INTAIRACT: Joint Hand Gesture and Fingertip Classification for Touchless Interaction

Xavier Suau, Marcel Alcoverro, Adolfo Lopez-Mendez,
Javier Ruiz-Hidalgo, and Josep Casas

Universitat Politècnica de Catalunya

Abstract. In this demo we present *intAIRact*, an online hand-based touchless interaction system. Interactions are based on easy-to-learn hand gestures, that combined with translations and rotations render a user friendly and highly configurable system. The main advantage with respect to existing approaches is that we are able to robustly locate and identify fingertips. Hence, we are able to employ a simple but powerful alphabet of gestures not only by determining the number of visible fingers in a gesture, but also which fingers are being observed. To achieve such a system we propose a novel method that jointly infers hand gestures and fingertip locations using a single depth image from a consumer depth camera. Our approach is based on a novel descriptor for depth data, the Oriented Radial Distribution (ORD) [1]. On the one hand, we exploit the ORD for robust classification of hand gestures by means of efficient *k-NN* retrieval. On the other hand, maxima of the ORD are used to perform structured inference of fingertip locations. The proposed method outperforms other state-of-the-art approaches both in gesture recognition and fingertip localization. An implementation of the ORD extraction on a GPU yields a real-time demo running at approximately 17fps on a single laptop.

1 Introduction

Until recent years, interaction between humans and computer systems has been driven through specific devices (*i.e.* mouse, keyboard). Such device-dependency turns interaction into a non-natural dialog between humans and machines. Hand gesturing is an interesting way to provide a more immersive and intuitive interaction. Recent consumer depth cameras provide pixel-wise depth information in real-time opening the door to new research directions in the field of Natural User Interfaces (NUI). Our proposal uses this kind of camera as input (*i.e.* Kinect), not requiring any other specific display nor hardware.

Combining a basic set of fingertip configurations with simple hand motion has proven to be successful with modern trackpad devices [2]. Our idea is to extend such paradigm to the touch-less world, providing a more immersive experience than physical trackpads.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 602–606, 2012.

The proposed demonstration enables the user to interact with virtual objects by means of combining easy hand motion with finger configurations and movements. Such approach renders a different interaction than recent systems based only on motion [3] or hand pose [4,5], which usually result in complex and difficult to memorize alphabets. For example, a *show menu* command may be performed by showing four fingers combined with a global vertical movement of the hand (as in [2]), while with the reference methods, a specific hand gesture must be assigned to the command.

We believe that exploiting hand gestures in combination with simple motions will have a much higher user acceptance, enabling more commands using an easy and small set of hand gestures. Such strategy allows a highly scalable and configurable interaction. Furthermore, this renders a more tractable hand analysis problem, as one does not necessarily need to estimate the full hand pose. However, fingertip localization must be performed, not being only a problem of detecting the *number* of fingertips in the current input image but *which* fingers (and fingertips) are visible and *where* they are located. Intra-gesture variations (*i.e.* rotation and translation) are also considered, strongly increasing the robustness of the system.

Quantitative results are obtained through evaluation with a recent 3D feature benchmark, revealing the convenience of using ORD for hand gesture classification. Fingertip localization results are compared to a state-of-the-art Random Forest approach.

Even if this demo is focused on interactivity with virtual objects, the system may be extended to a large number of applications. Gaming, creative design, control of CAD environments or musical applications are just some examples.

2 Technical Overview

We propose a novel use of the Oriented Radial Distribution (ORD) feature, presented by Suau *et al.* [1]. The 3D point cloud obtained from a Kinect sensor is our input data. The ORD feature characterizes a point cloud in such a way that its end-effectors are given an elevated ORD value, providing a high contrast between flat and extremal zones. Therefore, ORD is suitable to both globally characterize the structure of a hand gesture and to locally locate its end-effectors (generally fingers). A two-step method is proposed, namely hand gesture classification and fingertip localization, which are obtained with a single ORD calculation on GPU (see Fig. 1). The hand gesture classification step is performed using a k-Nearest Neighbors (k-NN) search on a template dataset. A graph-matching algorithm is used to infer finger locations from the fingertip annotation of the recognized gesture, taking advantage of the ORD structure of the hand under analysis. To automatically annotate the fingertip locations in the training images, we recorded several sequences using a colored glove. This procedure enables an easy extraction of the ground-truth fingertip locations during the training phase. Note that the glove is used only for annotation purposes, as in test time no glove is required.

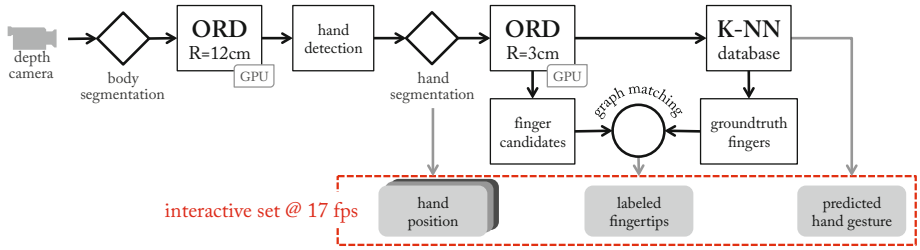

Fig. 1. Technical scheme of the *intAIRact* demo. An *interactive set* containing the last hand positions, hand gesture and fingertip locations is obtained at each frame (17 fps).

Fig. 2. Samples of the annotated dataset. We show two examples per gesture (columns), emphasizing that gestures are performed with rotations and translations, resulting in a challenging classification problem (for example, observe the variability of gesture 4). Label 0 corresponds to *no gesture* (*i.e.* other gestures, transitions). The colored glove is only used in the training phase.

Demonstration Operation. We design our system to trigger events as a function of the inferred hand gesture, the fingertip locations and hand trajectory at time t. As a result, we have a user-friendly and scalable touchless interaction, since different events can be triggered by rather subtle changes of any of the mentioned elements. As an example, in our application we define the events *Show/Hide Object Menu* as the set {Gesture 4, fingertips up, hand going up/down}, i.e., two different events are triggered by just a change of one element, the hand trajectory. However, the possibilities of this *interactive set* go beyond that; fingertip locations allow us to compute hand rotations for different gestures. Consequently, a user can trigger a high number of events by remembering 9 gestures and combining them with simple translations and rotations.

3 Quantitative Results

Besides qualitative results (see video), we provide some figures to point out the classification results evaluated against reference methods. A dataset consisting of 4 users performing 9 gestures is used (Fig. 2). Two recordings per user are provided for training purposes, each clip containing between 3000-6000 frames.

Hand Gesture Classification Results. A benchmark consisting of various 3D features (Depth, Curvature, 3DSC [6], VFH [7] and SHOT [8]) is considered in

order to evaluate the performance of ORD regarding the classification task. ORD achieves a classification F-Measure of 85.8%. The best result in the benchmark is achieved by the depth feature (67.7%) followed by VFH (49.9%). Therefore, the ORD feature largely outperforms the benchmark, also pointing that depth-based features (ORD and depth) are more convenient for analyzing depth data than 3D based features.

Classification with ORD is also evaluated with small training datasets, obtained as reduced versions of the full dataset by Euclidean clustering. The proposed method successfully tolerates drastic reductions of the training dataset, showing an F-Measure degradation of about 6% with a dataset reduction $\times 10$.

Fingertip Localization Results. To evaluate the proposed algorithm, we implement a fingertip localization method using Random Forests (RF). The RF method is based on the successful system for detecting body parts from range data proposed by Shotton et al. [9]. We use very similar depth-invariant features, but in addition to depth data, we include the ORD feature, which slightly increases the average finger localization accuracy from 58% to 60%. However, the proposed Nearest Neighbor + Graph Matching finger localization method improves the reference RF approach by 8%, achieving an accuracy of 68%.

Computational Performance. The demonstration is carried out on an Intel Core2 Duo CPU E7400 @ 2.80GHz. To calculate the ORD feature, we have coded a parallel implementation on a NVIDIA GeForce GTX 295 GPU, performing about $70 - 140\times$ faster than the implementation in [1]. The complete demonstration setup performs in real-time, at a frame-rate of about 17 fps. A frame-rate of 16 fps is achieved by [10]. However, our proposal delivers fingertip positions in addition to hand gestures.

References

1. Suau, X., Ruiz-Hidalgo, J., Casas, J.R.: Oriented Radial Distribution on Depth Data: Appication to the Detection of End-Effectors. In: ICASSP (2012)
2. Apple Inc.: Magic Trackpad (2012)
3. Suau, X., Ruiz-Hidalgo, J., Casas, J.R.: Real-Time Head and Hand Tracking based on 2.5D data. Transactions on Multimedia, 1 (2012)
4. Keskin, C., Kırac, F., Kara, Y.E., Akarun, L.: Real Time Hand Pose Estimation using Depth Sensors. In: ICCV-CDC4CV, pp. 1228–1234 (2011)
5. Minnen, D., Zafrulla, Z.: Towards robust cross-user hand tracking and shape recognition. In: ICCV-CDC4CV, Oblong Industries, Los Angeles, CA, USA (2011)
6. Frome, A., Huber, D., Kolluri, R., Bülow, T., Malik, J.: Recognizing Objects in Range Data Using Regional Point Descriptors. In: Pajdla, T., Matas, J. (eds.) ECCV 2004, Part III. LNCS, vol. 3023, pp. 224–237. Springer, Heidelberg (2004)
7. Rusu, R., Bradski, G., Thibaux, R., Hsu, J.: Fast 3D recognition and pose using the viewpoint feature histogram. In: IROS, pp. 2155–2162 (2010)

8. Tombari, F., Salti, S., Di Stefano, L.: Unique Signatures of Histograms for Local Surface Description. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 356–369. Springer, Heidelberg (2010)
9. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-Time Human Pose Recognition in Parts from Single Depth Images. In: CVPR, pp. 1297–1304 (2011)
10. Uebersax, D., Gall, J., Van den Bergh, M., Van Gool, L.: Real-time Sign Language Letter and Word Recognition from Depth Data. In: ICCV-HCI, pp. 1–8 (2011)