Visual Face Recognition using Bag of Dense Derivative Depth Patterns

Tomás Mantecón, Carlos R. del-Blanco, Fernando Jaurequizar and Narciso García

***Abstract*—A novel biometric face recognition algorithm using depth cameras is proposed. The key contribution is the design of a novel and highly discriminative face image descriptor called Bag of Dense Derivative Depth Patterns. This descriptor is composed of four different stages that fully exploit the characteristics of depth information: 1) dense spatial derivatives to encode the 3D local structure, 2) face-adaptive quantization of the previous derivatives, 3) multi-bag of words that creates a compact vector description from the quantized derivatives, and 4) spatial block division to add global spatial information. The proposed system can recognize people faces from a wide range of poses, not only frontal ones, increasing its applicability to real situations. Lastly, a new face database of high resolution depth images has been created and made it public for evaluation purposes.**

***Index Terms*—Depth camera, dense depth derivatives, quantized patterns, face recognition, depth face database, SVM, bag-of-words.**

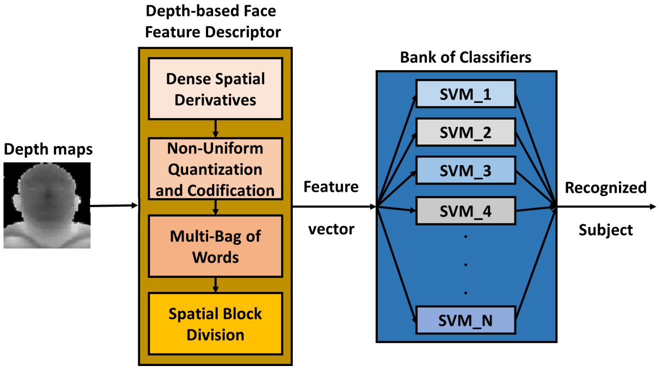

Figure 1. Stages of the proposed depth-based face recognition system.

I. Introduction

Most of the existing face recognition algorithms use color imagery as input data, which is usually processed by a machine learning framework. Since color information is strongly dependent on the illumination conditions, a feature extraction stage is needed to obtain a feature vector that robustly represents the face information. For this purpose, some of the most popular feature extraction techniques used in face recognition are: the Scale Invariant Feature Transform (SIFT) [1], the Histogram of Oriented Gradients (HOG) [2], the Local Binary Pattern (LBP) [3], the Principal Component Analisys (PCA), [4], the Local Gabor Binary Patterns (LGBP) [5], the Local Phase Quantization (LPQ) [6], and the Binarized Statistical Image Features (BSIF) [7]. Of special consideration is the LBP descriptor and its numerous variants [8], which have proven to be highly robust to even non-linear illumination variations. All these works have common strategies that try to be robust to changing illumination conditions. However, the performance of a color-based recognition algorithm is still greatly affected by varying or low level illumination conditions. This fact compromises the use of such systems for highly demanding applications, such as security and defense.

Depth-imagery based solutions have emerged in the last years to solve some of the problems posed by color imagery. The first advantage of depth imagery is to be almost immune to variations of the illumination conditions in indoor environments (they can work even in the absence of illumination). For example, [2] proposes to jointly use LBP and HOG descriptors with depth imagery using a Support Vector Machine (SVM) classifier. In [9], an LBP variation for 3D face recognition is proposed, and in [10] an LBP based-solution for the recognition of facial expressions using 3D curvature information is used. Nonetheless, these works have two main problems. The first one is that they propose slight variations of existing feature descriptors designed for color imagery, and therefore they do not fully exploit the specific characteristics of depth information. The second one is the restricted spatial and depth resolution of most of the current depth cameras. To alleviate these problems, some works have made use of pre-computed 3D-face models to be more robust and accurate in the recognition task [9] [11]. But, they have the disadvantage that the process of obtaining 3D face models is costly and complex.

To compensate the relatively low resolution of depth imagery, some works combine color and depth information. However, the use of depth information is merely secondary, just as a complement of the color-based recognition [12] [13], precisely due to the imbalanced situation between the resolution of both types of imagery.

Another lack of the current visual face recognition works is that the feature extraction techniques are not tailored to face structures. Although some of them have been specifically tested for face detection, such as the Kernel Based Local Binary Pattern [14] and the Semi-Local Structure Patterns [15], they do not explicitly exploit the 3D structure information of a human face.

In this paper, a new face recognition system that solves all those problems is proposed using only depth imagery. This

This work has been partially supported by the Ministerio de Economía y Competitividad of the Spanish Government under project TEC2013-48453 (MR-UHDTV), and by AIRBUS Defense and Space under project SAVIER.

Tomás Mantecón, Carlos R. del-Blanco, Fernando Jaurequizar and Narciso García are with the Grupo de Tratamiento de Imágenes, UPM, Madrid 28040, Spain (e-mail: {tmv,cda,fjn,narciso}@gti.ssr.upm.es).

system makes use of the latest generation of 3D cameras, which are capable to acquire images with higher spatial and depth resolution. These improved characteristics are fully exploited by a novel depth face descriptor, called Bag of Dense Derivative Depth Patterns (Bag-D3P). More precisely, the design of this descriptor considers: 1) the specific characteristics of depth images, 2) how to fully exploit the extended spatial and depth resolutions of the latest 3D cameras, and 3) the prior knowledge about the 3D structure of a human face to improve the recognition accuracy. Additionally, the proposed face recognition system has proven to be able to operate in the wild, this means that it can recognize human faces from a broad range of angles of view, unlike current existing algorithms that are almost limited to frontal faces. This work is a mayor evolution of [16] since the face descriptor has been re-designed and improved. Only the last stage, spatial block division, is common for both descriptors (that is also a common strategy for others descriptors such as LBP). In addition to offer a more formal and complete state-of-the-art evaluation.

The organization of the paper is as follow. Section II briefly describes the overall face recognition system. Section III focuses on the new designed depth face descriptor. Then, Section IV describes the depth-based face database, created to evaluate the proposed algorithm with the imagery provided by the latest generation of depth cameras. The obtained results are presented in Section V, and finally, conclusions are drawn in Section VI.

II. System overview

The proposed depth-based face recognition system can be divided into two main stages (see Fig.1). The first stage is the Depth-based Face Feature Descriptor, which processes the incoming flow of depth images to produce highly discriminative and robust face descriptors, called Bag-D3P descriptors. This stage, which represents the main contributions of this paper, is described in detail in Section III. The second stage, called SVM Bank of Classifiers, takes the flow of computed face descriptors to recognize the person. Each classifier is based on the SVM Pegasos algorithm [17], which is able to perform an online training, unlike the standard SVM formalization. This feature is essential for the system, since the computed Depth-based Face Feature Descriptor is quite demanding in memory (it produces large feature vectors), which can cause problems with a batch training framework. Additionally, a Hellinger kernel, more commonly known as Bhattacharyya distance [18], has been applied to compute non-linear decision boundaries that can improve the recognition capabilities. On the other hand, the multiple face recognition capability is achieved by a one-vs.-all configuration, which consists in that an independent SVM is trained for each human face. Positive samples are those Bag-D3P descriptors corresponding to the considered face, and the negative ones are obtained from the other faces in the database.

Once the bank of SVMs is trained, a face recognition result is obtained for a candidate depth image by assigning the face label of the SVM that has achieved the highest classification score.

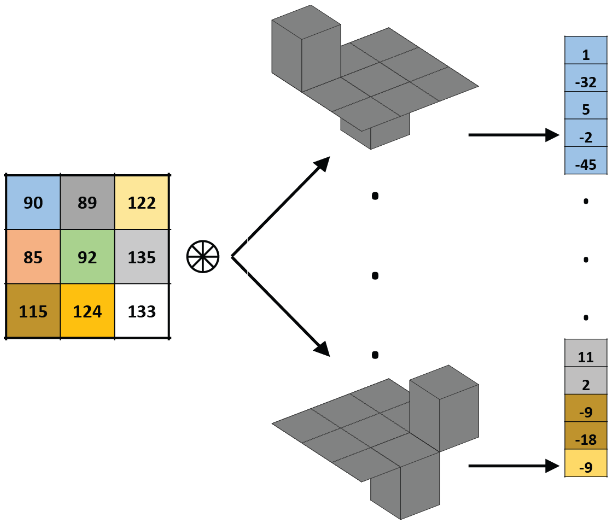

Figure 2. First stage of Bag-D3P computation: dense spatial derivatives.

III. Bag of Dense Derivative Depth Patterns

The novel Bag of Dense Derivative Depth Patterns (Bag-D3P) is a feature descriptor that 1) fully exploits the characteristics of depth imagery, 2) takes advantage of the extended spatial and depth resolution of depth sensors, and 3) is adapted to maximize the discrimination of human faces. The computation of the Bag-D3P can be divided into four stages: dense spatial derivatives, non-uniform quantization and codification, multi-bag of words, and spatial block division.

The first stage, dense spatial derivatives, characterizes each pixel in an depth image region by densely computing all the pairwise pixel differences in the pixel neighborhood. From a signal processing point of view, these differences represent a bank of 2D derivative filters at multiple scales and orientations. Fig. 2 illustrates this process for one pixel and a neighborhood of 8 pixels. As a result, a vector of dense spatial derivatives values is obtained, whose length is $N_{dsd} = \frac{N_{nh}^2+N_{nh}}{2}$, where N_{nh} is the number of pixel neighbors. For example, for a neighborhood of $N_{nh} = 8$, a value of $N_{dsd} = 36$ is obtained.

The second stage, non-uniform quantization and codification, takes into account prior information about human faces to translate the previous vector of dense spatial derivatives into a compact binary code. Fig. 3 illustrates this second stage. The first step of this stage quantizes each component vector by means of a non-uniform scheme that is adapted to the typical range values of a human face. For this purpose, eight

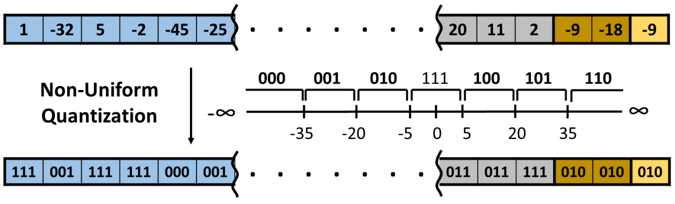

Figure 3. Second stage of Bag-D3P computation: non-uniform quantification and codification.

symmetric intervals of different widths are used: two intervals with low positive and negative derivative values, $(-5, 0)$ and $(0, 5)$ mm, representing small random depth variations (usually due to noise); four intervals with a range of positive and negative values, $(-35, -20)$, $(-20, -5)$, $(5, 20)$, and $(20, 35)$ mm that encode the most representative depth reliefs of a human face; and two open end intervals, $(-\infty, -35)$ and $(+35, +\infty)$ mm, which represent the large depth variations usually corresponding to object borders (such as face to background). The decision values of the previous intervals are adapted to both capture the most discriminative depth information of a human face, and exploit the higher depth resolution of the new depth sensors (i.e. for other depth sensors, the noise variations prevent the use of such fine-grain intervals). This fact contributes to dramatically increase the capability of distinguishing different depth-based patterns, and ultimately the target human faces.

The last step of the second stage is the coding of the quantized values. Since there is $N_{int} = 8$ intervals, $N_b = \log(N_{int}) = 3$ bits are required to represent every quantized value following a fixed-length encoding strategy. Finally, all binary numbers resulting from the quantization process of a pixel neighborhood are concatenated to form a binary word of $N_{bin} = 108$ bits ($N_{dsd} \times N_b = 36 \times 3 = 108$).

The third stage, multi-bag of words, uses an extension of the bag of words strategy to compactly represent the set of $N_{bin} = 108$-bits words corresponding to a pixel region. Without this stage, the generated high volume of data will be unmanageable. More specifically (see Fig. 4), this stage starts by dividing every binary word into smaller chunks of length $N_{div} = 9$ bits, which are then converted into decimal values. Next, a histogram per chunk is computed using all decimal values coming from the considered chunk in a set of binary words. As a result, multiple histograms are obtained, one per word chunk, which are finally concatenated to form the Bag-D3P descriptor of an image region. The final descriptor has a length of:

$$N_{dct} = N_h \times 2^{N_{div}} = 12 \times 2^9 = 6144 \tag{1}$$

The last stage, spatial block division, is a strategy to incorporate global spatial information to the descriptor. Although it is true that the multi-bag of words strategy is essential to manage the high volume of data (the set of binary words) from an image region, the global spatial structure information is lost. However, the discrimination capability can be improved by adding some global spatial information. For this purpose, a face region is divided into $N_s \times N_s$ non-overlapping blocks, and then a Bag-D3P descriptor is computed per block. Finally, all Bag-D3P descriptors are concatenated into one vector, which is delivered to the bank of SVMs for the recognition task.

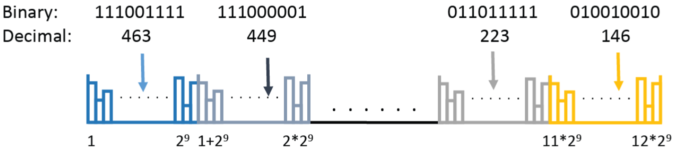

Figure 4. Third stage of Bag-D3P computation: multi-bag of words.

The resulting descriptor length is $N_T = N_s \times N_s \times N_{dct}$, which for a face block division of 4×4, and the previous parameter values, results in a vector of length:

$$N_T = 4 \times 4 \times 6144 = 98304 \tag{2}$$

IV. Depth-Based Face Database

As far as the authors' knowledge, there are no publicly available depth-based face databases acquired by the last generation of depth sensors, which provide extended spatial and depth resolutions. This has motivated the creation of such database using the new Kinect version 2, which has a significant higher depth resolution than the first Kinect version, and others depth sensors [19]. The database is composed of near 21.000 face images of 18 different male and female subjects from different perspectives. The database, called High Resolution Range based Face Database (HRRFaceD), is publicly available in the following address: http://www.gti.ssr.upm.es/data/HRRFaceD_database.html.

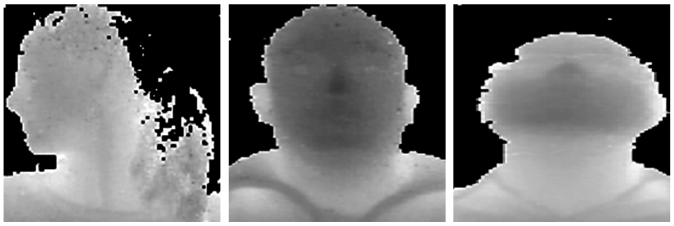

Figure 5. Example of depth face images obtained by the Kinect version 2.

Fig. 5 shows three different depth face images included in the HRRFaceD database, obtained by the Kinect version 2. Observe that the main characteristics of the face can be distinguished: eyes, mouth, ears and nose.

V. Results

The proposed face recognition system based on depth imagery, from now on called Bag-D3P (as the designed descriptor), has been compared with different state-of-the-art solutions, using the created HRRFaceD database, and also other public depth-based face databases (whose depth images have a noticeable lower depth resolution). These databases are briefly described as follows. The IIT-D RGB-D database [20] is composed of 106 male and female subjects with between 11 and 254 images per subject. The images are captured in color and depth, and contain different expressions per subject. The EURECOM Kinect database [21] is composed of 52 male and female subjects. Two acquisition sessions at different times are performed per subject, and 9 different states are available per session: neutral, smile, open mouth, left profile, right profile, occlusion eyes, occlusion mouth, occlusion paper, and light on. Also, color and depth imagery are acquired. The Biwi Kinect Head Pose database [19], which is composed of more than 15000 images (in color and depth) of 20 male and female subjects, comprises different head poses per subject.

The state-of-the-art approaches used for comparison purposes are: an LBP-based method [12] that uses the standard LBP technique; a SIFT-based method [8] that utilizes a dense SIFT feature extraction; a PCA-based method [4] that builds a low-dimensional face subspace; an LGBP-based method [5] that combines Gabor and LBP features; an LPQ-based method [6] that claims to be more robust than LBP to blurriness; an HOG-based method [22] that exploits the gradient information; a BSIF-based method [7] that is a combination of LBP and LPQ methods; and a DLQP-based method [23] that is an evolution of the LBP descriptor for depth imagery.

Notice that some of the previous methods use both color and depth imagery for the recognition task, which is an advantage for them from the recognition performance point of view.

The metric used to measure the recognition performance is the accuracy:

$$\text{Accuracy} = 100 \times \frac{\text{Total number of correct faces}}{\text{Total number of faces}} \quad (3)$$

which represents the percentage of faces per subject that are correctly assigned to that subject.

TABLE I presents the mean accuracy for the created HRRFaced database. The proposed Bag-D3P algorithm achieves the best results, increasing in at least a 20% the accuracy obtained by the others. This fact proves the high performance of the designed depth face descriptor using only depth imagery.

Table I
ACCURACY RESULTS USING THE HRRFACED DATASET

	Bag-D3P	DLQP	LBP	SIFT
HRRFaced	**94.30**	73.47	59.17	71.94

Table II
ACCURACY RESULTS USING THE EURECOM DATASET

		Bag-D3P	PCA	LBP	LGBP
EURECOM	Session 1	**96.16**	63.94	89.90	81.73
	Session 2	**86.54**	33.85	82.69	66.92

Table III
ACCURACY RESULTS USING THE BIWI DATASET

	Bag-D3P	LBP
Biwi	**94.13**	85.20

Table IV
ACCURACY RESULTS USING THE IIT-D RGB-D DATASET

	Bag-D3P	LBP	LPQ	HOG	BSIF
IITD	**84.72**	84.40	84.70	82.80	77.80

The other three tables show the results for the other public face databases containing color and depth imagery. TABLE II, TABLE III, and TABLE IV present the mean accuracy for the EURECOM, the Biwi, and IIT-D RGB-D databases, respectively. Again, the proposed Bag-D3P algorithm achieves the best recognition performance, despite that only depth information has been used. It is also worth to note that the difference in performance between the proposed face recognition algorithm and the others is not so high as in the case of the HRRFaced database. This is due to the fact that the depth data of those databases have a lower depth resolution because of the used camera sensor. This is specially significant for the IIT-D RGB-D database, in this case depth resolution is not only reduced by the use of a lower depth resolution sensor, but also due to a quantization stage is applied over its values to reduce the number of bits to codify its depth values. For this reason, the proposed solution does not considerably improve the results obtained by other solutions. Anyway, the recognition performance of the proposed solution is still the best, despite that the Bag-D3P algorithm has been specially designed for the latest generation of depth sensors. Noticeable are also the results obtained in the EURECOM database, where the results obtained in the session 2 are significantly worse than for the session 1. The reason is that the database setting requires to use the session 1 as training and the session 2 as testing. In any case, the results obtained by the proposed algorithm are quite satisfactory, improving by far the other algorithms.

VI. Conclusions

A novel depth-based recognition system of human faces has been proposed in this paper. The algorithm is based on a bank of SVM classifiers that can easily discriminate the human faces thanks to a novel depth face descriptor, called Bag-D3P. This descriptor exploits both, the extended range resolution of the latest depth cameras, and the specific 3D structure of a human face. In addition, the system is able to recognize face poses with different orientations, not only frontal ones. Excellent recognition results have been obtained over different depth-based face databases, included the HRRFaceD database. This one has been specifically created to validate the algorithm with images of extended depth resolution, acquired by the latest generation of camera sensors. The presented system has been also compared with other state-of-the art-approaches, achieving the best performance, specially with high resolution depth images.

References

[1] S. Anith, D. Vaithiyanathan, and R. Seshasayanan, “Face recognition system based on feature extration,” in *International Conference on Information Communication and Embedded Systems (ICICES)*, Feb. 2013, pp. 660–664.

[2] B. Jun, I. Choi, and D. Kim, “Local Transform Features and Hybridization for Accurate Face and Human Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1423–1436, Jun. 2013.

[3] G. Kayim, C. Sari, and C. Akgul, “Facial feature selection for gender recognition based on random decision forests,” in *Signal Processing and Communications Applications Conference (SIU)*, Apr. 2013, pp. 1–4.

[4] M. Turk and A. Pentland, “Eigenfaces for Recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[5] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, “Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition,” in *IEEE International Conference on Computer Vision (ICCV)*, vol. 1, Oct. 2005, pp. 786–791.

[6] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, “Recognition of blurred faces using Local Phase Quantization,” in *International Conference on Pattern Recognition (ICPR)*, Dec. 2008, pp. 1–4.

[7] J. Kannala and E. Rahtu, “BSIF: Binarized statistical image features,” in *International Conference on Pattern Recognition (ICPR)*, Nov. 2012, pp. 1363–1366.

[8] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, “Local Binary Patterns and Its Application to Facial Image Analysis: A Survey,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 765–781, Nov. 2011.

[9] H. Tang, B. Yin, Y. Sun, and Y. Hu, “3D face recognition using local binary patterns,” *Signal Processing*, vol. 93, no. 8, pp. 2190 – 2198, Aug. 2013.

[10] Y. Wang, M. Meng, and Q. Zhen, “Learning Encoded Facial Curvature Information for 3D Facial Emotion Recognition,” in *International Conference on Image and Graphics (ICIG)*, Jul. 2013, pp. 529–532.

[11] S. Elaiwat, M. Bennamoun, F. Boussaid, and A. El-Sallam, “3-D Face Recognition Using Curvelet Local Features,” *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 172–175, Feb. 2014.

[12] T. Gao, X. Feng, H. Lu, and J. Zhai, “A novel face feature descriptor using adaptively weighted extended LBP pyramid,” *Optik - International Journal for Light and Electron Optics*, vol. 124, no. 23, pp. 6286 – 6291, Dec. 2013.

[13] L. Ding, X. Ding, and C. Fang, “Continuous Pose Normalization for Pose-Robust Face Recognition,” *IEEE Signal Processing Letters*, vol. 19, no. 11, pp. 721–724, Nov. 2012.

[14] X. Li, W. Hu, Z. Zhang, and H. Wang, “Heat Kernel Based Local Binary Pattern for Face Representation,” *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 308–311, Mar. 2010.

[15] K. Jeong, J. Choi, and G.-J. Jang, “Semi-Local Structure Patterns for Robust Face Detection,” *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1400–1403, Sep. 2015.

[16] T. Mantecon, C. del Blanco, F. Jaureguizar, and N. Garcia, “Access control based on visual face recognition using Depth Spatiograms of Local Quantized Patterns,” in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2015, pp. 530–531.

[17] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, “Pegasos: primal estimated sub-gradient solver for SVM,” *Mathematical Programming*, vol. 127, no. 1, pp. 3–30, Oct. 2011.

[18] E. Choi and C. Lee, “Feature extraction based on the Bhattacharyya distance,” *Pattern Recognition*, vol. 36, no. 8, pp. 1703 – 1709, Aug. 2003.

[19] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, “Random Forests for Real Time 3D Face Analysis,” *International Journal on Computer Vision*, vol. 101, no. 3, pp. 437–458, Feb. 2013.

[20] G. Goswami, M. Vatsa, and R. Singh, “RGB-D Face Recognition With Texture and Attribute Features,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 10, pp. 1629–1640, Oct. 2014.

[21] R. Min, N. Kose, and J.-L. Dugelay, “KinectFaceDB: A Kinect Database for Face Recognition,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.

[22] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[23] T. Mantecon, C. del Blanco, F. Jaureguizar, and N. Garcia, “Depth-based face recognition using local quantized patterns adapted for range data,” in *IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 293–297.