# Bi-modal Face Recognition
## *How combining 2D and 3D Clues Can Increase the Precision*

Amel Aissaoui[1] and Jean Martinet[2]
[1]*USTHB, Bab Ezzouar, Algeria*
[2]*Lille 1 University, Villeneuve-d'Ascq, France*

Keywords: Face Recognition, Multimodal, 2D, 3D, LBP, RGB-depth.

Abstract: This paper introduces a bi-modal face recognition approach. The objective is to study how combining depth and intensity information can increase face recognition precision. In the proposed approach, local features based on LBP (Local Binary Pattern) and DLBP (Depth Local Binary Pattern) are extracted from intensity and depth images respectively. Our approach combines the results of classifiers trained on extracted intensity and depth cues in order to identify faces. Experiments are performed on three datasets: Texas 3D face dataset, BOSPHORUS 3D face dataset and FRGC 3D face dataset. The obtained results demonstrate the enhanced performance of the proposed method compared to mono-modal (2D or 3D) face recognition. Most processes of the proposed system are performed automatically. It leads to a potential prototype of face recognition using the latest RGB-D sensors, such as Microsoft Kinect or Intel RealSense 3D Camera.

## 1 INTRODUCTION

2D face recognition is a mature research domain, with typically high matching precision in specific contexts. It mainly deals with images – color or grey-level – acquired via a usual camera, representing faces' visual appearance. Because all human faces are similar in their configurations (inter-class similarity) and the appearance of a given face can greatly vary due to e.g. changes in pose, expression, or illumination (intra-class variation), unconstrained 2D face recognition is a hard problem. Indeed, using only face intensity images yields to a high sensitivity to intra-class variations. The past decades have witnessed tremendous efforts focused towards 2D face images (Zhao et al., 2003). Despite the great progress achieved so far within the field, 2D image is still not reliable enough (Abate et al., 2007), especially in the presence of pose and illumination changes (Phillips et al., 2000).

3D face recognition was proposed as a potential solution for face recognition. These approaches are based on 3D shape data from faces (Bowyer et al., 2006) obtained with specific equipments such as laser scanners. Using the face 3D shape allows a high invariance to illumination and pose variation; a high face recognition accuracy was reported in the literature using 3D approaches (Phillips et al., 2005). However, 3D data need expensive equipment and human cooperation, which limits their application field. Besides, they always require accurate registration before shape-based 3D matching. Matching methods of 3D scans are very expensive in terms of CPU resources and time processing since they are based on optimization of complex geometric equations. In order to bypass the time processing issue of the 3D matching methods, researchers now tend to focus on using face range images instead of 3D point clouds. Such kind of methods emerged recently specially with the rapid development in 3D imaging systems, such as time-of-flight cameras or the Microsoft infrared camera Kinect, which brought a major solution to use range images without dealing with expensive equipments.

Although 3D recognition does not suffer from light or pose changes, it does not benefit from visual information – such as textures. In recent years, the wide availability of RGB-D sensors has opened new research directions in face recognition and bi-modal 2D+3D face recognition using intensity and depth information received a high attention from researchers in the field. Depth and texture play complementary roles in the coding of faces as they typically represent different characteristics of the face to be recognized. The 2D image provides informations about face textured regions with little geometric structure (e.g. hairy parts, eyes, eyebrows), and the 3D data

Aissaoui A. and Martinet J..
Bi-modal Face Recognition - How combining 2D and 3D Clues Can Increase the Precision.
DOI: 10.5220/0005359605590564
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 559-564
ISBN: 978-989-758-090-1

provides informations regarding less textured regions (e.g. nose, chin, cheeks). Hence, using the shape information in addition to the intensity images allows a better face representation and therefore better precision and robustness in face recognition.

In this paper, we introduce a novel bi-modal framework for face recognition. Our objective is to use the intensity and depth information in order to obtain more accuracy and robustness in face recognition compared to mono-modal approaches. Our approach is based on local attributes calculated from intensity and depth face images. We use the well known LBP descriptor (Ojala et al., 2001; Ojala et al., 2002) for intensity feature extraction. Aside from its simplicity and compactness, the LBP have demonstrated a good discriminative power in face recognition literature (Huang et al., 2011). Unlike most bi-modal approaches that use the same descriptor for both modalities (Chang et al., 2003; Xu et al., 2009; Jahanbin et al., 2011; Huang et al., 2009), we use the DLBP, which is an extension of the LBP dedicated for range images, in order to extract the depth cues. This allows extracting of more discriminative features from depth images. Facial cues from the 2D and 3D images are used to train classifiers, and their decisions are combined in order to find the final identity decision.

The remainder of the paper is organized as follows. Section 2 describes related works on 2D and 3D face recognition methods and show the complementarity of both techniques. Section 3 introduces the proposed bi-modal approach for face recognition. Section 4 presents experimental results on three collections that demonstrates the performances of the proposed method. Section 5 gives a conclusion to our work.

## 2 RELATED WORKS

Many approaches have been proposed for face recognition from intensity images (Abate et al., 2007; Bowyer et al., 2006) and a multitude of 2D face descriptors was proposed in the literature (Zhao et al., 2003). Among these descriptors, Local Binary Patterns (LBP), first proposed by Ojala et al. (Ojala et al., 2001; Ojala et al., 2002), are widely used. They encodes pixel-wise information in a given image. The operator describes each pixel with the relative grey-level values of its neighboring pixels:

$$LBP_{(R,V)}(x,y) = \sum_{i=0}^{V-1} s(n_i - n_c)2^i, \qquad (1)$$

$$\text{with } s(k) = \begin{cases} 1 & \text{if} \quad k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where :

- $n_c$ is the grey level value of the central pixel from the local neighborhood;
- $n_i$ are the grey level values of the neighbor pixels around the central pixel with a radius $R$.

In order to acquire 3D data without using expensive laser scanners, an increasingly popular method consists in using depth sensors, such as time-of-flight cameras or the Microsoft Kinect infrared sensor. Besides, classic image descriptors seem to be convenient for face representation in such images. The advantage of using range images is that features can be easily extracted from raw data, which allows fast processing. They also lie in a 2D dimensionality, which avoid heavy computational cost. Moreover, depth data lie in the 2D domain (while representing 3D data), and can therefore be dealt with using standard 2D imaging techniques.

Huang et al. (Huang et al., 2006) proposed an extended LBP version, named 3DLBP, designed to describe face depth images. Beside the information provided by LBP, 3DLBP also considers the magnitude of the difference between the central pixel and its neighborhood. However, this coding scheme suffers from 3 drawbacks: the feature vector size is very large, the coding principle is very sensitive to the depth variations, and it is limited to small radius. Other extensions have been proposed such as Local Derivative Patterns (Zhang et al., 2010) and more recently Local Vector Patterns (Fan and Hung, 2014; Hung and Fan, 2014) for taking into consideration high-order local pattern variations than just the first order with usual LBP. Other descriptors dedicated to face range images were recently introduced such as Depth Local Quantized Pattern (DLQP) by Mantecon et al. (Mantecn et al., 2014). In our earlier works, we have proposed the Depth Local Binary Pattern (DLBP) (Aissaoui et al., 2014) as a powerful extension to LBP dedicated to face range images. DLBP is a compact descriptor, and the coding scheme is stable according to small changes in depth values of the neighborhood. Unlike 3DLBP and DLQP, it is designed for large radiuses in order to extract more discriminative features from smooth and low-contrast data, since it works on a multi-scale level.

Bi-modal 2D+3D face recognition makes use of both modalities to represent a face, with the objective of taking advantage of both 2D and 3D data complementarity. Combining 2D and 3D information can be performed using different strategies (Husken et al., 2005). The most used strategy is the late fusion which takes place at the decision stage. Therefore, we consider this strategy when combining depth and intensity cues in the proposed system. Both 2D and

3D methods are affected by face expression changes, which is an open issue that is not addressed in this paper.

# 3 PROPOSED APPROACH FOR BI-MODAL 2D-3D FACE RECOGNITION

Our approach consists in recognizing faces using two modalities: 2D (intensity image) and 3D (depth image). First, 2D and 3D features are extracted from intensity and depth face images using the LBP and the DLBP descriptor, respectively. Then, 2D and 3D classifiers are constructed using a Support Vector Machines based training algorithm. Finally, a late fusion consisting in combining decisions from 2D and 3D classifiers is applied.

## 3.1 Feature Extraction: LBP and DLBP

LBP is considered to be one of the simplest and most efficient local 2D face descriptors. For this reason, we use the conventional LBP (See Section 2) in order to extract features from intensity images.

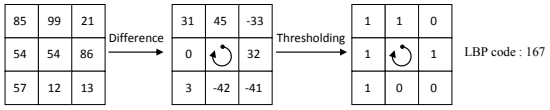


Figure 1: Illustration of LBP coding for a given pixel.

In order to extract depth features, we use the DLBP descriptor. The DLBP is founded upon two parts: the sign and the magnitude of the differences between a pixel and its neighbors. Therefore, a pixel $p(x,y)$ is represented with two codes $c_s$ and $c_m$. In order to code the sign of the neighborhood, the original LBP method is used: $c_s$ is obtained from neighborhood pixels by generating binary values according to the difference $DD$ between the central pixel and each neighbor pixel. Magnitude coding consists in assigning each neighbor pixel a binary value according to a threshold of absolute depth magnitude $T_m$. The binary sequence for $c_m$ is obtained in the same way as for the sign. Fig. 2 gives an illustration of sign magnitude coding with a radius $R = 1$, a neighborhood $N = 8$ and a magnitude threshold $T_m = 3$. The magnitude threshold is calculated automatically using the depth gradient. For more details about the DLBP principle, authors can refer to our previous work (Aissaoui et al., 2014).

After LBP and DLBP calculation, local histograms are then extracted from the LBP and DLBP

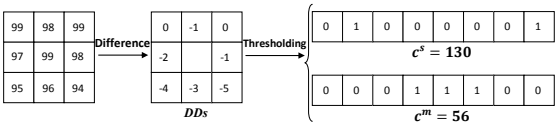


Figure 2: Illustration of $c^s$ and $c^m$ coding for a given pixel.

maps generated from intensity and depth image. Histograms offer more invariance to geometric translations, and reduce the descriptor size and consequently the processing time. That is why they are usually used in LBP-based face recognition methods. In order to preserve the spacial information, the map is divided into different regions from which histograms are extracted and then concatenated to form the LBP and the DLBP descriptor.

## 3.2 Bi-modal Recognition

After feature extraction, 2D and 3D modalities are processed independently. A face is represented by two vectors: LBP and DLBP histograms. An SVM based training is performed for each modality in order to obtain classifiers from both modalities. We choose the SVM because it is a very powerful classification method and a high accuracy was reported when using SVM for face recognition (Byun and Lee, 2002).

In order to identify a new face, the decisions obtained from the trained classifiers (from both modalities) are fused in order to find the face identity. Fusion at this stage is the most generally applicable strategy in 2D+3D face recognition task and good precision was reported using this fusion strategy. Moreover, intensity and depth information have different nature. The late fusion allows to perform different processing to each modality. The decision fusion is performed using the weighted majority vote (Xu et al., 1992). This fusion rule is among the simplest and easiest to implement but it allows us to investigate what using both intensity and depth information brings to the solution of face recognition.

# 4 EXPERIMENTS

In order to demonstrate the benefits from combining 2D and 3D modalities, we evaluate in this section the proposed bi-modal approach on three 3D face datasets: TEXAS (Gupta et al., 2010), FRGC (Phillips et al., 2005) and BOSPHORUS (Savran et al., 2008). The TEXAS 3D dataset consists of 1149 images of 118 persons. All faces are frontal with different expressions and illumination changes. The FRGC dataset contains 4007 images of 466 person with frontal views, minor pose variations and major

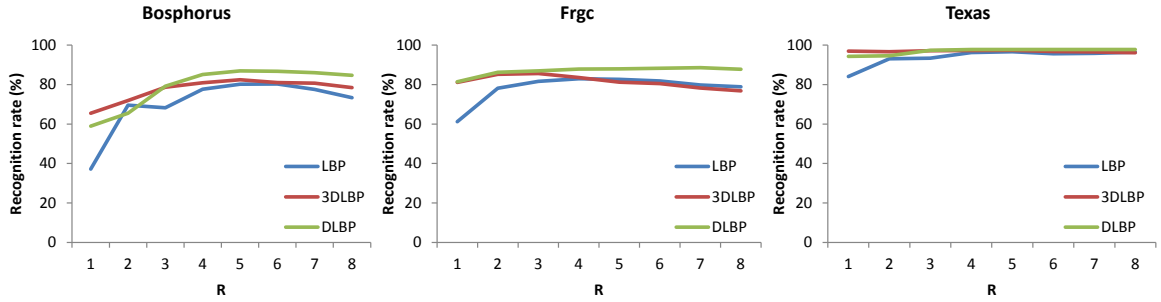


Figure 3: Comparison between LBP, 3DLBP, and DLBP for different radius values, for 3 collections.

expression and illumination variations. A number of images is removed from this dataset since the depth maps were unusable because of large holes due to the scanner artifacts. The BOSPHORUS dataset includes 4666 images of 105 persons with different expressions and pose variations. Images with large pose variations were not used from this dataset.

We have implemented the feature extraction process using different parameters for LBP and DLBP. Support Vector Machines (SVM) based on Radial Basis Function (RBF) kernels are used for classification. The precision is evaluated with a 10-fold cross validation.

We have first calculated the precision (recognition rate) for LBP, 3DLBP, and DLBP on range image. A $5 \times 5$ grid is used when calculating local histograms (i.e. 25 histograms per descriptor map). We apply different radiuses ($R \in [1,8]$) for each descriptor for the three collections. We can see in Fig. 3 that 3DLBP and DLBP generally perform better than LBP on range images which shows how important it is to take into account the magnitude of the local changes in range images. Besides, we also note a general increase in the precision when the radius is larger especially for DLBP, that successfully represent larger local patterns variations with more discriminative codes.

In Fig. 4, we report the recognition rates obtained using 2D, 3D and bi-modal face recognition. The recognition rate is calculated when varying the parameters values of both descriptors and the best rate reached by each approach is then reported. Results show two important points:

- In FRGC and Texas collections, the 3D recognition gives better precision then 2D recognition. This is because the illumination variation in these two collections is very high. Therefore, using depth allows a robust discrimination and thus better recognition rates are obtained. In BOSPHORUS collection, there is no illumination variation and the depth maps have a low quality (depth resolution). This is why 2D recognition yields better results then 3D approach in this collection.
- The bi-modal approach gives the best precision, specially in FRGC and TEXAS collections. This shows how combining both modalities enhances the precision of the mono-modal approaches. In BOSPHORUS, the bi-modal approach is as precise as the 2D approach. This can be explained by the fact that the improvement margin is very small. Indeed, only with 2D modality, a recognition rate of $99,35\%$ is reached.

## 5 CONCLUSIONS

We have presented a novel approach for bi-modal face recognition. It is based on two information sources: the LBP descriptor captures the 2D texture information, and the DLBP descriptor that represents the face 3D details. Our goal was to investigate how using both modalities (2D and 3D) can enhance the recognition precision. Experiments on different 3D face collections were conducted in order to evaluate our method. According to the obtained results, we can see that, in general, the bi-modal approach is a potential solution to face recognition in uncontrolled environment, allowing more precision and more robustness through the complementarity of 2D and 3D modalities.

Our future works are directed towards exploring other fusion strategies. Indeed, fusion at decision stage assume that modalities are independent. However, we can see that relative positions of face components (eyes, nose, etc.) are the same in images of both modalities. Moreover, 2D and 3D modalities might be affected similarly by certain acquisition conditions such as occlusion or pose variations. This shows that depth and intensity images are likely to be somehow dependent. This induces us to consider earlier fusion and other combinations in our future works.

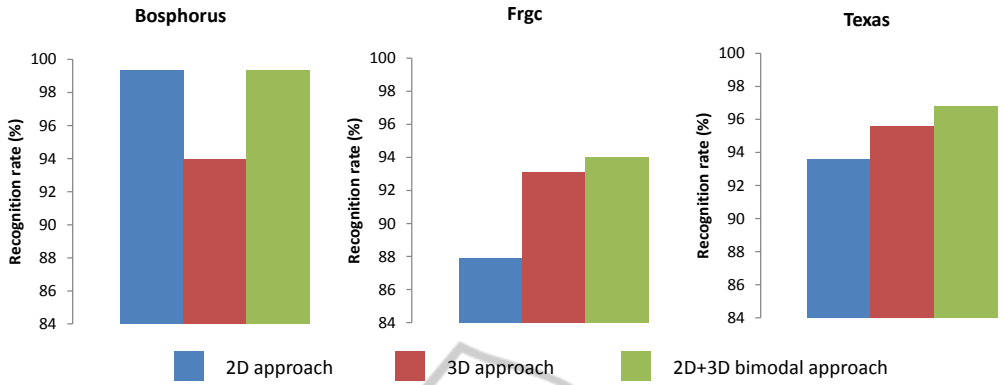


Figure 4: Comparison between 2D face recognition (LBP) with greyscale images, 3D face recognition (DLBP) with range images, and bi-modal recognition with a late-fusion, for 3 collections.

## REFERENCES


Abate, A. F., Nappi, M., Riccio, D., and Sabatino, G. (2007). 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885 – 1906.

Aissaoui, A., Martinet, J., and Djeraba, C. (2014). Dlbp: A novel descriptor for depth image based face recognition. In *Proceedings of the 21th IEEE international conference on Image processing*, pages 298–302.

Bowyer, K. W., Chang, K., and Flynn, P. (2006). A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. *Computer Vision and Image Understanding*, 101(1):1–15.

Byun, H. and Lee, S.-W. (2002). Applications of support vector machines for pattern recognition: A survey. In *Proceedings of the First International Workshop on Pattern Recognition with Support Vector Machines*, SVM '02, pages 213–236, London, UK, UK. Springer-Verlag.

Chang, K., Bowyer, K., and Flynn, P. (2003). Face recognition using 2D and 3D facial data. In *ACM Workshop on Multimodal User Authentication*, pages 25–32. Citeseer.

Fan, K.-C. and Hung, T.-Y. (2014). A novel local pattern descriptor – local vector pattern in high-order derivative space for face recognition. *IEEE Transactions on Image Processing*, 23(7):2877–2891.

Gupta, S., Castleman, K., Markey, M., and Bovik, A. (2010). Texas 3D face recognition database. In *Image Analysis and Interpretation. IEEE Southwest Symposium on*, pages 97–100. IEEE.

Huang, D., Ardabilian, M., Wang, Y., and Chen, L. (2009). Asymmetric 3D/2D face recognition based on lbp facial representation and canonical correlation analysis. In *Proceedings of the 16th IEEE international conference on Image processing*, ICIP'09, pages 3289–3292, Piscataway, NJ, USA. IEEE Press.

Huang, D., Shan, C., Ardabilian, M., Wang, Y., and Chen, L. (2011). Local binary patterns and its application to facial image analysis: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41(6):765–781.

Huang, Y., Wang, Y., and Tan, T. (2006). Combining statistics of geometrical and correlative features for 3D face recognition. In *Proceedings of the British Machine Vision Conference*, pages 879–888.

Hung, T.-Y. and Fan, K.-C. (2014). Local vector pattern in high-order derivative space for face recognition. In *Proceedings of the 21th IEEE international conference on Image processing*, pages 239–3243.

Husken, M., Brauckmann, M., Gehlen, S., and Von der Malsburg, C. (2005). Strategies and benefits of fusion of 2D and 3D face recognition. In *Computer Vision and Pattern Recognition-Workshops. IEEE Computer Society Conference on*, pages 174–174. IEEE.

Jahanbin, S., Choi, H., and Bovik, A. (2011). Passive multimodal 2-d+3-d face recognition using gabor features and landmark distances. *Information Forensics and Security, IEEE Transactions on*, 6(4):1287–1304.

Mantecn, T., del Blanco, C., Jaureguizar, F., and Garca, N. (2014). Dlbp: A novel descriptor for depth image based face recognition. In *Proceedings of the 21th IEEE international conference on Image processing*, pages 293–297.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 24(7):971–987.

Ojala, T., Valkealahti, K., Oja, E., and Pietikinen, M. (2001). Texture discrimination with multidimensional distributions of signed gray-level differences. *Pattern Recognition*, 34(3):727 – 739.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W. (2005). Overview of the face recognition grand challenge. In *Computer vision and pattern recognition. IEEE computer society conference on*, volume 1, pages 947–954. IEEE.

Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The feret evaluation methodology for face-

recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104.

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. (2008). Bosphorus database for 3D face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer.

Xu, C., Li, S., Tan, T., and Quan, L. (2009). Automatic 3D face recognition from depth and intensity gabor features. *Pattern Recognition*, 42(9):1895 – 1905.

Xu, L., Krzyzak, A., and Suen, C. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(3):418–435.

Zhang, B., Gao, Y., Zhao, S., and Liu, J. (2010). Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor. *Trans. Img. Proc.*, 19(2):533–544.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *Acm Computing Surveys*, 35(4):399–458.