Wild Wild Emotion: A Multimodal Ensemble Approach

John Gideon*, Biqiao Zhang*, Zakaria Aldeneh*,
Yelin Kim†, Soheil Khorram*, Duc Le*, Emily Mower Provost*
University of Michigan, Ann Arbor*
University at Albany, State University of New York†
gideonjn@umich.edu, didizbq@umich.edu, aldeneh@umich.edu,
yelinkim@albany.edu, khorrams@umich.edu, ducle@umich.edu, emilykmp@umich.edu

ABSTRACT

Automatic emotion recognition from audio-visual data is a topic that has been broadly explored using data captured in the laboratory. However, this data is not necessarily representative of how emotion is manifested in the real-world. Recent work aims to adapt these earlier methods to data collected in the wild. In this paper, we describe our proposed system for the 2016 Emotion Recognition in the Wild challenge. We use the Acted Facial Expressions in the Wild database 6.0 (AFEW 6.0), which contains short clips of popular TV shows and movies that contain more variability in the data than laboratory recordings. While previous submissions have concentrated on visual features, we explore a method which fuses together both the audio and visual modalities. We explore a set of features that incorporate information from facial expressions and speech, in addition to cues from the background music and overall scene. We design sets of pairwise and multiclass classifiers and fuse the resulting systems. Our fusion increases the performance from a baseline of 38.81% and 40.47% to 43.86% and 46.88% for validation and test sets, respectively. While the video features perform better than audio features alone, a combination of the two modalities achieves the greatest performance. This demonstrates the effectiveness of multimodal emotion fusion. Because of the flexible design of the fusion, it is easily adaptable to other multimodal learning problems.

CCS Concepts

•Computing methodologies → Artificial intelligence; Ensemble methods;

Keywords

Emotion Recognition, Ensemble Learning, Emotion in the Wild, Multimodal Learning

ICMI '16 November 12–16, 2016, Tokyo, Japan
© 2016 ACM. ISBN 123-4567-24-567/08/06. . . $15.00
DOI: 10.475/123_4

1. INTRODUCTION

Automatic emotion recognition from human vocal and facial expressions has received attention in a variety of fields ranging from computer science, to psychology and psychiatry [3, 12]. In particular, recent work has focused on developing automatic emotion recognition systems for more natural, spontaneous multimedia data [7]. In this work, we present a system that can identify human emotion 'in the wild', defined as emotion collected in variable settings. We propose a fusion of audio-visual emotion recognition systems that automatically classifies seven different emotions in short movie clips for the 2016 Emotion Recognition in the Wild challenge.

Emotion datasets collected in a controlled laboratory setting have been widely used in emotion recognition research [3, 4, 12]. These datasets provide insight into how humans express different types of emotion. However, open questions still remain whether techniques developed in these settings can be transferred to datasets collected in the wild. The 2016 Emotion Recognition in the Wild challenge provides an opportunity to test data in a real-world context. The Acted Facial Expressions in the Wild database 6.0 (AFEW 6.0), accompanying the challenge, presents short clips taken from popular movies. Each clip contains a multitude of different visual and audio cues that could be interpreted as indicators of emotion. Capturing and interpreting all these different types of cues in a single system is difficult. The successful detection of emotion in a setting outside the laboratory would allow for real-world applications, ranging from the recommendation of movies based on emotion content to aiding individuals with mental disorders, such as autism.

Previous work on the AFEW dataset focused on the video modality and considered audio of secondary importance [16, 20]. We present a unified system that is designed from the bottom up to combine information from the video and audio modalities. We present an ensemble approach that fuses different classification models. The fusion of multiple classifiers has been shown to be particularly effective for smaller data sets [17].

The overview of our proposed system is shown in Figure 1. We extract a variety of audio and video features that describe the overall emotional information of a video clip, such as speech prosody and energy features, HSV color histogram features, pixel intensity features, Action Unit (AU) features, and LBP-TOP features. We build two sets of emotion classification systems, one with Support Vector Machines (SVMs) and the other with Random Forests (RF), and use a classifier fusion approach to combine the two systems.

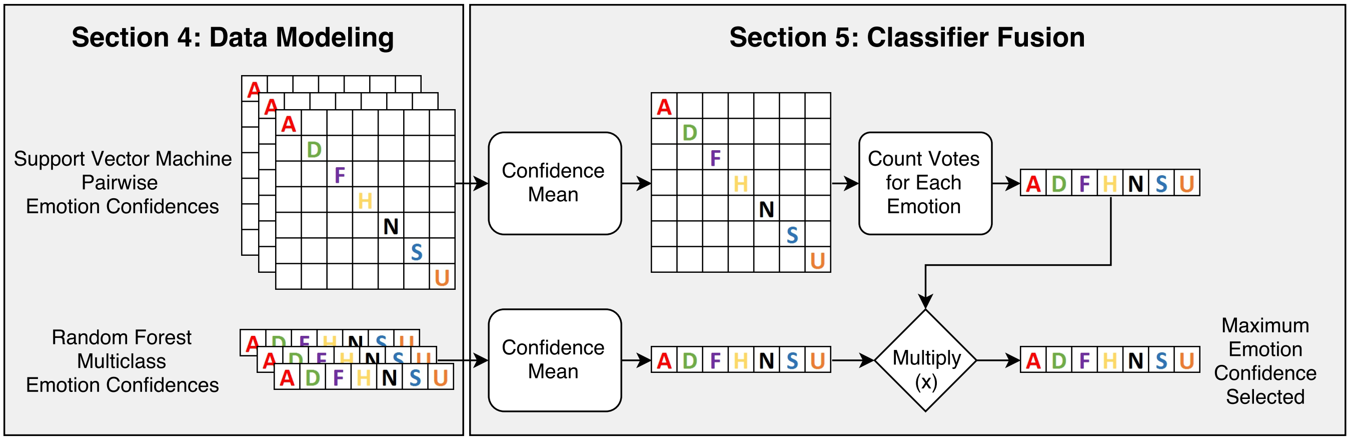

Section 4: Data Modeling
Section 5: Classifier Fusion
Support Vector Machine Pairwise Emotion Confidences
Random Forest Multiclass Emotion Confidences
A D F H N S U
Confidence Mean
Confidence Mean
Count Votes for Each Emotion
Multiply (x)
Maximum Emotion Confidence Selected

Figure 1: The fusion of pairwise and multiclass subsystems. The mean is taken over each subsystem type. Votes are counted using the pairwise mean confidences. These vectors are multiplied to get the final confidences. The emotion is selected with the maximum confidence.

Fold	Train	Val	Train+Val	Test
Angry	133	64	197	83
Disgust	74	40	114	36
Fear	81	46	127	66
Happy	150	63	213	135
Neutral	144	63	207	174
Sad	117	61	178	71
Surprise	74	46	120	28
Total	773	383	1156	593

Table 1: The number of utterances in each of the emotion classes and fold divisions.

Our experimental results on the challenge dataset show an improvement from a baseline performance of 38.81% and 40.47% to 43.86% and 46.88% for val and test sets, respectively. This demonstrates the effectiveness of including different modalities and learning paradigms for emotion detection. The key contributions of our system are as follows:

- We demonstrate that a wide range of audio-visual features can be used for emotion classification in the wild.
- We introduce a fusion method that utilizes different emotion classifiers and combines both pairwise and multiclass knowledge. This differs from prior work that focused on a single emotion classifier.

2. DATASET AND FOLDS

The AFEW 6.0 dataset contains popular TV and movie clips divided into seven categorical emotions including angry, disgust, fear, happy, neutral, sad, and surprise. The amount of utterances in each of these emotions is shown in Table 1. The clips have an average length of 2.46 seconds with a standard deviation of 1.00 seconds.

The dataset is divided into train, validation (val) and test sets. During the development of the component classifiers (called subsystems below) we combined the train and val sets and performed 10-fold cross-validation. This allowed us to get a better estimate of the actual test performance than only using val for performance measurement. We report val accuracy when the system is only trained using the train fold. Finally, the test accuracy is determined using a system trained on both the train and val sets.

3. FEATURE EXTRACTION

Interspeech 2010 Acoustic Feature Set (IS10). We use the Interspeech 2010 feature set, extracted using openSMILE [10]. This feature set contains a variety of statistics over frame-level acoustic features including loudness, Mel-frequency cepstrum coefficients (MFCCs), line spectral pairs (LSPs), fundamental frequency (F0), voicing, shimmer, and jitter. This results in 1592 utterance-level features.

Dimensional Emotion Estimates (VAD). We hypothesize that auxiliary emotion characteristics will be helpful for predicting categorical emotion labels. We train regressors for valence (positive vs. negative), activation (calm vs. excited) and dominance (dominant vs. submissive) on outside emotion corpora. We apply these models to the AFEW 6.0 data, resulting in a set of secondary features. AFEW 6.0 contains both speech and background music. Therefore, the regressors are trained on both speech and music emotion corpora, including: the improvisation part of IEMOCAP [3] (4784 utterances), the spontaneous and improvisation part of MSP-IMPROV [4] (7452 utterances), and a self-collected music corpus containing 200 30-second music clips ranging from classical music, film score to pop music. The two speech corpora have labels for valence, activation and dominance, while the music corpus only has labels for valence and activation. This results in 8-dimensional estimates for each utterance (valence × 3 corpora, activation × 3 corpora, dominance × 2 corpora). For the speech corpora, we train the regressor of each dimension (e.g., valence) using the multi-task feature learning method proposed in [1] to avoid overfitting to specific dataset, with each corpus as a task. This method assumes that there exists a common sparse feature representation, either on the original feature space or a transformed feature space, across tasks. In this work, we assume the shared representation is on the original feature space. For the music corpus, we use regularized linear regression. For both algorithms, the regularization parameter C is selected using 5-fold cross-validation on the training corpus, in the range between $\{10^{-6}, 10^{-5}, ..., 10^{6}\}$. The IS10 feature set described above is used.

HSV Color Histogram (HSV). We extract the color histogram in the HSV color space at frame-level, as in previous work on gif emotion recognition [14]. We set the quantization level to 8 for hue, and 2 for saturation and value. We calculate 8 statistics, including mean, standard deviation, max, min, range, upper quartile, lower quartile and interquartile range over the frame-level feature. This results in 256 (32 × 8) utterance-level features.

Pixel Intensity Change (PIC). We extract features that reflect the change between frames. We take the mean, standard deviation, max and min of the intensity image of each frame (converted from the color image), and the absolute difference between two consecutive frames. We also calculate the mean-squared error between each pair of consecutive frames. This results in a 9-dimensional frame-level feature vector. Again, we applied the above-mentioned 8 statistics to generate the 72 utterance-level features.

Action Unit Features (AU). We use Action Unit (AU) features extracted using CERT [15]. AU features capture anatomical movement of facial muscles related to emotion [9]. The CERT AU features include: (i) AU 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, 28, 45, (ii) fear brow (AU 1+2+4) and distress brow (AU 1, 1+4), and (iii) the left and right unilateral features of AU 10, 12, and 14. This results in 28-dimensional frame-level AU features in total. We applied 10 statistics to the frame-level features, which are: mean, standard deviation, max, min, range, upper quartile, lower quartile, interquartile range, skewness and kurtosis. Note that the statistics are only calculated for utterances where at least half of the frames had AUs extracted successfully. Only 929 out of the 1156 training and validation utterances, and 469 out of the 593 test utterances have valid utterance-level AU features.

Local Binary Pattern-Three Orthogonal Planes Features (LBP-TOP). We use Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) features [21] provided in the challenge dataset [8]. LBP-TOP features are histogram-based image features that describe texture of an image, and it has been widely used in facial emotion recognition (a comprehensive survey can be found in [19]). This results in 2832-dimensional LBP-TOP features calculated over each utterance. The LBP-TOP features are not available for 29 utterances in the train and val sets.

4. DATA MODELING

4.1 Support Vector Machines

We build a binary SVM for each of the 21 pairs of emotions, and use majority vote to decide the final predicted label. We construct different combinations of five sets of features in preliminary experiments, and only select the combinations (see Table 2) that produce 10-fold cross-validation accuracy higher than 35% for further classifier fusion (Section 5). For each feature combination, we train two versions of SVM: with Information Gain (IG) feature selection [6] and without. In the former version, we apply IG feature selection for each binary classification. Features with zero IG are removed. This results in 20 SVM subsystems each with 21 binary emotion comparisons.

Audio Feats.		Video Feats.			Accuracy	
IS10	VAD	HSV	PIC	AU	No IG	With IG
✓	✓				36.4%	35.8%
✓					36.9%	35.4%
✓	✓	✓			37.8%	35.3%
✓	✓	✓	✓		37.7%	38.3%
				✓	39.3%	37.2%
		✓		✓	40.5%	38.5%
		✓	✓	✓	41.3%	40.0%
✓	✓	✓	✓	✓	42.7%	44.4%
✓	✓			✓	42.5%	44.6%
✓	✓	✓		✓	43.2%	45.0%

Table 2: 10-fold train+val accuracy on the SVM subsystems. Only those greater than 35% accuracy are shown and used. These are sorted in order or performance.

Feature Sets	Accuracy
IS10+VAD+HSV	38.1%
IS10+VAD+HSV+AU	45.4%
LBP-TOP	38.3%

Table 3: 10-fold train+val accuracy on the random forest subsystems using different feature sets.

Leave-one-fold-out validation is performed on the training set to select the SVM hyper-parameters. When generating the final test predictions, the classifiers are built using the full training set. We implement the subsystems using LIBSVM [5], and adopt the radial basis function (RBF) kernel. The range of the kernel width parameter γ is $\{2^{-10}, 2^{-9}, ..., 2^{-1}\}$, and the range of the cost parameter C is $\{10^{-5}, 10^{-4}, ..., 10^{5}\}$.

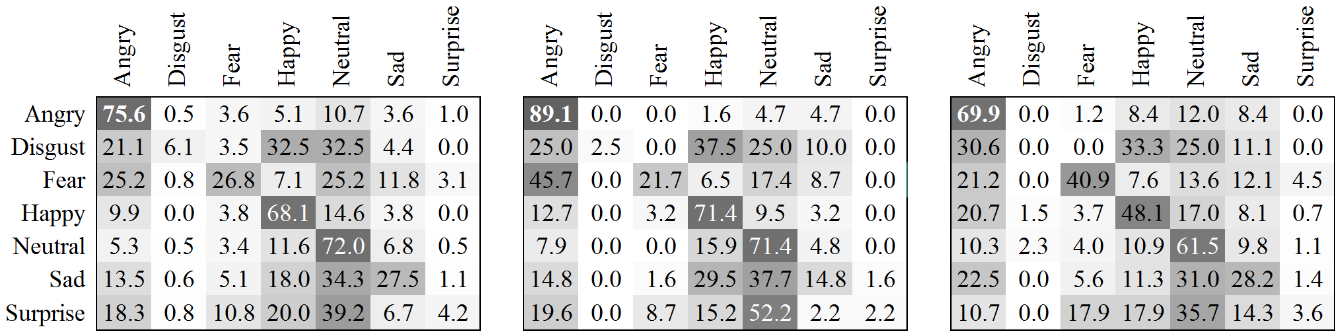
We report the accuracy of each subsystem using the 10-fold train+val set described in Section 2 in Table 2. This accuracy is used to evaluate the subsystems in the development phase. While AU features alone provide better performance than the others combined, the addition of audio features to AU features increases the subsystem performances from 37.2% to 44.6% when using IG.

To facilitate the overall classifier fusion, we calculate the sigmoid transformation of the absolute decision value to represent the confidence level of each binary prediction.

4.2 Random Forests

In addition to the pairwise modeling described above, we perform multiclass modeling using a Random Forest (RF) classifier with the feature sets described in Table 3. These include a combination of IS10, VAD, HSV, and AU features shown to be particularly effective in pairwise classification. Additionally, the high-dimensional LBP-TOP features are used as another source of video modality information. RF has been shown to work particularly well with small datasets of high dimensionality [2]. A RF classifier works by building a set of N trees using bootstrapped samples of the original dataset. Only a random subset of the features are used to create the splits at each node. Given a test example, a RF classifier computes the confidences by considering the proportion of trees that predict each class label.

The 10-fold train+val cross-validation accuracy of using RF on the different feature sets is shown in Table 3. We run cross-validation on the training data to pick the optimal number of trees, N. We fix the random number of features to be considered at each step to $\sqrt{d}$, where d is the dimensionality of the feature vector, as typically used for RF. Unlike the SVM-based approach described above, we



	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	75.6	0.5	3.6	5.1	10.7	3.6	1.0
Disgust	21.1	6.1	3.5	32.5	32.5	4.4	0.0
Fear	25.2	0.8	26.8	7.1	25.2	11.8	3.1
Happy	9.9	0.0	3.8	68.1	14.6	3.8	0.0
Neutral	5.3	0.5	3.4	11.6	72.0	6.8	0.5
Sad	13.5	0.6	5.1	18.0	34.3	27.5	1.1
Surprise	18.3	0.8	10.8	20.0	39.2	6.7	4.2

(a) Train+Val (Accuracy = 46.54%)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	89.1	0.0	0.0	1.6	4.7	4.7	0.0
Disgust	25.0	2.5	0.0	37.5	25.0	10.0	0.0
Fear	45.7	0.0	21.7	6.5	17.4	8.7	0.0
Happy	12.7	0.0	3.2	71.4	9.5	3.2	0.0
Neutral	7.9	0.0	0.0	15.9	71.4	4.8	0.0
Sad	14.8	0.0	1.6	29.5	37.7	14.8	1.6
Surprise	19.6	0.0	8.7	15.2	52.2	2.2	2.2

(b) Val (Accuracy = 43.86%)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	69.9	0.0	1.2	8.4	12.0	8.4	0.0
Disgust	30.6	0.0	0.0	33.3	25.0	11.1	0.0
Fear	21.2	0.0	40.9	7.6	13.6	12.1	4.5
Happy	20.7	1.5	3.7	48.1	17.0	8.1	0.7
Neutral	10.3	2.3	4.0	10.9	61.5	9.8	1.1
Sad	22.5	0.0	5.6	11.3	31.0	28.2	1.4
Surprise	10.7	0.0	17.9	17.9	35.7	14.3	3.6

(c) Test (Accuracy = 46.88%)

Figure 2: Confusion matrices and accuracies of the train+val, val, and test sets.

do not apply any feature selection to reduce the number of features when using RF classifier.

To facilitate the overall classifier fusion, the fraction of decision trees in the RF selected as each emotion is output as the seven-dimensional confidence vector.

5. CLASSIFIER FUSION

We fuse the above-mentioned pairwise SVM and multiclass RF models (Figure refFusionDiagram). Ensemble learning provides a process to optimally combine a set of classifiers based on the confidence of each classifier, individually [17]. Many different combination rules were examined including decision trees, stacked generalization, and logistic regression. However, simple averaging over the confidences of subsystems produced the highest cross-validation accuracy. Additionally, we develop a methodology for combining pairwise and multiclass confidences.

Each pairwise SVM subsystem outputs a 7x7 matrix of confidences for each utterance. In this matrix, the confidence at index (i, j) represents the confidence of selecting $emotion_i$ instead of $emotion_j$. This value ranges from -1 to 1 and can be negative if selecting $emotion_j$ is more confident than $emotion_i$. Additionally, each multiclass RF subsystem outputs a seven-dimensional vector of confidences that is a probability distribution for each utterance.

For each utterance, a gating function is applied to the subsystems. Only the outputs of subsystems trained including AU features are used when utterance AU features are properly extracted. Conversely, when AU extraction fails we only include subsystem outputs trained without AU features. However, the LBP-TOP random forest subsystem is always included to provide information from the video modality even when AU is not available.

Once the subsystems are selected using the above gating method, the confidence matrices of the pairwise subsystems are averaged. In particular, this method works well when each classifier is differentiated from one another [18]. This allows for more confident systems to have a stronger impact in the decision, as subsystems with low confidence on a particular utterance will be near zero. Each of the pairwise decisions are considered by tallying the votes for each winning emotion, as typically used for one-versus-one SVMs [13]. This results in a seven-dimensional vector of emotion votes between zero and six. This is multiplied by the mean confidence vector of multiclass subsystems to further decrease the likelihood of those emotions given low confidence in the multiclass context. This results in the final confidence scores. The emotion associated with the highest confidence score is selected.

6. RESULTS

The confusion matrix for the train+val, val, and test results can be seen in Table 2. Our fusion method results in an increase in accuracy for the best subsystem train+val 10-fold cross-validation (46.5%). This supports previous work demonstrating that classifier fusion can improve the performance of an audio-visual emotion recognition system [11]. Additionally, we achieve 43.86% on the val fold, a 5.05% increase from the baseline of 38.81%. Finally, the fusion has an accuracy of 46.88% on the test set, a 6.41% increase from the baseline of 40.47%.

Similar to previous Emotion in the Wild papers, our system performs best on the majority class emotions of angry, happy, and neutral [16,20]. This may be due to (i) the class imbalance in the training data and (ii) our choice of accuracy as the performance measure. Less importance is given to the minority classes when validating the system using accuracy instead of a measure such as unweighted average recall (UAR). We believe that this strong performance is also in part due to the acoustic feature sets that are effective at capturing the high energy speech present in anger and happiness. On the contrary, the lack of high-energy speech is likely an indicator of neutrality.

7. CONCLUSIONS

The 2016 Emotion Recognition in the Wild challenge has provided a collection of popular TV and movie clips spanning different emotions. The small size of the dataset (72 minutes) provides a difficult learning task well suited to ensemble learning. In this paper, we present a collection of subsystems trained using pairwise and multiclass methodologies. They are built on a variety of features designed to represent emotion present in the face and speech of actors, as well as the cues from the musical scores and overall scene. We created a fusion of these subsystems based on classifier confidence. We improve from the baseline performance of 38.81% and 40.47% to 43.86% and 46.88% for validation and test sets, respectively. We demonstrate that a combination of the audio and video modalities outperforms video alone with an improvement from 37.2% to 44.6% for the SVM subsystems. This demonstrates the effectiveness of leveraging a variety of features and models to detect emotion when working with data captured in highly variable settings.

8. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[4] C. Busso, S. Parthasarathy, A. Burmania, M. Abdel-Wahab, N. Sadoughi, and E. Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. 2015.

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[6] M. Cover Thomas and A. Thomas Joy. *Elements of information theory*. Wiley, 1991.

[7] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426. ACM, 2015.

[8] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 423–426, New York, NY, USA, 2015. ACM.

[9] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[10] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[11] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*, pages 359–368. Springer, 2011.

[12] S. Haq and P. J. Jackson. Multimodal emotion recognition. *Machine audition: principles, algorithms and systems*, pages 398–423, 2010.

[13] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[14] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated gifs. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 213–216, 2014.

[15] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.

[16] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 494–501. ACM, 2014.

[17] R. Polikar. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer, 2012.

[18] D. Ruta and B. Gabrys. An overview of classifier fusion methods. *Computing and Information systems*, 7(1):1–10, 2000.

[19] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[20] A. Yao, J. Shao, N. Ma, and Y. Chen. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 451–458. ACM, 2015.

[21] 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*, 29(6):915–928, 2007.