Face identification based on synergism of classifiers in rectified stereo images

Diedre Carmo, Raul Alves, Luciano Oliveira
Intelligent Vision Research Lab
Federal University of Bahia
Salvador, Bahia, Brasil
http://ivisionlab.dcc.ufba.br

***Abstract*—This paper proposes a method to identify faces from a stereo camera. Our approach tries to avoid common problems that come with using only one camera that shall arise while detecting from a relatively unstable view in real world applications. The proposed approach exploits the use of a local binary pattern (LBP) to describe the faces in each image of the stereo camera, after detecting the face using the Viola-Jones' method. LBP histogram feeds then multilayer perceptron (MLP) and support vector machines (SVM) classifiers to identify the faces detected in each stereo image, considering a database of target faces. Computational cost problems due to the use of dual cameras are alleviated with the use of co-planar rectified images, achieved through calibration of the stereo camera. Performance is assessed using the well established Yale face dataset, and performance is assessed by using only one or both camera images.**

I. Introduction

A face identification system is defined basically by a system capable of verifying the identity of a person from an image or video feed. One of the most used ways of achieving that is by comparing extracted *features* from the captured face with pre-loaded representations from each identification target. Face identification systems are commonly used in security systems, using the face of a person as a biometric information [1].

Some examples of past endeavors in frontal face identification are [2], [3], [4]. Schwartz *et al.* [2] used a set of feature descriptors and partial least squares (PLS) to perform feature weighting [5] as a tree search optimization for the one-to-all comparison method for identification; initially, Schwartz *et al.* [2] used only the histogram of oriented gradients (HOG) [6] and local binary pattern (LBP) [7], common descriptors for textures, later increasing the descriptors by a cluster of more than 7,000 features for better results [8]. Guillaumin *et al.* [3] computes the probability of two images to belong to the same class using a k-nearest-neighbour approach, achieving good results on hard datasets (for the time). Min *et al.* [4] registers a point-cloud model of frontal faces using active depth sensors. A comparison to a point cloud models of the person to be identified is realized, with minimum euclidean distance as a identification criteria. Even though near 100% accuracy is achieved, a small easy dataset is used and different angles and lighting were not tested properly.

These works show that, in controlled environments, the classical approach of detecting a face in one image view and comparing face features to a database reaches almost 100% accuracy in identification. However, due to the hazardous nature of real world environments, changes in face angle, occlusion, lighting and variations in distance to the camera are to be expected in real life applications. Many recent works have been trying to come with a solution to overcome those problems.

Recently, there is a growing trend in using neural networks and deep learning in face identification [9] [10] [11]. Part of this advancement is due to the availability of larger and more complex face datasets, such as [12], allowing the excellent training performance of convolutional neural networks (CNN) to shine. They have shown state-of-the-art performance in ignoring lighting, angle, and occlusions for face segmentation. Not only that, but performance ratings for recognition surpass previous approaches. Another growing area of interest in face identification is the use of multiple views and 3D reconstructions of the face to try and ignore the aforementioned problems, found in works such as [13] [14].

A. Proposal

This work proposes an approach, consisting of using facial stereo images to improve quality of the identification. A target face is detected on the left image, and translated to the second image via calibration parameters and rectification. Both face regions are analyzed with extraction of LBP features and an one-to-all comparison to each identification target is performed. The most likely person is found by taking into account: the output of multiple MLP networks and SVM classifiers trained to identify each person (see Fig. 1); the redundancy of the evaluation possibly returns different results from the stereo images.

II. Face Identification Based on Stereo Redundancy

This work implements stereo redundancy in the form of analyzing the same face from two slightly different views, using different methods for classification on each view. This way, we intend to improve the reliability of the identification. Figure 1 summarizes our proposed method depicting five main steps, which will be covered in detail in the following sections, as follows:

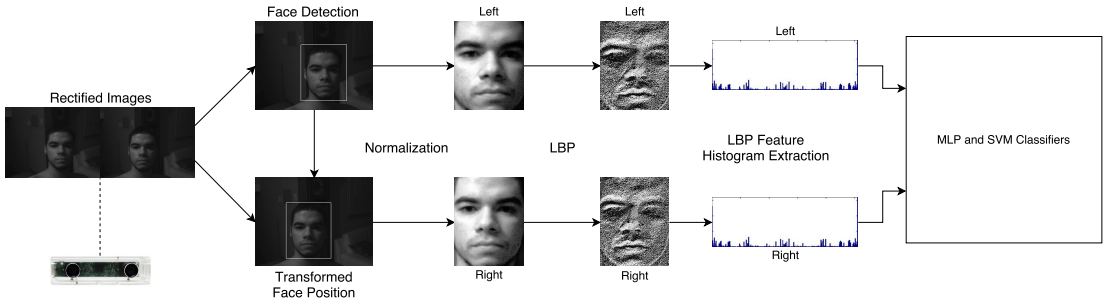


Fig. 1. Outline of our method: Left and right images are captured from the camera, and are rectificated to ensure the images are on the same plane, using pre calculated calibration parameters. LBP haar cascades [15] classifiers are used to detect a face in one image. The detected area is translated to represent the same face in the other image. Face regions are manually normalized to remove background and approach them to trained faces. LBP is extracted again, now only from the normalized face region, to be fed into neural networks and SVM classifiers trained in recognizing each target in the dataset, finally outputing a identification rank.

- Two images are acquired used an E-con Systems Tara Stereo Camera [16], which outputs two 752×480 resolution images, and provides calibration parameters for the camera along with an SDK for its use.
- Using the provided calibration parameters, rectification [17] of the image is performed to ensure the images are on the same plane, and to allow a transformation from one image to the other to only involve translation.
- Face detection is performed in one of the images using Viola-Jones' [15] detector. Due to the images being in the same horizontal plane, face regions can be translated from one image to another. This results in better timing performance for the algorithm, removing the need to recompute faces in the second image.
- Due to the expected different distances and angles for the detected face, faces are normalized with three operations: cropping the image to remove hair regions and background, resize to a fixed size of 168×192, the same used in the Yale dataset; and image histograms are also equalized to reduce the impact of lighting variations.
- LBP [7] descriptors are extracted only from the normalized face region, and fed into MLP and SVM classifiers.
- An MLP and an SVM are created in training phase for each identification target, trained over LBP histograms. Those classifiers return prediction rates for each target and each view of the camera. A sort of the results by the higher prediction issues an identification ranking.

A. Rectification

To allow correct redundancy between the captured faces in different cameras, a calibration and rectification process is needed. The used camera comes intrinsically and extrinsically calibrated with its parameters stored inside it, which are easily retrieved through its SDK. They are used to achieve horizontal rectification [17], causing the first and the second camera views to be shifted relative to each other, only along the X-axis.

Image rectification allows the transformation of points between images. Making it possible to realize detection of faces in both images using only one image. Computational cost for the face detection is effectively halved.

B. Feature Extraction

LBP is a feature extractor that operates in pixels and its neighbors. Considered a good texture descriptor and resistant to lightning variations, it is often used in classification problems where texture differentiation between observed objects is important [18]. Obtaining LBP occurs as follows:

- Divide the examined image into cells (for example, regions of 16
  16 pixels);
- For each window, compare a non-edge pixel to each of its neighbors, excluding the pixels from the window border;
- Construct a binary word for each comparison, where 0 is the center pixel is higher and 1 is the center pixel is smaller;
- Construct the histogram for each cell, which will contain information on the frequency of occurrence of each LBP value;
- Concatenate the histograms of all cells in a single descriptor for the image.

The size used for the cells of the first step can greatly influence the discriminant power of the Local Binary Pattern. The selection of small cells leads to the growth of the size of the final histogram in a quadratic way, increasing the computation time of the LBP. However, the use of larger cells may lead to poorer performance. In this work, we observed high computation times for a 16×16 windows. Increasing window size until 24×24 did not change our results and provided faster execution times.

C. Face normalization

After detecting the face, normalization takes places in the form of manual cropping of each face image to remove the area around the face and resize it to a default size (168×192). After that, an image histogram equalization [19] is performed to improve image illumination.

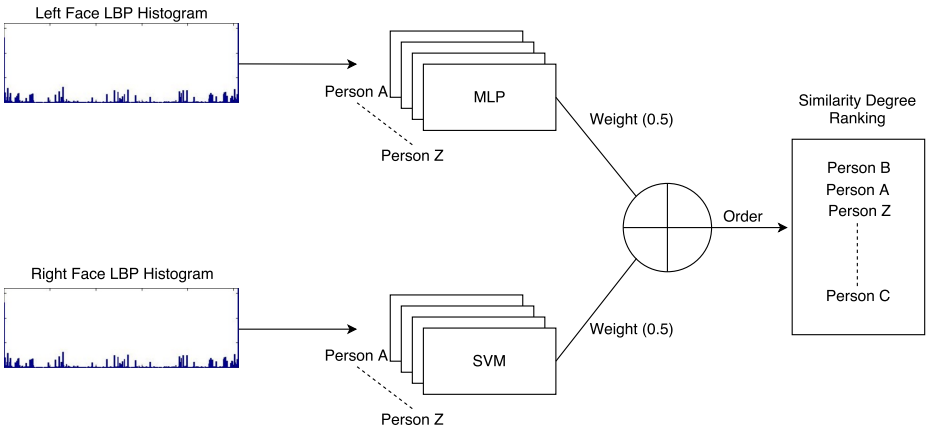


Fig. 2. Architecture of the face identification. Every target in the data set has an MLP and an SVM trained into identifying them. A probe input is evaluated by all classifiers, outputting a prediction number that, when sorted, constitutes a ranking of the targets most related to the input.

D. Face Identification

Identification is performed using LBP histograms of the input image as an input to one classifier for each camera view. For each target, MLP and a linear SVM classifiers are trained (see figure 2). Parameters used are as in table I. The choice of this classifiers is justified in our results in Section III.

To train the classifiers, we considered each person in the target database. For each one, the face of the person was considered the positive class, while the other people were considered the negative class.

Identification process occurs as follows: The image's LBP descriptor is given to all MLP or SVM classifiers that are trained in the system, having the left face LBP to feed to the MLPs and the right face LBP to feed to the SVMs. This operation returns a ranking of responses from the classifiers, showing the distance between the input probe face and every face registered in the target database. A "degree of similarity" is composed from those responses in the following way: the MLP and the SVM prediction operations return a floating point response equal to -1 or 1 (no or yes). These responses are normalized into a 0 to 100 general response, named degree of similarity (see Fig. 2). By sorting the scores of the classifiers for each person, it is possible to infer who are the top-N targets. It is easy to see that if the output of the SVM and MLP diverge, this degree of similarity is affected.

III. Experimental Analysis

The Yale dataset [20] [21] was used to validate the proposed identification method (results can be seen in Fig. 3b) First, we justify the classifier choice with a comparison between common supervised classifiers, performing on binary detection of a person from the Yale dataset, with a 5-fold cross validation approach, and analyzing the ROC curve for them (see Fig. 3). Second, a cumulative curve for the final identification was plotted. Probes for the identification test were two images from the same person in the dataset, one as the left image and another as the right image, in a 10% hold-off validation strategy. Also, performance is compared with and without the redundant identification (see Fig. 3b). Tests were run in a i5-4200U processor.

TABLE I
Parameters for evaluated classifiers

Classifier	Parameters
Decision Tree	Gini's diversity index split criterion; 4 maximum splits
SVM	Linear Kernel; Kernel Scale: 30
Neural Network	Multilayer Perceptron; backpropagation method; 0.0001 weight scale
Coarse KNN	100 neighbors; euclidean distance; equal weight
Gaussian SVM	Gaussian Kernel, Kernel Scale: 30
AdaBoost	30 learners; 0.1 learning rate; 20 splits; RUSBoost method

A. Discussion

Best performance was observed in the Linear SVM and MLP classifier, according to Fig. 3a), making them the choice for the generation of similarity degrees for our synergistic classification. The identification phase showed 84.2% perfomance on top-1 identification on the Yale dataset (see figure 3b), although with some identifications performing sub-optimally (outside of the top-1).

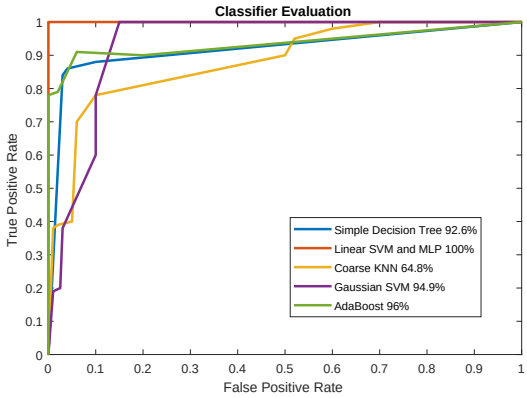
Our method presented surprisingly good timing performance of 3 to 5 identifications per second while running in real-time with 40 identification targets.

Most of the probes that were not identified in the top-1 result were identified in the top-2 or 3 as a similar person, as visible in the cumulative curve (Fig. 3b). This indicates that using this method with too many targets in the dataset will most likely deteriorate its accuracy and timing performance.

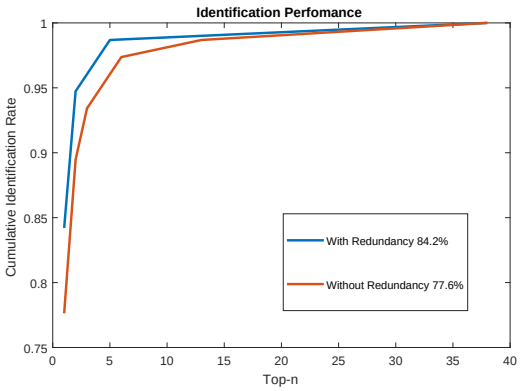
B. Why not use the depth image from the stereo view?

Preliminary evaluation showed that the accuracy in the depth map extracted from the used camera was not good enough to extract accurate face features from the image. This is supported by observations from many studies such as [22]. Usually, some additional techniques are used to achieve better precision with passive stereo vision.

Due to its low depth accuracy, the stereo vision was used as means of achieving redundancy instead of depth. Analysis



(a) Evaluation of the best classifier to be used when generating similarity degrees for a person.



(b) Cumulative results for identification with and without both classifiers. These results show that using both cameras improves perfomance on the Yale Dataset.

Fig. 3. Identification and Classification performance.

of the same face twice is performed but without the computational cost of face detecting twice, due to calibration and rectification of the images. The redundancy showed a small improvement in detection rate, as visible in our results (see Fig. 3).

IV. Conclusion

This paper presented a method for face identification using a stereo camera that showed 84.2% accuracy in face identification on the Yale Dataset results for a relatively simple face fronting dataset, with lighting and slight angle variations. This performance is not close to state-of-the-art face identification algorithms. More testing and fine tuning needs to be done in the future. We plan to apply this method into harder datasets that are more likely to represent real world performance.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[2] W. Schwartz, H. Guo, and L. Davis, "A robust and scalable approach to face identification," *Computer Vision–ECCV 2010*, pp. 476–489, 2010.

[3] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Computer Vision, 2009 IEEE 12th international conference on*. IEEE, 2009, pp. 498–505.

[4] R. Min, J. Choi, G. Medioni, and J.-L. Dugelay, "Real-time 3d face identification from a depth camera," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 1739–1742.

[5] H. Wold, "Partial least squares," *Encyclopedia of statistical sciences*, 1985.

[6] C. Shu, X. Ding, and C. Fang, "Histogram of the oriented gradient for face recognition," *Tsinghua Science & Technology*, vol. 16, no. 2, pp. 216–224, 2011.

[7] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[8] W. R. Schwartz, H. Guo, J. Choi, and L. S. Davis, "Face identification using large feature sets," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2245–2255, 2012.

[9] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[13] H. Drira, B. B. Amor, A. Srivastava, M. Daoudi, and R. Slama, "3d face recognition under expressions, occlusions, and pose variations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2270–2283, 2013.

[14] Y. Lei, M. Bennamoun, M. Hayat, and Y. Guo, "An efficient 3d face recognition approach using local geometrical signatures," *Pattern Recognition*, vol. 47, no. 2, pp. 509–524, 2014.

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[16] econSystems, "Usb stereo camera — 3d depth camera (oem)," 2017, https://www.e-consystems.com/3D-USB-stereo-camera.asp, acessado 5 de Jun de 2017.

[17] C. Loop and Z. Zhang, "Computing rectifying homographies for stereo vision," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1. IEEE, 1999, pp. 125–131.

[18] R. Rouhi, M. Amiri, and B. Irannejad, "A review on feature extraction techniques in face recognition," *Signal & Image Processing*, vol. 3, no. 6, p. 1, 2012.

[19] C.-h. Chen, *Handbook of pattern recognition and computer vision*. World Scientific, 2015.

[20] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[21] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[22] N. Uchida, T. Shibahara, T. Aoki, H. Nakajima, and K. Kobayashi, "3d face recognition using passive stereo vision," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2. IEEE, 2005, pp. II–950.