Review Article

Processing Facial Emotion Recognition on Forensic Type Sketches

Regep Sim Reyhan[*]

Doctoral Student, Doctoral School of Electronics, Telecommunications and Information Technology, University of Polytechnics Bucharest, Splaiul Independenței 313, București 060042, Romania

*Corresponding author: Regep Sim Reyhan

Abstract

The idea suggested in this paper consists in forensic sketches that have to be drawn and furthermore analyzed from an emotional face recognition point of view. The detection itself is more effective and useful to further identify the future behavior and possible reactions of the suspect based on the emotional analysis of the areas on the face. The study embraces this idea and experiments with forensic like sketch images within a system that is trained on a FER+ database, and executed using two different schemes, ML (Multi-Label learning) and CEL (Cross-Entropy Loss). The conducting experiment resulted that for forensic sketch images, CEL system is more efficient than ML training method 69.77% vs 69.05% for the best test accuracy, even if the experiment on the standard image equivalents proved otherwise.
Keywords: Sketch recognition, emotion, expression, face recognition, forensic science.

INTRODUCTION

Facial recognition is a topic that is researched thoroughly and yet with so much more potential left. In the forensic department, the analysis of the possible dangerous suspects is made through sketches afterwards associating the results to a database.

An important step could be made, and is prone to further research, by analyzing and drawing the sketch including the facial expression, given by the emotional state the person was in, during the last encounter with the eye witness.

In this way, the suspects behavior would be more predictable and the level of danger the suspect presents to the people around, can be better assessed, managed or even avoided by the authorities.

In comparison with the already developed systems, the one that can pursue such a job has to be able to evaluate forensic sketches, recognize emotion at a face level, in a small amount of time, due to the ongoing investigation.

Rosenberg and Ekman's findings were concluding that we can consider facial expression and some emotional experiences as being most of the time, congruent and the intensity that results from the emotional event can mediate the human responses cohesion [1]. The facial recognition concept was being researched since the 70s but the computational limitations and various illumination and occlusion issues were never entirely eliminated to the day.

Many systems have been implemented over the years to try pursuing this matter, yet the research for the idea of emotion recognition from forensic like sketch images is scarce.

One such algorithm that processes a large dataset, but needs to be tested on a forensic sketch must cover same important aspects of the analysis, since there is one sketch per individual – in case of forensic research, and the system can process the information only within its dataset due to lack of sketch imaging in real life.

The way it is trained, for example, has to cover, despite the multitude of methods that don't, the fact that a face in distress can present multiple emotions at once, therefore it should be able to test and identify those.

On the other hand, the identification must limit to a certain amount of emotions, since only some of them are relevant from a forensic point of view.

For classical HFR (Heterogenous Face Recognition) systems, the most common scenario, the face recognition based on a photo-sketch processing is also influenced by the occlusions that tends to appear on an evaluated image, leading to extreme variation in

the result, due to inaccuracies within the sketch analysis.

The work of forensic examiners consists in hours of research of a face in mugshot photos, or in surveillance videos, each of them, having different performances, based on the experience, fact which suggests the need of a standardized method of face and emotion recognition, which would be equally more faster and efficient having the capability of identifying individuals not only in image databases, but in videos, forensic sketches and live streaming surveillance records.

A facial and emotional system must be thoroughly checked and evaluated before being a reliant tool for such important jobs. The legal implications and the possible negative outcomes can lead to way more serious issues than use in another field of research regarding the facial recognition systems.

Since we are talking of images equally alike to official forensic drawings, their quality will be lower than a normal image taken on a street for example, and the ability of the drawer will also influence the results. This creates the eventuality of covering such inconveniences while experimenting with any algorithm on these kinds of images, hence the probable type of sketches must be covered.

In the field of research, there were some topics regarding the facial recognition at a forensic like images, as it will be presented in the following discussion.

Using a DCNN (Deep Convolutional Neural Network) to convey the algorithm, Galea, and Farrugia [2] used transfer learning in order for the network to be able to learn, method applied to emphasize the connections between the following two, sketches and photos. In their research, the VGG [3] system was utilized for training and also testing.

The performance of VGG and the number of needed multiple parameters to train lead to overfitting. Their method, DEEPS (Deep photo-sketch system), made use of a 3-dimensional morphable model to create a new dataset of images using variation. This gave the system the ability to be more accurate despite the dissimilarity of the sketches compared to their photographic counter parts.

DEEPS managed to outperform other relevant systems to the field, like LGMS, by managing to obtain the mean value of the rank retrieval of 325.02 for forensic sketches, while applied on forensic images a value of 398.27.

In another order of ideas, a group of researchers [4], approached the forensic matter by using both the human resource as well as the machine learning technique. Their idea consisted in creating a method by using groups of specialized people in face recognition as well as reviewers.

Their groups were composed of super-recognizers, forensic facial examiners and reviewers as a group, and fingerprint recognition specialist alongside with undergraduate students as the 2nd group which was the control one.

People's ability was to be evaluated in comparison with four FCR - DCNN "performers". There were used same pairs of images for FCR and specialist groups. Some researchers performed the analysis thoroughly with laborious methods and on larger amount of time, while the others tried to work faster but this involved, in the end, less research.

The deep recognition algorithms used in the experiment were both public and more performant, older and newer.

The results are referred as median performance value of each group. We will name the algorithms A1, A2, A3, A4 where A1 is the newest and A4 is the oldest. More detail about the specific algorithms used can be found in [4].

In the below Table-1, we can see the human performer type and the algorithm counterpart, as they were compared in pairs.

Table-1: Specialists vs CNN

Human Performer	Results - human	Results - algorithm	Algorithm used
Forensic Facial Examiner	0.93	0.96	A1
Facial Reviewers	0.87	0.85	A2
Super-recognizers	0.83		
Finger Print Examiners	0.76	0.76	A3
Students	0.68	0.68	A4

The outcome of the research was quite interesting, as seen above, the most performant was the newest developed algorithm that manages to even exceed a professional's performance, while the oldest one, still developed in the last decade, is as performant as an undergraduate student. Yet the best performance of value 1 was acquired by fusing both the Facial Examiner with the newest algorithm.

EXPERIMENTAL SECTION

The experimental solution on which the paper is based on, is a DCNN solution, a VGG system that pursues an image dataset FER to which Barsoum et al., [5] added a crowd sourced label system, resulting in FER+ dataset and the process itself. The originality comes from the idea of testing, analyzing and evaluating how a Neural Network core algorithm trained on a label-oriented dataset can perform on a forensic sketch type images, and what are the limitations.

Integrating the dataset itself is a laborious process due to its low-quality images of 2034 pixels that have to be refracted into a sketch test system. The image size has the great advantage of faster processing of numerous images, but arrives with its own limitations, especially for forensic imaging.

Compared to FER that uses a smaller range of emotions, the basic ones, FER+ approach is better equipped for forensic emotion recognition research, hence the choice for experiment.

Architecturally speaking, the experiment is developed using images from UTK dataset [6], drawn as sketch images, to resemble forensic imaging, paper drawn or digitally. The processing behind is made by a VGG that is oriented on GPU processing power.

Two schemas were taken into analysis for the evaluation, the ones fitting the best the area of expertise, Multi-Label learning, and Cross-entropy system.

Multi-Label learning is applied for the emotion choosing for every image in the training, based on the probability of occurrence of each emotion. Technically as long as the model output is in congruence with the learnt labeling, the loss function will have a small value.

On the other hand Cross-entropy system uses a standard loss function for label classification, where N = 1, j has values up to number of training examples existing, and k the number of emotions we use – the labels themselves (8 in this case).

$$L = -\frac{1}{N} \sum_j \sum_k y_k^j \log x_k^j \dots\dots\dots\dots\dots\dots (1)$$

After conducting the experiment we can see that the behavioral results of the ML vs the CE system, of the FER+ dataset using system developed by [5], as a reference for the next steps, in the Fig-1.

Scheme	Best Validation accuracy	Best test accuracy - for the validation	Best test accuracy
ML	86.56%	85.43%	85.75%
Cross-Entropy	87.59%	86.83%	86.83%

Fig-1: CE vs ML on FER +

Considering over the experiment that we use a maximum of 100 steps in which the system is training and testing, Cross Entropy managed to reach maximum accuracy for test and validation at epoch 87 while ML, needed further training to reach the maximum value of test at epoch 96.

While more emotions are available for analysis, in the forensic matter, the resulted sketch image can have serious variations, hence the expectation is to discover, no matter the description quality and drawing the person searched for, and as it is stated in the paper, the possibility to predict further steps, based on the expression recognition.

The test dataset used presents both higher quality drawings and very low-quality ones, hence the system needs to have no such limitations, as seen in Figure-2.

Fig-2: Face – sketch drawing association

Images were drawn at different qualities, so the system can be able to evaluate them with both CE and ML, in order to check the efficiency of an existing algorithm, trained on a popular big dataset, and tested on forensic sketch images. In comparison, for more experimental results, the tests were performed on the grayscale version of the original images.

RESULTS AND DISCUSSION

While executing the method with ML system we can see that the emotion recognition peaks at a value of 72.35% for its best test accuracy compared to the sketch result of 69.05%, first reaching its best test value calculation at epoch 27, while the latter at 51.

The highest validation performance is at 86.76% (epoch 96) with 71.35% test accuracy, but for the analysis of the sketch execution drops to test accuracy of 68.62 % congruent to the best validation performed at 62nd epoch with 86.34%.
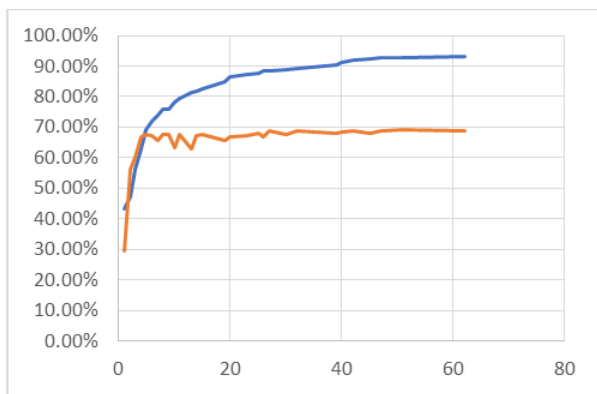

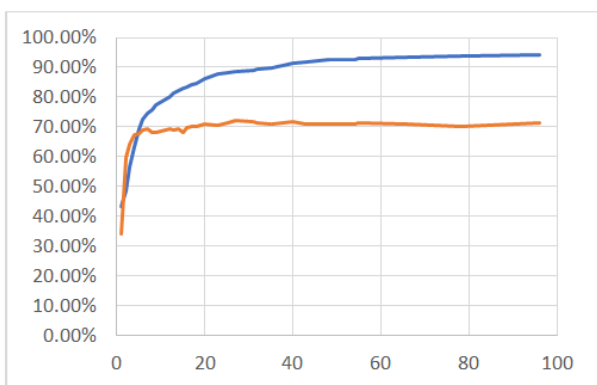**Fig-3: Multi-Layer Sketch Train Vs Test**

Fig-4: Multi-Layer Grayscale Train Vs Test

Multi-Label system seems to drop in performance and not reach a higher value after the 50-60 epochs for the sketch recognition system. This is due to the lack of detail in the images with lower resolution, and the training system not presenting sketch like images.

On the other hand, for the Cross-entropy evaluation the situation is different as seen below.
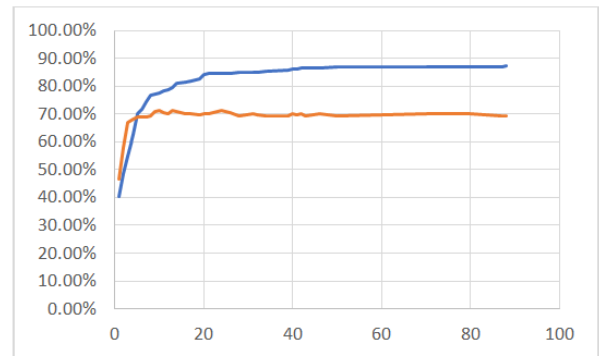

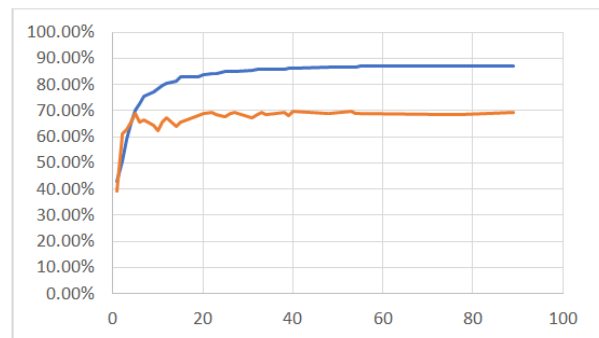**Fig-5: Cross-Entropy Grayscale Train Vs Test**

Fig-6: Cross-Entropy Sketch Train Vs Test

For the case in which Cross Entropy classification scheme is used, the grayscale dataset examination reaches 71.35% as the maximum test accuracy, interestingly enough the same as the test accuracy (based on the best validation) for ML, yet the validation performance it is 87.31% at an earlier epoch 88.

Sketch dataset performed slightly better than ML, peaking at 69.77% test accuracy and 69.48% equivalent to best validation. The validation performance was slightly lower than the grayscale experiment for cross entropy, but a little higher than both ML executions, 87.15% at epoch 89 for cross entropy sketch vs 86.78% ML – grayscale, 87.31% CE grayscale and the latter and the lower value for ML sketch – 86.34%.

CONCLUSION

The best performer of the systems overall was the CE analysis made on the sketch dataset, even if when the experiment executed on the grayscale normal image version of the dataset, resulted with slightly better accuracy in the favor of ML system.

The initial transformation of the FER2013 [7] dataset is improved by adding of the new labeling system that gives the algorithm the capacity to learn on evaluating more complex emotions that can also appear on the forensic line of work. In doing so, the images used validating and training in the FER+ Dataset have been labeled by the crowd taggers. In forensic field, this

can alter the outcome of the result, due to the lack of experience of the tagger.

The improvement of the system can be achieved in the forensic field, firstly by having a group of professionals to evaluate and label the future dataset of images, and the ones on which the system would be tested, so the same experience could be used at a maximum.

Further research is in place, with better labeling the dataset, and creating a new one by adding more forensic images so it can also train on such sketches.

ACKNOWLEDGMENT

REFERENCES

1. Ekman, P., & Rosenberg, E. L. (2016). What the Face Reveals - Basic and Applied Studies of Spontaneus Expression Using the Facial Action Coding System (FACS), Oxford, Oxford University Press, 63-85.
2. Galea, C., & Farrugia, R. A. (2017). Forensic face photo-sketch recognition using a deep learning-based architecture. IEEE Signal Processing Letters, 24(11), 1586-1590.
3. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
4. Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ... & Chen, J. C. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences, 115(24), 6171-6176.
5. Barsoum, E., Zhang, C., Ferrer, C., & Zhang, Z. (2016). Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label, ACM International Conference on Multimodal Interaction (ICMI).
6. Zhang, Z., Song. Y., & Qi, H. (2017). Age Progression/Regression by Conditional Adversarial Autoencoder, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
7. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Zhou, Y. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64, 59-63.