# You Only Feel Once

Assignment #3

Image Based Biometrics 2019/20, Faculty of Computer and Information Science, University of Ljubljana

Barbara Aljaž (63140002)

***Abstract*—Human emotions are something that cannot be described easily. It is difficult to model something that can be very specific for each of us but some common points had been established during the years of research. This vagueness is also the reason that automatic emotion detection from images is hard. In this paper we studied only the detection of facial expression which is one of the basis for further research of human emotion. We wanted to do it quickly and this is the reason we chose Yolov3-tiny architecture. Our model is not as accurate as current state-of-the-art detectors but it can run in real time.**

## I. Introduction

Automatic facial expression recognition offers us many possibilities for human-computer interaction. With recognition of human emotions we can expand the communication with additional information that were obtained non-verbally and also not with direct interaction. With detection of human emotions we can also improve security systems and assure higher level of safety for the person in question and others (for example recognition of possible suicide attempts on bridges). But recognition of human emotional state is complex and contains many aspects. It is hard to define it and many models had been proposed. One of the most basic one is the one proposed by psychologists Paul Ekman and Wallace V. Friesen in 1970s which describe 6 basic human emotions: happiness, sadness, disgust, fear, surprise, and anger. But the scientist don't agree about how many emotions are there really and which are they. Nevertheless this basic model is basis for many facial expression recognition algorithms and we too used it, with addition of facial expression for neutral emotion. Our intention was to learn the model that could detect facial expression of people on images in real time and that is the reason we used neural network Yolov3-tiny which is capable of high speed detections but with the cost of lower detection accuracy.

## II. Related Work

The problem of automatic face expression recognition has been addressed by many different methods. More traditional methods use geometric features, e.g. facial landmarks are calculated and then relations between them are studied like in [1], or appearance features, e.g. by using local binary pattern histograms, like in [2]. In addition to processing 2D images, some methods also used the depth information, e.g. in [3] they used the color image and depth image from Kinect sensor.

In computer vision there were large improvements at different tasks in last decade because of convolutional neural networks (CNN) and recurrent neural networks (RNN). In [4] both were used, they first implemented simple feed-forward convolutional neural network to recognize facial expression from images of faces but they also studied the possibility of facial expression recognition from video by detection of micro-expressions. For that part they used a long-short-term-memory recurrent neural network (LSTM). In [5] they used the combinations of two CNNs, one was trained on images, the other on geometric landmarks of faces. In [6] they constructed special network that uses the region layer that captures local appearance changes for different facial regions. Also human have sometimes problems with facial expression recognition from single image because emotions are dynamic and because of that systems that are based on processing of image sequences were developed. In [7] they proposed the hybrid RNN-CNN framework that outperformed the methods based solely on CNNs. The problem of the large amount of inter-personal variations such as expression intensity variation was addressed in [8], where they used LSTM to learn the characteristics of facial expressions. In [9] they proposed network that used covariance pooling because of larger importance of how facial landmarks are distorted than presence or absence of specific landmarks. Methods that are based on RNNs don't take into account the importance of frames and that was addressed in [10] where they created Frame Attention Networks constructed of self-attention kernels that were directly learned from frame features and relation-attention kernels that were learned from video-level features and frame features. When approaching the problem of facial expression detection, most methods first extract faces from image and then perform recognition, but in order to avoid the face detection and speed up the inference, the use of Faster R-CNN was proposed in [11]. Methods that perform detection and recognition in single network don't reach the accuracy of current state-of-the-art algorithms for facial expression recognition but allow the lower inference time.

## III. Methodology

We used neural network architecture Yolov3-tiny that is based on Yolov3 architecture described in [12]. Yolo (You Only Live Once) network is fully convolutional network but is different from other networks in a sense that single network is used to predict bounding boxes of detected objects and to do classification. In original Yolov3 network the feature extraction is done by deep network architecture called Darknet-53 that has 53 convolutional layers, each followed by batch normalization layer and Leaky ReLU activation. They don't use any pooling layer and downsampling is done by convolutional layer with stride 2. The prediction is then done by additional convolutional layers of size 1x1 and the result are possible detected boxes, each containg 4 bounding box offsets, detection score and classes predictions. This procedure is done on multiple scales. Coordinated of predicted bounding boxes are then extracted using 4 bounding box offset and anchor boxes that are precalculated for specific training data. This is done due to the possibility of unstable gradient during training if we use the absolute value of width and height of bounding box. In the training phase, the loss is calculated as combination of mean square error of position and size of bounding box, binary cross-entropy loss for score of present bounding box, binary cross-entropy loss for score of absent bounding box and binary cross-entropy loss for class predictions. Important step of extraction of predicted boxes is also non-maximum suppression that removes overlapping bounding boxes.

Yolov3-tiny implements the same idea, but the used network is much shallower. The feature extractor consists of only 9 convolutional layers and instead of convolution with stride 2, maxpool layers are use. The grid prediction is done by two additional convolutional layers of size 1x1 in two scales. The architecture of network is written in Table I, input of network are color images of size 416x416 pixels. As addition to convolutional and maxpool layers, this network consists also of Route Layer which takes one or more preceding layers and merge them, Upsample layer that doubles the dimensions of input and YOLO layer that does the prediction. The smaller complexity of this network in comparison to original network enables also shorter time of training which was a great for us because of our hardware limitations.

For training phase we used framework Darknet[1] and pretrained weights for Yolo v3 tiny available in same repository. We used standard Yolo v3 tiny architecture and trained for 14000 iterations, using batch size 64 and subdivisions of 16. We used learning rate of 0.001 and it was reduced twice during learning, at 11200 and 12600 iterations. This framework also uses data augmentations. Stochastic gradient descent with momentum of 0.9 and decay of 0.0005 was used as the optimization method. Used hardware for learning and inference phase was GeForce GTX 1050 Ti.

| Layer | Type | Filters | Size/Stride | Output |
|---|---|---|---|---|
| 0 | Convolutional | 16 | 3x3 | 416x416 |
| 1 | Maxpool | | 2x2/2 | 208x208 |
| 2 | Convolutional | 32 | 3x3 | 208x208 |
| 3 | Maxpool | | 2x2/2 | 104x104 |
| 4 | Convolutional | 64 | 3x3 | 104x104 |
| 5 | Maxpool | | 2x2/2 | 52x52 |
| 6 | Convolutional | 128 | 3x3 | 52x52 |
| 7 | Maxpool | | 2x2/2 | 26x26 |
| 8 | Convolutional | 256 | 3x3 | 26x26 |
| 9 | Maxpool | | 2x2/2 | 13x13 |
| 10 | Convolutional | 512 | 3x3 | 13x13 |
| 11 | Maxpool | | 2x2/1 | 13x13 |
| 12 | Convolutional | 1024 | 3x3 | 13x13 |
| 13 | Convolutional | 256 | 1x1 | 13x13 |
| 14 | Convolutional | 512 | 3x3 | 13x13 |
| 15 | Convolutional | 36 | 1x1 | 13x13 |
| 16 | YOLO | | | |
| 17 | Route 13 | | | |
| 18 | Convolutional | 128 | 1x1 | 13x13 |
| 19 | Upsample | | /2 | 26x26 |
| 20 | Route 19 8 | | | |
| 21 | Convolutional | 256 | 3x3 | 26x26 |
| 22 | Convolutional | 36 | 1x1 | 26x26 |
| 23 | YOLO | | | |

Table I: Architecture of Yolov3-tiny

## IV. Results

The problem at the start was to find appropriate dataset for training and testing the model as most of the datasets that contain labelled bounding boxes and appropriate labels for 6 main emotions and neutral expression are not freely available for download. We decided to use Expression in-the-Wild (ExpW) dataset[2] that was used in [13], [14] and had originally purpose of learning to recognize social relations from images but it also contains labels for facial expressions. It contains 91,793 images manually labeled with bounding boxes and expressions. Face detections in this dataset are not reliable and for each annotation there is also a score for face detection, so we used only images that had score higher that 50 %. That way we extracted 31989 images. The problem of this dataset is that classes are highly unevenly represented (for example there are 10272 faces annotated with label neutral but only 508 faces annotated with label anger) and that there is only one face annotated on every image, even though some images contain more than one face. Two examples from this database that show the described problem of missing annotations can be found on Figure 1.

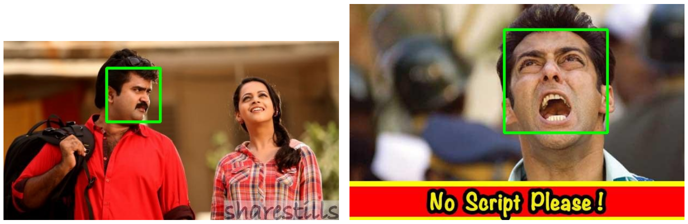


Figure 1: Examples from the first database, labelled as anger (left) and sadness (right)

We combated the problem of uneven distribution of classes with additional data from Expression part of Aff-Wild2 database[3] used in [15–20]. We used training and validation parts of the dataset (61 videos) because annotations for those two were available. For every frame of videos there is only label of emotion but no bounding box information so we needed to annotate faces first. For this we used RetinaFace face detector [21], [22] implementation[4] and saved coordinates of one or two (some videos contain multiple people) bounding boxes and corresponding emotion labels for each frame. With this we got more data, but in this dataset there were also many missing annotations (for example some videos were from Youtube and showed people reacting on other video that contained also faces and those faces were not annotated). Two example images from this dataset can be found in Figure 2

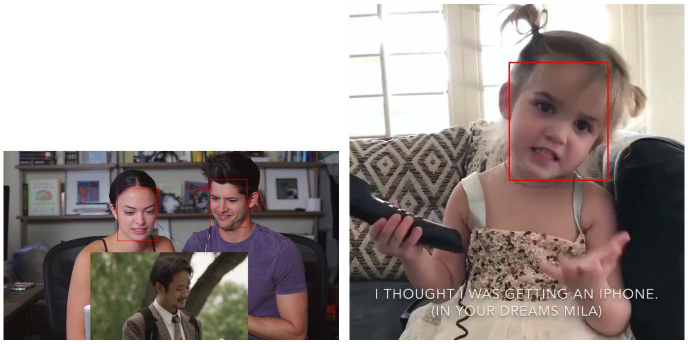


Figure 2: Examples from the second database, labelled as happiness for both annotations (left) and sadness (right)

After the training phase we evaluated model on 2500 random images from each of the datasets that were not used during training. Reported (by Darknet) mean average precision for single emotions and whole testing dataset are in Table II. We can see how there is a big difference in mAP for different classes.

[1] https://github.com/AlexeyAB/darknet

[2] https://riemenschneider.hayko.at/vision/dataset/task.php?did=381

[3] https://ibug.doc.ic.ac.uk/resources/aff-wild2/

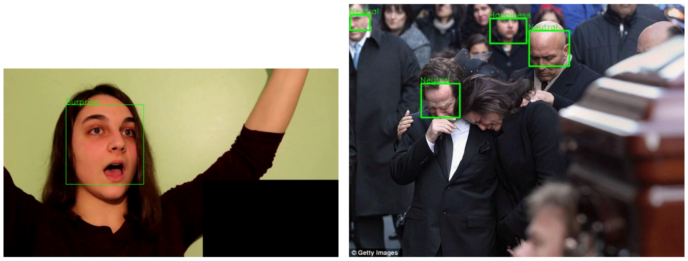
[4] https://github.com/deepinsight/insightface/tree/master/RetinaFace

It can be a consequence of distribution of classes in training data, but it is more probable that some emotions are harder to recognize or to distinct between them. For example, from single image it can be hard to recognize if the person on it is happy or just surprised even for a human. Learned model doesn't by far reach the accuracy that is achieved by current state-of-the-art algorithms for facial expression detection. Reasons for that are already mentioned missing annotations in dataset and that YOLO v3 Tiny model emphasis more on speed than high accuracy. But we think that maybe with more iterations of training phase we could reach higher accuracy but due to limited hardware resources and time this has not been checked yet.

| Class | mAP |
|---|---|
| Anger | 43.23% |
| Disgust | 6.91% |
| Fear | 67.66% |
| Happiness | 33.25% |
| Sadness | 31.18% |
| Surprise | 23.66% |
| Neutral | 31.64% |
| Combined | **33.93%** |

Table II: Mean average precision on testing data

To transfer detection to Python we converted learned model from Darknet format to format compatible with Python library Tensorflow[5]. For this part we used this repository[6]. We created the demo program that uses the detector or single image or camera input. On our hardware, demo on camera does indeed run in real time which was our preference. In Figure 3 are two examples of correct and incorrect outputs of our model.



© Getty Images

Figure 3: Sample output prediction of our learned model.

## V. Conclusion

In this paper we investigated a problem of automatic face expression detection using Yolov3-tiny network. This showed to be a difficult task and there are more possible paths that could be researched further. We would like to try model learning using dataset with more reliable annotations and try other architectures as Yolov3-tiny is perhaps not complex enough for given problem. Also there is a desire of further research of possibility of learning a model that combines YOLO and LSTM networks for facial expression detection in video.

## References

[1] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, pp. 7714–7734, 06 2013.

[2] S. L. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," 12 2012, pp. 1–5.

[3] W. Wei, Q. Jia, and G. Chen, "Real-time facial expression recognition for affective computing based on kinect," 06 2016, pp. 161–165.

[4] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," 05 2017.

[5] H. Jung, S. Lee, J. Yim, S. I. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," 12 2015, pp. 2983–2991.

[6] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," 06 2016, pp. 3391–3399.

[7] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," 11 2015.

[8] D. Kim, W. Baddar, J. Jang, and Y. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 04 2017.

[9] D. Acharya, Z. Huang, D. Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," 06 2018, pp. 480–4807.

[10] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," 09 2019, pp. 3866–3870.

[11] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, and L. Xun, "Facial expression recognition with faster r-cnn," *Procedia Computer Science*, vol. 107, pp. 135–140, 12 2017.

[12] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 04 2018.

[13] C. C. L. Zhanpeng Zhang, Ping Luo and X. Tang, "Learning social relation traits from face images," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[14] ——, "From facial expression recognition to interpersonal relation prediction," in *arXiv:1609.06426v2*, 2016.

[15] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," *arXiv preprint arXiv:1910.04855*, 2019.

[16] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, pp. 1–23, 2019.

[17] D. Kollias and S. Zafeiriou, "A multi-task learning & generation framework: Valence-arousal, action units & primary expressions," *arXiv preprint arXiv:1811.07771*, 2018.

[18] ——, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018.

[19] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1972–1979.

[20] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal 'in-the-wild'challenge," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1980–1987.

[21] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *CVPR*, 2016.

[22] J. Deng, J. Guo, Z. Yuxiang, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," in *arxiv*, 2019.

[5] https://www.tensorflow.org/

[6] https://github.com/mystic123/tensorflow-yolo-v3