This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Facial Expression Recognition from World *Wild* Web

Ali Mollahosseini[1], Behzad Hassani[1], Michelle J. Salvador[1],
Hojjat Abdollahi[1], David Chan[2], and Mohammad H. Mahoor[1,2]
[1] Department of Electrical and Computer Engineering
[2] Department of Computer Science
University of Denver, Denver, CO
Ali.Mollahosseini@du.edu, Behzad.Hasani@du.edu, Michelle.Salvador@du.edu)
habdolla@du.edu, davidchan@cs.du.edu, and mmahoor@du.edu

Abstract

Recognizing facial expression in a wild setting has remained a challenging task in computer vision. The World Wide Web is a good source of facial images which most of them are captured in uncontrolled conditions. In fact, the Internet is a Word Wild Web of facial images with expressions. This paper presents the results of a new study on collecting, annotating, and analyzing wild facial expressions from the web. Three search engines were queried using 1250 emotion related keywords in six different languages and the retrieved images were mapped by two annotators to six basic expressions and neutral. Deep neural networks and noise modeling were used in three different training scenarios to find how accurately facial expressions can be recognized when trained on noisy images collected from the web using query terms (e.g. happy face, laughing man, etc)? The results of our experiments show that deep neural networks can recognize wild facial expressions with an accuracy of 82.12%.

1. Introduction

The World Wide Web (aka the Internet) has become a vast abundant source of information and data. Especially with the growth and use of social media and the availability of digital cameras on smart phones, people can easily add data to the Internet by taking photos, writing a short description, and immediately uploading them to the social media. People add more information to each photo by doing a tag, like, dislike, or comment on photos posted by friends or others on the Web. It is estimated that over 430 million photos are uploaded to Facebook and Instagram servers every day [10, 4]. Among photos posted on the Web, facial images have the highest incidents (e.g. selfies or self-portrait images are very popular nowadays). These facial photos are often taken in the wild under natural conditions with varying and diverse parameters such as scene lighting, user's head pose, camera view, image resolution and background, subject's gender, ethnicity, and facial expressions among others. Furthermore, the labels given by users use a wide range of vocabulary that is commonly understood to describe emotions, facial attributes, and expressions, of the pictures' contents. These photos are truly Wild images both in terms of the image quality/conditions and the labels given by users. An interesting question that may arise is, how the labels given wildly to facial images on the web by general users are consistent with the six basic emotions defined by psychologists.

On the other hand, computer vision and machine learning techniques for facial expression recognition are finding their ways into the design of a new generation of Human-Computer Interfaces. In order to train a machine learning system, many researchers have created databases using human actors/subjects portraying basic emotions [9, 25, 16]. However, most of the captured datasets mainly contain posed expressions acquired in a controlled environment. This is mostly due to the fact that it is hard and time consuming to collect unposed facial expression data in lab settings. However, in real applications, the system needs to capture and recognize spontaneous expressions, which involve different facial muscles, less exaggeration/intensity and have different dynamics than posed expressions. Researchers who have created spontaneous expression databases have captured the human face spontaneously while watching a short video or filling a questionnaires [7, 17, 18]. However, the datasets are still captured in controlled lab settings (i.e. with the same illumination, resolution, etc.) or have a limited number of subjects, ethnicities, and poses poorly representing the environment and conditions faced in real-world situations. Existing databases in the wild settings, such as SFEW [3] or

FER2013 [6], are also either very small or have low resolution without facial landmark points necessary for preprocessing.

Moreover, state-of-the-art machine learning algorithms such as Deep Neural Network requires big data for training and evaluation of the core algorithms. Given all the aforementioned motivations, this paper presents the results of our recent study with the aim of resolving the following questions:

1. How consistent are the expression labels given by general web users compared to the six basic expression labels annotated by expert annotators on facial images?

2. How accurately can a state-of-the-art algorithm classify images when trained on facial images collected from the web using query terms (e.g. happy face, laughing man, etc)?

To address these questions, we created a database of in-the-wild facial expressions by querying different search engines (Google, Bing and Yahoo) We then annotated a subset of images using two human annotators and showed the general accuracy of the querying search engines for facial expression recognition. We trained two different deep neural network architectures with different training settings i.e. training on clean well-labeled data, training on a mixture of clean and noisy data, and training on mixture of clean and noisy data with a noise modeling approach using a general framework introduced in [38]. In other words, given the result of annotations, the noise level of each search engine is estimated as a prior distribution on the labels of our posterior set allowing for greater classification performance when we sample noisy labels and true labels in the same proportion. In order to achieve this, we learned a stochastic matrix where the entries are the probability of confusion in the labels. From this matrix, we can extract a posterior distribution on the true labels of the data conditioned on the true label given the noisy label, and the noisy label given the acquired data. For more information on the technique, see [38].

The rest of this paper is organized as follows. Section 2 reviews existing databases and state-of-the-art methods for facial expression recognition in the wild. Sec. 3 explains the methodology of automatically collecting a large amount of facial expression images from the Internet and procedure of verifying them by two expert annotators. Section 4 presents experimental results on training two different network architectures with different training settings, and section 5 concludes the paper.

2. Facial Expression Recognition in the wild

Automatic Facial Expression Recognition (FER) is an important part of social interaction in Human-Machine-Interaction (HMI) systems [22]. Traditionally, automatic facial expression recognition (AFER) methods consist of three main steps 1) registration and preprocessing, 2) feature extraction, and 3) classification. Preprocessing and registration form an important part of the AFER pipeline. Many studies have shown the advantages of using facial image registration to improve classification accuracy in both face identification and facial expression recognition [8, 27]. In the feature extraction step, many methods such as HOG [17], Gabor filters [14], Local binary pattern (LBP) [31], facial landmarks [12], pixel intensities [19], and Local phase quantization (LPQ) [43], or a combination of multiple features using multiple kernel learning methods [41, 42] have been proposed to extract discriminative features. Classification is the final step of most AFER techniques. Support vector machines [43], multiple kernel learning [41, 42], dictionary learning [20] etc. have been shown to have a great performance in classifying discriminative features extracted from the previous stage.

Although, traditional machine learning approaches have been successful when classifying posed facial expressions in a controlled environment, they do not have the flexibility to classify images captured in a spontaneous uncontrolled manner ("in the wild") or when applied to databases for which they were not designed. The poor generalizability of traditional methods is primarily due to the fact that many approaches are subject or database dependent and only capable of recognizing exaggerated or limited expressions similar to those in the training database. Many FER databases have tightly controlled illumination and pose conditions. In addition, obtaining accurate training data is particularly difficult, especially for emotions such as sadness or fear which are extremely difficult to accurately replicate and do not occur often in real life.

Recently, facial expression datasets with in the wild settings have attracted much attention. Dhall *et al.* [2] released Acted *Facial Expressions in the Wild (AFEW)* from movies by semi-automatic approach via a recommender system based on subtitles. AFEW addresses the issue of temporal facial expressions and it is the only temporal publicly available facial expression database in the wild. A static subset *Static Facial Expressions in the Wild (SFEW)* is created by selecting static frames which covers unconstrained facial expressions, different head poses, age range, and occlusions and close to real world illuminations. However, it contains only 700 images and there are only 95 subjects in the database. In addition, due to the wild settings of the database, the released facial location and landmarks do not capture the faces in all images correctly making some training and test samples unusable (See Fig. 1).

The Facial Expression Recognition 2013 (FER-2013) database was introduced in the ICML 2013 Challenges in Representation Learning [6]. The database was created us-

Figure 1. Sample of images from SFEW [3] and their original registered images published with the database.

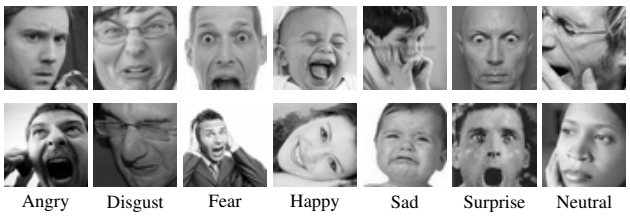

Figure 2. Sample of images from FER2013 database [6].

ing the Google image search API that match a set of 184 emotion-related keywords to capture the six basic expressions as well as the neutral expression. Human labelers rejected incorrectly labeled images. Images are resized to 48x48 pixels and converted to grayscale. The resulting database contains 35,887 images most of them in wild settings, yet only 547 of the images portray disgust. Figure 2 shows some sample images of FER2013. FER2013 is currently the biggest publicly available facial expression database in wild settings, enabling many researchers to train machine learning methods where large amounts of data are needed such as Deep neural networks. However, as shown in Fig. 2, the faces are not registered, and unfortunately most of facial landmark detectors fail to extract facial landmarks at this resolution and quality.

In addition, FER in the wild is really a challenging task both in terms of machine and human performance. Extensive experiments in [5] show that even humans are only capable of 53% agreement in terms of Fleiss kappa over all classes to classify AFEW video clips without listening to the audio track. State-of-the-art automated methods have achieved 35% accuracy on AFEW video clips by using audio modalities [44]. Even recognizing expression from still images or static frames using traditional machine learning approaches are not accurate and the best performance on SFEW 2.0 database is reported as 50% accuracy (with a baseline of 39.13%) [44].

Recently, deep neural networks have seen a resurgence in popularity. Recent state-of-the-art results have been obtained using neural networks in the fields of visual object recognition [13, 33], human pose estimation [36], face verification [34], and many more. Even in the FER field results so far have been promising [11, 21, 15, 11], and most of the facial expression recognition challenge winners have used deep neural networks [35, 40].

In the FER problem, however, unlike visual object databases such as imageNet [1], existing FER databases often have limited numbers of subjects, few sample images or videos per expression, or small variation between sets, making neural networks significantly more difficult to train. For example, the FER2013 database [6] (one of the largest recently released FER databases) contains 35,887 images of different subjects yet only less than 2% of the images portray disgust. Similarly, the CMU MultiPIE face database [9] contains around 750,000 images but is comprised of only 337 different subjects, where 348,000 images portray only a "neutral" emotion and the remaining images do not portray anger, fear or sadness.

In a recent study [21], the authors proposed a deep neural network architecture and combined seven well-known facial expression databases (i.e. MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013) to perform an extensive study on subject-independent and cross database. The results of the proposed architecture were comparable to or better than the state-of-the-art methods, However, the majority of data were still posed images and performance on wild databases (SFEW and FER2013) were only comparable to the state-of-the-art methods.

Considering the need to develop an automated FER in wild system, and issues with the current facial expression in wild databases, a possible solution is to automatically collect a large amount of facial expression images from the abundant images available on the Internet, and directly use them as ground truth to train deep models. However, consideration should be done to avoid false samples in the search engine results for expressions such as disgust or fear. This is due to the higher tendancy of people to publish happy or neutral faces that can be mislabeled or associated with disgust or fear by web users.

Nonetheless, semi-supervised [37], transfer learning [24], or noise modeling approaches [32, 38] can be used to train deep neural networks with noisy data by obtaining large amounts of facial expression images from search engines, along with a smaller subset of fully well-labeled images.

3. Facial expressions from the *wild* web

To create our database with the larger amount of images necessary for Deep Neural Networks, three search engines were queried by facial emotion related tags in six different languages. The search engines used where Google, Bing, and Yahoo. Other search engines where considered such as

Baidu and Yandex. However they either did not produce a high percentage of the intended images or they did not have accessible APIs for automatically querying and pulling image urls into the database.

A total of 1250 search queries where compiled in six languages and used to crawl Internet search engines for the image urls in our dataset. The first 200 urls returned for each query were stored in the database (258,140 distinct urls). Among the 258,140 urls, 201,932 images were available for download. OpenCV face recognition was used to obtain bounding boxes around each face. Bidirectional warping of Active Appearance Model (AAM) [23] and a face alignment algorithm via regressing local binary features [26, 39] were used to extract 66 facial landmarks. The employed facial landmark localization techniques have been trained using the annotations provided from the 300W competition [28, 30, 29]. Images with at least one face with facial landmark points were kept for the next processing stages. A total of 119,481 images where kept. Other attributes of the queries were stored if applicable such as; intended emotion, gender, age, language searched, and its English translation if not in English.

On average 4000 images of each queried emotions were selected randomly, and in total 24,000 images were given to two expert annotators to categorize the face in the image into nine categories (i.e. No-face, six basic expressions, Neutral, None, and Uncertain). The annotators were instructed to select the proper expression category on the face, where the intensity is not important as long as the face depicts the intended expressions. The *No-face* category was defined as images that: 1) There was no face in the image; 2) There was a watermark on the face; 3) The bounding box was not on the face or did not cover the majority of the face; 3) The face is a drawing, animation, painted, or printed on something else; and 4) The face is distorted beyond a natural or normal shape, even if an expression could be inferred. The *None* category was defined as images that portrayed an emotion but the expression/emotions could be categorized as one of the six basic emotions or neutral (such as sleepy, bored, tired, seducing, confused, shame, focused, etc.). If the annotators where uncertain about any of the facial expressions, images were tagged as *uncertain*. Figure 3 shows some examples of each category and the intended queries written in parentheses.

The annotation was performed fully blind and independently, i.e. the annotators were not aware of the intended query or other annotator's response. The two annotators agreed on 63.7% of the images. For the images that were at a disagreement, favor was given to the intended query i.e. if one of the annotators labeled the image as the intended query, the image was labeled in the database with the intended query. This happened in 29.5% of the images with disagreement between the annotators. On the rest of the images with disagreement, one of the annotations was assigned to the image randomly. Table 1 shows the number of images in each category in the set of 24,000 images that were given to two human annotators. As shown, some expressions such as Disgust, Fear, and Surprise have few images compared to the other expressions, despite the number of queries being the same.

Figure 3. Sample of queried images from the web and their annotated tags. The queried expression is written in parentheses.

Table 1. Number of annotated images in each category

Label	**Number of images**
Neutral	3501
Happy	7130
Sad	3128
Surprise	1439
Fear	1307
Disgust	702
Anger	2355
None	403
Uncertain	280
No-face	3755

Table 2 shows the confusion matrix between queried emotions and their annotations. As is shown, happiness had the highest hit-rate (68%) and the rest of emotions had hit-rates at less than 50%. There was about 15% confusion with *No-Face* category for all emotions, as many images from the web contained watermarks, drawings etc. About 15% of all queried emotions resulted in neutral faces. Disgust and Fear had the lowest hit rate among other expression with 12% and 17% hit-rates respectively and most of the result of disgust and fear are mainly happiness or *No-Face*.

Table 2. Confusion Matrix of annotated images for different intended emotion-related query terms

	Happy	**Sad**	**Surprise**	**Fear**	**Disgust**	**Anger**	**Neutral**	**No-Face**	**None**	**Uncertain**
Happy	**68.18**	2.66	1.23	0.74	0.33	1.59	5.67	18.54	0.74	0.33
Sad	16.5	**42.42**	1.52	1.88	0.57	4.73	16.55	13.31	1.57	0.98
Surprise	27.6	6.31	**20.11**	5.62	1.07	4.85	17.1	14.73	1.65	0.96
Fear	18.74	10.91	6.49	17.69	1.47	6.39	13.92	**20.49**	2.22	1.67
Disgust	**26.71**	7.47	4.48	4.53	12.61	9.62	17.34	12.41	2.99	1.84
Anger	22.28	7.39	2.31	2.11	1.19	**30.59**	16.21	14.43	2.34	1.14

4. Training from web-images

The annotated images labeled with six basic expressions as well as neutral faces are selected from 24,000 annotated images (18,674 images). Twenty percent of each label is randomly selected as a test set (2,926 images) and the rest are used as training and validation sets. A total of 60K of not annotated images (10K for each basic emotion) is selected as noisy training set.

As baselines, two different deep neural network architectures are trained in three different training scenarios: 1) training on well-labeled images, 2) training on a mixture of noisy and well-labeled sets, and 3) training on a mixture of noisy and well-labeled sets using a noise modeling approach introduced in [38]. The network architecture we used in these experiments are AlexNet [13] and a networks for facial expression recognition recently published in WACV2016 in [21], called WACV-Net in the rest of this paper. All networks are evaluated on a well-labeled test set.

AlexNet consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers. To augment the data, ten crops of registered facial images of size 227x227 pixels are fed to AlexNet. We have tried a smaller version of AlexNet with smaller input images of 40x40 pixels and smaller convolutional kernel sizes, but the results were not as promising as the original model. WACV-Net consists of two convolutional layers each followed by max pooling, four Inception layers, and two fully-connected layers. The input images are resized to 48x48 pixels with ten augmented crops of 40x40 pixels. Our version of AlexNet performed more than 100M operations, whereas the WACV-Net performs about 25M operations, due to size reductions in Inception layers. Therefore, WACV-Net trained almost four times faster than AlexNet and consequently it had faster evaluation time as well.

In the first scenario, the network is trained on only well-labeled set with random initialization. In the second scenario (mixture of noisy and well-labeled sets), the network is pre-trained with only well-labeled data, and then trained on the mixture of the noisy and well-labeled sets. This increased about 5% in accuracy compared with training on the mixture of the noisy and well-labeled sets from scratch. In the last scenario (mixture of noisy and well-labeled sets using the noise modeling), as the posterior computation could be totally wrong if the network is randomly initialized [38], the network components are pre-trained with the well-labeled data. In addition, we bootstrap/upsample the well-labeled data to half of the noisy data. In all scenarios, we used a mini-batch size of 256. The learning rate is initialized to be 0.001 and is divided by 10 after every 10,000 iterations. We keep training each model until convergence.

Table 3. Recognition accuracy of AlexNet and WACV-Net on well-labeled test set with different training settings

	AlexNet	WACV-Net [21]
Train on well-labeled	82.12%	75.15%
Train on mix	69.03%	67.04%
Train on mix with noise estimation [38]	81.68%	76.52%

Table 3 shows the overall recognition accuracy of AlexNet and WACV-Net on the test set in three training scenarios. As shown, in all cases AlexNet performed better than WACV-Net. Training on mixture of the noisy and well-labeled data were not as successful as training on only well-labeled data. We believe that this was due to the fact that facial expression images crawled from the web are very noisy and in most expressions, less than 50% of the noisy data portray the intended query. The noise estimation approach can improve the accuracy of the network trained on the mixture of noisy and well-labeled sets. The best result is achieved from training AlexNet on well-labeled data. This gives slightly better overall accuracy (1%) than training on the mixture of noisy and well-labeled sets using noise modeling.

Table 4 shows the confusion matrix of AlexNet trained on the well-labeled set. Table 5 shows the confusion matrix of AlexNet trained on the mixture of noisy and well-labeled sets with noise estimation [38]. As shown in these tables, the noise estimation approach can improve the recognition accuracy of sadness, surprise, fear and disgust expressions. The reason is that there are fewer samples of these expressions in the well-labeled sets compared with other labels, and therefore including noisy data increases the training samples if the posterior distribution is estimated well. However, in some cases such as neutral faces and angry, training on only well-labeled data has higher recognition accuracy,

Table 4. Confusion matrix of AlexNet Trained on well-labeled

		predicted						
		NE	**HA**	**SA**	**SU**	**FE**	**DI**	**AN**
Actual	**NE**	**79.12**	6.73	9.98	0.46	0	0	3.71
	HA	6.37	**91.63**	1.14	0.29	0.14	0.07	0.36
	SA	14.52	5.24	**73.10**	0.24	0.48	0.71	5.71
	SU	10.59	6.47	1.18	**76.47**	3.53	1.18	0.59
	FE	4.14	3.45	7.59	15.86	**60**	2.76	6.21
	DI	2.41	4.82	8.43	2.41	1.2	**57.83**	22.89
	AN	8.6	2.87	5.73	1.79	0.36	5.73	**74.91**

* NE, HA, SA, SU, FE, DI, AN stand for Neutral, Happiness, Sadness, Surprised, Fear, Disgust, Anger respectively.

Table 5. Confusion matrix of AlexNet Trained mixture of noisy and well-labeled sets with noise estimation

		predicted						
		NE	**HA**	**SA**	**SU**	**FE**	**DI**	**AN**
Actual	**NE**	**65.20**	10.67	20.19	0.23	0	1.62	2.09
	HA	3.29	**91.56**	3.72	0.21	0.21	0.43	0.57
	SA	7.62	3.33	**84.29**	0.24	0.95	1.43	2.14
	SU	5.29	5.88	4.12	**76.47**	5.29	1.76	1.18
	FE	0.69	3.45	11.72	13.79	**63.45**	3.45	3.45
	DI	2.41	4.82	6.02	2.41	1.2	**68.67**	14.46
	AN	6.45	2.87	12.54	2.87	0.36	4.66	**70.25**

* NE, HA, SA, SU, FE, DI, AN stand for Neutral, Happiness, Sadness, Surprised, Fear, Disgust, Anger respectively.

as the prior distribution on the well-label set may not fully reflect the posterior distribution on the noisy set.

Figure 4 shows a sample of randomly selected images misclassified by AlexNet trained on the well-labeled and their corresponding ground-truth given in parentheses. As the figure shows, it is really difficult to classify some of the images. For example, we were unable to correctly classify the images in the first row. Also, the images in the second row have similarities to the misclassified labels, such as nose wrinkle in disgust, or raised eyebrows in surprise. It should be mentioned that classifying complex facial expressions as discrete emotions, especially in the wild, can be very difficult and even there was only 63.7% agreement between two human annotators.

5. Conclusion

Facial expression recognition in a wild setting is really challenging. Current databases with in wild setting are also either very small or have low resolution without facial landmark points necessary for pre-processing. The Internet is a vast resource of images and it is estimated that over 430 million photos are uploaded on only social network servers every day. Most of these images contain faces, that are captured in uncontrolled settings, illuminations, pose, etc. In fact it is Word *Wild* Web of facial images and it can be a great resource for capturing millions of samples with different subjects, ages, and ethnicity.

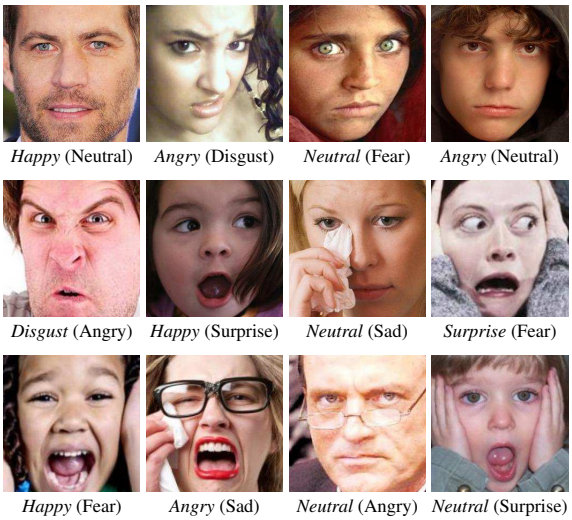

Figure 4. Samples of miss-classified images. Their corresponding ground-truth is given in parentheses.

Two neural network architectures were trained in three training scenarios. It is shown that, training on only well-labeled data has higher overall accuracy than training on the mixture of noisy and well-labeled data, even with the noise estimation method. The noise estimation can increase the accuracy in sadness, surprise, fear and disgust expressions, as there were limited samples in well-labeled data. But still training on only well-labeled data has a higher overall accuracy. The reason is that as annotations of web images showed, most of the facial images queried from the web have less than 50% hit-rates and even for some emotions such as disgust and fear, the majority of the results portrayed other emotions or neutral faces.

The whole database, query terms, annotated images subset, and their facial landmark points will be publicly available for the research community.

6. Acknowledgment

This work is partially supported by the NSF grants IIS-1111568 and CNS-1427872. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 3

[2] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Pro-*

ceedings of the 15th ACM on International conference on multimodal interaction, pages 509–516. ACM, 2013. 2

[3] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2106–2112. IEEE, 2011. 1, 3

[4] M. Drange. Why is this canadian hacker better than facebook at detecting gun photos? http://www.forbes.com/sites/mattdrange/2016/03/31/facebook-guns-beet_farmer-image-recognition/#23db8f4478ed, 2016. 1

[5] T. Gehrig and H. K. Ekenel. Why is facial expression analysis in the wild challenging? In *Proceedings of the 2013 on Emotion recognition in the wild challenge and workshop*, pages 9–16. ACM, 2013. 3

[6] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015. 2, 3

[7] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *EDM*, pages 43–50, 2013. 1

[8] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning. Local features based facial expression recognition with face registration errors. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE, 2008. 2

[9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 1, 3

[10] Instagram. Press page. https://www.instagram.com/press/?hl=en, 2016. 1

[11] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, 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*, pages 543–550. ACM, 2013. 3

[12] H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. In *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, volume 4, pages 3732–3737. IEEE, 1997. 2

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3, 5

[14] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Image processing, IEEE Transactions on*, 11(4):467–476, 2002. 2

[15] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013. 3

[16] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998. 1

[17] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013. 1, 2

[18] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, 2013. 1

[19] M. Mohammadi, E. Fatemizadeh, and M. H. Mahoor. Pca-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation*, 25(5):1082–1092, 2014. 2

[20] M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor. Intensity estimation of spontaneous facial action units based on their sparsity properties. 2015. 2

[21] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016. 3, 5

[22] A. Mollahosseini, G. Graitzer, E. Borts, S. Conyers, R. M. Voyles, R. Cole, and M. H. Mahoor. Expressionbot: An emotive lifelike robotic face for face-to-face communication. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, pages 1098–1103. IEEE, 2014. 2

[23] A. Mollahosseini and M. H. Mahoor. Bidirectional warping of active appearance model. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 875–880. IEEE, 2013. 4

[24] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, 2014. 3

[25] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005. 1

[26] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 4

[27] E. Rentzeperis, A. Stergiou, A. Pnevmatikakis, and L. Polymenakos. Impact of face registration errors on recognition. In *Artificial Intelligence Applications and Innovations*, pages 187–194. Springer, 2006. 2

[28] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 2015. 4

[29] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 4

[30] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 896–903, 2013. 4

[31] 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. 2

[32] S. Sukhbaatar and R. Fergus. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4, 2014. 3

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 3

[34] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 3

[35] Y. Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013. 3

[36] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014. 3

[37] J. Weston, F. Ratle, H. Mobahi, and R. Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012. 3

[38] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015. 2, 3, 5

[39] L. Yu. face-alignment-in-3000fps. https://github.com/yulequan/face-alignment-in-3000fps, 2016. 4

[40] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM, 2015. 3

[41] X. Zhang, M. H. Mahoor, and S. M. Mavadati. Facial expression recognition using {l} _ {p}-norm mkl multiclass-svm. *Machine Vision and Applications*, pages 1–17, 2015. 2

[42] X. Zhang, A. Mollahosseini, B. Kargar, H. Amir, E. Boucher, R. M. Voyles, R. Nielsen, and M. Mahoor. ebear: An expressive bear-like robot. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 969–974. IEEE, 2014. 2

[43] W. Zhen and Y. Zilu. Facial expression recognition based on local phase quantization and sparse representation. In *Natural Computation (ICNC), 2012 Eighth International Conference on*, pages 222–225. IEEE, 2012. 2

[44] Y. Zong, W. Zheng, X. Huang, K. Yan, J. Yan, and T. Zhang. Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis. *Journal on Multimodal User Interfaces*, pages 1–10, 2016. 3