Bimodal 2D-3D face recognition using a two-stage fusion strategy

Amel AISSAOUI1 and Jean MARTINET2

1 University of Science and Technologies Houari Boumediene Algiers, Algeria
Email: aaissaoui@usthb.dz
2 CRIStAL Lille 1 University Villeneuve d’Ascq, France
Email: jean.martinet@univ-lille1.fr

Abstract—This paper presents a novel approach for bimodal face recognition. In our approach, faces are represented by both texture and depth images. The well known Local Binary Pattern (LBP) is used to describe the texture images. The depth faces representation is based on the Depth Local Binary Pattern, which is an extension of the the LBP descriptor allowing more discriminative power of smooth depth images. In order to perform the bimodal face recognition, a two-stage fusion scheme is proposed. It allows to take advantage of the complementarity of range and texture modalities at both descriptor (early fusion) and decision (late fusion) levels. We have conducted extensive experiments on several datasets in order to evaluate our approach. The obtained results show that our combination of texture and depth descriptors yields higher results than when taken separately or using an early/late fusion scheme.

Keywords—Face recognition, LBP, DLBP, Bimodal fusion.

I. INTRODUCTION

As one of the few biometric methods that combine the merits of both high accuracy and low intrusiveness, face recognition technology has a variety of potential applications in information security, law enforcement and surveillance, smart cards, access control, among others. For this reason, this domain has received an increasing attention from both the academic and industrial communities during the past 20 years. However, all human faces are similar in their configurations and hence allow low distinctiveness. Moreover, intra-class variations, due to factors such as pose, expression, or illumination changes are often greater than inter-class ones. The past decades have witnessed tremendous efforts first focused towards 2D face images [1]. Despite the great progress achieved so far within the field, 2D image face recognition is still not reliable enough [2], especially in the presence of pose and illumination changes [3]. To deal with such unsolved issues in 2D face recognition, efforts are then focused on 3D face models or scans [4].

However, 3D data need expensive materials and human cooperation which limits their application field. Although 3D face scans capture facial surface structures, and thereby theoretically reputed to be robust to illumination variations, they are likely to be more sensitive to expression changes. 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 optimisation of complex geometric equations. In order to deal with time processing of the 3D matching methods, researchers tend to focus on using depth images of the face instead of 3D point clouds. Such kind of methods has more emerged with the rapid development in 3D imaging systems, such as the Microsoft infrared camera Kinect.

A step forward consists in combining 2D and 3D information in order to benefit from the strength and complementarity of both modalities. The 2D image provides informations about face textured regions with little geometric structure (e.g. hairy parts, eyes, eyebrows), and the 3D data provides informations regarding less textured regions (e.g. nose, chin, cheeks). Hence, merging both sources of information is likely to enhance the precision and robustness of face recognition.

Several merging strategies exist, depending on whether the merging is applied before or after classification [5].

- The early fusion (fusion of descriptors, before classification) consists in merging descriptors built from each modality separately, and the new descriptor is used for training and matching. Another less popular way consists in merging raw data from the sensors to make new data, before extracting the descriptors.
- The late fusion (fusion of decisions, after classification) comes after that separate classifiers are built for each modality. The classifier outputs are merged.

In this paper, we propose a bimodal face recognition approach, based on both the LBP (Local Binary Pattern) [6], [7], known as one of the simplest and most efficient descriptor in 2D face recognition, and the DLBP (Depth Local Binary Pattern) [8], that is an extension of LBP for the depth images. A two-stage fusion scheme is proposed to perform the bimodal face recognition. The proposed scheme aims at exploiting the complementarity of both modalities at the descriptor and decision stages.
The paper is organized as follows: Section II describes relevant 2D and 3D face recognition methods, and discusses the fusion strategies when combining both. Section III introduces the proposed bimodal face recognition approach based on a two-stage fusion. Section IV presents experimental results on six collections that demonstrate the performances of the proposed method. Section V gives a conclusion to our work.

II. RELATED WORK

Many approaches have been proposed for face recognition [2], [4]. 2D face recognition mainly deals with images representing faces’ visual appearance. Such images – color or grey-level – are generally acquired via a usual camera. Among popular descriptors for 2D face recognition, LBP, first proposed by Ojala et al. [6], [7], is considered to be one of the simplest and most efficient local 2D face descriptors. Due to its high computational efficiency and good 2D face discriminative property, the LBP descriptor has gained considerable attention and has already been used in many face image analysis fields [9]. The LBP method encodes pixel-wise information in a given image. The operator describes each pixel with the relative grey-level values of its neighboring pixels.

Due to its success in 2D face description [10], some researchers in 3D face recognition have investigated the LBP descriptor for depth image description. Authors in [11] used LBP directly for depth images to achieve 3D face recognition. In [12], the LBP descriptor was applied to amplify the details of 3D depth images in an asymmetric face recognition system.

Since the intrinsic nature of grey-level and depth face images are fundamentally different, other LBP-based descriptors dedicated to depth face image were recently introduced. Huang et al. [13] proposed an extended LBP version, named 3DLBP. Beside the information provided by LBP, 3DLBP also considers the magnitude of the difference between the central pixel and its neighborhood. In [14], authors proposed an LBP-based descriptor called Depth Local Quantized Pattern (DLQP) where a quantification step is introduced in order to increase its capacity to distinguish different depth patterns. We recently proposed a new versatile descriptor called DLBP [8]. Unlike 3DLBP and DLQP, DLBP is designed to consider large radiiuses (i.e. large neighborhoods around the central pixel) in order to extract more discriminative features from smooth and low-contrast data, since it works on a multi-scale level.

Bimodal 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 in a complementary manner. More specifically, combining 2D and 3D methods has enabled to reach higher results than usual 2D methods or 3D methods using depth images taken separately. Regarding the 2D-3D face recognition fusion, late fusion is the most used strategy in bimodal face recognition [15], [16], [17], [18], [19], [2], [20] – only a few methods based on early fusion are proposed [11], [21].

The late fusion is applied after the descriptor matching step. It can be done at the score-level (i.e. class-wise similarity measures) by using several rules such as a sum, a product, or a weighted sum, or at the decision-level (i.e. identity) by using a majority vote method, or a weighted majority vote [20], where confidence values are assigned to classifiers according their precision for a given decision. Late fusion is simpler since both modalities are considered independently, and only their outputs are taken into account for fusion. However, systems based on such approach are likely to give results similar to those obtained with a single modality. Indeed, a given classifier can dominate the others, and therefore be the unique responsible for the system precision.

Most methods based on this type of fusion consider that both modalities are independent. However, the independence hypothesis for 2D and 3D representations is arguable since the data is extracted from the same face. As stated by Husken et al. [16], the position of fiducial points (eyes, nose tip, etc.) are identical. Late fusion does not allow a synergetic processing since each modality is taken separately. On the contrary, early fusion is likely to perform better by exploiting the complementarity of 2D-3D face data.

The choice of the best fusion strategy is crucial in order to benefit from the complementarity of modalities, and therefore to enhance the results [22]. However, for most 2D-3D bimodal approaches, this choice is not obvious. Experimental studies have been carried out in order to compare several strategies [23], [11], but the results are not conclusive. Indeed, decision fusion (late) yields better results than descriptor fusion (early) in the works of Benabdellkader et al. [23], while the opposite result was reported in the works of Li et al. [11].

In the proposed approach, we consider both early and late fusion strategies by introducing a two-stage fusion scheme that benefits from both strategies.

III. PROPOSED BIMODAL FACE RECOGNITION SYSTEM

Our contribution is two-fold. First, we propose a bimodal representation of face data based on LBP extracted from grey-level image for the 2D part, and DLBP extracted from depth images for the 3D part. Second, we introduce a two-stage fusion strategy for recognition, by matching a sample test face to known faces in a collection.

A. Bimodal face representation

We use the standard LBP (see Eq. 1) for representing 2D information of the face texture.

$$\text{LBP}_{(R,V)}(x,y) = \sum_{i=0}^{V-1} s(n_i - n_c)2^i, \quad (1)$$

where:

- $n_c$ is the grey-level value of the central pixel from the local neighborhood;
\begin{itemize}
\item \(n_i\) are the grey-level values of the neighbor pixels around the central pixel with a radius \(R\). The position of the neighbor pixel is given by Eqs. 2 and 3 (with a bilinear interpolation in case the estimated position does not match a pixel), allowing several radius values and neighborhood sizes.
\end{itemize}

\begin{align}
    x_{pi} &= x_p + R \cos \left( \frac{i}{\sqrt{V}} \pi \right) \quad (2) \\
    y_{pi} &= y_p + R \sin \left( \frac{i}{\sqrt{V}} \pi \right) \quad (3)
\end{align}

Regarding the 3D information, we use DLBP [8], as described in Eq. 4.

\[
    \text{DLBP}_{(R,V)}(p) = \left( \frac{c_{(R,V)}(p)}{c_{(R,V)}^m(p)} \right) = \\
    \left( \frac{\sum_{i=0}^{V-1} s(p, p)^{2i}, s(k) = \begin{cases} 1 & \text{if } k \geq 0 \\ 0 & \text{otherwise} \end{cases}}{\sum_{i=0}^{V-1} m(p, p)^{2i}, m(k) = \begin{cases} 1 & \text{if } k \geq S_{(R,V)}^m \\ 0 & \text{otherwise} \end{cases}} \right) \quad (4)
\]

where:
\begin{itemize}
\item \(p_i\) is the \(i^{th}\) neighbor pixel whose position relatively to \(p\) is defined according to \(R\) and \(V\) (see Eqs. 2 and 3);
\item \(S_{(R,V)}^m\) is a magnitude threshold calculated automatically using the depth gradient. For more details about the DLBP descriptor, authors can refer to our previous works [8].
\end{itemize}

In Eq. 4, \(s(k)\) and \(m(k)\) correspond to the sign and magnitude of the value \(k\). In our approach, faces are represented with both LBP and DLBP descriptors.

\section{B. Two-stage fusion}

We describe here the proposed two-stage fusion strategy, as shown Fig. 1.

The first step consists in extracting the bimodal face representation (LBP and DLBP feature vectors) as described previously. Three classifiers are designed. The first classifier is obtained from the 2D modality (LBP vector extracted from grey-level images). The second classifier is obtained from the 3D modality (DLBP vector extracted from depth images). The last classifier is obtained from both modalities fused with an early fusion scheme (a new descriptor obtained by concatenating both LBP and DLBP feature vectors). Classifiers are trained separately, and they are used later for test faces identification.

In order to identify an unknown face, decisions returned from the three classifiers are merged (late fusion) to output the identity, as described below.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Fig_1}
\caption{Overview of the two-stage fusion scheme for bimodal 2D-3D face recognition.}
\end{figure}

Given a face represented by a grey-scale image and a depth image, three descriptors \(D_1\), \(D_2\) and \(D_3\) are extracted using LBP, DLBP and their fusion respectively. Three classifiers \(M_1\), \(M_2\) and \(M_3\) trained separately provide three decisions \(M_1(D_1), M_2(D_2)\) and \(M_3(D_3)\) respectively.

In order to get the final output, we propose to use a decision fusion method based on the Weighted Majority Algorithm (WMA) [24] principle, which is among the most powerful, the simplest, and the easiest to implement. Let \(M_j(D_j) = i\) be the output that classifier \(M_j\) assigns to class \(i, (i \in \{1, 2, ..., n\})\) to the descriptor \(D_j\), where \(j \in \{1, 2, ..., m\}\) and \(m\) and \(m\) is the modality number – i.e. three in our case with 2D, 3D and 2D-3D early fusion. An indicator function \(F\) (defined by the following Eq. 5) is assigned to each classifier.

\[
    F_j^i(D_j) = \begin{cases} 1 & \text{si } M_j(D_j) = i, \\ 0 & \text{otherwise}. \end{cases} \quad (5)
\]

The combination of both classifiers is hence written as:

\[
    F_i = \sum_{j=1}^{m} \alpha_{ij} F_j^i(D_j) \quad (6)
\]

for each \(i \in \{1, 2, ..., n\}\). The weight \(\alpha_{ij}\) represents the reliability of classifier \(j\) for a given decision (class \(i\)). These weights are given by the classifiers’ recognition rates obtained in the training step for each class. The final decision (identity) is therefore given by \(\text{argmax}_i(F_i)\).

\section{IV. Experiments}

We report here experimental results that demonstrate the benefits from combining 2D and 3D data with the two-stage fusion. We have used five datasets:
• **FRGC** [22]: the most used 3D faces dataset, composed of 4007 face color images from 446 persons, with corresponding depth images. Depth images presenting lot of missing data (big holes) are not considered in the experiments.

• **Texas** [25]: this dataset contains 1149 color images from 118 persons, and corresponding depth images.

• **Bosphorus** [26]: a rich dataset in terms of changes in face expressions and pose. It contains 4652 color images from 105 persons, with corresponding depth images. Images with large pose variations (i.e. rotations larger than about 30 deg ) are not considered in the experiments.

• **FoxFaces** [27]: a multi-purpose face dataset including color and depth images from 64 persons. This dataset contains two sub-datasets: FoxStereo and FoxKinect. The depth maps in FoxKinect are generated from the Kinect sensor data, and depth maps in FoxStereo are generated with stereo reconstruction method [28]. Images with large pose variation are not considered in these experiments.

Faces are extracted from grey-level and depth images using the annotations found in the dataset. Then they are normalized to 100 × 100 pixels. We have implemented the feature extraction process using different parameters for LBP and DLBP, and we used those yielding the best results. 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.

The results of a comparison between 2D recognition, 3D recognition, early/late/two-stage fusion strategies for the five collections are presented Fig. 2.

The obtained results show the following points.

• In FRGC and Texas datasets, 3D recognition gives a higher precision compared to 2D recognition, contrary to the other datasets. This is explained by the very high lighting variations in these two collections (see Fig. 3). Consequently, the use of depth information allows more robust discrimination, yielding higher recognition rates.

• Descriptor fusion (early) slightly enhances the system performance with individual modalities for FRGC and Texas datasets. This does not hold for the other datasets. In Bosphorus, for instance, descriptor fusion results in a kind of average value of both modalities taken separately. We notice that for three collections (Bosphorus, FoxKinect, FoxStereo), the 3D modality gives lower results than the 2D modality. Each time, it results in pulling down the early fusion results. This is explained by the lower quality of 3D data, compared to the two other collections FRGC and Texas. Therefore, merging the descriptors is likely to accumulate possible noise from both modalities vectors. As a consequence, if the data in one vector is very noisy, it impacts the whole final descriptor, decreasing the global system performance. Another explanation for the low results of early fusion for the same three datasets (Bosphorus, FoxKinect, and FoxStereo) lies the complementarity relation. Indeed, descriptor fusion is more suitable in case of complementary data, like in FRGC and Texas where the lighting variations are important as shown Fig. 3. In such case, the 3D modality, which is robust to light change, complements the 2D modality.

• Decision fusion does either enhance or preserve monomodal system performances for all collections. This demonstrates that in this case of bimodal face recognition, decision fusion is a better option than descriptor fusion. Each descriptor is processed separately, and if a descriptor is better than the other for a given class, the corresponding decision will not be influenced by the other descriptor.

• Two-stage fusion turns out to be better than the two other strategies taken separately. It yields a higher precision than those obtained with descriptor or decision fusion for most datasets, and it guarantees a precision identical to decision fusion when no improvement is brought (Bosphorus).

As a general comment, based on the experiments carried out with several datasets, we can observe that our proposed bimodal approach based on the two-stage fusion scheme allows to enhance the face recognition precision. The method guarantees the highest precision obtained by decision fusion when no improvement is possible. Therefore, even when the descriptor fusion does not perform well, no decrease in precision is observed. This is due to the important aspect that the decisions used in two-stage fusion (resulting from 2D modality, 3D modality, and descriptor fusion) are considered independently. If one of the descriptors is better than the other for a given class, the corresponding classifier decision will not be influenced by the other.

V. Conclusion and Future Works

We have introduced a bimodal 2D-3D face recognition method based on LBD and DLBP, that combines both modalities and that implements an original two-stage fusion strategy.
The objective is to benefit from the strengths and complementarity of texture and depth data at both descriptor and decision levels.

The comparative study carried out between mono-modal and bi-modal approaches, and also between several fusion schemes, shows that combining both modalities allows a noticeable improvement over mono-modal approaches. Moreover, the two-stage strategy helps benefiting from both fusion strategies (early and late), and yields the highest recognition rate.

Our future works are directed towards taking into account the face orientation in the training and matching processes, since it is a parameter that has a great impact on both greyscale and depth images.

VI. ACKNOWLEDGMENT

This work was supported by "Empathic Products", ITEA2 1105.

REFERENCES


