Recognition of Occluded and Lateral Faces Using MTCNN, DLIB and Homographies

Gustavo Bezerra* and Rafael Gomes†
Departamento de Informática e Matemática Aplicada, Universidade Federal do Rio Grande do Norte
*gustavowl@lcc.ufrn.br †rafaelbg@dimap.ufrn.br

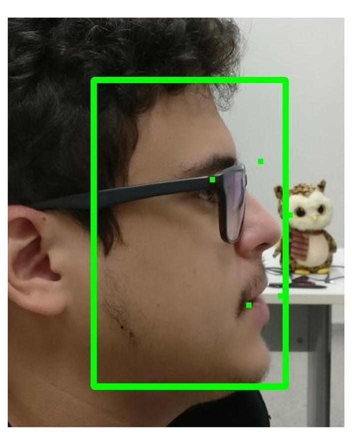
1. Introduction

Security is an issue since ancient times. Nowadays, police and cameras are used to achieve security. In addition, aiming a better performance, repetitive tasks that were performed by humans are now executed by robots.
The purpose of this work is to improve facial recognition, including occluded and lateral faces. Currently, it is desirable to detect new people on-the-fly via pattern matching.

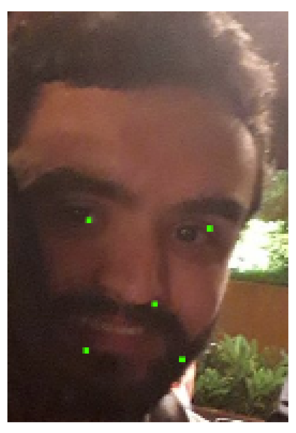
2. Background

2.1 MTCNN

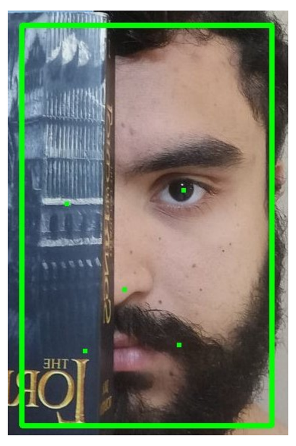
The Multi-task Cascaded Convolutional Network (MTCNN) developed by Zhang et al. [1] receives an image as input and outputs seven points for each detected face. The first two points correspond to the limits of the bounding box for a face. The remaining five points correspond to the position of the left and right pupils, nose tip, left and right mouth corners. The advantage of MTCNN is its accuracy regarding unfavorable poses, illumination and, specially, occlusion.



Unfavorable pose



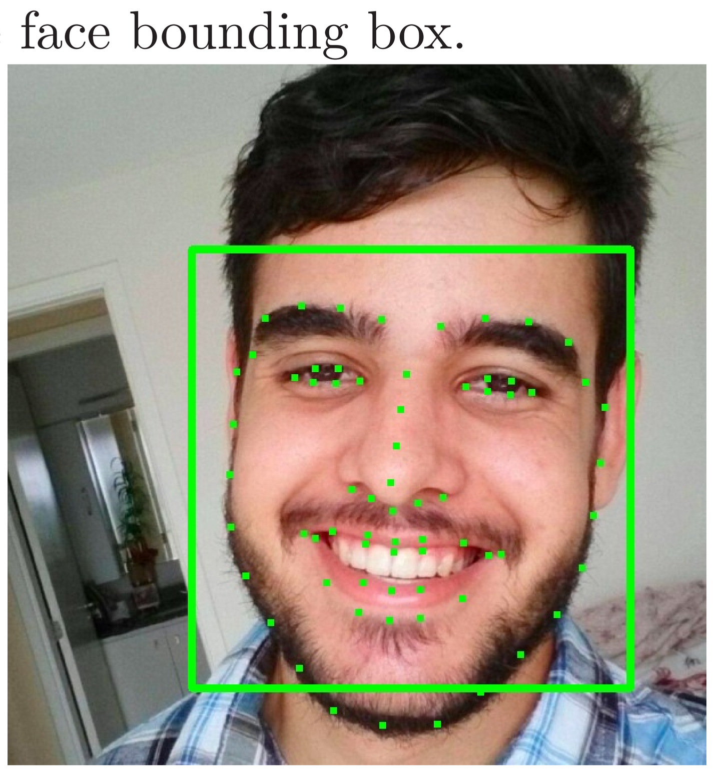
Unfavorable illumination



Occlusion

2.2 Shape Predictor

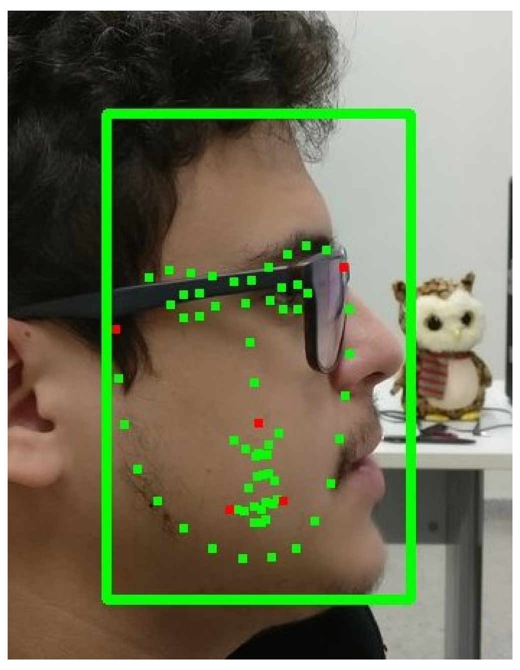
The DLIB library was originally developed in C++ with the purpose of supporting machine learning applications [2]. DLIB provides a shape predictor for frontal faces based on ensemble of regression trees ("shape_predictor_68_face_landmarks"). It returns 68 characteristic points for each person given the face bounding box.



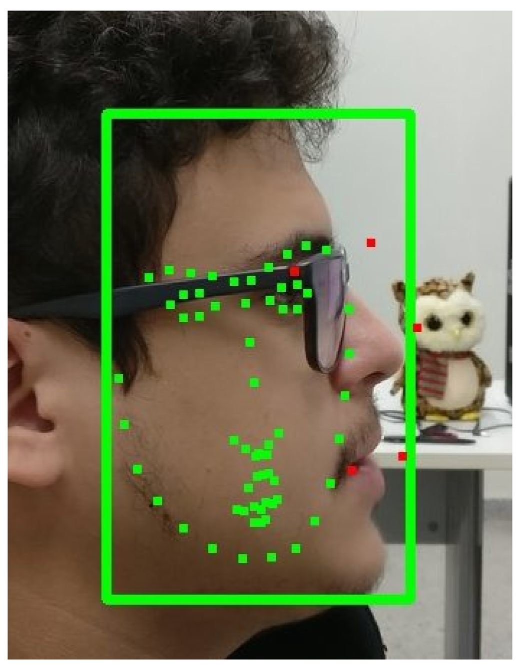
Extraction of 68 points using Shape Predictor

3. Hybrid Approach

MTCNN is used to detect the bounding boxes. The bounding boxes are sent to the Shape Predictor so it can extract the 68 characteristic points. However, a frontal face is expected. In order to diminish this inaccuracy, five Shape Predictor's points are substituted by the five MTCNN points (red points in the figures). The correspondence between DLIB and MTCNN's points are predefined.



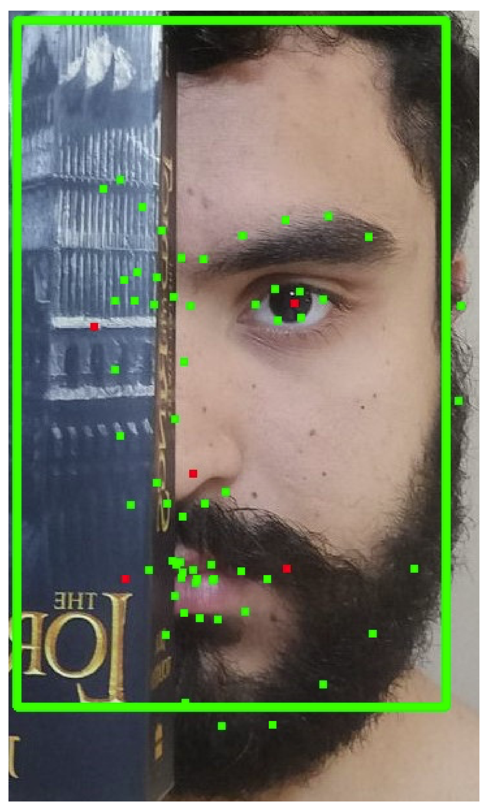
Shape predictor



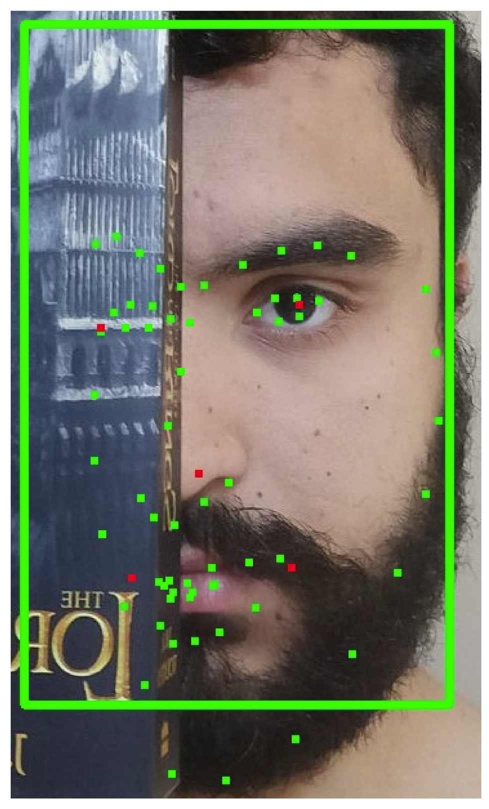
Hybrid approach

4. Homography

MTCNN's points are more accurate than DLIB's. Hence, they are used as the homography's destination points. In other words, DLIB's deformed plane will be mapped to MTCNN's more stable plane. For the homography, either five or four points are used: eyes, mouth corners and nose (optional). As a result, it is possible to verify in the right figure that the green points are surrounding the red ones; which did not happen in the left figure.



Hybrid approach



Applying homography

5. Results

In order to compare the different approaches, it is possible to use the difference between the obtained *norms* as a metric. A preliminary result can be seen in the next table. Positive numbers indicate better accuracy.

Image	mixed	5-points homography	4-points homography
in	-0.00672	0.0938	0.07216
out	-0.04372	-0.08464	-0.07186
occl	-0.00227	0.0283	0.08228
side	0.0068	0.14081	0.04475

However, executing the algorithm for a subset of the LFW dataset [3] [4] containing 785 images, the following result was obtained.

Technique	Tolerance	Correct	Wrong	Ratio (C/W)
5-points	0.4	163	622	0.26
5-points	0.5	272	513	0.53
5-points	0.6	275	510	0.54
4-points	0.4	134	651	0.21
4-points	0.5	237	548	0.43
4-points	0.6	275	510	0.54

Therefore, it is possible to notice that although applying homographies generated satisfactory results for a controlled dataset, it is necessary to refine the technique so the changes generated by the homographies do not prejudice DLIB and MTCNN's accuracy.

6. Concluding Remarks

The result achieved was very satisfactory for a limited image database, specially in lateral and occluded faces examples. On a larger dataset (containing only frontal faces), however, it was observed that using homographies inserted some noise that harmed the DLIB and MTCNN's accuracy and precision. Therefore, future works may focus on reducing the homographies' negative influences such as "chin elongation".

7. References

[1] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[2] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[3] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[4] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.