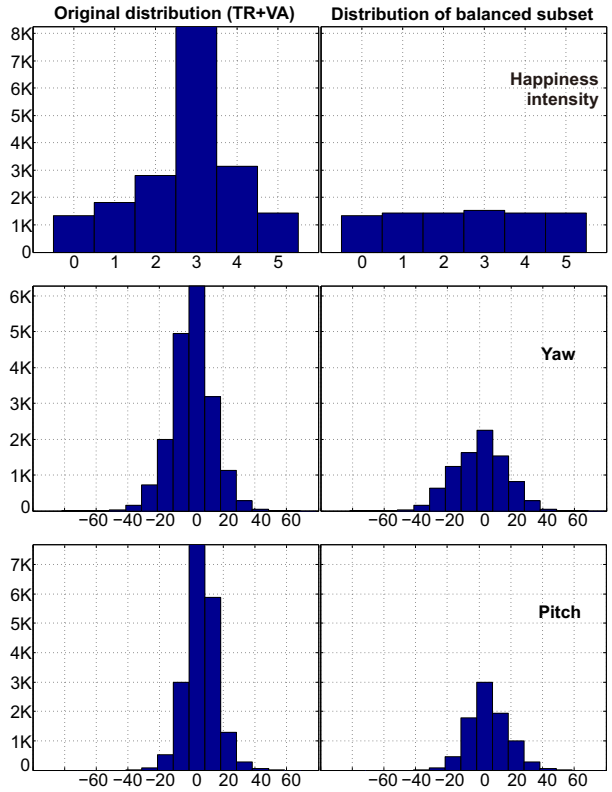

Figure 2: Left column: Distributions of the 18767 detected faces from the full HAPPEI dataset (TR+VA) for happiness intensity, yaw, and pitch. Right column: distributions of a more balanced subset of the HAPPEI dataset, as generated by the method in [23].

based on this approach. Finally, directly training with all 18767 HAPPEI faces was also tested (submission 4).

3.3 Training

The high dimensionality of the geometric feature vector can potentially be a limiting factor in the attempt to create a predictive system. The fact that all possible distances between the 49 facial points are considered, indicates that many among the 1176 distances may be very similar and even highly correlated, since they derive from neighboring facial points. This raises the danger of *multicolinearity* for typical regression-based systems. However, it is reasonable to assume that the observed data (displacement of facial points / change in distances) is generated by a system or process which is driven by a smaller number of not directly observed or measured variables. Such latent variables could be the facial muscles or Action Units described by the FACS [5].

This makes the use of Partial Least Squares (PLS) regression very appealing for this case, since it is particularly useful for predicting a response variable from a large set of highly correlated predictors, while at the same time making use of their common structure. More specifically, PLS projects the predictors to a set of orthogonal latent vectors, or *components*, which have the best predictive power to approximate the response variable. In essence, it combines characteristics of both Principal Component Analysis (PCA) (maximum variance of inputs) and Ordinary Least Squares (OLS) (maximum input-output correlation), by maximizing the covariance between the response and predictor variables. As such, it performs dimensionality reduction and prediction in a single step.

Assuming a set of predictor variables in the form of a matrix $\mathbf{X}$ (rows corresponding to observations) and a set of response variables $\mathbf{Y}$, the PLS framework decomposes them into the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \mathbf{E},$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^\top + \mathbf{F},$$

where $\mathbf{T}$ and $\mathbf{U}$ are matrices containing the extracted latent components, $\mathbf{P}$ and $\mathbf{Q}$ represent the loadings, and $\mathbf{E}$ and $\mathbf{F}$ the residuals. The PLS algorithm finds the weight vectors $\mathbf{w}$ and $\mathbf{v}$ by optimizing the following objective function to maximize the covariance between the latent components of the predictor and the response variables:

$$[cov(\mathbf{t},\mathbf{u})]^2 = max_{|\mathbf{w}|=|\mathbf{v}|=1}\,[cov(\mathbf{X}\mathbf{w},\mathbf{Y}\mathbf{v})]^2,$$

where $\mathbf{t}$ and $\mathbf{u}$ are the column vectors of $\mathbf{T}$ and $\mathbf{U}$, respectively, and $cov(\mathbf{t},\mathbf{u})$ is the sample covariance. With the estimated latent components $\mathbf{T}$ and $\mathbf{U}$, the regression coefficients between $\mathbf{X}$ and $\mathbf{Y}$ can be estimated by:

$$\mathbf{B} = \mathbf{X}^\top\mathbf{U}\left(\mathbf{T}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\right)^{-1}\mathbf{T}^\top\mathbf{Y}.$$

As such, the predicted response can be estimated by a simple matrix multiplication $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}$. All our submissions used the Matlab implementation of PLS called *plsregress*, which is based on the SIMPLS algorithm [1].

The number of latent components plays an important role in the success of the model and may act as a kind of regularization. Deciding on the number of components is very important. A large number of components will do a good job in fitting the current observed data, but will result in overfitting and thus poor generalization ability. For this reason we studied the impact of different numbers of latent components on the overall performance of 3 different approaches: MILP balancing, random balancing, and no balancing. Fig. 3 depicts the results.

It is clear that the MILP balancing technique achieves approximately the same minimum RMSE and percentage of explained variance with *half* the number of components compared to random balancing (40 vs. 80). This is an indication of its better generalization ability. As such, all the subsequent submissions (5-8) were based on this approach. It should be noted that the performance of the random balancing technique fluctuates according to the randomly selected subset (theoretically, although highly unlikely in practice, it could be the same as the MILP approach if *exactly* the same training examples were randomly selected). In all our experiments, the minimum RMSE for the random balancing was achieved with 60-100 components.

It is also evident that using the whole available dataset without balancing results in severe overfitting. RMSE drops monotonically with increased number of components, but the percentage of explained variance remains very low compared to the balancing approaches. This essentially means that the system overfits on the most over-represented happiness classes (2 and 3) and does not learn to recognize the under-represented cases (0 and 5). This highlights the importance of balancing: training with a more balanced subset (with less than half the number of training samples) results in better generalization characteristics.

Fig. 4 depicts the result of the proposed face-level happiness estimation algorithm, trained on the MILP balanced subset, for some faces from the Gallagher dataset [6]. It is evident that the algorithm can estimate a broad range of happiness expressions, even though in some of the cases the headpose is far from frontal. More

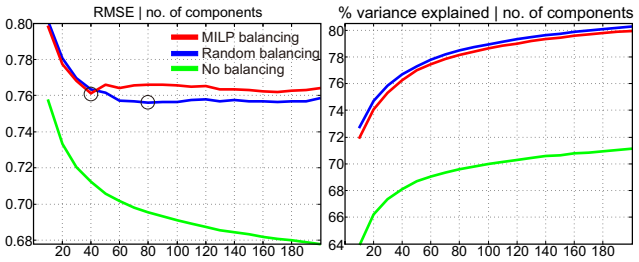

Figure 3: Impact of the number of PLS components on the RMSE and the % explained variance, for different balancing techniques. The MILP balancing approach achieves approximately the same minimum RMSE with half the number of components compared to random balancing (40 vs. 80).

importantly though, the ordinal relation between expressions is preserved, i.e. a face that appears less happy than another will have a lower score.

4. GROUP-LEVEL ESTIMATION

Having obtained face-level estimations of happiness intensity for each individual face in an image, we now combine them in order to get an estimation of the overall happiness of the image. In this section we discuss the features and training strategies we used in our submissions for this purpose.

4.1 Mean of Face-level Estimations

Using the mean of face-level estimations as a predictor for the overall group-level happiness is the simplest approach one can use. Undoubtedly, group-level happiness is more complicated than a simple mean of local happiness estimations, as discussed in [2]. However, averaging the provided face happiness annotations for each image gave a $RMSE_{TR} = 0.63$ for the TR set, a $RMSE_{VA} = 0.59$ for the VA dataset and a $RMSE_{TR+VA} = 0.61$ for the combined TR and VA datasets, which is considerably lower than the baseline of 0.78 for VA [3]. This prompted us not to dismiss this approach. As such, our submissions 1 to 4 use the mean $\bar{h}$ of our face-level happiness predictions to estimate the group-level happiness.

$$\bar{h} = \frac{\|\mathbf{h}\|_1}{k}, \tag{3}$$

where $\mathbf{h}$ is a vector containing all the face-level happiness estimations, and k is the total number of detected faces.

4.2 Happiness Distribution

Our objective is to follow a data-driven approach for using face-level predictions in order to estimate group-level happiness. Simply combining all face-level predictions in a feature vector is not ideal, since the number of people per image varies and thus, the vector length will change as well. For this reason the distribution of our face-level predictions was used; all face-level estimations of an image were rounded to the nearest integer and combined into a 6-bin histogram, ranging from 0 to 5. The histogram was then normalized with the total number of detected faces k, resulting in the following happiness distribution d.

$$d(i) = \frac{\sum_{j=1}^{k} \delta\left(i - \text{round}[h_j]\right)}{k}, \tag{4}$$

where δ is the Kronecker delta function, $i \in [0,5]$, h_j is the face-level happiness estimation for face j, and round$[\cdot]$ a function that rounds its argument to the nearest integer. This distribution was combined with the mean $\bar{h}$ of all face-level estimations in order to form a 7-element feature vector, which was then used for training.

A two-layer feed-forward neural network was used in order to learn the mapping from the 7-element feature vector to the given image annotations. It comprised 10 hidden sigmoid neurons and a single *linear* output neuron. Training was performed with the Levenberg-Marquardt backpropagation algorithm using Matlab's *nftool* and a MSE metric.

As with the individual faces, the image annotations in HAPPEI for the group-level estimations are highly unbalanced, with very few cases of high ($\approx 1.4\%$ with 5) and low ($\approx 3.5\%$ with 0) intensity. Fig. 5 depicts the distributions of HAPPEI for group-level happiness. This motivates us to use the combined TR+VA datasets in order to train our models for the final submissions, since we cannot afford to split the already very few training examples for labels 0 and 5.

Following the same approach as before, a randomly balanced subset of 498 images was created out of the combined TR+VA datasets. The subset had an approximately uniform distribution and included *all* the training annotations of low (0) and high (5) intensity images, with only a random subset of the over-represented classes 1 through 4. The network was trained in the combined TR+VA dataset and validated in the balanced subset. Our submission 5 was based on this training scheme, exhibiting the best performance in the test set. Fig. 6 depicts the error histogram, as well as the validation and training errors for each epoch.

4.3 Face Significance

In submissions 1-5, all detected faces contributed equally to the group-level estimation. However it has been shown that this is not generally the case [2]. Larger faces and ones that are closer to the group seem to have a greater impact on the overall mood of an image. Therefore we estimate the significance s_i of a face i using the following equation:

$$s_i = \frac{b_i}{\sum_{j=1}^{k} \left\|\mathbf{c}_i - \mathbf{c}_j\right\|_2}, \tag{5}$$

where b_i is the size of the bounding box of face i (in pixels), $\mathbf{c}_i$ is a vector containing the x and y coordinates of its bounding box center, and k the total number of detected faces in the image. Equation (5) essentially normalizes the size of a face by the sum of its Euclidean distances with all other faces. As a result, small faces which are located away from the group are penalized, while larger faces which are closer to all others are assigned a higher significance. If $k = 1$, then significance is set to 1.

The significance s of each face is used in order to adjust its contribution in the group mean and in the happiness distribution. More specifically, the estimation of the mean now becomes a weighted average $\tilde{h}$ based on the face-level estimations and the corresponding significance of each face, as follows:

$$\tilde{h} = \frac{\mathbf{h}^\top \mathbf{s}}{\|\mathbf{s}\|_1}, \tag{6}$$

where $\mathbf{h}$ and $\mathbf{s}$ are vectors containing all the face-level happiness estimations and the significance of each face, respectively. Using directly $\tilde{h}$ as an estimator for the group-level happiness results in $RMSE_{TR+VA} = 0.6408$, which is a slight improvement compared to $\bar{h}$ with an $RMSE_{TR+VA} = 0.6422$. Similarly, equation (4) is changed accordingly in order to utilize face significance in the happiness distribution:

$$\tilde{d}(i) = \frac{\sum_{j=1}^{k} \delta\left(i - \text{round}[h_j]\right) s_j}{\|\mathbf{s}\|_1}. \tag{7}$$

Figure 4: Results of the proposed face-level happiness estimation algorithm on unseen data from the Gallagher dataset [6].

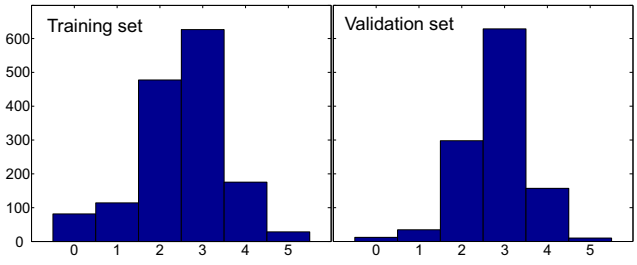

Figure 5: Distributions of group-level happiness intensity for the Training and Validation sets.

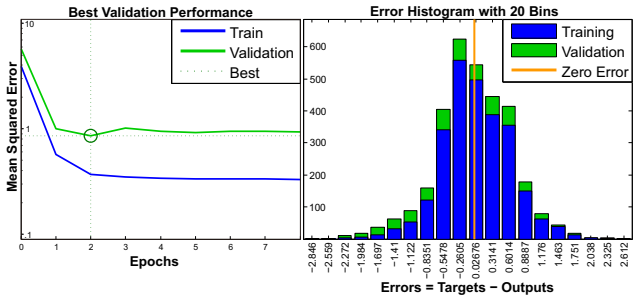

Figure 6: Best validation performance and error histogram for submission 5.

As before, $\tilde{d}$ is combined with $\tilde{h}$ to form a 7-element feature vector, which is used in submissions 6-8.

Submissions 6 and 7 follow the same approach as submission 5; training in the whole dataset and validating on a balanced subset. The difference is that the balanced subset was created with the MILP approach of [23]. More specifically, the balancing method was used in order to select a subset (out of the combined TR+VA datasets) with uniform distributions across 3 attributes: overall image happiness, average size of faces, and number of detected faces. The network characteristics were identical with submission 5, with the exception that submission 7 was based on a network with only 5 hidden neurons.

Finally, submission 8 used exactly the same features as submissions 6 and 7, but instead of a neural network, the PLS method with 6 latent components was used. The purpose of this was to explore the performance of a simpler linear model in addition to the non-linear neural networks which are more prone to overfitting.

5. DISCUSSION

Fig. 1 depicts the flowchart and the performance of all our approaches. None of our submissions are based on the provided TR and VA splits; instead we created our own in order to make better use of the available data, especially for the under-represented classes of 0 (neutral) and 5 (thrilled). The performance of our submitted models is depicted in the middle and right columns. However, to facilitate comparisons with other methods, we also include the performance of our submissions when trained according to the provided TR and VA splits in the left column. The legend on the lower left part of Fig. 1 provides details regarding the data splits.

5.1 Model Evaluation

One immediate observation from Fig. 1 is that when following the provided TR and VA splits, all models exhibit a lower RMSE by at least 0.05. This may be due to the fact that the provided TR and VA splits have a very similar distribution, as seen in Fig. 5. However, since the distribution of the test (TE) set may be different compared to TR and VA, this may become a limiting factor, which – in conjunction with the under-represented classes of 0 and 5 – may result in lower generalization characteristics.

Another insight from Fig. 1 is that the proposed regression-based face-level estimation of happiness is promising. The geometric features, although not popular in the facial expression analysis community, seem to estimate different intensities of happiness reliably, when the detection of facial points is accurate.

Having a diverse dataset with many different training examples across different headposes contributes to an increased performance. Even more, a balanced dataset will result to a more stable estimation across different headposes. Fig. 7 depicts the results of the proposed algorithm for different viewpoints of the same expression, for a face taken from the Multi-PIE dataset [8]. It can be seen that yaw has a moderate impact on the estimation of happiness intensity. Within $\pm 15^{\circ}$ from frontal, happiness may fluctuate up to 10%, whereas for higher deviations like $\pm 45^{\circ}$, happiness may change up to 17%. The combination of higher yaw *and* pitch values seems to have a greater impact on the happiness estimation exhibiting a difference of approximately 30% compared to frontal. However, such headpose combinations are unlikely to be encountered in personal photo-collections, where most of the photos feature frontal faces. The observed discrepancy between frontal and higher yaw/pitch values may be attributed to the lack of training examples for these cases. It is expected that a diverse training dataset with enough examples for non-frontal faces, in combination with the proposed training approach, would result in more headpose-invariant happiness estimations.

The dataset balancing strategy also exhibited promising results, at least for the face-level model. Training with a subset of 8592 faces (46% of the whole dataset), balanced across 3 different attributes (happiness intensity, yaw, pitch), proved to be better in terms of generalization capability, compared to training directly with 18767 faces. This indicates that more data is not necessarily better. When a training dataset is highly skewed and acquiring additional data is not possible, it may be preferable to compile a balanced subset, with as uniform a distribution as possible across all important dimensions. For this purpose, the MILP balancing approach appears to result in models that can generalize better with fewer latent components, compared to compiling a training subset randomly.

The simple averaging of face-level estimations also exhibited unexpectedly competitive results. Actually, submission 3, which is

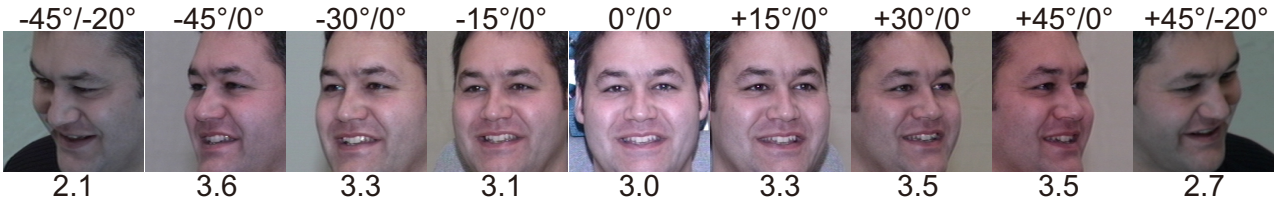

Figure 7: Results of the proposed face-level happiness estimation algorithm, on different yaw/pitch combinations of the same unseen expression (taken from the Multi-PIE dataset [8]).

based on a simple mean of the estimated face happiness intensities, exhibited the second best performance in the test set. This may imply that bottom-up techniques may be more competitive than top-down approaches for group-level happiness estimation.

The significance factor of each face, which takes into consideration face size and average distance with all the other faces, exhibited a modest performance improvement. Although the submissions in which it was included did not perform very well, this could be due to the training strategy rather than the type of feature. In fact, when compared to the simple mean, the addition of the face significance factor improves the overall RMSE in the combined TR and VA dataset from 0.6422 to 0.6408. Since no training method is involved in the use of a simple averaging, this improvement may not be the result of overfitting, but may have to do with the efficiency of the feature.

The inclusion of the happiness distribution feature did not seem to improve the results as initially expected. Although the best submission (5) included this feature, the improvement was marginal compared to the simple mean of submission 3, and it could easily be due to the training method used.

Finally, the training strategy used in the neural networks of submissions 5-7 indicates that overfitting was an issue. The performance for the combination of TR+VA increased, but performance in the test set was reduced. The non-balanced dataset, as well as the type of features used could be additional limiting factors here.

5.2 Feature Interpretation

The use of distances as geometric features may offer a direct semantic interpretation of facial deformations. The set of learned regression weights contained in matrix **B** gives a strong indication on the contribution of each facial distance to the overall estimation of a face's happiness. A similar approach has also been recently used for feature ranking/selection [9]. Fig. 8 depicts the top 100 distances with the highest positive and negative contributions to the estimation of happiness. Positive contributions represent a proportional relation, i.e. happiness increases with distance, and vice-versa for negative contributions.

The facial distances with the stronger positive impact to happiness are located in the mouth, connecting outer points of the upper part of the lower lip with the outer points of the upper part of the upper lip (points 49-33 and 47-37 in the Intraface numbering convention). This intuitively makes sense, since these 2 distances can capture both the opening movement of the mouth and the extension of the corners that occurs during smiling. The same also accounts for the second strongest positive distances (Intraface points 48-44 and 48-46), which form a V shape in the middle of the inner part of the mouth and can also capture these two motions. Many other positive distances capture the eccentricity of the mouth in relation to a stable point, e.g. the tip of the nose. Finally, the inner points of the eyebrows seem to play also an important positive role, both

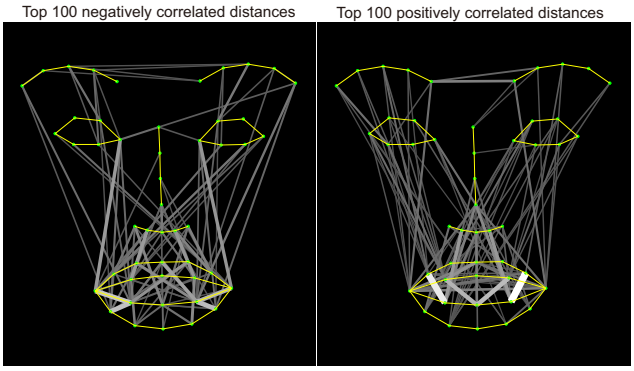

Figure 8: Top 100 distances that contribute negatively and positively to the model's estimation of happiness. Thicker and brighter lines indicate higher contribution.

in relation to each other as well as to the corners of the mouth, forming a Π shape.

The facial distances with the stronger negative impact to happiness are related mainly with the corners of the mouth. Interestingly, the distance of the mouth corners in relation to a distant facial point seems to affect the formation of happy/excited expressions both in a positive and negative way. However, the 'distant' points are different in each case. For the positive case, these are the inner parts of the eyebrows. This means that when the inner eyebrows as raised and the corners are extended, the mouth-eyebrow distance increases, contributing positively to a happy impression. On the other hand, the 'distant' points for the negative case are the inner corners of the eyes. These points are among the most stable ones, since they do not move during various facial expressions. As such, the only way that the eye-mouth distance can increase is by extending the corners of the mouth downwards, which is associated with sad expressions. Consequently, when distances between Intraface points 23-32 and 26-38 are increased, the impression of a face's happiness decreases.

6. CONCLUSIONS

This paper presented a series of techniques employed for the group-level emotion recognition sub-challenge of the EmotiW 2016 contest. Our approach is predominately based on bottom-up elements, in which the individual happiness level of each face is estimated separately. The proposed uses geometric features derived from 49 facial points. These features are used in order to train a model on a carefully compiled dataset using Partial Least Squares regression. The training dataset is a subset of the whole HAPPEI dataset, which we carefully balanced across emotion, yaw, and pitch. In addition to the face-level estimations, some top-down features are also employed. Various techniques are explored in

combining these estimations in order to achieve a good group-level prediction. Our best submission exhibits an RMSE of 0.8316 on the competition test set, which compares favorably to the RMSE of 1.30 of the baseline [3]. Our future work includes the addition of more complex top-down scene-related features, such as the provided CENTRIST descriptor [25], which may further improve the performance of the method.

7. ACKNOWLEDGMENTS

This study is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

8. REFERENCES

[1] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251 – 263, 1993.

[2] A. Dhall, R. Goecke, and T. Gedeon. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26, Jan 2015.

[3] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. EmotiW 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th International Conference on Multimodal Interaction*, ICMI '16. ACM, 2016.

[4] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe. The more the merrier: Analysing the affect of a group of people in images. In *Proceedings of Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8, May 2015.

[5] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978.

[6] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Proceedings of Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[7] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *Proceedings of Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 256–263, June 2009.

[8] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807 – 813, 2010. Best of Automatic Face and Gesture Recognition 2008.

[9] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller. CCA based feature selection with application to continuous depression recognition from acoustic speech features. In *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3729–3733, May 2014.

[10] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah. Contrasting and combining least squares based learners for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 459–466, New York, NY, USA, 2015. ACM.

[11] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541 – 558, 2011.

[12] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 443–449, New York, NY, USA, 2015. ACM.

[13] C. Orrite, A. Gañán, and G. Rogez. *HOG-Based Decision Tree for Facial Expression Classification*, pages 176–183. Springer, Berlin, Heidelberg, 2009.

[14] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63:104 – 116, 2015.

[15] C. Qu, H. Gao, E. Monari, J. Beyerer, and J. P. Thiran. Towards robust cascaded regression for face alignment in the wild. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–9, June 2015.

[16] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, pages 3–8, New York, NY, USA, 2015. ACM.

[17] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):944–958, May 2015.

[18] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, June 2015.

[19] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):258–273, Feb 2010.

[20] M. F. Valstar, J. Gratch, B. W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. *CoRR*, abs/1605.01600, 2016.

[21] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, ICMI '07, pages 38–45, New York, NY, USA, 2007. ACM.

[22] P. Viola and M. J. Jones. Robust real-time face detection. *Int'l Journal of Computer Vision*, 57(2):137–154, 2004.

[23] V. Vonikakis, S. Ramanathan, and S. Winkler. Shaping datasets: Optimal data selection for specific target distributions across dimensions. In *Proceedings of 2016 IEEE International Conference on Image Processing (ICIP)*, September 2016.

[24] V. Vonikakis and S. Winkler. Emotion-based sequence of family photos. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1371–1372, New York, NY, USA, 2012. ACM.

[25] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1489–1501, Aug 2011.

[26] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539, June 2013.