CHILDATTEND: A Neural Network based Approach to Assess Child Attendance in Social Project Activities

João Vitor Andrade Estrela[a] and Wladmir Cardoso Brandão[b]

Department of Computer Science, Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Hozizonte, Brazil

Keywords: Children Identification, Face Detection, Face Recognition, Image Alignment, Neural Network.

Abstract: Social project sponsors demand transparency in the application of donated resources. A challenge for non-governmental organizations that support children is to provide proof of children's participation in social project activities for sponsors. Additionally, the proof of participation by roll call or paper reports is much less convincing than automatic attendance checking by image analysis. Despite recent advances in face recognition, there is still room for improvement when algorithms are fed with only one instance of a person's face, since that person can significantly change over the years, especially children. Furthermore, face recognition algorithms still struggle in special cases, e.g., when there are many people in different poses and the photos are taken under variant lighting conditions. In this article we propose a neural network based approach that exploits face detection, face recognition and image alignment algorithms to identify children in activity group photos, i.e., images with many people performing activities, often on the move. Experiments show that the proposed approach is fast and identifies children in activity group photos with more than 90% accuracy.

1 INTRODUCTION

According to the United Nations (UN)[1], non-governmental organizations (NGOs) are task-oriented nonprofit organizations driven by people focused in performing a variety of service and humanitarian functions. In particular, Charity NGOs are operational organizations that meet the needs of disadvantaged people and groups, usually being funded by donations from sponsors. Similarly to all other organizations, NGOs need to be open about their goals, and donor sponsors expect that they demonstrate the same level of transparency and accountability as the private organizations. Thus, credibility is crucial for donations to continue to flow.

Some Charity NGOs are specialized in supporting children, many of them developing social projects that promote children's active participation in cultural and educational activities. A challenging problem for these associations is to provide sponsors with proof of children's participation in social project activities. The automatic verification of attendance by image analysis is much more convincing than a paper list or report. Therefore, the use of image analysis techniques, such as algorithms for detecting and recognizing faces, to assess the frequency of children in social project activities is fundamental to the transparency and credibility of NGOs.

Face detection and recognition are two different tasks that demand separate techniques (Hjelmås and Low, 2001; Zhao et al., 2003). Face detection, in particular, is an essential step for face recognition. It is a specific case of object-class detection which has been widely studied for decades. Consequently, several algorithms and methods have been developed over the years in order to address this task (Ming-Hsuan Yang et al., 2002). Face recognition, on the other hand, is applied to faces that have already been detected. This technique, in turn, is most commonly used for person identification and authentication (Zulfiqar et al., 2019). Yet, it can be used for a variety of tasks (Zhao et al., 2003), such as unlocking devices (Patel et al., 2016), paying for services and products (Li et al., 2017), identifying criminals (Sharif et al., 2016) and assessing people attendance (Kawaguchi et al., 2005).

Face identification is the process of comparing selected facial features of one's face against a preexisting set of faces (Guillaumin et al., 2009). In other words, it is the attempt to answer the question "Who are you?". This process can also be referred to as a

[a] https://orcid.org/0000-0001-9348-0215

[b] https://orcid.org/0000-0002-1523-1616

[1] https://research.un.org/en/ngo

Estrela, J. and Brandão, W.
CHILDATTEND: A Neural Network based Approach to Assess Child Attendance in Social Project Activities.
DOI: 10.5220/0010400406020609
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 602-609
ISBN: 978-989-758-509-8

1:N relation, or one-to-many matching. Face authentication, in turn, is the process of comparing one face to another single face, consisting of a 1:1 relation, or one-to-one matching. This process is used to validate a claimed identity based on the image of a face, i.e., the objective is to answer the question ”Is this really you?”. Although the two processes are different, both can be referred to as face recognition for simplicity.

There are many issues that must be addressed to detect and recognize faces in an image (Hjelmås and Low, 2001; Ming-Hsuan Yang et al., 2002; Jafri and Arabnia, 2009; Malikovich et al., 2017). One problem is lighting, i.e., the input images can present different lighting conditions depending on the quality of the device used to take the photo and the environment lighting condition (Peixoto et al., 2011). Another problem is the angle of the faces (Zhu and Ramanan, 2012). Ideally, every face should have a frontal arrangement. Nonetheless, usually faces are positioned slightly sideways, which many times hinders the process of detection. For that reason, the extraction of high-quality facial features is crucial for maximizing the accuracy of the face recognition algorithm.

In this article we propose CHILDATTEND, a neural network based approach that exploits face detection and recognition and image alignment algorithms to assess child attendance in social project activities by identifying children in activity group photos. Particularly, the proposed approach uses a set of images to train a neural network in a labeling task. Experimental results show that CHILDATTEND is efficient and effective, being able to label children in activity group photos within a few seconds while providing accuracy of approximately 90%. The main contributions of this article are:

- A novel neural network approach to assess child attendance in social project activities, which uses an efficient and effective face detector, along with a face recognition algorithm based on the Euclidean distance.
- A thoroughly evaluation of the proposed approach using a comprehensive image dataset within 4 different face detection algorithms.

In the remaining of this article, Section 2 presents theoretical background on face detection and recognition, including three research questions derived from previous experiments reported in literature. Section 3 presents the related work. Section 4 presents CHILDATTEND, our proposed approach. Sections 5 and 6 present the experimental setup and results, respectively. Finally, Section 7 presents the conclusions and directions for future work.

2 BACKGROUND

Many face detection and recognition algorithms have been proposed in the past years (Jafri and Arabnia, 2009; Kumar et al., 2019). The Viola-Jones algorithm (Viola and Jones, 2001) demands full view frontal upright faces, performing face detection in four stages: Haar feature selection, creation of an integral image, Adaboost training and cascading classifiers. In particular, it uses a small number of features and a “cascade” model which allows false positive regions of the image to be quickly discarded while spending more computation on promising object-like regions. Thus, it is very fast and presents high detection rate. In addition, it does not require much computational work, for this being widely used in real-time applications. However, it performs poorly for rotated or occluded faces, and under exploits issues that may hinder the detection of the face.

In feature-based approaches for face recognition, the input image must be processed to identify and extract distinctive facial features such as the eyes, mouth and nose, the geometric relationships among facial points must be computed, and the facial image must be reduced to a vector of geometric features. In this process, it is crucial to evaluate and test data normalization strategies (Hazim, 2016) to improve the quality of feature extraction. Additionally, statistical pattern recognition techniques must be used to match faces using the geometric measurements.

The Euclidean Distance (Liwei Wang et al., 2005) is a very common metric that is broadly used in several applications. It is a fast and efficient way to calculate the distance between two vectors. In face recognition, the vectors represent features of the faces that are being compared. Thus, the greater the Euclidean distance, the less faces are similar. Usually, one single 128D embedded vector representing the face in a Euclidean space of 128 dimensions are required to calculate the Euclidean distance. However, it is possible to re-sample the face before extracting the vector. The number of re-samples is called “number of jitters”. It is intuitive to consider that the increase in the number of jitters might increase the accuracy of the face recognition algorithm, since it might create a more general representation of the face.

3 RELATED WORK

There are four categories of face detection methods (Ming-Hsuan Yang et al., 2002): i) knowledge-based where human knowledge on what constitutes a face is encoded and the relationship between facial features is captured; ii) feature-invariant to look for possible existent structural features of a face even when there are variations in different conditions such as viewpoint, lighting and pose; iii) template matching where many patterns are stored to describe the face as a whole or the facial features separately; iv) appearance-based using a set of training images to capture and learn the representative variability of faces.

A reliable algorithm for object detection (Dalal and Triggs, 2005) uses a feature extractor to obtain image descriptors and a Linear Support Vector Machine (SVM) (Cortes and Vapnik, 1995) to train highly accurate object classifiers. Different from the Viola-Jones algorithm (Viola and Jones, 2001), it counts the occurrences of gradient vectors that represent the light direction to select image segments. On top of that, overlapping local contrast normalization is used to improve accuracy.

Multi-task cascaded convolutional neural networks (CNN) were trained to make three types of predictions: face classification, bounding box regression and facial landmark localization (Zhang et al., 2016). First, the image is re-scaled to a range of different sizes (called an image pyramid). Then, the P-Net uses a shallow CNN, which proposes candidate facial regions. Next, the R-Net filters the bounding boxes by refining the windows to reject a large number of non-faces windows through a more complex CNN. Finally, the O-Net proposes five facial landmarks, which are: left eye, right eye, nose, left mouth corner and right mouth corner. Different from the other aforementioned algorithms, this algorithm softens the impact of many problems that weren't previously taken into account by other detectors, e.g., pose variation and bad lighting.

The YOLO algorithm (Redmon and Farhadi, 2018) uses a softmax function along with multi-label classification for face classification. Considering the accuracy of the detection, this algorithm presented a much higher detection rate compared to that of other detectors. Previous experiments show that conditions such as occlusion, poor lighting, variation in pose and rotation no longer hinder the facial detection process. Additionally, the YOLO detector performs substantially faster than the previously cited detectors.

Some systems to take the attendance of students in class have been proposed in the past years (Kawaguchi et al., 2005). In such systems, the images of the students' faces are stored with their names and ID codes in a database. In addition, the data of students observed during 79 minutes were used, yielding multiple faces of the same person. Different from these systems, our proposed approach only have one image of each person available, i.e., it identifies faces with a single instance of training data. In other words, CHILDATTEND is a one-shot face recognition system.

Recently, a novel approach detects and identifies faces in images with multiple people (Bah and Ming, 2019). Particularly, the Haar face detector is used to detect faces, which are then used as input to a face recognition mechanism. Additionally, considering the accuracy of the algorithms, the authors evaluate the Haar classifier, the Local binary patterns (LBP) classifier and its improved versions. Different from this approach, we evaluate the accuracy of four different face detectors and we use a single face recognition approach with Resnet to extract face encodings and the Euclidean distance to measure the similarities between the faces.

4 CHILDATTEND APPROACH

In this section we present our proposed approach. In particular, we use two face detectors, Haar and YOLO. Our choice was based on the experimental results in order to maximize the accuracy of the process while maintaining a reasonably high speed. It is important to remember that to recognize a face, we first need an "example" of that face. Figure 1 shows the insert strategy to add a new person. This strategy was adopted to test and evaluate the following research questions:

- How do gray-scale images affect the accuracy of the face recognition in different types of images?
- Will the accuracy of the face recognition increase if we align the faces beforehand?

4.1 The Insertion Strategy

From Figure 1, we observe that the first step of the insertion strategy is face detection. The Haar detector receives the input image and if it fails to locate a face in the image, we use the second YOLO detector. If neither detector locates a face in the image, the process ends, else we extract the face coordinates (left, top, width and height).

In the next step, we align the cropped face using face landmarks. First, we compute the center of mass for each eye, then we compute the angle between the eye centroids. Next, we calculate the correct position of the eyes and set the scale of the new resulting image by taking the ratio of the distance between eyes in the current image to the ratio of distance between eyes in the desired image. Finally, we compute the center

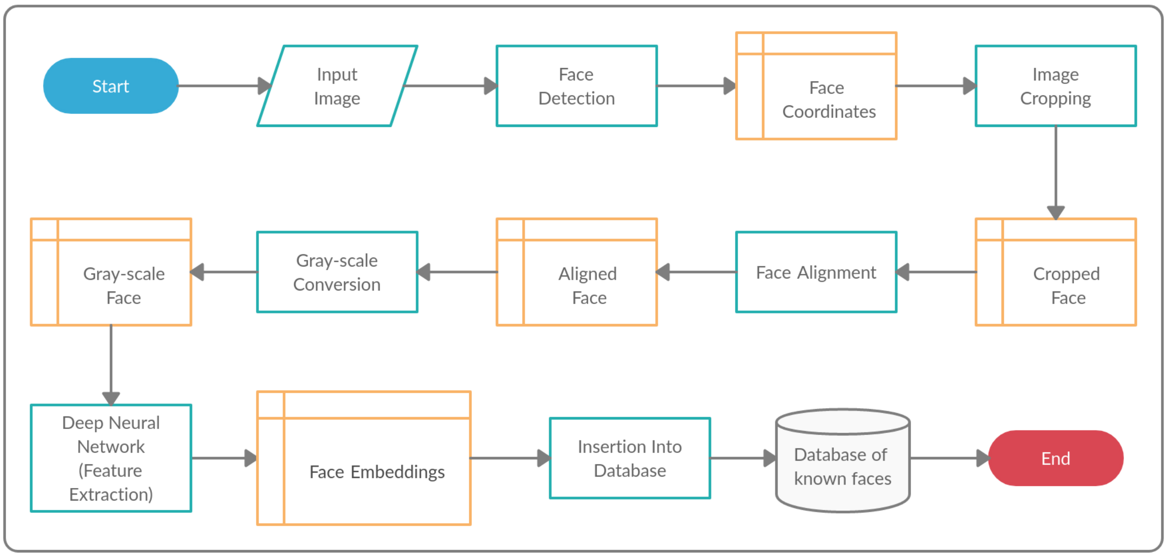

Figure 1: General workflow of our insertion strategy.

coordinates, the median point between the two eyes in the input image. After this process, we have our aligned and resized face of 256x256 pixels. Resizing is a form of normalization to guarantee that all images have the same resolution. Additionally, a smaller resolution makes the algorithms run faster without losing accuracy. After the image is aligned, we convert it from RGB to gray-scale.

The final step is feature extraction, where we extract the facial features using a Resnet that maps human faces into 128D vectors. The model is a ResNet with 29 convolutional layers. It is essentially a version of the ResNet-34 network (He et al., 2016) with a few layers removed and the number of filters per layer reduced by half. After we have the 128D vector that represents the face, we insert it into our database of known faces.

4.2 The Recognition Strategy

Different from the insertion strategy, in the recognition strategy we only use the YOLO detector, since it outperforms Haar in images with lighting issues and with several people in many different positions. Particularly, in the detection step it receives a completely new image as input. Next, it detects the faces, yielding a list of face coordinates that can contain one or more face locations. The next steps are exactly the same as in the insertion process. For each face located in the image, the algorithm crops, aligns and resizes it. This gives us a list of aligned cropped faces, each with a resolution of 256x256 pixels. We then convert each aligned face to gray-scale. Now that we have a list of aligned faces in gray-scale that was outputted by the face detector, we enter the recognition stage, in which a neural network is used to extract the features of each face. After the process of feature extraction is complete, we have a brand new list of 128D vector representations (embeddings) of the faces found in the new input image.

In the last stage of the recognition, we compare the faces by their levels of similarity. The 1:N comparison is done using the Euclidean distance. For each vector in the new list of vectors, we calculate its distance to every other vector previously inserted into our database. And, as previously noted, the smallest distance is used to find the label of the recognized person. However, the smallest Euclidean distance must be less or equal than a previously chosen threshold, otherwise the algorithm will label the face as ’Unknown’.

5 EXPERIMENTAL SETUP

In this section, we describe the experimental setup that supports our investigation. Particularly, we address the following research questions:

- How does the change in the number of jitters impact the accuracy of the face recognition?
- How does the color of the image impact the accuracy of the face recognition?
- Will the accuracy of the face recognition increase if we align the faces beforehand?

5.1 Dataset

Our data consists of two different sets. The first set contains 18 pictures of individual children's front faces with good lighting conditions and the second set contains 23 images, most of them characterized by groups of people (including adults and children), and the same people who are present in the first set of images of individual faces are also present in the images in the second set. Each of the images of groups of people in the second set was analysed manually. Thus, we created a third set by combining the two previous sets, selecting some images from the second set and discarding others. This third new set is described as follows:

- It contains 36 images of 18 people, with 2 distinct images of each person in different conditions.
- One image contains the person's front face, with good lighting and without occlusion. The other image is a photo of the same person, only in different conditions, which include occlusion, bad lighting and different poses. Additionally, some looked older because it is natural for children's faces to change considerably over the years.
- The images have different resolutions, ranging from 283x565 to 3888x6912 pixels. However, after the process of alignment described in Section 4.1, they present the same resolution of 256x256 pixels.

5.2 Experimental Procedures

Figure 2 shows the workflow of the experimental procedures we used to evaluate our proposed approach.

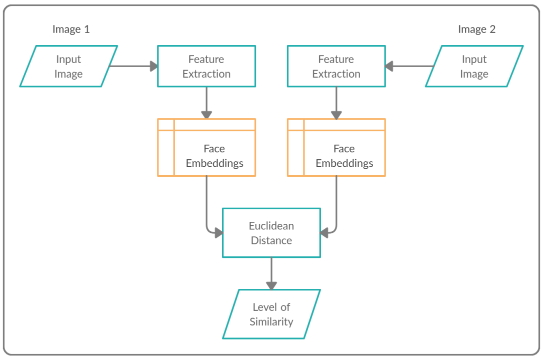

Figure 2: Workflow of our experimental procedures.

From Figure 2 we observe that the leftmost input image (*Image 1*) represents a front face with good lighting and without occlusion. The rightmost input image (*Image 2*) represents an angled face with poor lighting and occlusion. As usual, we first detect the faces in the images. Four different face detectors were evaluated in our experiments: i) Haar[2]; ii) HOG[3]; iii) MTCNN[4], and; iv) YOLO[5].

Additionally, we extract the face embeddings by using ResNet[6] from each face in each image and compare them using the Euclidean distance. The output of this process is a number between 0 and 1, which gives us the level of similarity between the faces, that is, the higher the value (closer to 1), the more different the two faces are. Therefore, a reduction in the Euclidean distance would mean an increase in the accuracy of the face recognition algorithm, considering that the comparison is done using two different photos of the same person in different circumstances. So, we used these facts to verify the impact of the change in the number of jitters, the color of the image and the alignment of the faces in the recognition process.

We changed the parameters and analysed the Euclidean distance in different iterations. In the first iteration, we used the original RGB images with no face alignment. In the next iteration, we used gray-scale images with no face alignment. In the final iteration, we used gray-scale images with face alignment. In addition, we also changed the number of jitters in each iteration (only during the process of insertion). Thus, we test whether a more general vector representation of the face would significantly improve the accuracy of the recognition or not.

As stated in the previous sections, we also tested the accuracy of all 4 face detectors (Haar, HOG, MTCNN and YOLO) using the images from the second set. We manually labelled the images, counting the number of faces in each one. Then, we ran each face detector to compare their results and measure their detection, as well as their execution time. Finally, we also used the second set to test the face recognition algorithm based on the recognition rate. For each image, we ran the YOLO detector (since it is the best of all 4 detectors) to find the faces. Then, the detected faces were used as input to the face recognizer. Considering the face recognizer itself, we evaluated the recognition rate in different iterations, since the face recognizer requires a threshold for identifying faces. Therefore, we varied its value from 0.30 to 0.70, since values below 0.30 and above 0.70 did not show an improvement in the recognition.

[2] https://docs.opencv.org/

[3] http://dlib.net/face_detector.py.html

[4] https://pypi.org/project/mtcnn/

[5] https://pjreddie.com/darknet/yolo/

[6] https://face-recognition.readthedocs.io/en/latest

6 EXPERIMENTAL RESULTS

In this section, we describe the experiments we performed to evaluate our approach. Significance is verified with the ANOVA test with a confidence level of 95%. Table 1 show the averages of the Euclidean distances in each different circumstance. The ”Jitters” column shows the number of times the images were re-sampled in an attempt to create a more generalized representation of the faces. The columns ”RGB”, ”Gray-scale” and ”Gray-scale aligned” show the averages of the Euclidean distances using color images without face alignment, gray-scale images without face alignment and gray-scale images with the face alignment algorithm applied, respectively.

Table 1: Averages of the Euclidean distances.

Jitters	RGB	Gray-scale	Aligned Gray-scale
1	0.5231	0.5008	0.5447
2	0.5108	0.4931	0.5411
3	0.5066	0.4887	0.5381
4	0.5055	0.4888	0.5374
5	0.5064	0.4882	0.5383
10	0.5049	0.4863	0.5386

From Table 1 we can observe that the smallest absolute distance corresponds to gray-scale images without face alignment and 10 jitters. The greatest absolute distance corresponds to gray-scale images with face alignment and 3 jitters. Thus, the null hypothesis was rejected, considering the f-ratio value of 129.2294 and the p-value of 0.00001. Since the result is significant at $p < 0.05$, it is possible to conclude that there was, indeed, significant difference between the groups. As a result, the best configuration that minimizes the Euclidean distance, maximizing the accuracy of the face recognition algorithm, is the use of gray-scale images without alignment and 10 jitters. Therefore, we answer the first research question stated in Section 5, so tne number of jitters impact the accuracy of the face recognition. Additionally, experiments also answer the second research question stated in Section 5, so converting the images to gray-scale does increase the accuracy of the face recognition. Also, a higher number of jitters gives us a higher accuracy, to a certain extent. On the other hand, the alignment of the faces does not improve the accuracy of the algorithm. In fact, in our case it hinders the process. Moreover, the process of face alignment is costly and hurts the speed of the overall process.

Regarding the accuracy of the detectors, 23 photos were analyzed using each detector. We kept record of the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Those values were then used to calculate the accuracy of the algorithms, using the following formula:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

Table 2 shows the level of accuracy of each face detection algorithm. We can observe that the Haar detector has the lowest absolute accuracy among the 4 algorithms. However, if we consider the confidence intervals of the Haar, HOG and MTCNN detectors, there is no statistical difference between them.

Table 2: Face detection accuracy.

Algorithm	Accuracy(%)
Haar	23.8213 ± 12.9518
HOG	34.3635 ± 14.5136
MTCNN	41.4500 ± 13.3559
YOLO	90.4978 ± 8.4235

However, it is clear that the YOLO detector is significantly more accurate than the others. Its absolute accuracy is greater than Haar, HOG and MTCNN by 66.67%, 56.13%, 49.04%, respectively. In addition, considering the confidence intervals, YOLO beats the Haar detector by at least 45.30%, the HOG detector by at least 33.19% and the MTCNN detector by at least 27.26%.

We also evaluated the speed of the algorithms. The following values presented in Table 3 are the result of 10 iterations (the same 23 photos were analyzed in each iteration) with confidence level of 95%.

Table 3: Average detection time per photo.

Algorithm	Time (seconds)
Haar	1.0056 ± 0.0245
HOG	2.6129 ± 0.0067
MTCNN	5.3393 ± 0.0618
YOLO	1.7322 ± 0.0101

From Table 3 we observe that there is a significant difference between the 4 algorithms in terms of speed, with the Haar detector having a faster execution time, with an average detection time of approximately one second. Next, we have the YOLO detector with an average detection time of 1.7322 seconds. Finally, we have the 2 slowest detectors, HOG and MTCNN, with an average detection time of approximately 2.61 seconds and 5.33 seconds, respectively. In addition, if we evaluae the total detection time for 23 photos, the results are even more significant, as shown in Table 4. From Table 4 we observe that the Haar detector has a total average detection time of approximately 23.14

Table 4: Total detection time.

Algorithm	Time (seconds)
Haar	23.1405 ± 0.5652
HOG	60.1057 ± 0.1547
MTCNN	122.8183 ± 1.4247
YOLO	39.8513 ± 0.2349

seconds, while the HOG and the MTCNN detectors have a total average detection time of approximately 60.10 seconds and 122.81 seconds, respectively. We can also observe that the YOLO detector is faster than the HOG and the MTCNN detectors, with an average detection time of approximately 39.85 seconds. And although it is slower than the Haar detector, it presents an outstanding balance between accuracy and speed compared to the other detectors.

Considering the performance of the facial recognition, we evaluated its accuracy using the same metric as (Bah and Ming, 2019), based on the recognition rate (RR):

$$RR = \frac{TotalFaces - TotalFalseRecognitions}{TotalFaces}. \quad (2)$$

We used the YOLO detector to find the faces and we varied the threshold values. A total of 248 faces were found in all images in each iteration. Based on the results presented in Table 5 we can conclude that the best threshold value that maximizes the recognition rate is 0.50, which gives us a recognition rate of 93.14%.

Table 5: Recognition rate (RR).

Threshold	False Recognition	RR
0.30	36	0.8548
0.35	30	0.8790
0.40	22	0.9112
0.45	21	0.9153
0.50	17	0.9314
0.55	44	0.8225
0.60	94	0.6209
0.65	174	0.2983
0.70	215	0.1330

Note that our model is based on a one-shot face recognition approach. Yet, this issue can be easily solved if we have more than one image of each person for training. Then, we can simply apply a classifier, e.g., K-Nearest Neighbors (KNN), and identify the person. Nevertheless, it is also possible to observe that a high threshold (close to 1.00) causes a lower recognition rate, which was already expected, since a higher threshold means that the algorithm becomes more flexible. Furthermore, YOLO detector has a slight negative impact in the face recognition process, since the faces in our dataset are angled, making it difficult to align the faces that are correctly detected (the true positives).

7 CONCLUSIONS

In this article we propose CHILDATTEND, a neural network based approach to assess child attendance in social projects, which exploits face detection, recognition and alignment to find, identify and label children's faces in digital images. Experimental results showed that our approach is fast and identifies children in group photos with more than 90% accuracy.

Additionally, we thoroughly evaluated face detection algorithms, and experimental results showed that the YOLO detector performs better than the other ones, with an average detection rate of more than 90%. Moreover, in terms of detection speed, the Haar algorithm performs better than the other detectors, although providing a low detection rate. Thus, we combined the two algorithms to produce a balanced (fast and accurate) result. The best threshold value that maximizes the performance of the recognition algorithm was 0.50, providing a one-shot recognition rate of 93.14%. We believe that using a classifier with multiple photos of the same person will bring an even better result.

In future work we plan to carry out a more extensive assessment of image processing techniques and strategies to improve the accuracy of the facial recognition. We also plan to use other datasets with multiple instances of the same face in order to assess the impact of different classifiers on the recognition rate.

ACKNOWLEDGEMENTS

The present work was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Financing Code 001. The authors thank the partial support of the CNPq (Brazilian National Council for Scientific and Technological Development), FAPEMIG (Foundation for Research and Scientific and Technological Development of Minas Gerais), and PUC Minas.

REFERENCES

Bah, S. and Ming, F. (2019). An improved face recognition algorithm and its application in attendance management system. *Array*, 5:100014.

Chandrappa, D. N., Ravishankar, M., and RameshBabu, D. R. (2011). Face detection in color images using skin color model algorithm based on skin color information. In *Proceedings of the ICECT '11*, pages 254–258.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the CVPR '05*, page 886–893.

Guillaumin, M., Verbeek, J., and Schmid, C. (2009). Is that you? metric learning approaches for face identification. In *Proceedings of the ICCV '05*, pages 498–505.

Hazim, N. (2016). Improve face recognition rate using different image pre-processing techniques. *American Journal of Engineering Research*, 5:46–53.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '16, pages 770–778.

Heisele, B., Poggio, T., and Pontil, M. (2001). Face detection in still gray images. Technical Report 1687, Massachusetts Institute of Technology.

Hjelmås, E. and Low, B. (2001). Face detection: A survey. *Computer Vision and Image Understanding*, 83:236–274.

Jafri, R. and Arabnia, H. (2009). A survey of face recognition techniques. *Journal Of Information Processing Systems*, 5:41–68.

Kanan, C. and Cottrell, G. (2012). Color-to-grayscale: Does the method matter in image recognition? *PloS one*, 7:e29740.

Kawaguchi, Y., Shoji, T., Lin, W., Kakusho, K., and Minoh, M. (2005). Face recognition-based lecture attendance system. *International Journal of Engineering Research & Technology*, 2(4).

Kumar, A., Kaur, A., and Kumar, M. (2019). Face detection techniques: A review. *Artificial Intelligence Review*, 52.

Li, Z., Feng, W., Zhou, J., Dan, C., and Peiyan, Z. (2017). Research on mobile commerce payment management based on the face biometric authentication. *International Journal of Mobile Communications*, 15:278.

Liwei Wang, Yan Zhang, and Jufu Feng (2005). On the euclidean distance of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1334–1339.

Malikovich, K. M., Ugli, I. S. Z., and O'ktamovna, D. L. (2017). Problems in face recognition systems and their solving ways. In *Proceedings of the ICISCT '17*, pages 1–4.

Ming-Hsuan Yang, Kriegman, D. J., and Ahuja, N. (2002). Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58.

Patel, K., Han, H., and Jain, A. K. (2016). Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283.

Peixoto, B., Michelassi, C., and Rocha, A. (2011). Face liveness detection under bad illumination conditions. In *Proceedings of the ICIP '11*, pages 3557–3560.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the CVPR '16*, pages 779–788.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767.

Rein-Lien Hsu, Abdel-Mottaleb, M., and Jain, A. K. (2002). Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706.

Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the ACM CCS '16*, page 1528–1540.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the CVPR '01*.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys*, 35:399–458.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the CVPR '12*, pages 2879–2886.

Zulfiqar, M., Syed, F., Khan, M. J., and Khurshid, K. (2019). Deep face recognition for biometric authentication. In *Proceedings of the ICECCE '19*, pages 1–6.