# Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild

Heysem Kaya[*]
Department of Computer Engineering
Boğaziçi University
34342, İstanbul, Turkey
heysem@boun.edu.tr

Furkan Gürpınar
Department of Computational Science and Engineering
Boğaziçi University
34342, İstanbul, Turkey
gurpinarfurkan@gmail.com

Sadaf Afshar
Department ofComputational Science and Engineering
Boğaziçi University
34342, İstanbul, Turkey
sa.afshar.sa@gmail.com

Albert Ali Salah
Department of Computer Engineering
Boğaziçi University
34342, İstanbul, Turkey
salah@boun.edu.tr

## ABSTRACT

This paper presents our contribution to ACM ICMI 2015 Emotion Recognition in the Wild Challenge (EmotiW 2015). We participate in both static facial expression (SFEW) and audio-visual emotion recognition challenges. In both challenges, we use a set of visual descriptors and their early and late fusion schemes. For AFEW, we also exploit a set of popularly used spatio-temporal modeling alternatives and carry out multi-modal fusion. For classification, we employ two least squares regression based learners that are shown to be fast and accurate on former EmotiW Challenge corpora. Specifically, we use Partial Least Squares Regression (PLS) and Kernel Extreme Learning Machines (ELM), which is closely related to Kernel Regularized Least Squares. We use a General Procrustes Analysis (GPA) based alignment for face registration. By employing different alignments, descriptor types, video modeling strategies and classifiers, we diversify learners to improve the final fusion performance. Test set accuracies reached in both challenges are relatively 25% above the respective baselines.

## Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition-Signal processing; I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

## General Terms

Human-Computer Interaction

## Keywords

audio-visual emotion corpus, audio-visual fusion, feature extraction, emotion recognition in the wild, SFEW, AFEW

## 1. INTRODUCTION

Audio and video based emotion recognition in the wild is challenging, because of noise, large idiosyncratic variance and sensor-related differences. Fixed-protocol challenges in this field provide a unique opportunity to push forward the state-of-the-art and to compare many approaches under very similar conditions. The Emotion Recognition in the Wild (EmotiW) challenge provides out of laboratory data -Acted Facial Expressions in the Wild (AFEW)-, collected from videos that mimic real life [4, 3, 5]. In 2015, the EmotiW campaign introduced a static facial expression challenge based on in-the-wild images collected from videos [6]. In this paper we propose several systems based on combinations of learners for both static and video-based emotion recognition, and report results with the standard challenge protocols.

Our contributions to the EmotiW Challenge are manifold: i) We employ a General Procrustes Analysis (GPA) based alignment method to have improved face registration, ii) we extract and combine a set of visual descriptors such as Scale Invariant Feature Transform (SIFT) [17], Histogram of Oriented Gradients (HOG) [2], Local Phase Quantization (LPQ) [11, 13], Local Binary Patterns (LBP) [18] and its Gabor extension (LGBP), as well as hand-crafted geometric features computed from fitted landmarks of GPA alignment, iii) we use the popular Three Orthogonal Planes (TOP), summarizing functionals (FUN) and Fisher Vector encoding [19] (FV) on low level descriptors for video modeling, iv) we contrast feature and score level fusion strategies using PLS and Kernel ELM classifiers.

The remainder of this paper is organized as follows. In the next section we provide background on the signal pro-

[*]Corresponding author

*ICMI '15,* November 09-13 2015, Seattle, WA, USA

ACM 978-1-4503-3983-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2823327.2823334.

cessing and machine learning methods used in the study. In Section 3 we briefly introduce the corpora and baseline feature sets. In Sections 4 and 5, we give experimental results for the static and video based challenges, respectively. Finally, Section 6 concludes.

## 2. BACKGROUND AND METHODOLOGY

Both challenge tasks require detection and registration on facial images. In this section, we provide brief background on the components of our proposed method.

### 2.1 Facial Registration

Prior to visual feature extraction, we analyzed the set of aligned faces in the challenge dataset subjectively and observed that the images were well-aligned for frontal faces, but the alignment was generally poor in the case of rotations. Moreover, the challenge dataset, which contains face detection annotations that were generated automatically, contains a number of false positives that may effect the subsequent processing adversely. Subsequently, we have decided to i) preprocess the annotated challenge set to remove false positives (purification), ii) to employ an alternative generalized Procrustes analysis (GPA) based alignment on the original images and videos. Procrustes analysis takes a set of landmark points, encoded by geometric coordinates (either in 2D or in 3D), and computes a transformation to align them to a reference face [9].

For the automatically aligned challenge set, we manually remove poorly aligned images from the training set. For the validation set, we use a subspace method to automatically remove false detections. For the test set, all instances are retained.

For the video challenge, the images are resized to $64 \times 64$ pixels to cope with the curse of dimensionality. For the static expression recognition challenge, we obtained better results with $128 \times 128$ pixel images in our preliminary experiments, therefore used this setting in all our subsequent experiments.

On the validation set, we use principal component analysis (PCA) based data purification as shown to be effective in [16, 25, 15]. The idea is to measure the mean reconstruction error per image $x_i \in \mathbb{R}^D$ with $Err_i = \frac{1}{D}||(x_i - \mu) - W_{pca}^T W_{pca}(x_i - \mu)||$, where $\mu \in \mathbb{R}^D$ is the training set mean vector, and $W_{pca}$ is the reduced PCA projection coefficient matrix learned from the training set. We discard the frames with a high reconstruction error, as these are probably poorly detected or aligned images. In our study, we use the $L_1$ norm and remove the videos that have less than three valid images from the validation set.

Next, we briefly explain the GPA based alignment method used in the study.

#### 2.1.1 *Generalized Procrustes Analysis*

To obtain a good alignment and face representation, translation, rotation and scale effects should be filtered out. This objective is achieved by constructing a face reference model to which all the faces are aligned. For this we use the procedure that is known as the generalized Procrustes analysis [9]. For this approach, shapes are represented by sets of points, called landmarks. The following iterative approach is applied to obtain a face reference model and at the same time to align all the faces to a unique model:

1. Landmarks of the faces are extracted by a supervised descent based method [29].
2. All the landmarks are centered by subtracting the mean of the landmark coordinates across training samples.
3. An initial approximation of the mean shape, which consists of the centered coordinates of the landmarks, is computed.
4. In order to avoid any shrinking, the landmarks and the mean shape is scale-normalized.
5. All faces are aligned to the current mean shape.
6. The mean shape is re-calculated by using the current set of aligned faces.
7. If the Procrustes distance between the mean shape and the re-calculated mean shape is above a threshold, the mean shape is updated, and the procedure is repeated until convergence.

### 2.2 Visual Descriptors

We extract SIFT, HOG, LBP, LPQ, and LGBP as visual descriptors in both image and video emotion classification challenges. For LPQ, LBP, and LGBP, the TOP extension is popularly used in video modeling. This extension applies the relevant descriptor on $XY$, $XT$ and $YT$ planes (where $T$ represents time) independently and concatenates the resulting histograms. Also in our implementation, we divide the video into two equal length volumes over the time axis and extract spatio-temporal TOP features from each volume to further enhance temporal modeling. In the following, we provide brief explanation of LPQ, LBP and LGBP descriptors. We then introduce our hand-crafted geometric features extracted from landmarks provided by Xiong and De La Torre's face detection/tracking method [29].

#### 2.2.1 *Local Phase Quantization*

The LPQ features are computed by taking 2-D Discrete Fourier Transform (DFT) of M-by-M neighborhoods of each pixel in the gray scale image. 2D-DFT is computed at four frequencies $\{[a, 0]^T, [0, a]^T, [a, a]^T, [a, -a]^T\}$ with $a = 1/M$, which correspond to four of eight neighboring frequency bins centered at the pixel of interest. The real and imaginary parts of resulting four complex numbers are separately quantized using a threshold of zero, which gives an eight bit string. This string is then converted into an integer value in the range of [0-255]. The pixel based values are finally converted into a histogram of 256 bins. Since this histogram representation does not keep the structural information of facial features, the face is divided into non-overlapping regions and an LPQ histogram is computed per region and concatenated for the final image representation. Here, we use an LPQ version from [11].

#### 2.2.2 *Local Binary Patterns*

After face alignment and conversion to gray scale, LBP computation amounts to finding the sign of difference with respect to a central pixel in a neighborhood, transforming the binary pattern into an integer and finally converting the patterns into a histogram. Uniform LBP clusters 256 patterns into 59 bins, and takes into account occurrence statistics of common patterns [18]. As in LPQ, the face is divided into non overlapping regions and an LBP histogram is computed per region.

### 2.2.3 Local Gabor Binary Patterns

In LGBP, the images are convolved with a set of 2D complex Gabor filters to obtain Gabor-pictures, then LBP is applied to each Gabor-picture (or Gabor-video). A 2D complex Gabor filter is the convolution of a 2D sinusoid (carrier) having phase $P$, spatial frequencies $u_0$ and $v_0$ with a 2D Gaussian kernel (envelope) having amplitude $K$, orientation $\rho$, and spatial scales $a$ and $b$. For simplicity, and in line with [1], we take $a = b = \sigma, u_0 = v_0 = \phi$ and $K = 1$ to obtain:

$$G(x, y) = e^{-\pi\sigma^2((x-x_0)_\rho^2+(y-y_0)_\rho^2)} e^{j(2\pi\phi(x+y)+P)}, \quad (1)$$

where the subscript $\rho$ stands for a clockwise rotation operation around reference point $(x_0, y_0)$ such that:

$$\begin{aligned}(x - x_0)_\rho &= (x - x_0)cos\rho + (y - y_0)sin\rho \\ (y - y_0)_\rho &= -(x - x_0)sin\rho + (y - y_0)cos\rho.\end{aligned} \quad (2)$$

Note that the effect of the phase is canceled out, since only the magnitude response of the filter is used for the descriptor. In this work, we use an open source script for Gabor picture extraction [10] and implement our own LGBP-TOP routine.

### 2.2.4 Geometric Features

Geometric features represent the shape of certain landmarks of the facial image. We used the fitted landmarks provided by the GPA based alignment. The landmark indices are shown in Figure 1.

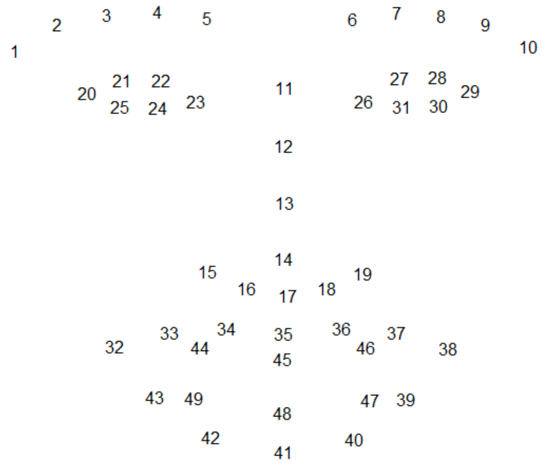

**Figure 1: Fiducial points**

Detailed explanation of the geometric features extracted from each image (for both challenges) is given in Table 1. As we shall see later on, this small set of hand-crafted geometric features reach the performance of appearance descriptors with much higher dimensionality.

In addition to frame level features, for the video challenge we also included Zero Crossing Rate (ZCR) of z-normalized coordinates of some landmarks over time. These geometric features are especially intended for the improved recognition of the "Disgust" class, whose expression sometimes includes a head oscillation in the yaw axis.

## 2.3 Video Modeling

In this study, we enhance diversity of learners by incorporating functionals on frame level features and by Fisher vector encoding of low level descriptors.

### 2.3.1 Functionals

In the state-of-the-art acoustic feature extraction pipeline, it is common to use a large set of summarizing functionals over the low level descriptor contours. In video modeling, on the other hand, few functionals such as mean and standard deviation are used. In this work, in addition to mean and standard deviation, we use three functionals based on polynomials, fit to each descriptor contour. The first is curvature, which is the leading coefficient of the second order polynomial. The other two are the slope and offset, computed from the first order polynomial, respectively. Therefore, when we have functional-based encoding, the video feature vector dimensionality is five times the frame-level dimensionality.

### 2.3.2 Fisher Vector Encoding

The Fisher vector (FV) provides a supra-frame encoding of the local descriptors, quantifying the gradient of the parameters of the background model with respect to the data. Given a probability model parametrized with $\theta$, the expected Fisher information matrix $F(\theta)$ is the expectation of the second derivative of the log likelihood with respect to $\theta$:

$$F(\theta) = -E[\frac{\partial^2 \log p(\mathcal{X}|\theta)}{\partial\theta^2}]. \quad (3)$$

The idea in FV in relation to $F(\theta)$ is taking the derivative of the model parameters and normalizing them with respect to the diagonal of $F(\theta)$ [19]. To make the computation feasible, a closed form approximation to the diagonal of $F(\theta)$ is proposed [19]. As a probability density model $p(\theta)$, Gaussian Mixture Models (GMM) with diagonal covariances are used. A K-component GMM is parametrized as $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$, where the parameters correspond to zeroth (mixture proportion), first (mean), and second order (covariance) statistics, respectively. It has been shown that using the zeroth order statistics is equivalent to the Bag of Words (BoW) model [24], however in FV, they were found to have a negligible effect on performance [19]. Therefore, only gradients of $\{\mu_k, \Sigma_k\}_{k=1}^K$ are used, giving a $2 \times d \times K$ dimensional super vector, where $d$ is the low level descriptor (LLD) dimensionality.

In order to efficiently learn a Background Probability Model (BPM) using GMM with diagonal covariances, we first need to decorrelate the data gathered from all instances. Principal Component Analysis (PCA) is applied on the data for this purpose. To reduce the computational cost, we take LLDs from every second frame to learn PCA and GMM. In our preliminary tests, this sub-sampling did not decrease the performance. Once the parameters of PCA projection and GMM are learned, we use all speech frames from each utterance without sub-sampling to represent them as a FV.

## 2.4 Model Learning

To learn a classification model, we employ kernel extreme learning machine (ELM) [12] and Partial Least Squares (PLS) regression due to their fast and accurate learning capability.

ELM proposes a multilayer perceptron architecture, but unsupervised, even random generation of the hidden node

**Table 1: Hand-crafted geometric features. LR indicates that the features are averaged over left and right parts of the face. Distance based features are always normalized with the face height. Features 18 through 23 are motivated by [22].**

| Feature # | Explanation | Landmarks Involved | Feature Type |
|---|---|---|---|
| 1 | Eye aspect ratio (LR) | [20:25], [26:31] | Distance |
| 2 | Mouth aspect ratio | 32, 35, 38, 41 | Distance |
| 3 | Upper lip angles (LR) | 32, 35, 38 | Angle |
| 4 | Nose tip - mouth corner angles (LR) | 17, 32, 38 | Angle |
| 5 | Lower lip angles (LR) | [32, 42] , [38, 40] | Angle |
| 6 | Eyebrow slope (LR) | [1, 5] , [6, 10] | Angle |
| 7,8 | Lower eye angles (LR) | [20, 23, 24, 25], [26, 29, 30, 31] | Angle |
| 9 | Mouth corner - mouth bottom angles | 32, 38, 41 | Angle |
| 10 | Upper mouth angles (LR) | [32, 34], [36, 38] | Angle |
| 11 | Curvature of lower-outer lips (LR) | [32, 43, 42], [38, 39, 40] | Curvature |
| 12 | Curvature of lower-inner lips (LR) | [32, 42, 41], [38, 40, 41] | Curvature |
| 13 | Bottom lip curvature | [32, 38, 41] | Curvature |
| 14 | Mouth opening / mouth width | 45, 48, 32, 38 | Distance |
| 15 | Mouth up/low | 35, 41, 45 | Distance |
| 16 | Eye - middle eyebrow distance (LR) | [3, 20, 23], [8, 26, 29] | Distance |
| 17 | Eye - inner eyebrow distance (LR) | [5, 20, 23], [6, 26, 29] | Distance |
| 18 | Inner eye - eyebrow center (LR) | [3, 23], [8, 26] | Distance |
| 19 | Inner eye - mouth top distance | 23, 26, 35 | Distance |
| 20 | Mouth width | 32, 38 | Distance |
| 21 | Mouth height | 35, 41 | Distance |
| 22 | Upper mouth height | 32, 38, 35 | Distance |
| 23 | Lower mouth height | 32, 38, 41 | Distance |

output matrix $\mathbf{H} \in \mathbb{R}^{N \times h}$, where $N$ and $h$ denote the number of instances and the hidden neurons, respectively. The actual learning takes place in the second layer between $\mathbf{H}$ and the label matrix $\mathbf{T} \in \mathbb{R}^{N \times L}$, where $L$ is the number of classes. $\mathbf{T}$ is composed of continuous annotations in case of regression, therefore is a vector. In the case of $L$-class classification, $\mathbf{T}$ is represented in one vs. all coding:

$$\mathbf{T}_{t,l} = \begin{cases} +1 & \text{if } y^t = l, \\ -1 & \text{if } y^t \neq l. \end{cases} \tag{4}$$

The second level weights $\beta \in \mathbb{R}^{h \times L}$ are learned by least squares solution to a set of linear equations $\mathbf{H}\beta = \mathbf{T}$. The output weights can be learned via:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{5}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse [20] that gives the minimum $L_2$ norm solution to $||\mathbf{H}\beta - \mathbf{T}||$, simultaneously minimizing the norm of $||\beta||$. To increase the robustness and generalization capability, the optimization problem of ELM is reformulated using a regularization coefficient on the residual error $||\mathbf{H}\beta - \mathbf{T}||$. The learning rule of this alternative ELM is related to Least Square SVMs (LSSVM) via the following output weight learning formulation:

$$\beta = \mathbf{H}^T(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{T}, \tag{6}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix, and $C$, which is used to regularize the linear kernel $\mathbf{H}\mathbf{H}^T$, corresponds to the complexity parameter of LSSVM [26]. This formulation is further simplified noting that the hidden layer matrix need not be generated explicitly given a kernel $\mathbf{K}$, which can be seen identical to Kernel Regularized Least Squares [12, 21]:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}. \tag{7}$$

PLS regression between two sets of variables $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{Y} \in \mathbb{R}^{N \times p}$ is based on decomposing the matrices as $\mathbf{X} = \mathbf{U}_x\mathbf{V}_x + r_x$, $\mathbf{Y} = \mathbf{U}_y\mathbf{V}_y + r_y$, where $\mathbf{U}$ denotes the latent factors, $\mathbf{V}$ denotes the loadings and $r$ stands for the residuals. The decomposition is done by finding projection weights $\mathbf{W}_x, \mathbf{W}_y$ that jointly maximize the covariance of corresponding columns of $\mathbf{U}_x = \mathbf{X}\mathbf{W}_x$ and $\mathbf{U}_y = \mathbf{Y}\mathbf{W}_y$. For further details of PLS regression, the reader is referred to [28]. PLS is applied to classification in one-versus-all setting between the feature matrix $\mathbf{X}$ and the binary label vector $\mathbf{Y}$, then the class giving the highest regression score is taken as prediction. The number of latent factors is a hyper-parameter to tune via cross-validation.

## 2.5 Fusion

We contrast two classical fusion schemes, namely early (feature level) and late (decision level) fusion. It has been shown that late fusion gives better results compared to feature level fusion. Furthermore, we apply weighted score fusion for each model and class, where optimal fusion weights are searched over a pool of randomly generated matrices. This scheme is shown to be successful on previous EmotiW Challenge corpora, considering different confusion matrices of each sub-system [14, 15].

We next briefly introduce the Challenge data [5], baseline features and the experimental protocol.

# 3. CORPORA AND EXPERIMENTAL PROTOCOL

In line with the previous two challenges, EmotiW 2015 presents video clips and images collected from movies representing close-to-real-world conditions [5, 7]. The challenge datasets are partitioned into training, development and test sets. The distribution of instances over partitions based on modality (audio, video) and alignment types is given in Table 2.

**Table 2: Instance distribution over partitions, alignment and modality types. GA: Given alignment, GPA: General Procrustes Analysis based alignment**

| **SFEW** | | | |
|---|---|---|---|
| # | Train | Val | Test |
| Images | 958 | 436 | 372 |
| GA | 891 | 427 | 372 |
| GPA | 724 | 343 | 290 |
| **AFEW** | | | |
| # | Train | Val | Test |
| Clips | 711 | 383 | 539 |
| Video-GA | 698 | 369 | 539 |
| Video-GPA | 663 | 340 | 455 |
| Audio | 707 | 383 | 539 |

The baseline video features consist of Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) compacted via uniform LBP [18]. For SFEW, feature level combination of Pyramid of HOG and LPQ is used in the baseline.

The baseline audio features are extracted via openSMILE tool [8] using INTERSPEECH 2010 Paralinguistic challenge baseline set [23]. We use the baseline acoustic features to model audio modality.

For comparability and reproducibility, we opt to use open-source tools for feature extraction. Particularly, VLFeat library [27] is used for both feature extraction (e.g. LBP, HOG), GMM training, and FV encoding.

In both challenges, we use three preprocessing alternatives: no-normalization (particularly for histogram-type descriptors that are normalized during the process), min-max normalization into [0,1] range, and z-normalization. Here, we report the best results over three alternatives.

The audio and video features are kernelized by means of Linear and Radial Basis Function (RBF) kernels, prior to model learning by classifiers. We optimize the model hyper-parameters on the challenge validation set. These are RBF kernel parameter $\gamma$, regularization parameter $\tau$ for ELM and number of latent components for PLS. Similarly, the number of PCA eigenvectors $p$ and GMM components $K$ used in the FV encoding are optimized using cross-validation.

The pipeline used in both challenges are very similar, save that we apply spatio-temporal modeling and combine the visual sub-systems with the audio sub-system in AFEW. The experimental results of each challenge are presented in the respective sections.

## 4. STATIC FACIAL EXPRESSION RECOGNITION CHALLENGE

The pipeline of the proposed multi-level score fusion system used in the SFEW Challenge is given in Figure 2. Since the images detected by different alignment methods do not overlap, we first optimize the fusion weights for each alignment individually then apply a second level fusion, where scores of missing images in an alignment are represented with a zero vector.

We extract 128 dimensional SIFT features from 32-by-32 pixel patches with 16 pixel shifts in both spatial directions. We extract two versions of HOG. One is extracted from the same overlapping patches as SIFT, where each patch is further divided into four non-overlapping sub-regions. This is denoted as HOG-OL. The other HOG version is obtained by dividing the image into $8 \times 8$ non-overlapping patches (HOG-NOL). LPQ and LBP are extracted from non-overlapping $6 \times 6$ and $8 \times 8$ patches, respectively. The validation set performances of extracted visual features from given alignment and GPA based alignment are given in Tables 3 and 4, respectively. On the overall, we observe higher performance of GPA based alignment compared to given aligned images. However, it is important to note that GPA detects face in 343 out of 436 images, whereas given alignment has 427 detected faces. Therefore, the improvement can partially be attributed to clean/less posed images that the GPA aligned systems cast their predictions.

**Table 3: Performance of various descriptors extracted from images with challenge alignment (GA).**

| Descriptor | ELM | PLS | Dimensions |
|---|---|---|---|
| HOG-NOL | 36.77 | **37.47** | 1984 |
| HOG-OL | **39.11** | **39.34** | 6076 |
| SIFT | **39.34** | 38.64 | 6272 |
| LBP | 36.30 | **38.64** | 3712 |
| LPQ | 34.43 | **38.17** | 9216 |

**Table 4: Performance of various descriptors extracted from GPA aligned images.**

| Descriptor | ELM | PLS | Dimensions |
|---|---|---|---|
| GEO | **46.06** | 43.44 | 23 |
| HOG-NOL | 46.94 | **48.98** | 1984 |
| HOG-OL | **48.40** | 46.06 | 6076 |
| SIFT | **46.06** | 43.15 | 6272 |
| LBP | 45.48 | **46.36** | 3712 |
| LPQ | 42.86 | **43.44** | 9216 |

In terms of feature types, we see that HOG variants give the best performance. SIFT and LBP follow the performance of HOG. An important observation in GPA alignment is that 23 dimensional geometric features have better performance than 9216 dimensional LPQ, and are on-par with 6272 dimensional SIFT features.

We next compare the feature and decision level fusion in pairwise manner. Here, decision fusion is done by simple weighing of score vectors $v_1$ and $v_2$ as $\alpha * v_1 + (1 - \alpha) * v_2$. The $\alpha$ parameter is searched over [0,1] range with steps of 0.05. The performances of early and late fusion alternatives for GPA aligned visual models are shown in Table 5. We see that: i) simple weighted fusion gives higher performance compared to feature level fusion, and ii) the combination of geometric features with an appearance descriptor (e.g. LBP) reaches the highest performance.

We finally apply multi-level, model- and class-weighted score fusion. Here, we test alternative combinations and observe that LPQ does not contribute to overall performance on the validation set. Therefore, we exclude it from the system. We test HOG-NOL and HOG-OL alternatives and make two test set submissions to see their performances. The fusion system with GPA aligned faces reaches 53.06% accuracy, whereas its given aligned counterpart gives 42.84% on the validation set. The second level fusion of the optimized alignment-specific systems give around 49.5%, where

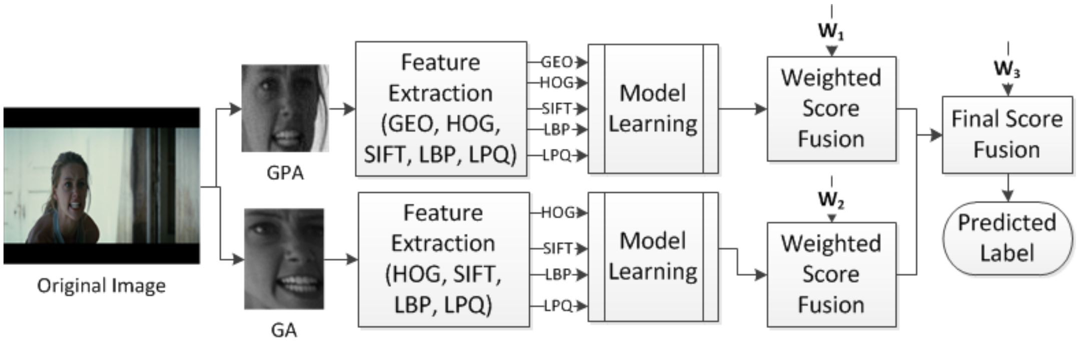

**Figure 2: Pipeline of the proposed system for SFEW**

**Table 5: Comparison of early and late fusion with GPA aligned images.**

| Feature Level Fusion | | | | |
|---|---|---|---|---|
| | HOG | SIFT | LBP | LPQ |
| GEO | **47.81** | 46.94 | **47.81** | 46.94 |
| HOG | | 45.77 | 47.23 | 44.31 |
| SIFT | | | 45.77 | 45.77 |
| LBP | | | | 42.86 |
| **Weighted Score Fusion** | | | | |
| | HOG | SIFT | LBP | LPQ |
| GEO | 49.25 | 48.10 | **50.15** | 47.23 |
| HOG | | 49.56 | 48.40 | 48.40 |
| SIFT | | | 48.40 | 46.06 |
| LBP | | | | 47.23 |

we see the GPA based system has higher face miss rate, but improves on the baseline alignment. The best validation set results and corresponding test performances are depicted in Table 6. Note that the validation and test set performances of the submitted systems are very similar, which indicates little effect of overlearning. Our top test set accuracy (49.46%) outperforms the challenge baseline (39.13%) by a large margin. The test set confusion matrix corresponding to the best results is depicted in Figure 3. We see that in the static image based expression recognition, recall of disgust and fear classes fail totally; while recall of happy, angry, neutral and surprise are reasonable.

**Table 6: Validation and test set performances of two submitted systems.**

| GA-Val | GPA-Val | Level 2-Val | Test |
|---|---|---|---|
| 42.15 | 53.06 | 49.53 | 49.46 |
| 42.84 | 52.48 | 49.30 | 48.92 |

## 5. VIDEO BASED EMOTION RECOGNITION CHALLENGE

In addition to frame-level feature extraction, in the AFEW challenge, we carried out video modeling via TOP extension, FV encoding and functionals (FUN). Here, we also benefited from the audio modality to improve the fusion performance. Similar to SFEW, we analyzed the individual and pairwise combined performances of video features. Out of all possible descriptor-video modeling combinations we eliminated the alternatives whose best performance is below 40%. A summary of the resulting pruned validation set performances is given in Table 7.

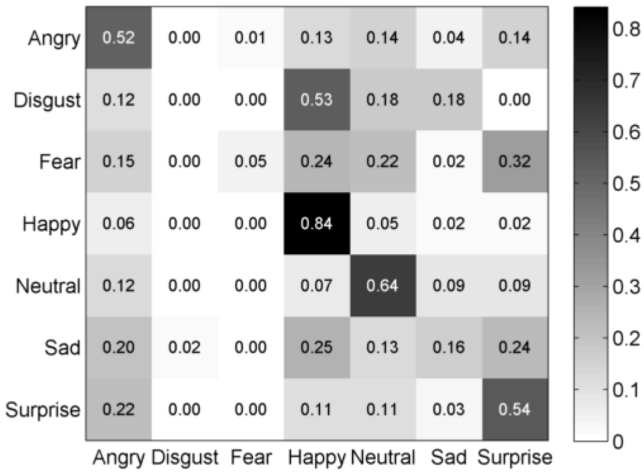

**Figure 3: Confusion matrix of the best submission in SFEW.**

The most remarkable performance is obtained with FV encoded geometric features. This is followed by LGBP-TOP and HOG-FUN. Zero Crossing Rate (ZCR) is applied to z-normalized contours of x coordinate of nose, y coordinate of lower lip, left and right eye heights, vertical distance of left and right eye brows to eye centers. When these six features are combined with GEO-FUN, they gave a considerable boost.

Pairwise feature and decision level fusion resulted in a pattern similar to SFEW. We obtained relatively better results with late fusion and the best results are obtained using combinations of appearance features with FV encoded geometric features. Simple weighted fusion of HOG-FUN with GEO-FV reached a validation set accuracy of 48.24%.

Our first test set submission consisted of weighted fusion of audio, LBP-TOP and LGBP-TOP. This setup is inspired from [15], and is intended to provide us a competitive benchmark. This system reached a test set accuracy of 50.28%. We then combined seven GPA aligned visual models (using GEO-FUN+ZCR and excluding of GEO-FUN) with the

**Table 8: Validation and test set accuracies of submitted systems. WF: weighted fusion, FF: feature level fusion.**

| Submission | System | Val | Test |
|---|---|---|---|
| 1 | WF (GA: LBP-TOP. LGBP-TOP & Audio) | 50.14% | 50.28% |
| 2 | WF (GPA: 7 Visual sub-sys & Audio) | 51.70% | 44.71% |
| 3 | WF (System 1 & 2) | 58.22% | 42.64% |
| 4 | WF (System 1 & GEO-FV, HOG-FUN) | 52.74% | 48.61% |
| 5 | WF (Audio, GA: LGBP-TOP, LBP-TOP, FF(SIFT-FUN & GEO-FV), LPQ-TOP with PLS) | 52.30% | 49.17% |
| **6** | **WF (Audio, GA: LGBP-TOP, HOG-FUN & SIFT-FV, LBP-TOP, LPQ-TOP with PLS)** | **52.30%** | **53.62%** |
| 7 | WF (Audio, GA: LGBP-TOP, HOG-FUN & SIFT-FV, LBP-TOP, LPQ-TOP, GPA: GEO-FV with PLS) | 54.47% | 52.69% |

**Table 7: Best validation set performances of two alignment systems.**

| | **Lin** | | **RBF** | |
|---|---|---|---|---|
| **Feature** | **ELM** | **PLS** | **ELM** | **PLS** |
| Audio | **36.59** | **36.29** | 35.51 | 34.73 |
| **Visual Features with Given Alignment** | | | | |
| HOG-FUN | 39.02 | **44.99** | 41.46 | 42.01 |
| SIFT-FUN | 40.92 | **43.63** | 42.28 | 41.19 |
| LBP-TOP | 41.19 | 41.46 | **42.01** | 40.65 |
| LGBP-TOP | 43.63 | 43.90 | **44.44** | 44.17 |
| LPQ-TOP | 40.65 | 41.19 | 40.65 | **42.01** |
| SIFT-FV | 40.11 | **41.73** | 40.38 | 40.38 |
| **Visual Features with GPA based Alignment** | | | | |
| HOG-FUN | 40.59 | **43.24** | 42.06 | 40.59 |
| SIFT-FUN | 38.53 | **42.94** | 40.59 | 39.12 |
| LBP-TOP | 41.76 | **42.65** | 42.06 | **42.65** |
| LGBP-TOP | 40.88 | **44.12** | 40.29 | 42.06 |
| GEO-FUN | 39.12 | 40.59 | **42.65** | 40.59 |
| GEO-FUN + ZCR | 40.00 | 40.00 | **44.12** | 42.06 |
| GEO-FV | 42.65 | **45.59** | 44.12 | 44.12 |
| SIFT-FV | 42.35 | 41.47 | **43.24** | 41.76 |

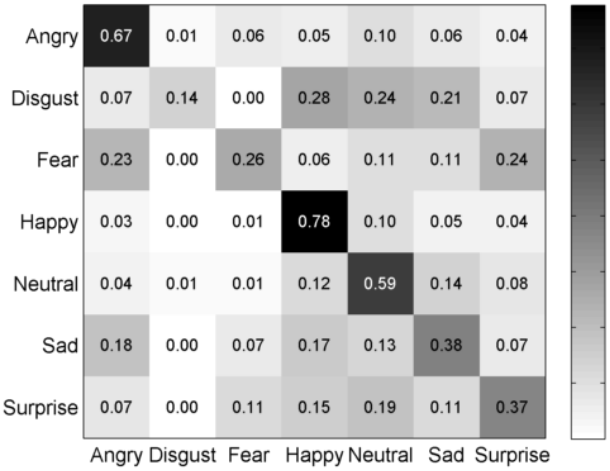

**Figure 4: Confusion matrix of the best submission in AFEW.**

audio model reaching a validation set accuracy of 51.7%. When this is combined with the first submission system, it reached 58.22% on the validation set. However, their test set performances were 44.71% and 42.64%, respectively. We attributed this dramatic difference to high number of missing test set videos (by GPA alignment) that are represented as zero score vectors during late fusion. During our analyses, we also observed that fusing best scores from PLS and ELM might result in lowered performance due to varying score ranges of these classifiers. For these reasons, we opted to fuse best models from given aligned images with PLS classifier. Our top test set result (53.62%) is obtained using weighted score fusion of audio model with five visual models (see Table 8 and Figure 4 for details). Compared to static expression challenge, the confusion matrix here shows higher recall in disgust and fear classes, which can be attributed to modeling visual dynamics and utilizing audio, respectively.

# 6. CONCLUSIONS

In this study, we diversify learners using alternative alignment, descriptor type and classifier options. We then use a least squares based classifier and weighted fusion. The multi-level fusion gives good validation set results in both challenges, and generalizes well in SFEW. In AFEW, the best test set results are obtained with single level weighted fusion scheme, although on the validation set better results are be achieved with multi-level version. The reason of the validation-test set performance discrepancy in AFEW is the high number of missing videos when aligned with GPA. Thus, the results on the validation set indicate that if it would be possible to use geometric features from the given alignment, or the number of valid videos in GPA aligned alternative would have been higher, better results could have been reached.

Comparing the confusion matrices of the best test set results, we see that in both challenges, the recall of disgust and fear classes is very low. In SFEW, the recall of these classes fails totally. In audio-only models, it is possible to obtain higher recall for the fear class. Therefore, more effort should be spent to modeling audio modality. On the other hand, disgust can be distinguished with visual cues, especially with dynamics of facial structure. We should note that the class distribution of the challenge data is also responsible from the low recall of this under-sampled class.

In future works, we will focus on the multi-level fusion of various alignments, which have different resolutions. We will also exploit ways for better alternative modeling of audio modality, which is not elaborated thoroughly in this study.

# 7. REFERENCES

[1] T. R. Almaev and M. F. Valstar. Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 356–361. IEEE, 2013.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '05)*, volume 1, pages 886–893. IEEE, 2005.

[3] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proc. of the 16th ACM Intl. Conf. on Multimodal Interaction (ICMI 2014)*. ACM, 2014.

[4] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proc. of the 15th ACM Intl. Conf. on Multimodal Interaction (ICMI 2013)*, pages 509–516. ACM, 2013.

[5] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, July 2012.

[6] A. Dhall, R. Goecke, L. S, and T. Gedeon. Static facial expression in though conditions: Data, evaluation protocol and benchmark. In *ICCV BEFIT Workshop)*. IEEE, 2011.

[7] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proc. of the 17th ACM Intl. Conf. on Multimodal Interaction (ICMI 2015)*. ACM, 2015.

[8] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. of the intl. conf. on Multimedia*, pages 1459–1462. ACM, 2010.

[9] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[10] M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb. Identification using encrypted biometrics. In *Computer Analysis of Images and Patterns*, pages 440–448. Springer, 2013.

[11] J. Heikkilä, V. Ojansivu, and E. Rahtu. Improved blur insensitivity for decorrelated local phase quantization. In *20th International Conference on Pattern Recognition (ICPR '10)*, pages 818–821, 2010.

[12] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(2):513–529, 2012.

[13] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *Cybernetics, IEEE Transactions on*, 44(2):161–174, 2014.

[14] S. E. Kahou, C. Pal, X. Bouthillier, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 543–550, 2013.

[15] H. Kaya and A. A. Salah. Combining modality-specific extreme learning machines for emotion recognition in the wild. *Journal on Multimodal User Interfaces*, 2015.

[16] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen. Partial least squares regression on Grassmannian manifold for emotion recognition. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 525–530, 2013.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[19] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, Minnesota, USA, 2007.

[20] C. R. Rao and S. K. Mitra. *Generalized inverse of matrices and its applications*, volume 7. Wiley New York, 1971.

[21] R. Rifkin, G. Yeo, and T. Poggio. Regularized least-squares classification. *NATO Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003.

[22] A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi. Effective geometric features for human emotion recognition. In *IEEE 11th International Conference on Signal Processing (ICSP)*, volume 1, pages 623–627, 2012.

[23] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH*, pages 2794–2797, 2010.

[24] J. Sivic and A. Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606, 2009.

[25] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 481–486, 2014.

[26] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[27] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[28] H. Wold. Partial least squares. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, pages 581–591. Wiley New York, 1985.

[29] X. Xiong and F. De la Torre. Supervised Descent Method and Its Application to Face Alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pages 532–539, 2013.