Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning

Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, Stefan Winkler
Advanced Digital Sciences Center (ADSC)
University of Illinois at Urbana-Champaign, Singapore
{hongwei.ng, vietdung.n, bbonik, stefan.winkler}@adsc.com.sg

ABSTRACT

This paper presents the techniques employed in our team's submissions to the 2015 Emotion Recognition in the Wild contest, for the sub-challenge of Static Facial Expression Recognition in the Wild. The objective of this sub-challenge is to classify the emotions expressed by the primary human subject in static images extracted from movies. We follow a transfer learning approach for deep Convolutional Neural Network (CNN) architectures. Starting from a network pre-trained on the generic ImageNet dataset, we perform supervised fine-tuning on the network in a two-stage process, first on datasets relevant to facial expressions, followed by the contest's dataset. Experimental results show that this cascading fine-tuning approach achieves better results, compared to a single stage fine-tuning with the combined datasets. Our best submission exhibited an overall accuracy of 48.5% in the validation set and 55.6% in the test set, which compares favorably to the respective 35.96% and 39.13% of the challenge baseline.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*Neural nets*

General Terms

Algorithms

Keywords

Emotion Classification, Facial Expression Analysis, Deep Learning Networks

1. INTRODUCTION

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

The majority of existing techniques focus on classifying 7 basic (prototypical) expressions, which have been found to be universal across cultures and subgroups, namely: neutral, happy, surprised, fear, angry, sad, and disgusted [27]. More detailed approaches follow the Facial Action Coding System (FACS), attempting either to classify which Action Units (AU) are activated [24] or to estimate their intensity [18]. Fewer works follow the dimensional approach, according to which facial expressions are treated as regression in the Arousal-Valence space [29]. A very detailed and recent review can be found in [21].

The Emotion Recognition in the Wild (EmotiW) contest, and its Static Facial Expression Recognition in the Wild (SFEW) sub-challenge, follow the categorical approach of the 7 basic expressions. Images are selected from movies, in a semi-automated way, via a system based on subtitles [5, 6]. The challenging characteristics of SFEW are twofold. First, imaging conditions are close to real-life, including low and uneven illumination, low resolution, occlusions, non-frontal head-poses, and motion blur. Second, the size of the dataset is relatively small (≈1K/0.5K/0.3K for training/validation/testing), which makes it difficult to train large-scale models and thus, is prone to overfitting.

In order to overcome these challenges, researchers in the previous EmotiW contests have mainly used *fusion* of multiple features, coupled with different machine learning approaches. In [22], different kernels were learned for LPQ-TOP, audio, gist and SIFT features, and were combined in an SVM classifier. In [14], the optimal fusion of classifiers for HOG, dense SIFT, and deep convolutional features was learned based on a Riemannian manifold. In [23] audio, LPQ-TOP, LBP-TOP, PHOG and SIFT features were used along with a hierarchical classifier fusion method. In [3] HOG-TOP and audio features were fused using multiple kernel learning. In [13] convolutional and audio features were fused using Partial Least Squares and multiple classifiers. Finally, in [11] multiple Deep Convolutional Neural Network (CNN) were introduced for different data modalities (video frames, audio, human actions, mouth analysis), and different combination techniques for these models were explored.

In this paper we follow a *transfer learning* approach for deep CNN architectures, by utilizing a two-stage supervised fine-tuning, in the context of the SFEW sub-challenge. Starting from a generic pre-training of two different deep CNN architectures based on the ImageNet dataset, a first-stage fine-tuning is applied using the FER-2013 facial expression dataset [10], which comprises 28K/32K low resolution images of facial expressions, collected from the Internet using a set of 184 emotion-related keywords. A second-stage fine-tuning then takes place, based only on the *training* part of the EmotiW dataset, adapting the network weights to the characteristics of the SFEW sub-challenge. Both architectures were found to improve their performance through each of the fine-tuning stages,

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

DOI: http://dx.doi.org/10.1145/2818346.2830593.

while the cascade fine-tuning combination resulted in the submission with the highest performance.

The rest of the paper is organized as follows. Section 2 describes previous works which are related to our study. Section 3 describes in detail the proposed approach. Section 4 discusses the experimental results. Finally, concluding remarks are presented in Section 5.

2. RELATED WORK

Deep Learning-based approaches, particularly those using CNNs, have been very successful at image-related tasks in recent years, due to their ability to extract good representations from data. Judging a person's emotion can sometimes be difficult even for humans, due to subtle differences in expressions between the more nuanced emotions (such as sadness and fear). As a result, efficient features, finely-tuned and optimized for this particular task are of great importance in order for a classifier to make good predictions. It comes as no surprise that CNNs have worked well for emotion classification, as evidenced by their use in a number of state-of-the-art algorithms for this task, as well as winning related competitions [10], particularly previous years' EmotiW challenge [11, 14].

However, due to the small dataset size for the EmotiW 2015 image based static facial expression recognition challenge, it is easy for complex models like CNNs to overfit the data. To work around this problem of training a high-capacity classifier on small datasets, previous works in this area have resorted to using transfer learning across tasks, where the weights of the CNN are initialized with those from a network trained for related tasks before fine-tuning them using the target dataset [8, 9, 17, 26]. This approach has consistently achieved better results, compared to directly training the network on the small dataset, and is the one that we adopt in this paper as well.

3. EXPERIMENTAL SETUP

3.1 Data Preparation

For the EmotiW dataset, all faces were detected with OpenCV's Viola & Jones face detector (frontal and profile) [25]. The Intraface library [4] was used in order to detect 49 facial points. The fit of the alignment model, provided by Intraface, was used to discard false positives faces; any detection with a fit lower than 0.3 was considered a non-face. The average of the 6 points of the eyes (center of the eyes) was used in order to align the image in terms of rotation and scale.

The face bounding box was defined as a ratio of the eye-to-eye distance (e2e): The side boundaries were $0.62 \times$ e2e (counted from the respective eye corner); the upper and lower boundaries were $0.9 \times$ e2e and $1.35 \times$ e2e, respectively. These particular values were selected for three main reasons. First, they result in an approximately square bounding box. This is important, since the DNN training platform requires images to be square. If this condition is not met, images will be stretched, thus affecting their aspect ratio. Second, they ensure that a considerable part of the forehead is included, which exhibits potentially richer visual information for facial expressions, while discarding pixels located below the mouth, with limited visual information. Third, this particular approach resembles the cropping of faces in the FER-2013 dataset, which was used for pre-training, thus increasing the consistency between the two datasets. Finally, all EmotiW images were converted to grayscale, re-sized to 256×256 and normalized using min-max intensity normalization, stretching their intensity values to [0,255]. Fig 1 depicts our cropping method in comparison with the one provided by the organizers.

Figure 1: Comparison between our cropping method and the one provided by the organizers.

Regarding the FER-2013 dataset, the small size of its images (48×48 pixels) prevented the reliable detection of facial points. As a result, no alignment was used. Nevertheless, even without aligning these faces, and with their size being much smaller than those in the EmotiW target dataset (48×48 vs. 256×256), we observed a significant performance boost when using them for pre-training (see section 4). Figure 2 depicts a comparison between these two datasets for the 7 classes. Note the similarity in the cropping between the FER-2013 and our EmotiW images.

3.2 Architectures

The success of CNN for face emotion recognition motivated us to base our models on two representative CNN architectures, which we chose because of their nice tradeoffs between speed and accuracy [2]:

1. The ILSVRC-2012 [19] winning entry of [12] (AlexNet).
2. The CNN-M-2048 model from [2] (VGG-CNN-M-2048), which is a variant of the model introduced in [28].

3.3 Dataset and Training

As noted above, it is challenging to train a complex model such as a CNN using only a small amount of training data without overfitting. Our approach to tackling this problem follows recent works [2, 8, 9, 26], which consistently show that supervised fine-tuning with a relatively small dataset on a network pre-trained with a large image dataset of generic objects (e.g., ILSVRC) can lead to significant improvement in performance.

Specifically, we used the the FER-2013 face expression dataset introduced in the ICML 2013 workshop's facial expression recognition challenge [10] as auxiliary data to fine-tune the respective CNNs that were trained using the ILSVRC-2012 data for the architectures mentioned above.

The FER-2013 dataset comprises 3 parts; **a.** The *Original Training Data* (OTD – 28709 images), **b.** the *Public Test Data* (PTD – 3589 images), used when the competition was ongoing to provide feedback on the accuracy of participant's models, and **c.** the *Final Test Data* (FTD – 3589 images), used at the end of the competition to score the final models. We generated 3 variants using these parts of the FER-2013 dataset:

1. **FER28**. This consists of the OTD for training and the PTD for validating our fine-tuned models.
2. **FER32**. This consists of a combination of both the OTD and PTD (28709+3589=32298 total images) for training and the FTD for validating our fine-tuned models.

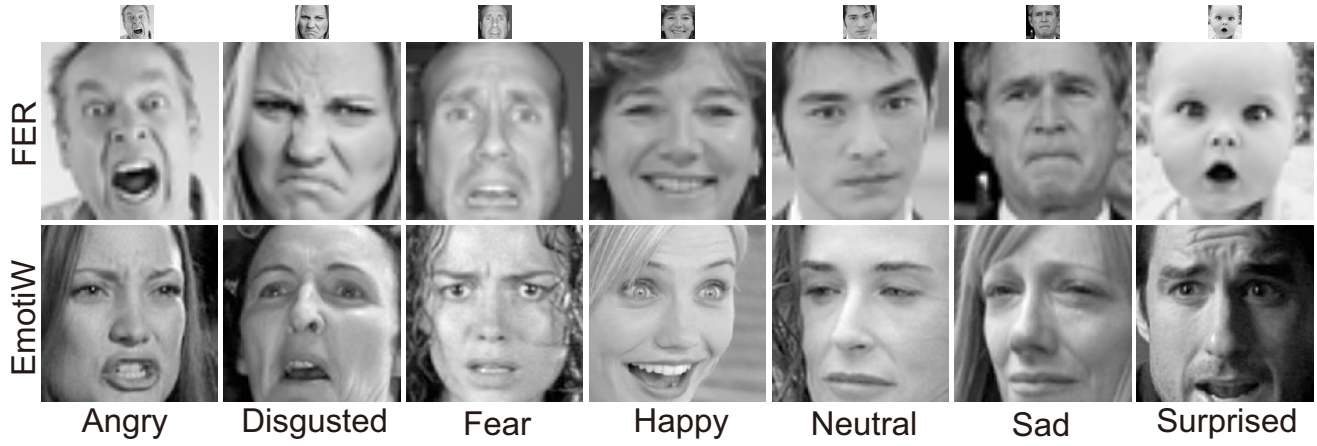

Figure 2: Comparison between the FER-2013 and EmotiW datasets. Top row: original size of the FER-2013 dataset (48×48 pixels). Middle row: upsampled FER-2013 dataset to 256×256 pixels. Bottom row: EmotiW dataset (256×256 pixels).

3. **FER32 + EmotiW**. This consists of a combination of both the FER32 and the EmotiW training data for training and *only* the EmotiW validation data for validating our fine-tuned models.

We experimented with different schemes for fine-tuning the base pre-trained CNN model using these datasets in combination with the EmotiW training data. These included directly fine-tuning the EmotiW dataset on the CNNs pre-trained on ILSVRC-2012, as well as a "staged" fine-tuning, where we fine-tuned first using data from the FER-2013 dataset before fine-tuning again with the target EmotiW dataset.

The training procedure for our CNNs closely follows that of [12]. They are trained using stochastic gradient descent with hyperparameters (momentum=0.9, weight decay=0.0005, initial learning rate=0.001). Note that since we are fine-tuning CNNs pre-trained on the ILSVRC-2012 dataset [19] using much smaller datasets, we set an initial learning rate of 0.001, which is lower than the typical 0.01 [2, 12], so as not to drastically alter the pre-trained weights. We found that doing so delays the onset of overfitting (as observed from the point where validation loss and training loss start to diverge), allowing the CNN to achieve higher accuracy on the validation data. The learning rate is dropped by a factor of 10 following every 10 epochs of training.

4. RESULTS AND DISCUSSION

The results of our submissions to the EmotiW 2015 SFEW challenge for the validation and test set, as well as some of our experiments that were not submitted, are summarized in Figure 3. The corresponding confusion matrix generated from the predictions of our classifier on the *test* set is shown in Figure 4.

4.1 Effects of Supervised Fine-tuning

Fine-tuning our CNN models using the auxiliary FER-2013 dataset in general led to a 10% increase in accuracy over the baseline method [7] on the test set (the exception being submission 2 with a gain of 7%). A further round of fine-tuning these models using the target EmotiW dataset typically improves the accuracy on the test data by another few percentage points. For our best model (submission 3), this was an overall increase in accuracy of more than 16% over the baseline method (55.6% vs 39.13%) on the test set.

Directly fine-tuning the networks with the EmotiW training data gave an accuracy of 37.8%. While this is slightly worse than that achieved by the baseline method, it is much lower than what we obtained using the auxiliary FER-2013 dataset, which hints at the difficulty of fine-tuning deep neural networks with small datasets, and the importance of using auxiliary data.

However, we note that if we were to train a CNN using data combined from the FER-2013 and EmotiW training set (submission 9), the performance on the test set will only be around 1.5% lower than if we had separated the two datasets and fine-tuned them successively (submission 8). This suggests that the relatively small EmotiW training set (which consists of 921 images) only had a marginal effect in improving performance of a CNN that had already been fine-tuned it on a fairly large dataset such as the FER-2013. We discuss this observation further in the next section.

4.2 Effects of More Labeled Data

Interestingly, even though some of our models were fine-tuned only on the FER-2013, without any data from the target EmotiW dataset, their results on the test set were competitive with those that underwent a second round of fine-tuning on the target dataset. Often the difference in accuracy is within a few percentage points (Figure 3: Submission 1 vs. 3, 6 vs. 8). Furthermore, for our best model that did not use any data from the target EmotiW dataset (Submission 1), its performance on the test set was close to 15% higher than the baseline method. This is surprising because the images in the FER-2013 dataset were thought to be less than ideal as they were not aligned, unlike the target EmotiW dataset, and much smaller in size (48×48 vs 256×256). Yet, models trained using only the FER-2013 dataset were able to give good predictions on the EmotiW dataset. This suggests that the *quantity* rather than quality of the data could be more important for improving the performance of CNN-based face expression classifiers at the early stages of building such classifiers.

The observation that fine-tuning with a low resolution (but larger) dataset has a positive impact on the overall performance is also inline with the latest neuroscientific findings. Mounting evidence suggests that facial expression recognition in humans is *holistic* rather than feature-based, especially in the case of photographs (but not for schematics) [15, 16]. As a result, even people with impaired high frequency vision can distinguish successfully between different expressions, even though they rely only on low frequency visual information [1]. In our case, the lower frequency (but larger) FER-2013 dataset seems to tune the network weights in a way that allows

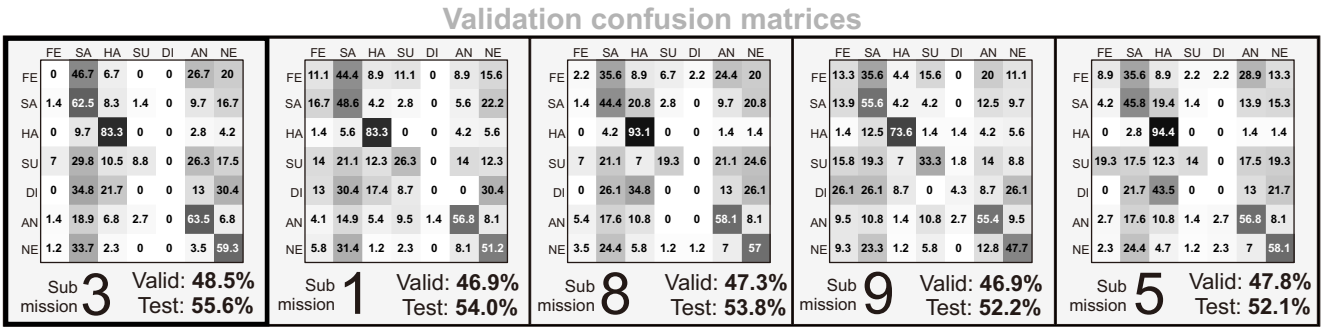

Figure 5: Confusion matrices generated from the EmotiW *validation* set for our 5 best submissions.

a better generalization over data with higher visual frequency, such as the EmotiW dataset.

4.3 Difficulties With Certain Expressions

We observe that some classes appear to be "harder" to train in the sense that, (1) none of our models were able to score highly for them, and (2) our model predictions for those classes tend to fluctuate depending on our training scheme and architecture. This appears to be the case for classes such as "disgust", "fear", "surprise", and "sad". We believe the reasons for this could be twofold. First, these classes tend to have much fewer training samples compared to classes such as "happy", "angry", and "neutral", making it harder to train the CNN to recognize them. Figure 6 depicts the size of the 7 classes on the EmotiW training set, where disgust, fear and surprise are the 3 smallest classes in terms of training samples. Second, these expressions can be very nuanced, making it difficult even for humans to agree on their correct labeling [20].

We suspect that the inherent difficulty in assigning labels to some of the samples may have caused them to be "mislabeled", thereby affecting the models that were trained on them. If this was the case, it would explain why our models (and the organizer's [7]) consistently perform better on the test data than on the validation data. A possible explanation is that samples for the "tricky" cases might (for some reason) be less ambiguous in the test data than they are in the validation data, which would make it less likely that one emotion will be "wrongly" predicted for another, hence causing the accuracy obtained for the test data to be higher than those obtained for the validation data. We arrived at this explanation by comparing the corresponding confusion matrices generated by our top 5 submissions for the EmotiW validation (Figure 5) and test data (Figure 4). Comparing Figures 5 and 4, we observe a general increase in accuracy for the "tricky" classes for all 5 models except that for submission 5, usually at the expense of the "happy" class. Furthermore, we note that "surprise" class, in general, appears to have the largest gain in accuracy (between 7.4% to 33.2% improvement). We also note that this improvement in accuracy from the validation to the test data is also observed in the organizer's baseline paper [7], where their approach improved from an accuracy of 35.96% on the validation data to 39.13% on the test data.

Lastly, we note that all except one of our models were unable to predict a single sample labeled "disgust" correctly. A reason for this could be an imbalance in the training datasets. Indeed, as Figure 6 indicates, "disgust" has the fewest samples in the EmotiW training dataset (comprising 7% of the whole dataset) and in our FER28 and FER32 training datasets (comprising 2% of each of these two datasets). The imbalance in the number of training samples for each class of emotions most likely caused our models to overfit on the emotions with more samples (e.g., "happy") at the expense of this class. Furthermore, the expression of disgust is very subtle, which means that it will be hard for a our CNN models to discover features to robustly distinguish this emotion from other similarly nuanced expressions such as sad, fear and neutral. This can be verified by examining the confusion matrices in Figure 4, which indicates that the "happy" class is often the highest scoring class, and that the classes "disgust", "sad", "fear", "neutral" are often mistaken for each other by our models. The combination of these two factors makes it even harder to train models to predict this emotion accurately. The above observations highlight the difficulty in training CNNs using a small unbalanced dataset with classes that are not visually distinctive.

5. CONCLUSIONS

We have shown that it is possible to obtain a significant improvement in accuracy (up to 16%) over the baseline results for expression classification on the EmotiW dataset using CNNs fine-tuned initially on auxiliary face expression datasets, followed by a final fine-tuning on the target EmotiW dataset. We also showed that, at least for the EmotiW dataset, its small size does not favor it for being used for training CNNs. However, CNNs trained on sufficiently large auxiliary face expression datasets alone can be used to obtain results much better than the baseline, without using any data from the EmotiW dataset. Furthermore, any additional improvement from using the EmotiW dataset, when a sufficiently large face dataset such as FER-2013 is available, whether by adding it to the auxiliary dataset or another round of fine-tuning, is likely to be marginal owing to its small size. This suggests that if we were to exploit deep neural networks such as CNN for face expression recognition to achieve the significant gains seen in other domains, then having bigger datasets is crucial.

Lastly, we also noted the inherent difficulty in assigning correct labels to faces depicting some of the more nuanced emotions, and how that can affect the performance of our models. This suggests that a cost-sensitive performance measure that penalizes a model mistaking samples from one class to another in "similar" classes less harshly might be more appropriate than the binary accuracy measure used for this challenge.

6. ACKNOWLEDGMENTS

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

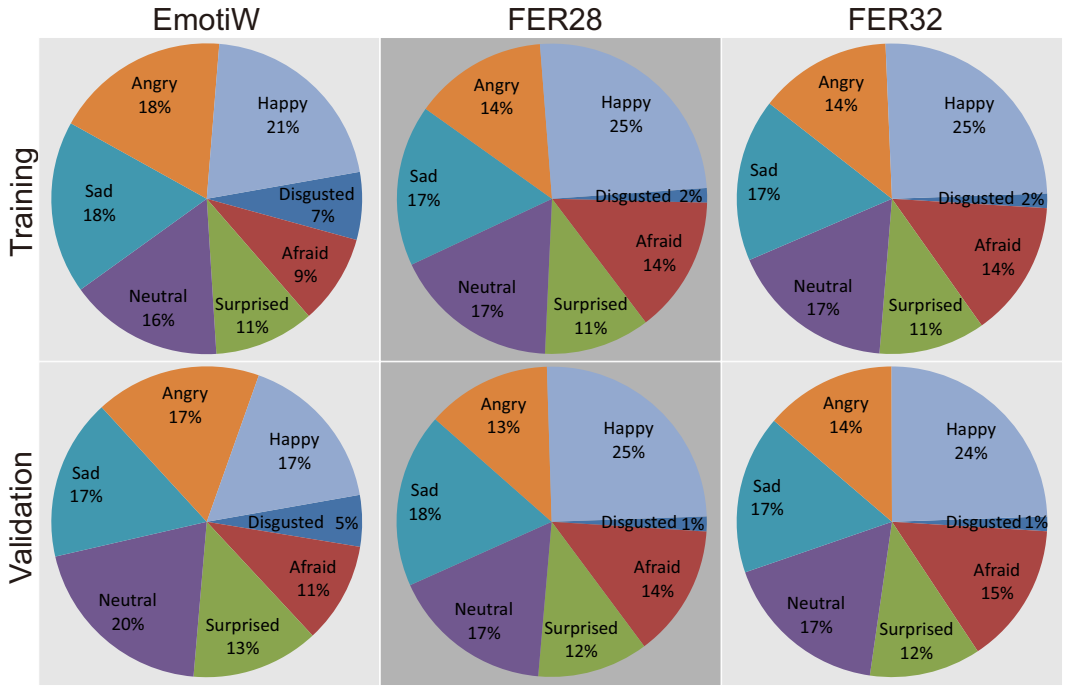

Figure 6: Distributions of the 7 classes on the EmotiW, FER28 and FER32 training and validation sets.

7. REFERENCES

[1] M. Boucart, J.-F. Dinon, P. Despretz, T. Desmettre, K. Hladiuk, and A. Oliva. Recognition of facial emotion in low vision: A flexible usage of facial features. *Visual Neuroscience*, 25(4):603–609, 2008.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[3] J. Chen, Z. Chen, Z. Chi, and H. Fu. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 508–513, New York, NY, USA, 2014. ACM.

[4] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn. Intraface. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pages 1–8, May 2015.

[5] 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, Nov 2011.

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

[7] A. Dhall, R. Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 17th International Conference on Multimodal Interaction*, ICMI '15. ACM, 2015.

[8] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition, 2014.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[10] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59 – 63, 2015. Special Issue on 'Deep Learning of Representations'.

[11] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu. 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, New York, NY, USA, 2013. ACM.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[13] 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, New York, NY, USA, 2013. ACM.