Computer Vision Pipeline for Facial Expression Recognition

Kamil Akhmetov
B17-DS-01
Innopolis University
Innopolis, Republic of Tatarstan, Russia
k.ahmetov@innopolis.ru

Vadim Solovov
B17-DS-02
Innopolis University
Innopolis, Republic of Tatarstan, Russia
v.solovov@innopolis.ru

I. Introduction

Rapid development of neural network models has made deep learning frameworks a standard approach for vast amount of machine learning tasks with applications in various branches of human life. Significant part of problems require solutions that are both effective and interpretable. This includes solutions specific for crucial fields as psychology and healthcare. Neural models while having exhaustive expressiveness are weakly interpreted [6], [7], meaning users may not rely on inferences of untrusted opaque algorithms. Traditional machine learning methods are more transparent, providing comprehensible results, however they require additional stages in the learning pipeline in comparison with neural networks that may facilitate end-to-end approaches. Capturing social signals from human beings is essential in order to empathize computer systems. Important information comes from facial expressions.

In our project, we have built a Computer Vision pipeline for Facial Expression Recognition, which includes data preprocessing, feature extraction and classification stages. One of the main goals of the project was to find most representative features, and, as a consequence, improve the existing solutions, by both accurately determining facial emotions and saving interpretability so that we could be able to reason the features that contribute to the recognition of a specific expression. We contribute to the field of Facial Expression Recognition in Computer Vision Science with applications in critical areas of human life, for example, sociology, driver safety, virtual reality, cognitive sciences, etc. A solution we achieved has shown **83.7% accuracy** score on a colored subset of Taiwanese Facial Expression Images Database (**TFEID**) [3].

II. Related Work

Deep Learning is currently a dominant approach for machine learning in general and computer vision in particular. Convolutional neural networks (**CNN**) and recurrent neural networks (**RNN**) are the deep learning based methods used for tasks as feature extraction, classification, recognition. The main advantage of CNN is that it completely (or highly) reduces the need in preprocessing steps and enables so called 'end-to-end' learning directly from the raw inputs [8]. In order to capture temporal changes CNN were applied in combination with RNNs or long short term memory (LSTM) networks. Commonly, deep learning approaches extract features with the desired characteristics directly from the raw data by the application of deep neural networks (mostly CNN), however, they require much higher computational power, comparing with traditional methods [9]–[14]. The other and often a more crucial drawback of deep learning approaches is the lack of interpretability, that still leaves most researches in the fields as physics or medicine using simplistic linear models. [6], [7]

Given the drawbacks of Deep Learning approaches, traditional methods remained in the spotlight of computer vision researches. There are several preprocessing, feature extracting and classifying techniques that are both powerful and comprehensible. For geometric features, relationship between facial components, information about positions and angles of different landmarks were used for constructing a training vector [15], [16]. Classifiers used to be either AdaBoost [4] or SVM [5]. The appearance features were usually extracted from the global face region, using, for example, local binary pattern (LBP) histogram for feature vectors and principal component analysis (PCA) for classifications [17]. But still this approach does not reflect changes in local mini-regions, that is why appearance features were obtained from different domains with different levels of information [18], [19]. Combining the benefits of previously described approaches [19], [20], hybrid features emerged in order to complement the weaknesses and provide even better results. Many systems [20], [21] were used to construct features in temporal domain extracting the geometric and appearance features tracked later frame-by-frame. Commonly speaking, classic FER approaches determine features and classifiers by experts. According to these findings, an improvement is needed.

III. Idea Explanation

Giving value to interpretability and simplicity, we rely on traditional models, thus the main focus is dedicated to preprocessing and feature extraction stages of the pipeline. We have tested several basic approaches and implemented a framework for experiments in the first part of the project. For the second iteration we expanded the amount of approaches used in experiments.

A. Preprocessing

Step of preprocessing converts input data points to a representation that is suitable for further feature extraction. This stage may include such operations as resizing, denoising, segmentation.

We ended by resizing only for classic HOG and LBP based approaches we implemented first. Further we applied landmark extraction for retrieving face and different parts as features.

B. Feature extraction

After the data is formatted and structurized, we are ready to extract apply algorithms for feature extraction and data vectorization. Finding and tuning these algorithms are the main part of our project. Currently we have implemented Histogram of Oriented Gradients [1] and Local Binary Pattern [2].

As an alternative we decided to apply facial landmarks extraction (such as mouth, eyes, nose, eyebrows, etc.) with their combination and normalization them. We tried different algorithms such as Viola Jones, Haar Cascade classification, and ensemble of regression trees for extracting specific facial regions to use for feature extraction.

C. Classification

Feeding crafted feature vectors to a classifer is the final stage of the facial expression recognition pipeline. AdaBoost with Decision Trees and polynomial Support Vector Classifier models are well known to solve such kind of tasks. We provide results for both models.

D. Code

Code for the project is available on GitHub: https://github.com/kamilkoduo/cv-fer

IV. Experiments & Results

As we have stated above, for the first iteration we have tried several configurations constructed from the previously used methods. Our start point was to test LBP and HOG algorithms in combination with AdaBoost and SVC models. For the second iteration we decided to extract facial landmarks in two forms: as intensity valued images and as shapes with points outlining these features.

A. Dataset

There are many datasets for human facial expressions such as CK+, JAFFE, TFEID, KFEID. For the sake of our experiments we have selected a relatively small facial expression dataset: Taiwanese Facial Expression Image Database. We have selected a subset of 336 samples uniformly distributed across 20 female and 20 male models and contains 8 facial expressions (including neutral).

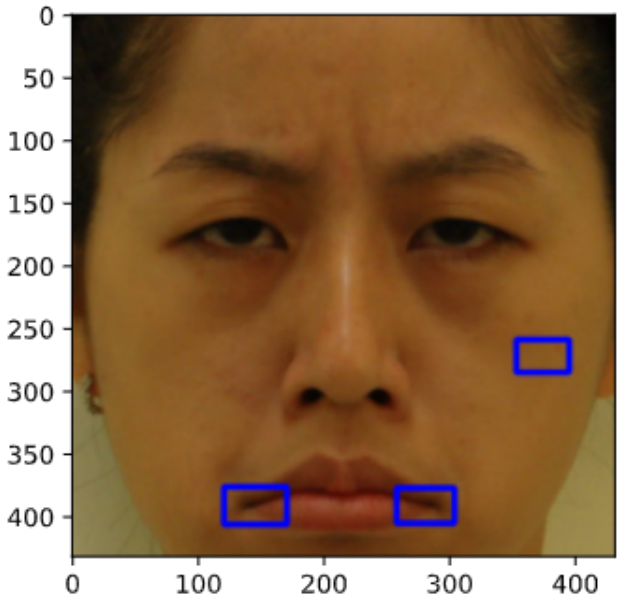
B. Performance evaluation

The data was split in following manner: randomly chosen 80% of data points of each expression are used for training and the remaining 20% used for testing, thus classes are uniformly distributed across both splits. Based on these data we have trained 4 configurations.

After the previous iteration we changed HOG parameters and the accuracy had improved by 8% (from 37% to **54%**) on SVC.

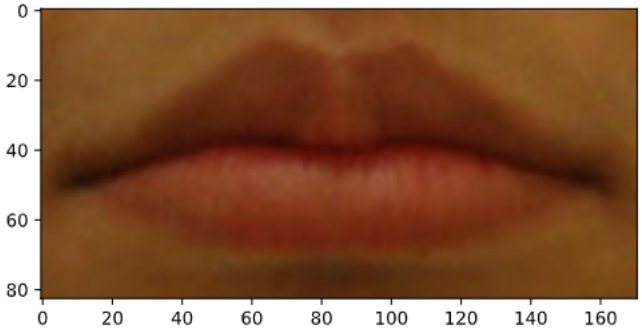
Given the unsatisfying results received from pure application of LBP and HOG feature extractors, we decided to find other ways to receive valuable information about facial expressions. For this we decided to extract feature landmarks such as eyes or eyebrows in two ways: as an image and as points outlining the feature. To extract facial landmarks as images we tried two approaches: Viola-Johnes and ensemble of regression trees. Viola-Johnes performed well on eyes, but the detection of mouth was inconsistent.

Fig. 1. Example of the extracted facial landmark using Viola-Johnes



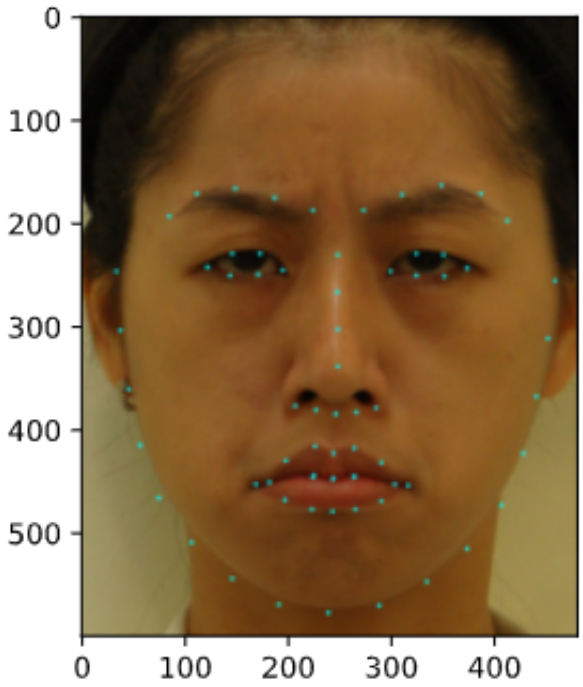
At the same time, the ensemble of regression trees proved to be great at extracting facial features.

Fig. 2. Example of the extracted facial landmark using regression trees



After extracting the images of features we used LBP and HOG for vectorization and further feature extraction. In this scenario LBP gave much better results of 72% and 75% on SVM and AdaBoost, respectively, while HOG gave results similar to results received after applying HOG to the face without extracting landmarks.

Fig. 3. Feature points on the original image



The second approach of using the points outlining facial landmarks proved to be the most efficient: after normalization of the coordinates and vectorization we received 81.7% and 83.7% accuracy on classifiers.

	LBP	HOG	FL Images	FL Points
SVC	0.34	0.54	0.72	0.817
AdaBoost	0.43	0.46	0.75	0.837

TABLE I: Accuracy Score

As we can see from the Table 1, while HOG feature extractor in combination with Ada-Boost classifier shows better accuracy score than LBP, it is much inferior to both approaches of retrieving facial landmarks.

V. Analysis

During this project we found out, that it is possible to get sensible results for the task of facial expression recognition without using Convolutional Neural Networks. We also found which ways of extracting features are more suitable for this task and which are less suitable: using feature extractors such as HOG or LBP on the whole image of the face does not give good results, however using LBP with the extracted images of landmarks gives much better results and using points that outline facial landmarks proved to be the most efficient way.

Our work can be improved in several dimensions:

More Feature Extractors: Given that there is a big amount of feature extraction algorithms and we have tried only several of them, there may exist better feature extraction algorithms.

Better data: While the TFEID is one of the largest datasets for the task of facial expression recognition, it is still very limited (there are only 40 Asian models) and may introduce bias to the model.

VI. Conclusion

To conclude, we have prepared a framework for our experiments and tested several configurations. We have seen, that the best approach to facial expression recognition is extracting feature landmarks as points, while using feature extraction methods on the image without facial landmark identification gives unsatisfactory results.

References

[1] McConnell, R K. Tue . "Method of and apparatus for pattern recognition". United States.

[2] T. Ojala, M. Pietikainen and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," Proceedings of 12th International Conference on Pattern Recognition, Jerusalem, Israel, 1994, pp. 582-585 vol.1, doi: 10.1109/ICPR.1994.576366.

[3] L. F. Chen and Y. S. Yen, "Taiwanese facial expression image database," Ph.D. dissertation, Brain Mapping Lab., Inst. Brain Sci., Nat. Yang-Ming Univ., Taipei, Taiwan, 2007

[4] Schapire, R. E. (2013). Explaining adaboost. In Empirical inference (pp. 37–52). Springer.

[5] Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[6] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, Klaus-Robert Mddotuller, Interpretable Deep Neural Networks for Single-Trial EEG Classification, ¡![CDATA[Journal of Neuroscience Methods]]¿ (2016), http://dx.doi.org/10.1016/j.jneumeth.2016.10.008

[7] Dong, Yinpeng Bao, Fan Su, Hang Zhu, Jun. (2019). Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples.

[8] Walecki, R.; Rudovic, O.; Pavlovic, V.; Schuller, B.; Pantic, M. Deep structured learning for facial action unit intensity estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, HI, USA, [2017]; pp. 3405–3414

[9] Kahou, S.E.; Michalski, V.; Konda, K. Recurrent neural networks for emotion recognition in video.In Proceedings of the ACM on International Conference on Multimodal Interaction, Seattle, WD, USA,9–13 November 2015; pp. 467–474.

[10] Kim, D.H.; Baddar, W.; Jang, J.; Ro, Y.M. Multi-objective based Spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affect. Comput. [2017],PP.

[11] Chu, W.S.; Torre, F.D.; Cohn, J.F. Learning spatial and temporal cues for multi-label facial action unit detection. In Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, [2017]; pp. 1–8.54.

[12] Hasani, B.; Mahoor, M.H. Facial expression recognition using enhanced deep 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Hawaii, HI, USA, [2017]; pp. 1–11.55.

[13] Graves, A.; Mayer, C.; Wimmer, M.; Schmidhuber, J.; Radig, B. Facial expression recognition with recurrent neural networks. In Proceedings of the International Workshop on Cognition for Technical Systems,Santorini, Greece, 6–7 October 2008; pp. 1–6.56.

[14] Jain, D.K.; Zhang, Z.; Huang, K. Multi angle optimal pattern-based deep learning for automatic facial expression recognition. Pattern Recognit. Lett. 2017,1, 1–9. [CrossRef]57.

[15] Suk, M.; Prabhakaran, B. Real-time mobile facial expression recognition system—A case study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, [2014]; pp. 132–137.23.
[16] Ghimire, D.; Lee, J. Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sensors 2013,13, 7714–7734
[17] Happy, S.L.; George, A.; Routray, A. A real time facial expression classification system using local binary patterns. In Proceedings of the 4th International Conference on Intelligent Human Computer Interaction,
[18] Khan, R.A.; Meyer, A.; Konik, H.; Bouakaz, S. Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recognit. Lett. [2013], 34, 1159–1168.
[19] Ghimire, D.; Jeong, S.; Lee, J.; Park, S.H. Facial expression recognition based on local region specific features and support vector machines. Multimed. Tools Appl. [2017], 76, 7803–7821.
[20] Benitez-Quiroz, C.F.; Srinivasan, R.; Martinez, A.M. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, [2016]; pp. 5562–5570.
[21] Torre, F.D.; Chu, W.-S.; Xiong, X.; Vicente, F.; Ding, X.; Cohn, J. IntraFace. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Ljubljana, Slovenia, [2015]; pp. 1–8