Rule-based Hand Posture Recognition using Qualitative Finger Configurations Acquired with the Kinect

Lieven Billiet[1], Jose Oramas[1], McElory Hoffmann[1,2], Wannes Meert[1] and Laura Antanas[1]

[1] *Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium*

[2] *Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa*

Keywords: Hand Posture Recognition, 3D Data, Model-based Recognition, Rule-based Model.

Abstract: Gesture recognition systems exhibit failures when faced with large hand posture vocabularies or relatively new hand poses. One main reason is that 2D and 3D appearance-based approaches require significant amounts of training data. We address this problem by introducing a new 2D *model-based approach* to recognize hand postures. The model captures a high-level *rule-based representation of the hand* expressed in terms of finger poses and their qualitative configuration. The available 3D information is used to first segment the hand. We evaluate our approach on a Kinect dataset and report superior results while using less training data when comparing to state-of-the-art 3D SURF descriptor.

1 INTRODUCTION

A variety of consumer devices using gestures as means of communication have been released in the recent years (e.g., the Microsoft Kinect). A factor that negatively still influence the user experience while using such gesture-based devices is the gesture recognition accuracy. To become part of everyday life, these systems need to have high accuracy and to adapt quickly to large body vocabularies. In this paper we focus on hand gestures and we introduce a new model-based approach to hand posture recognition. In contrast to purely appearance-based techniques, which use no structural information about the hands and thus, require a significant amount of training data, our simple rule-based and user-defined model can reliably recognize hand postures by estimating few parameters from little training images.

A hand posture recognition system involves (1) segmenting the hand and (2) extracting the hand pose description, two challenging problems for which many methods exist (Barczak and Dadgostar, 2005). Similar to (Ren et al., 2011a; Ren et al., 2011b; Izadi et al., 2011), in this work we use the Kinect and 3D depth information to solve the first problem. However, differently from these, we do not rely on wrist belts, external media or color-markers to assist the hand segmentation step. Furthermore, we find the location of the hand independently of its posture and our approach is able to discriminate between the cases when none, one or two hands are used for gestures.

We address the second problem by employing a qualitative hand model. In contrast to well anchored appearance-based techniques using either 2D (Liwicki and Everingham, 2009; Van den Bergh and Van Gool, 2011; Pugeault and Bowden, 2011; Altun and Albayrak, 2011) or 3D (Suryanarayan et al., 2010; Darom and Keller, 2012; Knopp et al., 2010) information, our approach is *model-based* and uses the 2D image data to encode a rule-based hand description. This has two main advantages over purely appearance-based approaches: a lower computational demand and any potential loss of discriminative power due to limited amounts of training data is countered by the use of structural information of the hand. Our set of rules are based on finger poses and properties and can robustly capture the hand posture configuration. Each rule uses finger poses such as *stretched* or *closed* and qualitative relations between them. Our model has only 9 degrees of freedom, which makes it easily applicable in practice. This is an important advantage compared to non-structural methods, for which data acquisition is often a high cost. Moreover, it offers an elegant alternative to a more demanding full kinematic hand model (Erol et al., 2007), which implies a more difficult recovery of all its parameters from a single video stream.

Related work (Keskin et al., 2011) proposes hand

Billiet L., Oramas J., Hoffmann M., Meert W. and Antanas L. (2013).
Rule-based Hand Posture Recognition using Qualitative Finger Configurations Acquired with the Kinect.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 539-542
DOI: 10.5220/0004230805390542

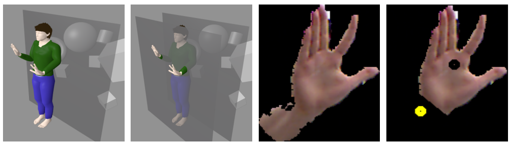

Figure 1: Hand segmentation with additional refinement. From left to right: body segmentation, original hand segmentation, hand segmentation after refinement, palm and wrist positions (black and yellow circles, respectively).

posture recognition approaches by fitting a 3D skeleton to the hand. Although semantically it follows the same direction as our work, it is still a low-level representation of the hand. Closely related are also the approaches of (Mo and Neumann, 2006; Holt et al., 2011; Ren et al., 2011a; Ren et al., 2011b). They use depth cues to build a more qualitative representation of the hand model based on hand parts and their configuration. Yet, different from these, we use depth information only for hand segmentation and employ a rule-based model to recognize hand postures.

Although hand posture recognition is a popular problem, recognition results presented in the literature are often based on self-gathered datasets which are not available. Exceptions are the recently introduced gesture recognition benchmarks (Guyon and Athitsos, 2011; Guyon et al., 2012). However, they mainly focus on hand motion or involve all body parts. Differently, our current work focuses on the hand posture recognition problem. Thus, we collected a dataset to evaluate our rule-based approach and, as a second contribution, we make this dataset publicly available at `http://people.cs.kuleuven.be/~laura.antanas/Kinectdata.zip`. We compare experimentally against an appearance-based model which employs the recently introduced 3D SURF (Knopp et al., 2010) descriptor. We show that starting from the same hand segmentations, our rule-based system achieves better results than 3D SURF. Our model-based approach demonstrates the advantage of using rules in case of few training data over more expensive 3D descriptors.

2 HAND SEGMENTATION

We use the Kinect and depth information to localize and segment the hand by detecting the point closest to the visual sensor and thresholding the depth image from this point. This is similar to (Mo and Neumann, 2006), however, we extend this work to deal with none, one, or two hands.

Our procedure includes several steps performed on the depth image: front-most body point detection, body segmentation by thresholding the depth, hand segmentation by re-thresholding the depth and finally, refined hand segmentation. The body depth threshold is estimated as the median depth of the detected objects area when an initial depth threshold from the front-most body point is considered. As potential hands have to obey certain size criteria, we consider as hands (at most) two objects lying at least T_{hands} cm before the body in the direction of the visual sensor. The thresholds were empirically established based on training data. As our experiments on new data show, the chosen thresholds allow for generalization. Figure 1 illustrates the two segmentations. Because hands may be extended quite far-away from the body, which implies that part of the arm might be considered as part of the hand, we include a segmentation re-estimation step. We use the closeness to the wrist and palm positions to successfully filter long wrists and refine the segmentation of the palm.

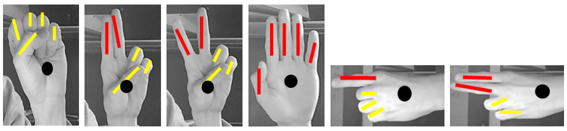

Figure 2: Hand postures and visualization of their models.

3 RULE-BASED RECOGNITION

To recognize hand postures we propose a model-based approach which we represent using rules. It is based on a fixed number of hand components. Each component is a finger group with its associated finger pose. Thus, each hand posture is a rule which captures the hand configuration. Because the rules are user-defined, training involves only the estimation of few parameters that are general finger properties and hold across all hand postures. Our hand model is inspired by (Mo and Neumann, 2006), however, differently, we consider possible finger poses as *stretched* or *closed*. Also the relations between the fingers can be either *joined* or *separated*. Figure 2 shows possible configurations of finger poses and relations between them. The black dots are markers defining the global position of the hand. Red lines represent stretched fingers, while yellow ones closed fingers.

This model can be extended, yet even in this simple form, it is able to distinguish between 512 poses without explicit training. Since the model needs only 9 degrees of freedom it is feasible to learn the parameters from a single video stream. Our representation allows the introduction of a second hand, extra finger orientations or even finger depth. This makes our ap-

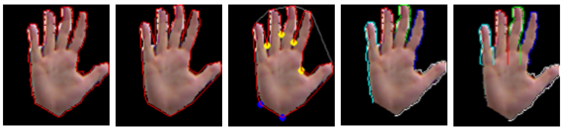

Figure 3: Phases of finger groups extraction process. From left to right: full contour, reduced contour, convexity analysis, finger groups, alignment.

proach general enough with respect to posture types.

From Hand Contour to Finger Groups. Finger groups are obtained from a convexity analysis of the hand contour as illustrated in Figure 3. In a first step, it is simplified to a polycontour. This removes many random small convexity defects whilst remaining the contour's characteristic form. Next, groups of fingers are found as parts of the contour, in-between subsequent convexity defects (marked as yellow circles). We use the width to determine the number of fingers in a group and empirically estimated thresholds on the training data to make the distinction between group sizes. Additionally, we impose that exactly 5 fingers are found across the groups. A final step aligns the bases of all fingers, except for the thumb, at the palm level. We estimate each finger group pose according to its length. Based on the lengths, we also determine if fingers are stretched or closed. Joined or separated fingers are decided based on the minimum of the distances of their joining point to the tips.

Rule-based Representation of the Hand Model. The hand model encodes the finger configuration that characterizes a specific hand posture. A different rule is associated to each hand posture category, such that the model is represented by the set of rules for the entire posture vocabulary. For example the third posture in Figure 2 is modeled using the following rule:

```
if orient=vertical,thumb=0,$f^1$=1,$f^2$=1,$f^3$=0,$f^4$=0,
($f^4$,$f^3$)=l,($f^3$,$f^2$)=l,($f^2$,$f^1$)=v,($f^1$,thumb)=l
then posture third,
```

where the two closed fingers f^3, f^4 and the closed thumb are expressed as '0' and the stretched fingers f^1, f^2, as '1'. Joined pairs of fingers (f^i,f^j) are indicated by '*l*' and the separated pair of fingers (f^2,f^1) by 'v'. This rule, except the orientation encoded as a separate rule test, is practically represented in our system using the string structure '0l1v1l0l0'. As another example, the rule for the second hand posture is encoded as '0l111l0l0'. We overcome the restriction in the original approach of (Mo and Neumann, 2006) that the palm must always face the camera by including a second global orientation. The global hand orientation is treated separately and it extends the space of possible configurations. This is encoded as an extra test in our rule-based model.

Hand Posture Recognition. Starting from the detected finger groups, we could learn the hand posture models using, for example, decision trees (Breiman et al., 1984). However, the goal of this work is to show the advantages and benefits of a rule-based system. Therefore, we assume a user-defined rule-based model and we use it to recognize hand postures by directly comparing the encoding of a newly extracted posture s_1 from a test image with the rule encoding of each posture s_2 in the vocabulary. We quantify the quality of a match as the number of characters that match. Because a finger configuration has nine characters, the similarity is a score in the interval $[0,9]$, given by the formula $9 - dist_h(s_1, s_2)$; $dist_h$ is the Hamming distance between the two strings.

The number of degrees of freedom makes our proposed model a sparse representation when the number of hand poses to be recognized is small with respect to all possible encodings. As a result, we propose also a nearest neighbor approximation, in which every hand posture reference rule encoding is replaced by its nearest neighbor encoding found in the training data. As an alternative, the sparseness problem can be solved by estimating the rules directly from the data, such that only meaningful posture models are learned.

4 EXPERIMENTS

We evaluated our approach on a real-world dataset which contains 8 different hand postures and was obtained from a Kinect device. The postures are illustrated in Table 1. The dataset was collected from 8 different persons. A first subset was used for parameter estimation and validation, in both segmentation and recognition steps; the remaining part is the test data. The training and validation set contains 1100 frames, while the test set 400 frames for all postures.

Evaluation. We evaluate the recognition performance of individual hand postures using the nearest neighbor approximation. We report results in terms of recall (R), precision (P) and accuracy (Acc). The confusion matrix obtained is shown in Table 1. Performance results are shown in Table 2.

As the confusion matrix shows, Posture 3 from left to right is often confused with Posture 4, Posture 8 with Posture 7 and Posture 6 with Posture 7, respectively. The false positive rate is explained by our chosen model which does not consider depth information. For example, Posture 3 (encoded '1v1v11111') is confused with Posture 4 (encoded '0l0l0l0l0') because of non-accurate finger length estimations. If the fingers in Posture 3 are considered folded instead of stretched, the encoding of Posture 3 becomes

Table 1: Confusion matrix for nearest neighbor approx.

		Predicted							
Correct		90%	10%	0%	0%	0%	0%	0%	0%
		4%	96%	0%	0%	0%	0%	0%	0%
		2%	4%	56%	38%	0%	0%	0%	0%
		0%	6%	0%	94%	0%	0%	0%	0%
		0%	0%	2%	0%	96%	0%	2%	0%
		0%	0%	0%	0%	0%	70%	30%	0%
		0%	0%	0%	0%	0%	2%	98%	0%
		0%	0%	0%	0%	4%	42%	2%	52%

Table 2: Performance results for nearest neighbor approx.

R	90%	96%	56%	94%	96%	70%	98%	52%
P	94%	83%	97%	71%	96%	61%	74%	100%
Acc	98%	97%	94%	95%	99%	91%	96%	94%

‘1v0v0l0l0’. This is, indeed, much closer to Posture 4. Also, the data used for training showed that estimating folded vs. stretched fingers is not completely possible using 2D information solely. This can be improved by considering also depth information.

Comparison to Related Work. We compare against recently introduced 3D SURF (Knopp et al., 2010), which has not been investigated yet for hand gesture recognition. Our aim is to show that, although depth information is essential for robust hand segmentation and may provide benefits for the posture recognition on its own, 3D descriptors, in their current state, do not pay-off for certain problems. We show that using our approach we obtain better results than using a more expensive 3D descriptor, which requires both more data and computational time to train a model. We consider a bag of words approach together with a multi-class SVM classifier, similarly as in (Knopp et al., 2010). We obtain the following results for 3D SURF: $R = 64.5\%$, $P = 71.0\%$ and $Acc = 87.88\%$, as apposed to $R = 81.5\%$, $P = 84.5\%$ and $Acc = 95.5\%$ for our rule-based approach.

REFERENCES

Altun, O. and Albayrak, S. (2011). Turkish fingerspelling recognition system using generalized hough transform, interest regions, and local descriptors. *Pattern Recognition Letters*, 32(13):1626–1632.

Barczak, A. L. C. and Dadgostar, F. (2005). Real-time hand tracking using a set of cooperative classifiers based on haar-like features. In *Research Letters in the Information and Mathematical Sciences*, pages 29–42.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth.

Darom, T. and Keller, Y. (2012). Scale-invariant features for 3-d mesh models. *IEEE Transactions on Image Processing*, 21(5):2758 –2769.

Erol, A., Bebis, G., Nicolescu, M., Boyle, R. D., and Twombly, X. (2007). Vision-based hand pose estimation: A review. *CVIU*, 108(1-2):52–73.

Guyon, I. and Athitsos, V. (2011). Demonstrations and live evaluation for the gesture recognition challenge. In *ICCV Workshops*, pages 461–462.

Guyon, I., Athitsos, V., Jangyodsuk, P., Hamner, B., and Escalante, H. J. (2012). Chalearn gesture challenge: Design and first results. In *CVPR Workshop on Gesture Recognition and Kinect Demonstration Competition*.

Holt, B., Ong, E.-J., Cooper, H., and Bowden, R. (2011). Putting the pieces together: Connected Poselets for Human Pose Estimation. In *Workshop on Consumer Depth Cameras for Computer Vision*.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. (2011). Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *ACM symposium on User interface software and technology*, UIST.

Keskin, C., Kira, F., Kara, Y. E., and Akarun, L. (2011). Real time hand pose estimation using depth sensors. In *Computational Methods for the Innovative Design of Electrical Devices*, pages 1228–1234.

Knopp, J., Prasad, M., Willems, G., Timofte, R., and Van Gool, L. (2010). Hough transform and 3d surf for robust three dimensional classification. In *ECCV*.

Liwicki, S. and Everingham, M. (2009). Automatic recognition of fingerspelled words in British sign language. In *IEEE Workshop for Human Communicative Behavior Analysis*, pages 50–57.

Mo, Z. and Neumann, U. (2006). Real-time Hand Pose Recognition Using Low-Resolution Depth Images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2(c):1499–1505.

Pugeault, N. and Bowden, R. (2011). Spelling It Out: Real–Time ASL Fingerspelling Recognition. In *Workshop on Consumer Depth Cameras for Computer Vision*.

Ren, Z., Meng, J., Yuan, J., and Zhang, Z. (2011a). Robust hand gesture recognition with kinect sensor. In *ACM*, MM, pages 759–760, New York, NY, USA. ACM.

Ren, Z., Yuan, J., and Zhang, Z. (2011b). Robust hand gesture recognition based on finger-earth mover’s distance with a commodity depth camera. In *ACM*, MM, pages 1093–1096, New York, NY, USA. ACM.

Suryanarayan, P., Subramanian, A., and Mandalapu, D. (2010). Dynamic hand pose recognition using depth data. In *ICPR*, pages 3105–3108.

Van den Bergh, M. and Van Gool, L. J. (2011). Combining rgb and tof cameras for real-time 3d hand gesture interaction. In *WACV*, pages 66–72.