Real-Time Hand Gesture Recognition in a Virtual 3D Environment

David Ferstl Mathias Rüther Horst Bischof

{ ferstl, ruether, bischof }@icg.tugraz.at
Institute for Computer Graphics and Vision
Infeldgasse 16, A-8010 Graz

We present a novel technique using a range camera for real-time recognition of the hand gesture and position in 3D. Simultaneously the user's hand and head pose are tracked and used for interaction in a virtual 3D desktop environment.

As human gestures provide a natural way of communications between humans, gesture recognition is a major field of research used also for human-computer communication. Most existing hand interaction systems are restricted to a 2D touch-sensitive plane, or track and recognize the hand gesture on color images. Due to the lack of depth information, these systems are limited to the 2D space only, provide little depth invariance and are sensitive to rotation and segmentation errors, whereas we built a system for bare-hand 3D interaction, independent from hand rotation and depth.

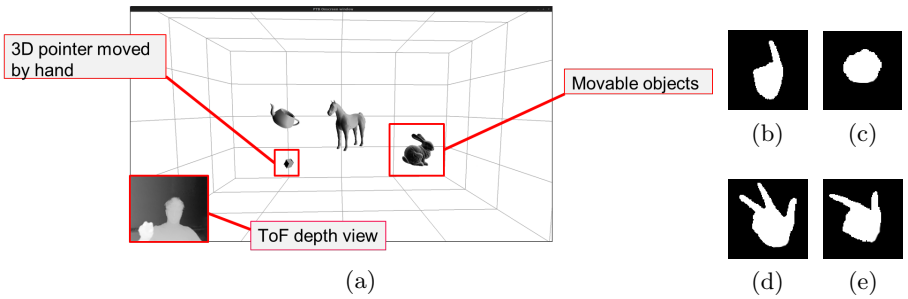

Fig. 1: Screenshot of the virtual 3D desktop environment (a). The pointer is moved by the hand in 3D and its function (moving (b), grabbing (c), rotating (d), scaling (e)) is changed according to the hand gesture

A presented gesture is recognized by a comparison of the segmented hand to a set of templates by *chamfer distance matching* (CD). In contrast to traditional CD matching [2] we reduce the score calculation to a pixel summation and normalization on the *distance transformed image* (DT). The CD score for each template T_g to a segmented hand candidate A is calculated through

$$S_g = \sum_{i=0}^{P_A} T_{DT,g}(A_{pts}(i))^2 + \frac{1}{P_{Tg}} \sum_{i=0}^{P_{Tg}} A_{DT}(T_{pts,g}(i))^2, \quad (1)$$

where P_A and P_{Tg} is the number of edge points of the candidate A respectively of the template T_g. This increases not only the calculation speed but also the recognition accuracy through a normalization by the number of edge points.

In our template vocabulary we defined 7 different gestures (e.g. Figure 1b-1e). Each gesture is represented in 25 different rotations ($-60°... + 60°$ in $5°$ steps) and 8 different postures resulting in a set of 1400 templates.

Using an off-the-shelf Time-of-Flight range camera the hand and the body can be segmented through a distance threshold independent from the environmental lighting conditions and background. In contrast to previous systems [1, 3] we use the depth information additionally to scale the segmented hand in real world metrics to become scale invariant. Hence, the system gets robust against hand pose variations in space and inaccurate segmentation. For example, when the segmentation also contains a part of the arm, the hand can still be recognized because the hand candidate and the hand templates have the same dimensions. Because the CD score between the candidate and each template is calculated separately, this becomes and ideal task for the GPU where we fully parallelized the recognition to get gesture, hand rotation and position in real-time.

We apply the gesture recognition framework in a virtual desktop environment, where the hand position in 3D is used to move inside our desktop. Simultaneously the hand pose is tracked accurately for a specific gesture. The gesture is used to switch between different interaction modes of moving, grabbing, rotating or scaling of objects. A screenshot of the desktop is shown in Figure 1a.

Similar to numerous so called "fish tank" virtual reality (VR) systems, also the head position is tracked and the perspective view of the virtual 3D environment is correctly rendered for the observer. While other fish tank VR systems require the user to wear physical markers [4, 5] our system relies on pure depth camera tracking.

To summarize, we created a real-time interaction system which is able to recognize and track the hand gesture and the 3D head / hand pose simultaneously. The recognition method incorporates rotation and scale invariance as well as a variability in gesture articulations and hand appearances to be applicable for different users, all without the need of complex trained classifiers or shape context methods. The experiments show that we achieve an average recognition accuracy of 93% at more than 15 fps on a standard consumer PC tested for 6 different datasets.

References

1. Van den Bergh, M., Van Gool, L.: Combining rgb and tof cameras for real-time 3d hand gesture interaction. In: Appl. Comp. Vis. (WACV), IEEE Workshop on. pp. 66–72 (2011)
2. Borgefors, G.: Hierarchical chamfer matching: a parametric edge matching algorithm. Pattern Anal. Mach. Intell., IEEE Trans. 10(6), 849–865 (1988)
3. Li, Z., Jarvis, R.: Real time hand gesture recognition using a range camera. In: Australasian Conf. Robot. and Automat. (ACRA). Proc. on. pp. 529–534 (2009)
4. Mulder, J., Van Liere, R.: Enhancing fish tank vr. In: Virt. Real. (VR), Proc. IEEE Int. Conf. pp. 91–98 (2000)
5. Ware, C., Arthur, K., Booth, K.S.: Fish tank virtual reality. In: Proc. of the INTERACT and CHI. Conf. on Human factors in computing systems. pp. 37–42. CHI '93, ACM, New York, NY, USA (1993)