Multi-modal Descriptors for Multi-class Hand Pose Recognition in Human Computer Interaction Systems

Jordi Abella
Computer Vision Center
jabella@cvc.uab.es

Raúl Alcaide
Computer Vision Center
ralcaide@cvc.uab.es

Anna Sabaté
Computer Vision Center
asabate@cvc.uab.es

Joan Mas
Computer Vision Center
jmas@cvc.uab.es

Sergio Escalera
Computer Vision Center
University of Barcelona
sergio@maia.ub.es

Jordi Gonzàlez
Computer Vision Center
Univ. Autònoma de Barcelona
jordi.gonzalez@uab.cat

Coen Antens
Computer Vision Center
coen@cvc.uab.es

ABSTRACT

Hand pose recognition in advanced Human Computer Interaction systems (HCI) is becoming more feasible thanks to the use of affordable multi-modal RGB-Depth cameras. Depth data generated by these sensors is a very valuable input information, although the representation of 3D descriptors is still a critical step to obtain robust object representations. This paper presents an overview of different multi-modal descriptors, and provides a comparative study of two feature descriptors called Multi-modal Hand Shape (MHS) and Fourier-based Hand Shape (FHS), which compute local and global 2D-3D hand shape statistics to robustly describe hand poses. A new dataset of 38K hand poses has been created for real-time hand pose and gesture recognition, corresponding to five hand shape categories recorded from eight users. Experimental results show good performance of the fused MHS and FHS descriptors, improving recognition accuracy while assuring real-time computation in HCI scenarios.

1. INTRODUCTION

An increasing interest in multi-modal data fusion is currently arising in the fields of Computer Vision thanks to affordable RGB-D cameras like $Kinect^{TM}$. Several applications benefit from these new sensors, such vision surveillance, face detection, object recognition, eHealth systems and Human Computer Interaction (HCI) systems [11].

Regarding HCI, the use of the depth sensor for hand detection and gesture recognition allows vision-based interfaces to use the user hands to interact and communicate with a computer, thus providing intuitive means of navigation and control. However, robust and accurate 3D hand pose recognition for real-time computing is still an open problem, and several research papers have been recently devoted to this issue.

Descriptors for 2D object recognition have been extensively studied in literature [14, 10]. In this sense, many authors have defined new 3D descriptors as an extension of well-known 2D image features, such as SIFT, HOG or SURF [15, 6]. Although successful results were obtained for some 3D object recognition tasks, methods are still computationally expensive for real-time performance [13]. In addition, several others depth descriptors for 3D object recognition have been recently proposed [1, 19, 16, 2]. The SHOT descriptor [12] uses normal points from a 3D grid surface and calculates the angle between them and their feature point normals. However, the lack of warping for non-rigid surfaces yields to poor results. Other 3D descriptors include spherical harmonics features [5] or the relative curvature of a vertex [3], though they become computationally expensive for real-time performance. Towards this end, the Fast-Point Feature Histogram (F-PFH) [21] computationally improves the original PFH (in essence, 3D geometric primitives based on linked points [22]) by codifying the 3D environment angles relation. Lastly, the Viewpoint Feature Histogram (VFH) [23] appears as a combination of F-PFH and surface normals. The common drawback in all previous descriptors is the generation of blurred point clouds, loosing partial 3D details, and thus reducing final recognition accuracy.

The aim of this paper is to present a comprehensive study and comparison of existing multi-modal descriptors applied to 3D hand pose recognition. We present a real-time hand pose recognition system based on two feature descriptors called Fourier-based Hand Shape (FHS) and Multi-modal Hand Shape (MHS). The FHS descriptor is based on the descriptor presented in [24], here extended to the use of 3D segmented depth data. In addition we reduce the descriptor size to 44 sub-sampled coefficients while enclosing relevant information for hand pose classification. The MHS descriptor is based on the work presented in [17], enhanced with an equally relevant features calculation by applying data

ICMI'13, December 9–13, 2013, Sydney, Australia
Copyright 2013 ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2532596.

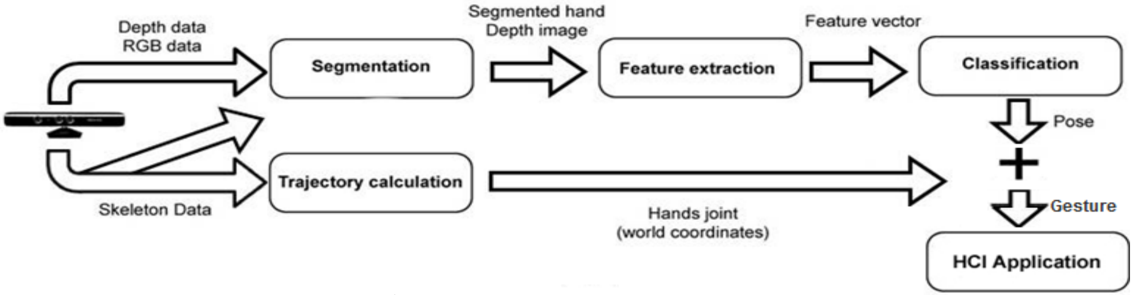


Figure 1: HCI system flowchart.

normalization. Finally, we combine the FHS and MHS descriptors which constitute the core of a real-time HCI system for hand pose detection and gesture recognition. For classification, we compare support vector machines (SVM) [7], randomized decision forests (RF) [4] and non-parametric K-Nearest Neighbors (KNN) [9], obtaining successful results.

The rest of the paper is organized as follows. Section 2 presents the proposed system for HCI. Section 3 includes hand pose classification results and HCI application examples. Section 4 describes the main conclusions obtained from the experimental results.

2. SYSTEM

The flowchart of the system being implemented is shown in Figure 1. First, both color and depth images are acquired from the Kinect^{TM} sensor and the user skeleton is computed based on Random Forest and Mean Shift approach of [25]. From that, hand segmentation is applied using skeleton and depth data only, yielding to a 3D segmented hand depth map. Then, features are calculated by generating a FHS/MHS descriptor feature vector as input to the pose classification module. These features codify local and global depth and silhouette statistics from the segmented used masks, meanwhile aligned RGB data is used for monitoring purposes within the HCI application interface. Finally, gesture recognition is calculated by combining the skeleton data trajectory and the classified pose for real-time interaction.

For current HCI experiments, the Kinect^{TM} motion sensing input device designed by Microsoft is utilized. The main advantage for the user is the control and interaction with the console by gesturing and speaking. The Kinect^{TM} contains an RGB color camera, a depth sensor and a multi-array microphone. Both color and depth sensors acquire at 640×480 pixels resolution and 30 fps acquisition rate. The depth sensor works within a range distance from 0.7 to 5.0 meters. It comes with built-in software, which is able to acquire 3D full-body in motion, face and hand recognition, and voice interpretation.

The advantage of Kinect^{TM} being multi-modal allows depth and color acquisition simultaneously. Segmentation process utilizes depth and skeleton data, and color image is utilized to combine both color and depth information by 3D calibration to visualize depth segmented data on color image at the HCI screen. With reference to gesture recognition, the SDK that comes with Kinect^{TM} is only able to recognize two hand poses, open and close hand. This limitation has challenged researchers to broaden the gesture recognition set. Therefore, a methodology for creating an extended gesture dictionary is presented, which include hand segmentation, descriptors computation, pose classification, gesture recognition and an integration step in a HCI system.

2.1 Hand Segmentation

Hand detection is based on depth segmentation. First, joints are detected using the original skeleton analysis libraries of Kinect^{TM} based on Random Forest segmentation and Mean Shift skeletal joint estimation [25]. From that, a 3D depth window centered at hand joint position is selected, thus generating the final depth hand segmentation. Since segmented hands size may vary from different individuals, image normalization by scaling to a model size is applied in order to obtain similar-sized hands segmented depth images. Results from hand segmentation are shown in Figure 2.

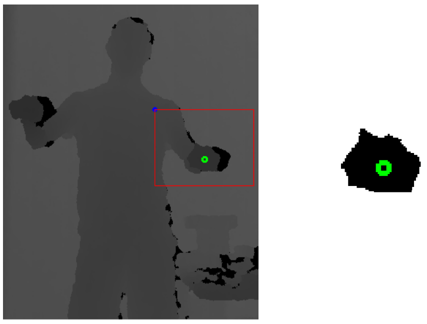


Figure 2: Example of depth hand image segmentation.

2.2 Hand Pose Descriptors

The following step is to generate an appropriate feature vector to describe a hand pose class, which is thereafter used as input for pose classification. This work presents a comparative study of feature descriptor representation using Fourier-based descriptors (FHS) and multi-modal hand shape descriptors (MHS) by five features descriptor sets, and classification comparing support vector machines (SVM) [7], the randomized decision forest (RF) [4] and the non-parametric K-Nearest Neighbour classifier(KNN) [9]. The five descriptor sets are presented next.

2.2.1 *Fourier based Hand Shape descriptors*

The FHS used in this study is based on the Fourier-based descriptor presented in [24], which considers the Fourier coefficients of a 2D segmented image. Contour points $(x_i, y_i), i = 0, .., N-1$ are represented by complex numbers, z_i:

$$\forall i \in [0, .., N-1], (x_i, y_i) \Leftrightarrow z_i = x_i + \mathrm{j}y_i. \quad (1)$$

In this study, FHS is applied to the 3D segmented image hand [24]. As an extension, we use 3D segmented depth data instead while reducing the descriptor to 44 sub-sampled coefficients, enclosing relevant information for pose classification as described next.

A discrete sub-sampled set from the hand contour is used as input to the Fast Fourier Transform (FFT). For the subsample, 64 points are selected, and thereafter the Fourier coefficients are calculated as follows:

$$C_k = \sum_{i=0}^{N-1} z_i \mathrm{e}(\frac{-2\pi \mathrm{j} ik}{N}),\ k = 0,..,N-1. \quad (2)$$

Then, the inverse Fourier Transform restores z_i as:

$$z_i = \sum_{k=0}^{N-1} C_k \mathrm{e}(\frac{2\pi \mathrm{j} ik}{N}),\ k = 0,..,N-1. \quad (3)$$

For a more intuitive interpretation, the zero frequency is moved to the center of the vector C, and the input image is multiplied by $(-1)^k = e^{j\omega}$, where ω is the value of $2\pi(k/2)$. As a result, the spectrum calculated is sampled by half the frequency.

Since rotation and scale invariant results are required, the coefficients are normalized by C_1, yielding $N-2$ Fourier descriptors I_k:

$$I_k = \frac{|C_k|}{|C_1|}, k = 2,..,N-1, \quad (4)$$

$$\widehat{C_k} = \sum_{i=0}^{M-1} z_i \mathrm{e}(\frac{-2\pi \mathrm{j} ik}{N}),\ k = 0,..,N-1. \quad (5)$$

The low-frequency components define the global shape of the boundary while high-frequency components represent fine details. C_0 corresponds to the image position. Hence, the selection of 44 coefficients is considered sufficient and representative of the current Fourier-based descriptor study using 3D segmented data. In order to determine the selection of Fourier frequencies to represent a shape, Figure 3 shows examples of shape poses reconstruction.

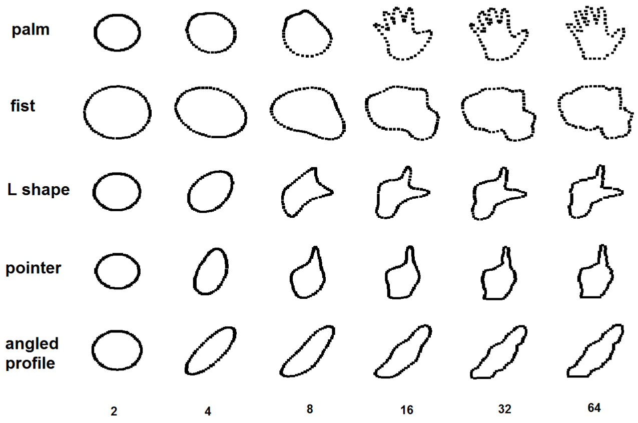


Figure 3: Example of reconstruction with FD: The image is sampled to obtain 64 frequencies. The output image is reconstructed using from 2 to 64. Sixteen frequencies are sufficient to identify a shape, where low ones represent the contour and high ones fine details.

2.2.2 *Multi-modal Hand Shape descriptors*

The MHS used in this study is extended from the descriptor presented in [17]. In particular, we apply an equally relevant features calculation step by means of data normalization.

The features of the descriptor are divided into three subsets A, B and C, in order to identify and characterize segment hand patches. The feature set MHS-A performs a global image statistics like the percentage of pixels that is covered by the blob contour, number of fingertips detected, the mean angle from the blob's centroid to those fingertips and the Hu moments. A second descriptor MHS-B is built from the number of pixels covered by every possible rectangle contained in the blob's bounding box and then normalized by its total size. Finally, a third feature set MHS-C uses a similar grid as the second set. However, instead of analyzing the coverage within different rectangles, it is composed from the difference between the mean depth for each pair of individual cells. In order to generate equally relevant features, the coefficients of the features descriptor are normalized. This features descriptor creates a 535 size vector.

2.2.3 *Combining FHS and MHS descriptors*

Five features descriptor sets were proposed for this work, as a combination of the above described FHS and MHS strategies. The first feature descriptor set, MHS, accounts for the complete 535 normalized MHS features. FHS accounts for the 44 FHS features described above. The third descriptor set, MHS-A, comprises the subset A of MHS, containing the first 10 MHS features. The fourth set, FHS-MHS-A, contains the complete 44 FHS features and the first 10 features corresponding to subset A of MHS. Finally, the fifth features descriptor set corresponds to the complete FHS descriptor and 525 features from the B and C subsets of MHS. Table 1 shows a brief description of the five set combinations. All features are normalized by each individual feature of the training set, thus yielding equally relevant information data for training each dimension of the feature space.

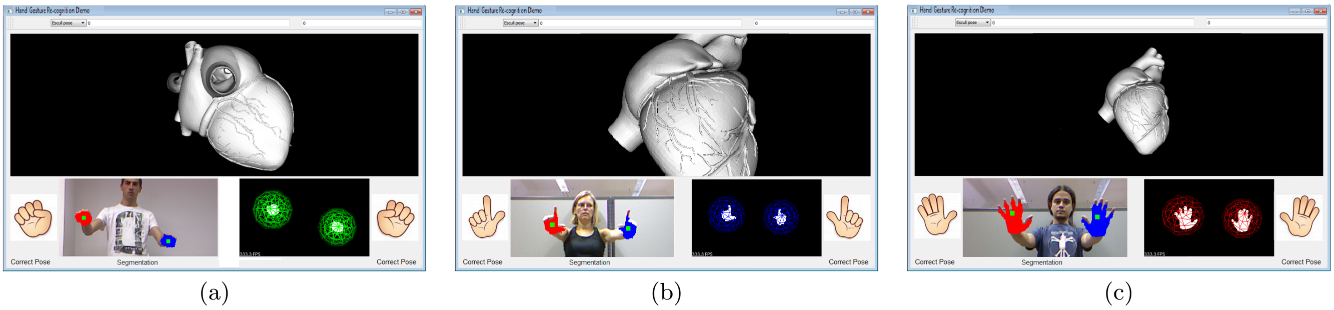
Descriptor set	Combination	# Features
1	MHS	535
2	FHS	44
3	MHS A set	10
4	FHS and MHS A set	54
5	FHS and MHS B and C sets	569

Table 1: Summary of the feature combination sets and descriptor lengths.

2.3 Gesture Recognition for HCI

The different multi-modal descriptors and classifiers were integrated as part of a HCI design for recognizing multiple user hand poses and track their trajectories. This pose-trajectory estimation was used in order to interact and navigate with volumes by means of different actions like zooming, rotation or translation. First, the HCI was utilized to acquire the dataset to train and to test the classifier in real-time.

The HCI system was developed in C++, using Qt[20], PCL[19] and VTK[26] for the graphical user interface. The Kinect^{TM} acquisition module was implemented by Windows



(a) (b) (c)

Figure 4: Human computer interaction system for gesture recognition: gestures are aimed to navigate through a 3D object. (a) Close hand for rotating, (b) L-pose for translation, and (c) open hand for zoom.

Kinect SDK, and image analysis by OpenCV[18] libraries. The Model-view-controller design (MVC) [8] was followed to structure the software application and separate user interface from the rest of implementation.

Figure 4 shows the HCI windows interface and examples of human interaction and gesture recognition. In this application, the gestures of the user are identified and directly linked to an action of the 3D object: zoom, rotation and translation. In the image, segmented hand regions are visualized (bottom-left), 3D hand point clouds are displayed (bottom-right), and the hand pose recognition is applied and showed for each of the to user hands. For example, in Fig. 4(a) one can see that the classification is correctly performed as 'close hand' pose, which in our application stands for rotation. Then the hand pose in combination with the real distance movements of the hands defines the input interaction for rotating the 3D volume. The same is applied for translation and zoom in Fig. 4(b) and (c), respectively.

3. EXPERIMENTAL RESULTS

This section describes the dataset generated by an integrated HCI system, and explains the experiments that were carried out by all fifteen combinations of descriptors and classifiers. Additionally, the HCI integration is also described as a result of a real-time gesture recognition HCI system.

The dataset for the experiments was acquired using the integrated HCI system, where both depth and skeleton data were extracted from the Kinect^{TM}, and processed to generate the descriptor feature sets. A total of 38K labeled instances from eight individuals were obtained during acquisition. Five poses were recorded from both right and left hands, using 24K instances for training and keeping 13K for testing. Table 2 shows the five hand shape categories corresponding to the training poses.

All experiments reported in this paper were performed on a DELL Precision T3500, with Intel Xeon CPU W3530 2.80GHz processor and 6GB RAM, running under Windows 7 Enterprise 64 bits SP1.

A separate training was considered for each training set, utilizing the support vector machine (SVM), the randomized decision forest (RF) and the non-parametric the non-parametric K-Nearest Neighbor classifier (KNN). SVM was trained with LIBSVM [7], using a radial basis function (RBF) kernel and optimizing parameters to maximize accuracy. RF was trained on 50 trees with depth of 15.

	palm	fist	L shape	pointer	angled profile
R					
L					

Table 2: Examples of the five hand pose categories evaluated in this paper. First row shows right hand and second row shows left hand.

3.1 Performance evaluation of the proposed descriptors

The combination of descriptors and classifiers produced a total of 15 experimental setups. Testing results are shown in Figure 5 in terms of accuracy percentage of true positive classification. The first columns stand for the success classification using the KNN classifier. The second ones correspond to the success rate under the SVM classifier. The third ones account to the success rates with the RF classifier. Each three-columns set correspond to, from left to right, the five descriptor training set employed at each classification. The highest success rate stands for the SVM classifier for all five descriptor sets, with a maximum success rate of 92.6%. The second one, the RF with a maximum success rate of 88.6%, and far from best results is the KNN with a maximum success rate of 83.5%. An explanation for the lower results when comparing the RF with the SVM may be the depth of 15 during RF training, showing the need of a larger depth which may increase the number of features to use.

Analyzing results, the best performance corresponds to the FHS by KNN, the FHS&MHS-A by SVM and the FHS by RF. The combination of FHS and MHS-A improves results with respect to FHS alone. A reason may be de addition of information which complements descriptor features. However, FHS&MHS-B-C results worsen slightly since MHS B-C does not provide relevant information to the FHS features descriptor.

Overall, the best performance of pose recognition accounts for the FHS&MHS-A set trained by SVM. To further analyze results from this best performance, the confusion matrix is evaluated, as shown in Table 3. Predicted output

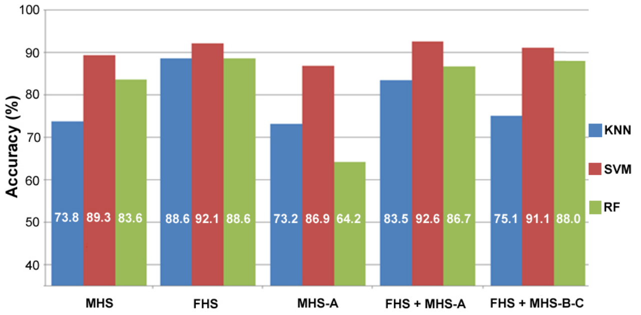


Figure 5: Testing results in accuracy percentage for each feature set and considered classifiers.

is compared against current results to calculate the error distribution. The true positives (TP) rate is significant on palm and fist recognition, while low on angled profile pose. On this case, L-shape pose may be recognized instead of angled profile. Because the angled pose is manually difficult to perform, there is more posture variability between different individuals, thus showing L-shaped rather than angled one.

		Predicted				
		Palm	Fist	L shape	Pointer	Angled profile
Output result	Palm	98,4%	0,7%	1,3%	0,1%	5,0%
	Fist	0,8%	98,8%	5,9%	0,7%	5,9%
	L shape	0,1%	0,3%	91,4%	1,7%	11,4%
	Pointer	0,6%	0,1%	1,1%	97,4%	0,6%
	Angled profile	0,1%	0,2%	0,3%	0,1%	77,0%

Table 3: Confusion matrix of the best performance (FHS&MHS-A by SVM), which quantifies the error distribution and success rate.

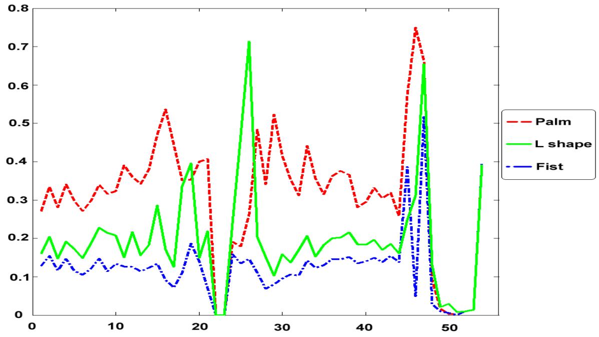


Figure 6: Features of best descriptor set: FHS & MHS-A set computing a total of fifty-four features. Comparison of three poses by their features values.

Concerning FHS&MHS-A analysis, the chart of Figure 6 shows a comparison between three feature vectors, palm, fist and L shape. Each vector value corresponds to the mean of trained features. For each class, the mean vector is calculated, as mean and median values are similar. Such intraclass variability may be related to the normalization process during segmentation as explained in section 2.1. Each line corresponds to the 54 feature values of palm, L-shape and fist poses. It may be observed that three classes are clearly distinguished, being the palm and fist the most ones.

A comparison between classes is shown in Figure 7. Each chart represents the difference between the mean of two classes. Changes between classes are represented by non-zero values. Similarly, if the same feature has zero values in all charts, the feature has no information. Therefore, this feature may be excluded from the vector, thus reducing training time. The additional 10 FHS features provide more information in all classes, yielding better results than FHS alone.

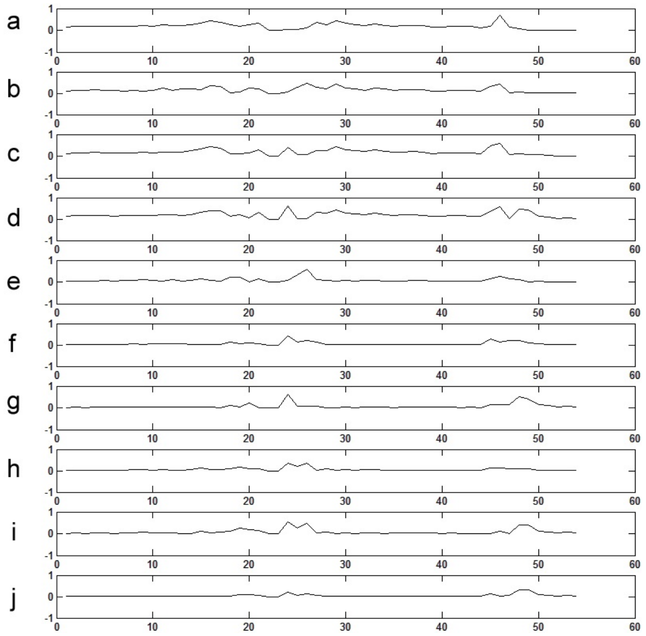


Figure 7: Class vectors mean differences: (a) palm vs fist; (b) palm vs L-shape; (c) palm vs pointer; (d) palm vs angled profile; (e) fist vs L-shape; (f) fist vs pointer; (g) fist vs angled profile; (h) L-shape vs pointer; (i) L-shape vs angled profile; (j) pointer vs angled profile.

3.2 Performance Evaluation on a HCI System

The dataset extraction and experiments were done using a HCI system, as all methods were part of the integration of a HCI system.

Best performed descriptor set FHS&MHS-A was benchmarked using the previously mentioned testing dataset. The processing time for descriptor step processing time was $4.527 msec$ and $0.271 msec$ for classification. These values contributed to the HCI system the achievement of real-time execution and a smooth look&feel.

In addition to the real-time and accurate classification performance, the HCI interface allows for visual inspection of user hand segmentation. Our visual inspection was very useful in order to evaluate the hand segmentation procedure for several users with different hand physiognomy, analyzing deviations of hand joint estimation, bad orientation of the subject and strange postural poses, allowing for better fitting of methods parameters for accurate user-independent 3D hand point cloud segmentation.

4. CONCLUSIONS

We presented a comprehensive study of existing multi-modal descriptors for multi-class hand pose recognition in Human Computer Interaction systems, comprising FHS and MHS. From them, five descriptor sets were proposed. Further training was undertaken by KNN, SVM and RF classifiers. For the experiments, a representative five poses dataset was created including both left and right hands. Images were acquired from eight individuals, thus yielding 38K frames, where 24K were used for training and the rest for testing. Results indicated higher accuracy on FHS than MHS, though FHS with the inclusion of MHS-A improved final recognition results. MHS by itself did not provide robust solutions, but its subset A contributed to improve accuracy when fused with FHS.

In particular, the proposed descriptors extend previous works by their application on 3D depth data as input to FHS instead of 2D image, and by the features-based normalization to equally train all data. Additionally, a fully functional HCI application was developed integrating the real hand pose segmentation, hand pose multi-classification, and gesture recognition technology for real-time navigation and manipulation of 3D volumes.

5. ACKNOWLEDGMENTS

The authors thank Miguel Ángel Blanco and Agnés Borràs from the Computer Vision Center, for their support during the development of the HCI application. This work has also been partially supported by the Spanish projects TIN2009-14501-C02-02 and TIN2012-39051.

6. REFERENCES

[1] L. Alexandre. 3D descriptors for object and category recognition: a comparative evaluation. In *Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[2] K. K. B Steder, R Rusu and W. Burgard. Narf: 3d range image features for object recognition. *IROS*, 2010.

[3] M. Ben-Chen and C. Gotsman. Characterizing shape using conformal factors. *Eurographics conference on 3D object retrieval (3DOR)*, pages 1–8, 2008.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] G. Burel and H. Henoco. Determination of the orientation of 3d objects using spherical harmonics. *Graph Models Image Process*, 57(5):400âĂŞ408, 1995.

[6] F. A.-R. C Redondo-Cabrera, R J Lopez-Sastre and S. Maldonado-Bascon. Surfing the point clouds: Selective 3d spatial pyramids for category-level object recognition. *CVPR*, 2012.

[7] C. Chand and C. Lin. Livsvm: A library for support vector machines. *Trans. on Intelligent System and Technology*, 2(1):1–27, 2011.

[8] E. Curry and P. Grace. Flexible self-management using the model-view-controller pattern. *IEEE Software*, 25(3):84–90, May 2008.

[9] R. Duda and P. Hart. *Pattern Classification (Pt.1)*. John Wiley and Sons (2nd Ed), New York, 2000.

[10] S. Escalera, A. Fornés, O. Pujol, J. Lladós, and P. Radeva. Circular blurred shape model for multiclass symbol recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions On*, 41(2):497–506, 2011.

[11] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopés, I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ChaLearn Multi-modal Gesture Recognition Grand Challenge and Workshop, 15th ACM International Conference on Multimodal Interaction*, 2013.

[12] S. S. F Tombari and L. Stefano. Unique signatures of histograms for local surface description. *Proceedings of the 11th European conference on computer vision conference (ECCV)*, page 356âĂŞ369, 2010.

[13] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. Recognizing objects in range data using regional point descriptors. *European Conference on Computer Vision (ECCV)*, 2004.

[14] R. Gonzalez and R. Woods. *Digital Image Processing.* Addison-Wesley Publishing Company, Inc, New York, 1992.

[15] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Gool. Hough transform and 3d surf for robust three dimensional classification. *ECCV*, 2010.

[16] M. S. M Ruggeri, G PAtane and D. Saupe. Spectral-driven isometry-invariant matching of 3d shapes. *IJCV*, 89:248–265, 2010.

[17] D. Minnen and Z. Zafrulla. Towards robust cross-user hand tracking and shape recognition. In *IEEE International Conference on Computer Vision Workshops*, pages 1235–1241, 2011.

[18] (OpenCV). Open source computer vision library. http://docs.opencv.org/, 2013.

[19] (PCL). Point cloud library. http://docs.pointclouds.org, 2013.

[20] (QT). Qt project. http://qt-project.org/doc/, 2013.

[21] N. B. R Rusu and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. *Int. Conference on Robotics and Automation*, pages 1848–1853, 2009.

[22] N. B. R Rusu, Z Marton and M. Beetz. Learning informative point classes for the acquisition of object model maps. *Control, Automation, Robotics and Visions*, pages 643–650, 2008.

[23] R. T. R Rusu, G Bradski and J. Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. *Int. Conference on Robotics and Automation*, pages 2155–2162, 2010.

[24] S. B. S Conseil and L. Martin. Comparison of fourier descriptors and hu moments for hand posture recognition. In *European Signal Processing Conference (EUSIPCO)*, 2007.

[25] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011.

[26] (VTK). Visualization toolkit. http://www.vtk.org/vtk/help/documentation.html, 2013.