Leveraging large face recognition data for emotion classification

Boris Knyazev[‡*], Roman Shvetsov[‡], Natalia Efremova[§] and Artem Kuharenko[‡]
[‡]*NTechLab, Russia*
[§]*University of Oxford, UK*
Email: borknyaz@gmail.com, r.shvetsov@ntechlab.com, natalia.efremova@gmail.com, a.kuharenko@ntechlab.com

***Abstract*—In this paper we describe a solution to our entry for the emotion recognition challenge EmotiW 2017. We propose an ensemble of several models, which capture spatial and audio features from videos. Spatial features are captured by convolutional neural networks, pretrained on large face recognition datasets. We show that usage of strong industry-level face recognition networks increases the accuracy of emotion recognition. Using our ensemble we improve on the previous year's best result on the test set by about 1%, achieving a 60.03% classification accuracy without any use of visual temporal information, showing a top-2 result in this challenge.**

***Keywords*-emotion recognition; video processing; image processing; large data; convolutional networks; classification**

I. Introduction

Emotion recognition potentially has many applications in academia and industry, and emotional intelligence is an important part of artificial intelligence. However, in contrast to such tasks as face recognition (FR), emotion recognition has not yet become so widespread. We believe that the reason for this is the fact that emotion recognition is much harder and requires more research and efforts to gain success. Face recognition is also hard, but training data with clean ground truth labels can be collected easier and benchmarks are usually objective (i.e. we know the identity). In emotion recognition, there is a lack of understanding and the agreement of what the labels should be. This can be proved by recent appearance of datasets with compound emotions [4] or with dominant and complementary emotions [12]. There is also a lack of training data due to difficulty of collecting rare emotions (how often do you clearly show fear?).

Emotion recognition from video is also more difficult than general video recognition. State of the art methods such as (Improved) Dense Trajectories [18] or 3D convolutional neural networks, which typically show a 70-90% accuracy on video datasets, fail to provide results above 40% on emotion datasets [19], [9].

The emotion recognition challenges, in particular EmotiW 2017 [6] and its predecessors [5], [7], allow to boost the progress in this area by providing data and benchmarks for training and evaluating novel methods. Once emotions can be recognized reliably and well understood, it can provide the same or even more benefits than face recognition. Due to the presence of concealed and deceptive emotions, it can lead to even more benefits, because humans need expert and rare knowledge to recognize concealed emotions, while machines could potentially perform this task easily [14], opening up new research areas as well as new privacy challenges analogously to face recognition.

In this work, we attempt to further contribute to the field of emotion recognition by presenting our solution to the fifth Emotion Recognition in the Wild Challenge (EmotiW) 2017, in particular to its audio-video emotion recognition sub-challenge.

Recent NIST reports show that face recognition networks of NTechLab are state-of-the-art in the wild images benchmark [16] and, in this work, we were able to employ them for emotion recognition by fine-tuning them on the emotion datasets.

Similarly to other video recognition challenges [1], we make features extracted from facial images of all frames publicly available. The code to reproduce our results using these features is also made publicly available on `https://github.com/bknyaz/emotiw`.

II. Methods

A. Networks

We experiment with four deep convolutional neural networks: VGG-Face [15] and three proprietary state of the art face recognition networks which we notate as FR-Net-A, FR-Net-B and FR-Net-C. Compared to VGG-Face, which is trained on 2600 individuals with around 3 million images, those networks are trained on a much larger data volume (Table I), which make them more powerful both for face and emotion recognition tasks (Table II).

B. Pipeline

The pipeline of our emotion recognition system is illustrated on Figure 1 and outlined below.

1) Face detection: To extract and align faces both from images of FER2013 and EmotiW video frames we use the dlib face detector [11]. Since videos in the EmotiW dataset are taken from movies and reality shows, some of them are challenging for face detectors due to poor light conditions, severe occlusions, variations in pose, makeup,

*Currently at the University of Guelph, Canada.

978-1-5386-2335-0/18/$31.00

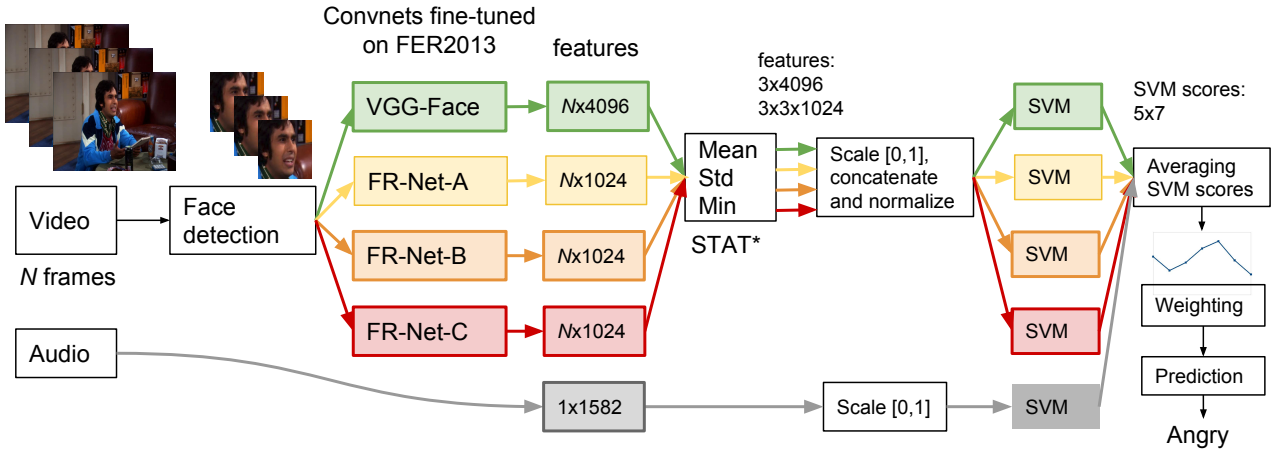

Figure 1. The pipeline of our emotion recognition system. For each video frame we first detect a face and feed the face image to each of the four networks, so that activations from one of the last layers are used as frame features. Then, for each network given its features for all video frames, we compute statistics STAT* (Section II-B3), which is invariant to the number of frames N, then each of the STAT* feature vectors is scaled to the range $[0, 1]$, followed by concatenation and normalization. For each network and for audio features we train a linear SVM with probabilistic calibration. During the test phase we average scores from all 5 SVMs, weight them according to the emotion distribution from Table V and predict the emotion with the highest probability.

facial accessories and other factors. Therefore, if a face was not found on the frame, the entire frame is passed to the network for two reasons. First, the network still could capture some contextual cues given an entire frame, and second, for a few videos all face detections were false positives because of the challenges mentioned above. To limit the number of cases when an entire frame is fed to the network, we apply a low face detection threshold.

2) Frames feature extraction: In all experiments we follow the same pipeline to obtain results on EmotiW validation videos. First, features for all frames are computed using all four networks. For VGG-Face, we empirically choose 4096 dimensional *fc6* features (after the first fully connected layer), while for other networks we use 1024 dimensional features of the layer before the last one.

3) Frame-level feature aggregation: Motivated by results of the previous year using statistical encoding (STAT) [3], we compute features of videos with one or several aggregation functions (e.g., mean or standard deviation) followed by rescaling to range $[0, 1]$ and concatenation (Table III). For instance, in case of VGG-Face if we compute *mean*, *std* and *min* features, which we denote as STAT*, then features are 12288 dimensional. In this work, *rootsift* normalization $(\text{sign}(\mathbf{x})\sqrt{|\mathbf{x}|/\|\mathbf{x}\|_1})$ [2] and global standardization is also applied to concatenated features. STAT and STAT* return a video representation invariant to the number of frames, so that we can feed it to an SVM, which requires features of the same length for all training and test data.

4) Classification: A linear SVM was trained on training data (one SVM per network) in case of reporting validation accuracy and, as in [9], on training plus validation data in case of test accuracy. The regularization constant of SVMs is found by 5-fold cross-validation.

5) Ensembling with audio features: We compute 1582 dimensional audio features using the Opensmile library [8]. We train a linear SVM in this case as well, so that our ensemble averages scores of 5 SVMs in total (Figure 1).

Table I
TRAINING DATASETS USED IN THIS WORK. FOR NTECHLAB'S FACE RECOGNITION DATA WE ONLY PROVIDE THE NUMBER OF IMAGES.

Dataset	#classes	#images
FER 2013	7 emotions	35k
EmotiW 2017 (video)	7 emotions	50k frames
VGG-Face data	2600 persons	3m
NTechLab FR data	-	50m

Table II
MODEL COMPARISON USING FRAMES FEATURE AVERAGING ON THE VALIDATION SET BEFORE AND AFTER FINE-TUNING (FT) ON FER2013. BEST CLASSIFICATION ACCURACIES IN EACH COLUMN ARE **IN BOLD**.

Model	Before FT	After FT	# features
VGG-Face	**37.9**	48.3	4096
FR-Net-A	33.7	44.6	1024
FR-Net-B	33.4	**48.8**	1024
FR-Net-C	37.6	45.2	1024
Audio features	35.0	-	1582

Table III
FEATURE COMPARISON ON THE VALIDATION SET FOR MODELS FINE-TUNED ON FER2013. STAT* IS STAT WITHOUT *max*. ** - OUR BEST SUBMISSION. F - FOURIER FEATURES, A - AUGMENTATION. ENSEMBLE: VGG-FACE + FR-NET-A+B+C + AUDIO. BEST CLASSIFICATION ACCURACIES IN EACH ROW ARE **IN BOLD**.

Model	Mean	STAT	STAT*	STAT*+A	STAT*+f	STAT*+f+A
VGG-Face	48.3	52.2	52.5	50.4	**53.0**	50.7
FR-Net-A	44.6	47.8	47.5	49.3	48.3	**49.6**
FR-Net-B	48.8	52.5	52.2	52.5	53.3	**53.5**
FR-Net-C	45.2	45.2	45.7	**53.0**	45.2	52.5
Ensemble	52.7	55.1	54.8	56.4**	56.4	**56.7**

III. EXPERIMENTS

A. Datasets

In this work, we use two emotion datasets to train the models (Table I): FER2013 [10] and data of this challenge EmotiW 2017. The FER2013 dataset was also used in the previous year's winning method [9]. It consists of 28709 training, 3589 validation and 3589 test images. We fine-tune the models using only a training set.

In the audio-video emotion recognition sub-challenge of Emotiw 2017 there are the same training (773) and validation (383) videos as in the EmotiW 2016 version, but this year 60 new test videos were added, which makes a total of 653.

B. FER2013 fine-tuning

All four networks are fine-tuned on FER2013 by replacing the last few fully connected layers with new ones (for VGG we only replace the classification layer) and then training all layers for 30k iterations with Nesterov SGD [13], [17] with the following parameters: learning rate 0.0001-0.0005, weight decay 0.0005, momentum 0.9, batch size 32 and polynomial learning rate decay. The networks achieved 70-72% accuracy on FER2013 validation data, but we did not analyze these results in depth, because the relationship between the validation accuracies on FER2013 and EmotiW was not always direct, perhaps due to differences between the two datasets and noisy labels in the former one.

As expected, fine-tuning (FT) on FER2013 boosts performance on EmotiW in all cases (Table II). Before FT better FR models usually have worse performance in the emotion recognition task, because face recognition should be emotion invariant, but after FT the results invert. For instance, FR-Net-A and FR-Net-B have the same architecture, but the latter was pretrained on much larger face recognition data. As a result, FR-Net-B is the worst before fine-tuning, but the best afterwards. This experiment confirms that pretraining on larger FR data positively affects emotion recognition accuracy.

C. Feature comparison

To improve STAT features used in [3], which include mean, standard deviation, minimum and maximum features we first performed an ablation study and removed the *max* features (Table III), denoting it as STAT*. This appeared to be important for improving generalization on the test set.

Afterwards, we found that frame-based augmentation in the form of horizontal flipping, rotation and scaling usually improves features except for VGG-Face. We compute 18 transformations per frame and average features of these transformations.

We then tried to add spectral features by computing the one dimensional Fourier transform (fft) for each neuron and then taking the average of that. For instance, for VGG-Face for one video we have 4096 dimensional fft features. These features significantly improved performance on the validation set (Table III), however, due to the limit of test submissions, we were not able to evaluate this and several other features.

D. Class distribution

In the EmotiW 2017 data we noticed that fitting the validation set did not contribute to the test performance (in some case it was even harmful), because the validation and tests sets are significantly different. Specifically, the validation set is relatively balanced, i.e. each emotion has about the same number of samples, while according to the confusion matrix (provided for submissions) of the test set it is imbalanced with many more samples of happy, neutral and angry emotions (Table V). Therefore, if during fitting of the validation set the model that predicts all emotions equally well will be chosen, that would decrease performance on the test set, because on the test set not all emotions are equally important.

This observation enables us to significantly improve results by weighting scores of emotions according to the square root of the observed test set frequencies. This distribution was available to all challenge participants and could be implicitly or explicitly exploited, which could be hard to check. We believe it is fair to use all information available to participants, because others can use it. Once such a distribution or any other useful information about test data is known, it is typically considered to be not test data anymore.

Due to weighted scores, our model started to have a high false negative error for disgust and surprise emotions, but a low false positive error for happy and neutral facial expressions - these two emotions make up more than 50% of test data (Figure 2). The proposed weighting enables the model to achieve a higher test classification accuracy while making a validation accuracy very low (Table IV).

E. Frame shuffle augmentation for LSTM

In this auxiliary experiment, conducted after the final test submission, we trained a Long Short-Term Memory (LSTM)

Table IV
COMPARISON OF OUR ENSEMBLES WITH OTHER RESULTS. ENSEMBLE: VGG-FACE + FR-NET-A+B+C + AUDIO. A - AUGMENTATION.

Model	Val acc	Test acc
Baseline [6]	38.81	41.07
EmotiW 2016 best results	59.42 [3]	59.02 [9]
Ensemble	54.83	-
Ensemble + A	56.40	54.98
Ensemble + A + class weights	48.30	60.03

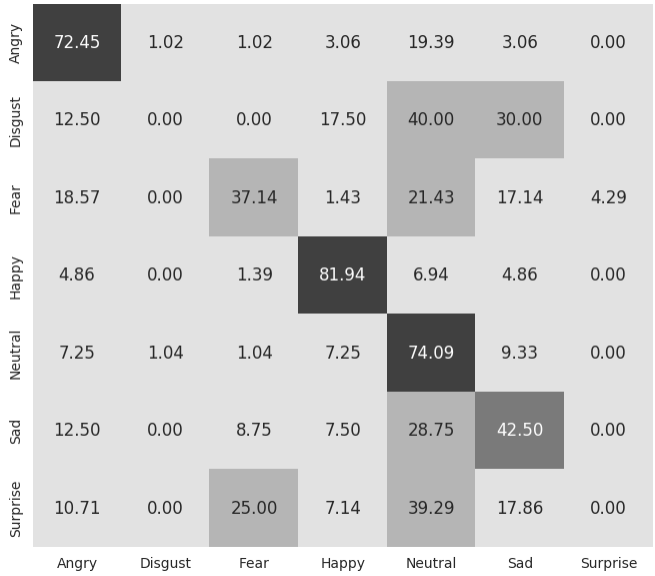

Figure 2. The confusion matrix of our final predictions on the test set. Rows are ground truth labels, columns are predictions. 28.75% for row 'Sad' and column 'Neutral' means that among all true sad emotions in the test set, 28.75% of them were classified as neutral by our model.

by randomly shuffling the order of frames during training (Table VI), which can be seen as a form of augmentation. Although it sounds counterintuitive, we achieve considerable accuracy gain compared to training on frames in the original order (during inference the order is fixed). Our results imply that each video in this task is not a *sequence* of video frames, but rather a *set* of frames. Manual examination of videos was consistent with our finding, because after shuffling frames of some video we still could (or still couldn't) recognize the correct emotion. LSTMs are trained on top of the FR-Net-B network on full sequences analogously to [9]. However, we were not able to improve our performance on the test set by adding LSTMs.

Table V
TRAINING, VALIDATION AND TEST SET EMOTION DISTRIBUTION IN EMOTIW 2017.

	An	Di	Fe	Ha	Ne	Sa	Su	Total
Train	133	74	81	150	144	117	74	773
Validation	64	40	46	63	63	61	46	383
Test	98	40	70	144	193	80	28	653
Class weights	0.15	0.10	0.13	0.19	0.21	0.14	0.08	1

Table VI
FOR THIS TASK THE ORDER OF FRAMES IS NOT IMPORTANT, SO THE VIDEOS CAN BE SEEN AS UNORDERED SETS OF FRAMES AND RANDOM SHUFFLING DURING TRAINING INCREASES VALIDATION ACCURACY.

Model	Val acc
LSTM + FR-NET-B	46.48
LSTM + FR-NET-B + frame shuffling	50.39

IV. CONCLUSIONS

In this work, we present an ensemble of models that achieves better emotion classification accuracy than the previous year's winner. We rely on strong face recognition convolutional networks which can be easily fine-tuned to perform the emotion recognition task. Audio features are also used to complement the visual models with an additional modality. We make frame level features computed with the convolutional networks publicly available to help the research community by reducing the task of emotion recognition from video to learning from high level features.

We showed that by leveraging large data good results can be obtained relatively easily in contrast to complex methods proposed earlier. In our view, one of the primary direction for the emotion recognition research community could be collection of larger and better data rather than developing complex methods tuned for a particular dataset. We hope that our findings and results will be useful to further develop and improve the emotion recognition field.

ACKNOWLEDGMENT

The authors would like to thank other members of the NTechLab team for useful advice and technical support.

REFERENCES

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2911–2918. IEEE, 2012.

[3] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 433–436. ACM, 2016.

[4] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*, 2017.

[5] A. Dhall et al. Collecting large, richly annotated facial-expression databases from movies. 2012.

[6] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017.

[7] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 427–432. ACM, 2016.

[8] 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.

[9] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450. ACM, 2016.

[10] 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. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[11] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[12] I. Lüsi, J. C. J. Junior, J. Gorbova, X. Baró, S. Escalera, H. Demirel, J. Allik, C. Ozcinar, and G. Anbarjafari. Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 809–813. IEEE, 2017.

[13] Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2).

[14] I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari. Automatic recognition of deceptive facial expressions of emotion. *arXiv preprint arXiv:1707.04061*, 2017.

[15] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[16] K. H. Patrick Grother, Mei Ngan. Ongoing face recognition vendor test (frvt). August 8 2017.

[17] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[18] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[19] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen. Holonet: towards robust emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 472–478. ACM, 2016.