Temporal Pyramid Matching of Local Binary Subpatterns for Hand-Gesture Recognition

Ana I. Maqueda, Carlos R. del-Blanco, Fernando Jaureguizar, and Narciso García

***Abstract*—Human–computer Interaction systems based on hand-gesture recognition are nowadays of great interest to establish a natural communication between humans and machines. However, the visual recognition of gestures and other human poses remains a challenging problem. In this paper, the original volumetric spatiograms of local binary patterns descriptor has been extended to efficiently and robustly encode the spatial and temporal information of hand gestures. This enhancement mitigates the dimensionality problems of the previous approach, and considers more temporal information to achieve a higher recognition rate. Excellent results have been obtained, outperforming other existing approaches of the state of the art.**

***Index Terms*—Color cameras, hand-gesture recognition, spatiotemporal descriptors, support vector machine (SVM) classifier, vision based.**

I. Introduction

In the last decades, there has been a great interest in human–computer interaction (HCI) systems to achieve enhanced interfaces that make the communication with computers as natural as the interaction among humans. There is a tendency toward interfaces that are more natural, intuitive, and user friendly, minimizing the learning process required by the user. In this sense, hand gestures play an important role since they can be seen as the most intuitive way to establish a communication with a computer. Interfaces based on hand-gesture recognition represent an attractive and natural alternative to traditional HCI devices [1], since they are less intrusive [2] and more convenient to interact with 3-D spaces [3].

To recognize hand gestures or body gestures in a vision-based environment, most popular approaches adopt a feature extraction stage to address several challenges, such as illumination changes, rotations, different viewpoints, occlusions, and so on. Many spatiotemporal descriptors have been proposed for dynamic gesture recognition. Most of them are spatiotemporal extensions of image descriptors, such as 3-D-Histogram of Oriented Gradients (HOG) [4], 3-D-Scale Invariant Feature Transform (SIFT) [5], 3-D-Speeded-Up Robust Feature (SURF) [6], volume local binary pattern (VLBP), and local binary pattern (LBP) from three orthogonal planes (LBP-TOP) [7].

For classification, two of the most popular approaches for hand-gesture recognition are those based on machine learning techniques and template matching. Thus, in [8], the hand pose is described by its contour shape, and then, classified by using template matching through a shape distance metric called finger-earth movers distance (FEMD). The work in [9] transforms the depth map of a hand pose into a point cloud, which is characterized by the ensemble of shape function (ESF) descriptor, and classified by a multilayered random forest (MLRF). In [10], the hand shape features are based on Gabor filters computed over intensity and depth images, and the classification task is carried out by a multiclass random forest. A 2-D volumetric shape descriptor along with a support vector machine (SVM) is presented in [11] for hand posture classification using depth imagery. More recently, the work in [12] utilizes depth and intensity channels with 3-D convolutional neural networks.

Unlike template matching, machine learning algorithms require a training stage to create a classification model from feature samples. In general, a very large number of training samples is needed for the classifier to appropriately learn the input features. This fact poses practical issues since the memory requirements for the training stage can be prohibitive. And this is even worse for high-dimensional feature vectors. To deal with this problem, methods such as principal component analysis [13], and random projection [14] have been employed. Other works just try to design efficient descriptors that achieve a tradeoff between recognition rate and feature dimensionality.

In this study, a new descriptor for hand-gesture recognition in video sequences is proposed. It extends and solves the problems of the hand-gesture recognition system described in [15] that introduced the volumetric spatiograms of LBP (VS-LBP). This feature descriptor could be impractical for some applications due to its excessive large dimensionality. The new descriptor, called temporal pyramid matching of local binary subpatterns (TPM-SLBsP), has two main advantages. First, it achieves a significant reduction in the descriptor dimensionality by introducing the concept of local binary subpatterns (LBsP). The second advantage is the adaption of the spatial pyramid matching (SPM) concept [16] to the temporal domain, which has been called temporal pyramid matching (TPM). This strategy encodes the temporal information in a very compact fashion, and allows to recognize gestures of different temporal length. As a result, the final video descriptor efficiently combines local and global spatial information to achieve a high discriminative hand-pose representation, also including multiresolution

temporal information to be able to recognize both static and dynamic gestures.

II. TPM-LBsP Video Descriptor

The VS-LBP descriptor implemented in [15] suffers from a dimensionality problem that could limit its practical application due to memory requirements. If less spatial and temporal information is considered, the dimensionality of the VS-LBP descriptor decreases, but at expense of a lower discriminatory ability.

To overcome these disadvantages, the VS-LBP descriptor has been improved by globally reducing its dimensionality, and considering more temporal information. This new approach consists of three steps. In the first one, a segmented video sequence is analyzed frame by frame to extract local spatial features, which are compactly represented by a histogram of LBsP (H-LBsP). The second step is based on generating global spatial histograms containing information about the location of the previously computed local features. This global spatial information is represented by the spatiograms of LBsP (S-LBsP). In the third step, the video sequence is analyzed in the time domain to introduce temporal information. TPM is applied to generate a collection of multitemporal histograms containing both spatial and temporal information from different subsequences. The concatenation of these histograms is the final representation of the video sequence.

A. H-LBsP

A new extension of the LBP descriptor [17] is proposed to extract local spatial information from each frame. It is called LBsP, and its main objective is to reduce the dimensionality of the final histogram (H-LBsP). Both LBP and LBsP are described later to better understand the principal differences between them.

1) LBP: This descriptor thresholds the neighborhood of a pixel by the intensity value of the center pixel, and forms a binary number that is finally converted to decimal. The resulting $\mathrm{LBP}_{P,R}$ pattern is defined as

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{1}$$

where g_c corresponds to the gray value of the center pixel of the local neighborhood, $g_p (p = 0, ..., P-1)$ corresponds to the gray values of P equally spaced sampling points on a neighborhood with radius R, and $s(x)$ is the sign function defined as

$$\mathrm{s(x)} = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0. \end{cases} \tag{2}$$

Finally, all the LBP patterns from a given image region are used to generate a histogram of 2^P labels or LBP types. An illustrative example is shown in Fig. 1, where the LBP computation is carried out in a 3×3 neighborhood, where the radius is $R = 1$, and the number of neighbors is $P = 8$.

2) LBsP: The LBsP descriptor follows the same strategy. It thresholds the neighborhood of each pixel and concatenates them to form a P-bit binary number. However in this case, the P-binary number is divided into n subpatterns of P/n bits, which are defined as

$$\mathrm{LBsP}^i_{P,R,n} = \sum_{p=\frac{P}{n}(i-1)}^{(\frac{P}{n}i)-1} s(g_p - g_c)2^{(p \bmod \frac{P}{n})} \tag{3}$$

where $i = 1, ..., n$, and $a \bmod b$ is the modulus operator or remainder after division between a and b.

Then, every $\mathrm{LBsP}^i_{P,R,n}$ subpattern contributes to a different histogram so that n histograms of $2^{\frac{P}{n}}$ bins are generated for the given image region. Finally, they are concatenated to form the H-LBsP. Fig. 2 shows the computation of the LBsP descriptor for a 3×3 neighborhood, $P = 8$ neighbors, and $n = 2$ subpatterns.

To summarize, while the LBP descriptor extracts only one pattern per pixel, and can generate 2^P different patterns per image region, the LBsP descriptor extracts n patterns per pixel, and can generate $2^{\frac{P}{n}}$ different patterns per image region. Therefore, this approach reduces the dimensionality of the final histogram from 2^P bins down to $n2^{\frac{P}{n}}$ bins. This is particularly interesting when using neighborhoods with a large number of sampling

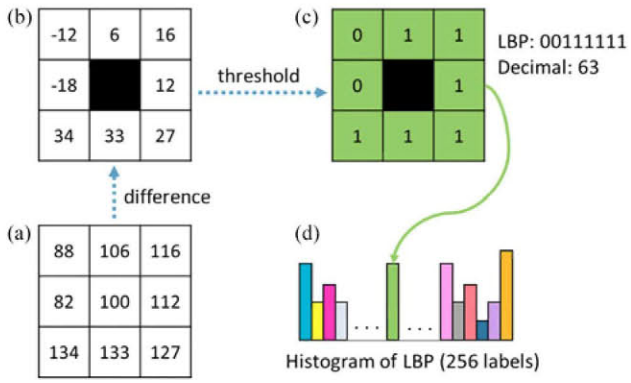

Fig. 1. Computation of the LBP operator. (a) 3×3 gray scale neighborhood. (b) Differences between the center pixel and its neighbors. (c) Thresholded neighborhood, binary pattern representation, and decimal conversion to obtain one LBP. (d) H-LBPs obtained from the considered region.

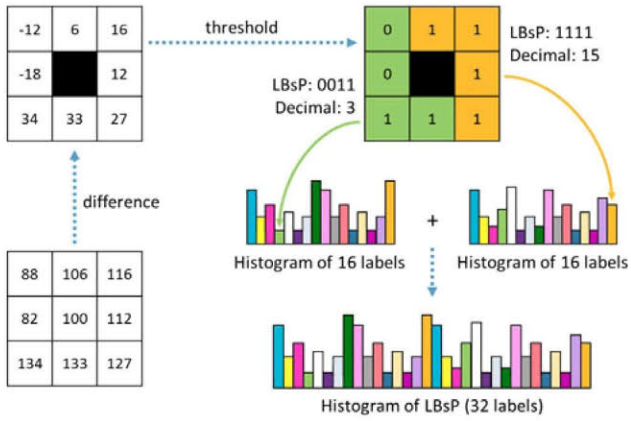

Fig. 2. LBsP computation. The LBP type 00 111 111 is divided into the LBsP types 0 011 and 1 111, each one contributing to a different histogram. The two resulting histograms are concatenated to form the H-LBsP histogram.

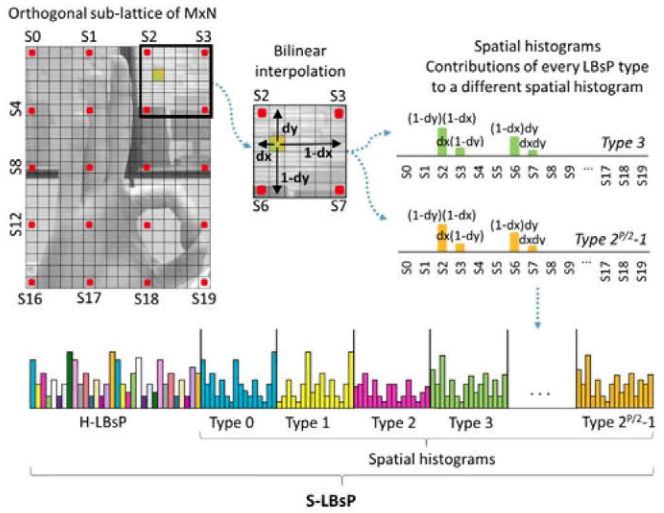

Fig. 3. S-LBsP computation. Red-colored dots correspond to the $M \times N$ lattice points. The colored pixel is an example, from which $n = 2$ LBsP are extracted: types 3 and $2^{\frac{P}{2}} - 1$. The coordinates of each LBsP type contribute to four bins within its respective spatiogram. Finally, all of them are concatenated along with the H-LBsP to generate the S-LBsP descriptor.

points, since the length of the histograms grows exponentially with P.

B. S-LBsP

To complement the local information obtained in the previous step, the "spatiogram" (spatial histogram) concept [15] gathers global spatial information from each frame by representing how the LBsP patterns are spatially distributed.

A new set of $2^{P/n}$ spatial histograms (one per each LBsP type) is computed as follows. For each frame, an orthogonal sublattice of M rows by N columns is overlaid on the image pixels (red dots in Fig. 3), and is represented by the spatial histograms with $M \times N$ bins. Lattice points serve as gathering knots for the spatial information so that the location of every previously computed LBsP pattern is defined by them. In particular, to keep a reasonable compromise between rich description and efficiency, the location of every LBsP pattern is defined by its four closest lattice points, which contributes to only four bins within the corresponding spatial histogram. This contribution is weighted using a bilinear interpolation among the four closest lattice points, as shown Fig. 3. This approach increases the robustness against slight image translations, and the grid effect.

In this process, a spatial histogram per each LBsP pattern is obtained. The final feature descriptor is obtained by concatenating all the spatial histograms with the H-LBsP histogram, which is called Spatiograms of LBsP, and whose notation is $S-\text{LBsP}_{P,R,n,M,N}$. The dimensionality of this new S-LBsP descriptor is $2^{\frac{P}{n}}(n + M \times N)$, lower than in the original S-LBP descriptor [15], which is $2^P(1 + M \times N)$.

C. TPM

The last step consists of adding temporal information to the previously computed spatial features. To that end, the SPM concept [16] has been extended to the time domain, and called TPM. To do so, the video sequence is represented as a time-ordered collection of S-LBsP image descriptors, one per frame. The TPM method (see Fig. 4) represents a video sequence as a multiresolution temporal pyramid, where each pyramid level is partitioned into 2^l subsequences. For each subsequence a feature vector is computed by applying the average pooling function to the set of S-LBsP descriptors derived from the frames of the considered subsequence. Finally, all feature vectors z_{lj} from all subsequences and pyramid levels are concatenated to form a multitemporal representation of the video sequence, whose notation is $\text{TPM} - \text{SLBsP}_{P,R,n,M,N,l}$.

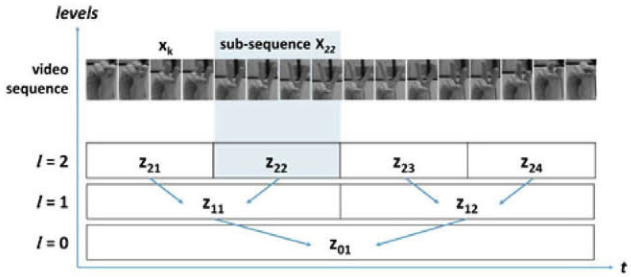

Fig. 4. TPM. For each level l, the video sequence is partitioned into 2^l subsequences. S-LBsP descriptors extracted from each subsequence are averaged by means of the average pooling function. The resulting z_{lj} histograms are concatenated to form the final feature vector representing the video sequence.

The average pooling function is defined as

$$z_{lj} = \frac{1}{K_l} \sum_{k \in X_{lj}} x_k \tag{4}$$

where K_l is the number of frames of each sub-sequence in the level l, X_{lj} is the set of frames belonging to the jth subsequence in pyramid level l, and x_k is the S-LBsP descriptor extracted from the kth frame in its corresponding X_{lj} set.

In the previous approach [15], the temporal information was extracted from a few specific temporal instants to keep the dimensionality of the final VS-LBP descriptor low. However, the new approach extracts temporal information from all the instants without discarding any frame. This way, not only the most representative states are considered, but also the whole evolution of the hand gesture along time. In addition, the multitemporal resolution scheme allows to detect hand gestures with different lengths in time. Moreover, as S-LBsP descriptors are averaged, histograms z_{lj} keep the same dimensionality as S-LBsP, offering a scalable solution.

III. Experiments

The proposed approach has been validated on three datasets: hand-gesture database [15], created for HCI, and American Sign Language (ASL) Fingerspelling [10] and Nanyang Technological University (NTU) [8], related to sign language. For classification, a bank of SVM classifiers was used under the one-versus-all strategy. The metric used to compare the different approaches is the *Average accuracy* metric, defined as $\frac{\textit{Total number of correct gestures}}{\textit{Total number of gestures}}$.

TABLE I
COMPARISON OF THE ACCURACY AND DIMENSIONALITY AMONG DIFFERENT STATE-OF-THE-ART METHODS USING THE HAND-GESTURE DATABASE

Method	Seq_1	Seq_2	Seq_3	Seq_4	Seq_5	Seq_6	Mean Accuracy	Dimension
$VLBP_{4,1,1}$ [7]	0.720	0.765	0.711	0.695	0.689	0.770	0.725	16384
$LBP\text{-}TOP_{8,8,8,1,1,1}$ [7]	0.507	0.521	0.535	0.487	0.485	0.510	0.507	768
$VS\text{-}LBP_{8,1,4,10,5}$ [15]	0.949	0.961	0.935	0.923	0.897	0.959	0.937	52480
$VS\text{-}LBsP_{8,1,2,10,8,9}$	0.984	0.972	0.946	0.934	0.908	0.970	0.952	11808
$TPM\text{-}SLBP_{8,1,10,8,3}$	0.974	0.961	0.976	0.960	0.926	0.975	0.962	145152
$TPM\text{-}SLBsP_{8,1,2,10,8,3}$	0.980	0.957	0.981	0.972	0.924	0.974	0.965	9184

TABLE II
COMPARISON ON THE ASL FINGER-SPELLING DATASET

Method	Accuracy	Dimension
DL + SVM [11]	0.962	256
SCF [18]	0.978	1000
ESF descriptor + RF [9]	0.850	640
ESF descriptor + MLRF [9]	0.870	640
GR + RF (depth) [10]	0.690	1024
GR + RF (color) [10]	0.730	1024
GR + RF (depth + color) [10]	0.750	2048
VS-LBP + SVM [15]	0.975	52480
VS-LBsP + SVM	0.985	11808
TPM-SLBP + SVM	-	145152
TPM-SLBsP + SVM	0.995	9184

TABLE III
COMPARISON ON THE NTU DATABASE

Methods Based on Template Matching	Accuracy
Shape context with bending cost [8]	0.791
Shape context without bending cost [8]	0.832
Skeleton matching [8]	0.786
Thresholding Decomposition + FEMD [8]	0.932
Near-convex Decomposition + FEMD [8]	0.939
Methods Based on Classifiers	**Accuracy**
DL + SVM [11]	0.971
S-LBP + SVM [15]	0.973
S-LBsP + SVM	0.979

Table I shows a comparison with other methods on the hand-gesture database [15], where optimum parameter configurations have been selected for each one. The $VS\text{-}LBsP_{P,R,n,M,N,s}$ and $TPM\text{-}SLBP_{P,R,M,N,l}$ show independently the efficiency of the LBsP and TPM concepts. The first one employs the spatial descriptor S-LBsP and the temporal sampling scheme proposed in [15]. The second one combines the spatial descriptor S-LBP proposed in [15] with the TPM approach. The proposed approach outperforms those in the state of the art, that is VLBP, LBP-TOP, and VS-LBP, in terms of recognition rate. Regarding dimensionality, only LBP-TOP would be the shorter one followed by TPM-SLBsP, however, it has a poor recognition rate. Therefore, TPM-SLBsP is still the best candidate.

Notice that both TPM-SLBsP and TPM-SLBP achieve comparable recognition rates. However, the dimensionality of TPM-SLBsP is much lower (a reduction of 93.67%). Comparing the temporal schemes (TPM-SLBsP and VS-LBsP), TPM achieves a slightly better improvement in accuracy. Moreover, it generates a shorter feature vector. This is attributed to the fact that TPM takes into account all frames in the video sequence for temporal analysis in a very compact fashion, while temporal sampling approach only considers a few of them.

Table II compares different state-of-the-art approaches with published results on the ASL fingerspelling dataset. In this dataset, 3 subjects perform 24 signs from the ASL language, where 250 frames have been collected for every sign. Therefore, this is an example of a large dataset, where half of the data have been considered for training and the other half for testing. As can be observed, the proposed approach outperforms all the methods. Notice that some of them use depth and intensity information, unlike the proposed one, that only use intensity. In terms of dimensionality, the length of the feature vectors for most of the descriptors are lower than for TPM-SLBsP. These descriptors use high-level features for representing the hand gestures, mainly related to the hand shape, leading to shorter feature vectors. However, this also implies lower recognition rate, especially in [9] and [10]. In the case of the TPM-SLBP descriptor, it has not been possible to perform the training process with optimum parameter configuration due to its high memory requirements.

Table III shows the results for the NTU dataset, which only contains single-image static gesture samples. For this reason, only the S-LBsP strategy is used and compared with other approaches, since TPM (or other temporal strategy) is not applicable. For this dataset, half of the data have been considered for training and the other half for testing as well. In general, both template matching and machine learning techniques achieve high recognition rates for static hand gestures since it is an easier problem than with dynamic gestures. Once again, the proposed approach, that is S-LBsP + SVM, is slightly better in spite of using only intensity information. It is closely followed by S-LBP+SVM that only uses intensity information as well, and DL+SVM that uses both intensity and depth information.

IV. CONCLUSION

A new descriptor for hand-gesture recognition has been proposed to increase the discriminative power and mitigate its dimensionality problems. The combination of both local and global spatial features makes it highly discriminative at frame level. Moreover, as the dimensionality of the final image descriptor that encodes these spatial features has been reduced, a new scheme for extracting more temporal information has been considered, which also introduce multiresolution temporal support. This increases the discriminatory ability of the final video descriptor to recognize dynamic hand gestures, and be more flexible regarding memory requirements.

REFERENCES

[1] Q. Feng, C. Yang, X. Wu, and Z. Li, "A smart TV interaction system based on hand gesture recognition by using RGB-D sensor," in *Proc. Int. Conf. Mechatronic Sci., Elect. Eng. Comput.*, Dec. 2013, pp. 1319–1322.

[2] A. Mewes, P. Saalfeld, O. Riabikin, M. Skalej, and C. Hansen, "A gesture-controlled projection display for CT-guided interventions," *Int. J. Comput. Assisted Radiol. Surg.*, vol. 11, pp. 157–164, May 2015.

[3] M. Billinghurst, T. Piumsomboon, and H. Bai, "Hands in space: Gesture interaction with augmented-reality interfaces," *IEEE Comput. Graph. Appl.*, vol. 34, no. 1, pp. 77–80, Jan. 2014.

[4] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-Gradients," in *Proc. 19th Brit. Mach. Vis. Conf.*, Sep. 2008, pp. 275-1–275-10.

[5] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th ACM Int. Conf. Multimedia*, Sep. 2007, pp. 357–360.

[6] G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. 10th Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 650–663.

[7] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[8] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1110–1120, Aug. 2013.

[9] A. Kuznetsova, L. Leal-Taix, and B. Rosenhahn, "Real-time sign language recognition using a consumer depth camera," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 83–90.

[10] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2011, pp. 1114–1119.

[11] Y. Wang and R. Yang, "Real-time hand posture recognition based on hand dominant line using kinect," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2013, pp. 1–4.

[12] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2015, pp. 1–7.

[13] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *Proc. Eur. Signal Process. Conf.*, Aug. 2012, pp. 1975–1979.

[14] J. Wang, X. Sun, P. Liu, M. F. She, and L. Kong, "Sparse representation of local spatial-temporal features with dimensionality reduction for motion recognition," *Neurocomputing*, vol. 115, pp. 150–160, 2013.

[15] A. I. Maqueda, C. R. del Blanco, F. Jaureguizar, and N. García, "Human-computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns," *Comput. Vis. Image Understanding*, vol. 141, pp. 126–137, Dec. 2015.

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, vol. 2, Jun. 2006, pp. 2169–2178.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[18] C. Keskin, F. Kira, Y. E. Kara, and L. Akarun, "Randomized decision forests for static and dynamic hand shape classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2012, pp. 31–36.