Randomized Decision Forests for Static and Dynamic Hand Shape Classification

Cem Keskin, Furkan Kıraç, Yunus Emre Kara and Lale Akarun
Boğaziçi University
Computer Engineering Department, 34342, Istanbul, Turkey
keskinc@cmpe.boun.edu.tr, {kiracmus, yunus.kara, akarun}@boun.edu.tr

Abstract

This paper proposes a novel algorithm to perform hand shape classification using depth sensors, without relying on color or temporal information. Hence, the system is independent of lighting conditions and does not need a hand registration step. The proposed method uses randomized classification forests (RDF) to assign class labels to each pixel on a depth image, and the final class label is determined by voting. This method is shown to achieve 97.8% success rate on an American Sign Language (ASL) dataset consisting of 65k images collected from five subjects with a depth sensor. More experiments are conducted on a subset of the ChaLearn Gesture Dataset, consisting of a lexicon with static and dynamic hand shapes. The hands are found using motion cues and cropped using depth information, with a precision rate of 87.88% when there are multiple gestures, and 94.35% when there is a single gesture in the sample. The hand shape classification success rate is 94.74% on a small subset of nine gestures corresponding to a single lexicon. The success rate is 74.3% for the leave–one–subject–out scheme, and 67.14% when training is conducted on an external dataset consisting of the same gestures. The method runs on the CPU in real–time, and is capable of running on the GPU for further increase in speed.

1. Introduction

Hand gestures are a natural part of human interaction. They play a complementary role for speech and a primary role for sign languages. Therefore, attempts to use the hand gesture modality in human computer interaction (HCI) has intensified research efforts for hand pose tracking and gesture recognition in the last decade.

Hand gesture recognition studies have initially relied on 2D models [5]. Although pose variability and occlusion limit the success of 2D approaches, successful models relying on partial models have been defined [15]. Approaches using articulated 3D models have relied on color images [3, 1, 13, 16, 4], as well as the use of multiple cameras or time-of-flight sensors [10, 9]. These approaches have achieved good performances even in the presence of occlusions and pose changes, though their time performances have limited their application in real time HCI applications [11].

Two developments have recently accelerated implementations of HCI using human body and hand gestures: The first is the release and widespread acceptance of the Kinect depth sensor. With its ability to generate depth images in very low illumination conditions, this sensor makes the human body and hand detection and segmentation a simple task. The second development is the use of fast discriminative approaches using simple depth features coupled with GPU implementation; enabling real time human body pose extraction [14, 6].

The approaches for human body pose detection using the Kinect camera use a variety of techniques: Shotton *et al.* [14] use a large amount of labeled synthetic images to train an RDF [2] for the task or body part recognition. In a later study, Girschick *et al.* [6] use the same methodology, but let each pixel vote for joint coordinates; and learn the voting weights from data. [18] relies on pre–captured motion exemplars to estimate the body configuration as well as the semantic labels of the point cloud. [8] uses an upper body model, and tracks it using a hierarchical particle filter. Although these ideas may be extended to extracting the 3D pose of the hand, the problem is made more difficult by the increased pose variability and self-occlusion.

In our previous work, we adapted the methodology of body pose estimation used in [14] to the hand [7]. In this work, a large synthetic dataset was generated using a realistic hand model, and RDFs were trained to assign each pixel a hand part label. Finally, mean shift algorithm was used to estimate the centers of hand parts to form the hand skeleton. Additionally, SVM and ANN were used to classify extracted skeleton parameters into hand shapes. However, the

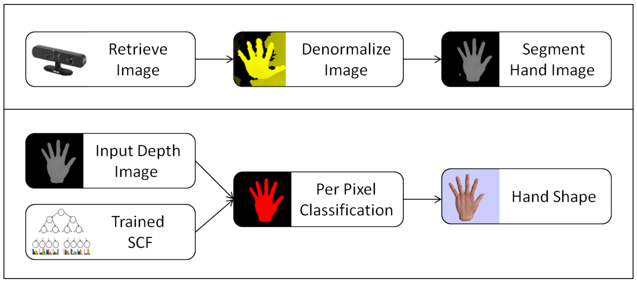


Figure 1. Flowchart for the hand shape classification process. The upper row illustrates the hand detection process. The lower row shows the shape classification process.

success rate of this approach is dependent on the accuracy of the skeleton, which can only be retrieved using synthetic images. In this work, we follow a more direct approach and alleviate the need for a synthetic model. Namely, a novel RDF that we call a Shape Classification Forest (SCF) is used to classify each depth pixel into a hand shape, instead of a hand part. We show that this classifier is significantly easier to train using real depth images, does not depend on the accuracy of hand pose estimation, and evaluates faster. The flowchart for the framework is depicted in Figure 1. The system retrieves the image, filters it for noise and compression artifacts around missing values, detects, tracks and segments the hand using motion cues, and classifies each pixel into a shape label. The final shape label is determined via voting using all the pixels.

The performance of the novel shape classification method is evaluated on the publicly available ASL letter dataset of [12] and is shown to achieve a success rate of 97.8%. Multi-user ASL letter recognition is a difficult task, and comparable good results to ours have been reported in the literature on other datasets. [17] provides a good review of ASL letter recognition on depth data. We also test the system on the ChaLearn Gesture Dataset (CGD2011). This dataset consists of several lexicons of gestures, each of which are intended for different kinds of applications, such as games, sign language, driving and dancing. For the experiments, a lexicon consisting of nine hand gestures is selected. The hands are detected using motion cues, and segmented from each video sample with a precision rate of 87.88% when there are multiple gestures, and 94.35% when there is a single gesture in the sample. Hand shape classification accuracy is measured using three different schemes. In the first case, training is conducted on half of the dataset, and the model is tested on the other half, reaching a success rate of 94.74% on a small subset of nine gestures corresponding to a single lexicon. In the second scheme, a subject is left out from the training set, and tests are done on the samples from this subject, which is called the leave–one–subject–out scheme. The success rate is 74.3% for this challenging case. To further test the generalization power of the model, we collected a dataset consisting of 4500 images by performing the same gestures in the target lexicon. While the model achieved a recognition rate of 99.5% on this simpler dataset, it scored 67.14% when tested on CGD2011. About 70% of the error is attributable to the mix-up of two similar gestures in this case.

The paper is organized as follows: In Section 2 we describe the novel hand shape classification method. Section 3 discusses the experiments conducted on the datasets used in evaluation. Finally, we conclude the paper and discuss future work in Section 4.

2. Hand Shape Classification

Hand shape classification is the act of assigning a class label c to an input image I, representing a certain configuration of the hand. The method described here uses depth and translation invariant features extracted from a depth image to infer the class label.

The approaches used to estimate body and hand pose in [14] and [7] respectively, use RDFs to assign a part label to each pixel in the input image. Inspired by these methods, we formulate an RDF for hand shape recognition, in which every pixel votes for a hand shape label instead of a hand part label. The final class label is determined by majority vote.

2.1. Decision Trees

Decision trees consist of split nodes, which are the internal nodes used to test the data, and leaf nodes, which are the terminal nodes used to infer the posterior probability of the data, based on statistics collected from past data. Each split node sends the incoming input to one of its children according to the test result. The test associated with a split node is usually of the form:

$$f_n(F_n) < T_n \tag{1}$$

where $f_n(F_n)$ is a function of a subset of features and T_n is a threshold, at split node n. The input is injected at the root node, which is forwarded by the split nodes according to the test results, and the posterior probabilities associated with the leaf node that is reached are used to infer the class label. Hence, the training of a decision tree involves determining the tests and collecting statistics from a training set. A decision tree is depicted in Figure 2.

In the case of a randomized decision tree, the features are randomly selected. Each nodes uniformly samples multiple feature parameters from a large feature space, and the corresponding test that provides the best split is chosen.

2.2. Shape Classification Forest

The input to a Shape Classification Tree (SCT) is a depth image I, and a pixel location $\boldsymbol{x}$, describing a pixel and its

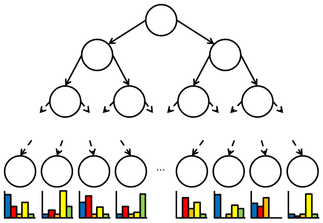


Figure 2. A Shape Classification Tree is essentially a Randomized Decision Tree with shape label histograms at the leaves.

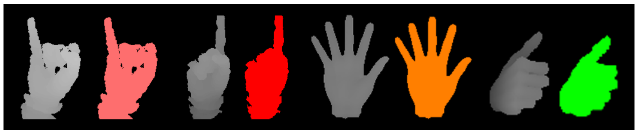


Figure 3. SCT training images: The first four images are real depth images and their labels, and the rest of the images are synthetic depth images and their labels.

local context. The output is a set of posteriors for the shape class label C_k assigned to the pixel. SCTs can be trained with both real and synthetic data, consisting of a set of depth images and their respective class labels. An example is given in Figure 3. Typically, images from the same hand shape and different angles should be assigned the same shape class label to ensure viewing angle independence, unless the shape itself it angle dependent. The features used in [7] and [14] proved to be fast and efficient, which are also used fr SCT. Given a depth image $I(\boldsymbol{x})$, where $\boldsymbol{x}$ denotes location, we define a feature $F_{\boldsymbol{u},\boldsymbol{v}}(I,\boldsymbol{x})$ as follows:

$$F_{\boldsymbol{u},\boldsymbol{v}}(I,\boldsymbol{x}) = I(\boldsymbol{x} + \frac{\boldsymbol{u}}{I(\boldsymbol{x})}) - I(\boldsymbol{x} + \frac{\boldsymbol{v}}{I(\boldsymbol{x})}) \quad (2)$$

The offsets $\boldsymbol{u}$ and $\boldsymbol{v}$ are relative to the pixel in question, and normalized according to the depth at $\boldsymbol{x}$ to ensure depth invariance. The depth of background pixels and the exterior of the image are taken to be a large constant.

Each split node is associated with a pair of offsets $\boldsymbol{u}$ and $\boldsymbol{v}$ and a depth threshold τ. The data is split into two sets as follows:

$$\begin{aligned} C_L(\boldsymbol{u},\boldsymbol{v},\tau) &= \{(I,\boldsymbol{x})|F_{\boldsymbol{u},\boldsymbol{v}}(I,\boldsymbol{x}) < \tau\} \quad (3) \\ C_R(\boldsymbol{u},\boldsymbol{v},\tau) &= \{(I,\boldsymbol{x})|F_{\boldsymbol{u},\boldsymbol{v}}(I,\boldsymbol{x}) >= \tau\} \quad (4) \end{aligned}$$

Here, C_L and C_R are the mutually exclusive sets of pixels assigned to the left and right children of the split node, respectively.

In the training phase, each split node randomly selects a set of features, partitions the data accordingly and chooses the feature that splits the data best. Each split is scored by the information gain:

$$S(\boldsymbol{u},\boldsymbol{v},\tau) = H(C) - \sum_{s\in\{L,R\}} \frac{|C_s(\boldsymbol{u},\boldsymbol{v},\tau)|}{|C|} H(C_s(\boldsymbol{u},\boldsymbol{v},\tau)) \quad (5)$$

where $H(K)$ is the Shannon entropy estimated using the normalized histogram of the labels in the sample set K. The process ends when the leaf nodes are reached. Each leaf node is then associated with the normalized histogram of the labels estimated from the pixels reaching it.

Starting at the root node of each SCT, each pixel $(I,\boldsymbol{x})$ is assigned either to the left or the right child until a leaf node is reached. There, each pixel is assigned a posterior $P(c_i|I,\boldsymbol{x})$ for each hand shape class c_i. For the final decision, the posterior probabilities estimated by all the trees in the ensemble are averaged:

$$P(c_i|I,\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} P_n(c_i|I,\boldsymbol{x}) \quad (6)$$

where N is the number of trees in the ensemble, and $P_n(c_i|I,\boldsymbol{x})$ is the posterior probability of the pixel estimated by the tree with index n. We call this ensemble a Shape Classification Forest (SCF). To determine a final hand shape label, the posterior probabilities of every pixel in the input image are averaged, and the label that maximizes this term is selected:

$$c^* = \arg\max_{c_i} \frac{1}{M}\sum_{m=1}^{M} P(c_i|I,\boldsymbol{x_m}) \quad (7)$$

where M is the number of foreground pixels in the input image, and c^* is the determined hand shape class label.

2.3. Shape Classification Pipeline

SCF is meant to assign posteriors to the pixels that are known to belong to the hand. Therefore, the hand needs to be detected and segmented before the evaluation step. Corresponding pipeline is illustrated in Figure 1.

First, the depth images are denormalized to retrieve the original depth values in the metric system. This step is required, when compressed media is used instead of a depth sensor. Compression also creates blending artifacts around the missing pixels caused by interpolation. Therefore, we dilate the depth shadows and assign them a large depth value to push them to the background. The denormalized images are then used to calculate the difference images between consecutive frames. The effect of the depth noise is cancelled by ignoring smaller differences at this step. The rest of the differences are accumulated for each image to calculate the total activity in the scene. A typical plot is given in Figure 4.

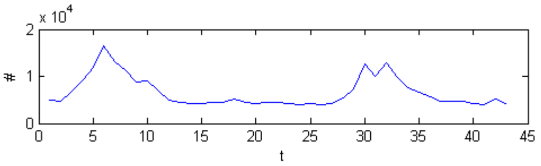


Figure 4. A typical activity plot corresponding to a gesture. The two peaks correspond to the arm movements from and to the resting position. The actual gesture is performed between the peaks in the case of static or dynamic hand shape based gestures.

The peaks in the activity plot correspond to the harsh movements of the hand, in particular to raising and lowering it. We detect the hand during the first high activity period (determined empirically by examining many samples), where it is assumed that the largest moving object is the hand. To isolate moving objects, a depth–aware connected component method is employed. When the activity is low, it is assumed that the hand is stationary, and the last known position of the hand is regarded as its current position.

A box around the hand, with each side equivalent to 40 cm calibrated to the current depth of the hand, is cropped from the depth image. Finally, background is segmented and eliminated from the cropped image by depth thresholding. Some samples are given in the first row of the Figure 6.

This method is designed to locate and track the hand when it is fast, and look for it in the last known locations when the hand stops moving. False positives have a larger impact on the recognition accuracy, as it leads to misclassification. Therefore, precision is favored over recall rate, and only images that are most likely to be the hand are retrieved.

Once the hand is segmented, each pixel is evaluated using SCF, and the final hand shape is determined by voting as described in Section 2.2.

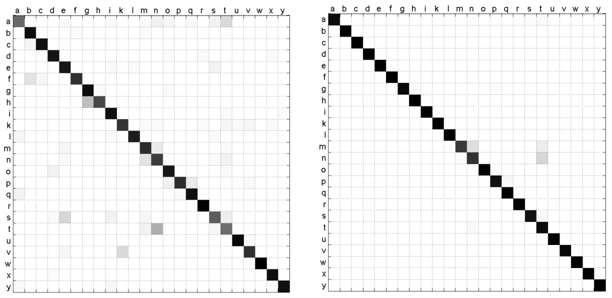
3. Experiments

Several experiments are conducted on a publicly available ASL dataset and on CGD2011 to test the accuracy and efficiency of the SCF.

3.1. ASL Dataset

The accuracy of the SCF is first tested on a dataset consisting of 65k depth and color hand images corresponding to 24 of the 26 ASL letters (omitting non–static letters j and z) performed by five subjects [12]. We disregard the color images, and further segment the hands in the depth images from their backgrounds.

Pugeault *et al.* reported their results on this dataset using both leave–one–subject–out cross–validation and by using half of the set for training and half for validation. For the former validation technique, we employed four trees of depth 20, and sampled 1000 features at each node. SCF achieved a recognition rate of 84.3%, while [12] report 47%. For the latter, an SCF consisting of a single tree reached 97.8%, compared to 69% using only depth features, and 75% using both depth and color features [12]. We provide the confusion matrices in Figure 5. Evidently, the main source of error is the similarity of the poses for the letters M, N and T in ASL. Disregarding such similar hand shapes or using depth sensors with higher resolution can further increase the success rate. Table 1 lists these values.



(a) (b)

Figure 5. Confusion matrix for the ASL letter classification task using SCF on the Pugeault dataset with 24 letters and five subjects [12]. a) Leave–one–subject–out with a success rate of 84.3%. b) Half training–half validation, with a success rate of 97.8%. The main source of error is the similarity of the poses for the letters M, N and T in ASL.

	Half vs. Half	Leave–one–out
Pugeault Depth	69%	N/A
Pugeault Color + Depth	75%	47%
SCF	97.8%	84.3%

Table 1. Comparison of the proposed method with the approach of Pugeault *et al.* [12].

3.2. ChaLearn Gesture Dataset

The system is also tested on an unnamed lexicon of the public CGD2011 dataset, which consists of static or dynamic hand shapes only. In the case of dynamic hand gestures, each image is independently labeled, ignoring temporal information. We cannot provide comparisons with other work, as this dataset was released recently.

Unlike the former ASL dataset, CGD2011 does not provide extracted hand images, as the dataset contains body gestures as well. The depth image streams are compressed in the form of AVI files. Hence, a denormalization step is necessary, and we retrieve the cropped hand images as explained in Section 2.3.

Hand detection precision is 94.35% for the samples that contain a single instance of a static or dynamic hand shape.

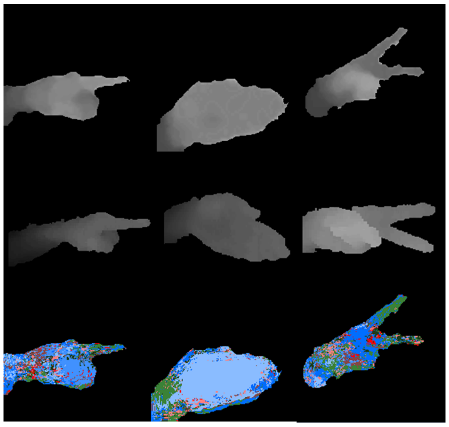


Figure 6. Pixel classification results. The first row corresponds to the depth images of the hand as extracted from videos in CGD2011. The second row consists of samples from the external dataset collected by performing the same gestures. Spatial and depth resolution is better for these images. The third row gives the per pixel classification results. Each pixel is colored according to the shape label that corresponds to the mode of their posterior probabilities.

For the samples that contain one or more gestures, in which the performer returns to the resting position between each gesture, the precision falls to 87.88%. Even though the recall rate is lower, the system manages to retrieve multiple relevant frames from each gesture sequence, thereby enabling correct classification.

Three types of tests are conducted to test the accuracy and generalization power of the system. In the first type of test, an SCF is trained on half of the samples in the dataset, and tested on the other half. The system achieves a success rate of 94.74% on a small subset of nine gestures corresponding to a single lexicon in CGD2011. In the second type of test, the models are trained on the dataset samples by leaving one subject out. Then, the model is tested on the samples from the subject that is left out. This scheme is considered to be more challenging, as it requires generalization to subjects that do not exist in the dataset. The success rate is 74.3% in this case, using five subjects. As expected, the score is lower than that of the similar scheme on the ASL dataset, which is estimated to be 84.3%. The main reason is the quality difference of both datasets. Unlike CGD2011, the ASL dataset provides higher spatial and depth resolution, and the hand images are cropped manually.

In the final type of test, a new dataset is collected by performing the same gestures as in the lexicon. We collected a total of 4500 images from a single subject, with 500 images for each of the nine hand shapes. Some samples are given in the second row of Figure 6. The model is trained on this external dataset, and is tested on CGD2011. This scheme is more challenging than the leave–one–out scheme, because now the model needs to generalize from a single subject, and the quality and resolution of the data differs from that of CGD2011. SCF achieved a recognition rate of 99.5% on this simpler dataset, when half of it is used for training and half of it is used for testing. Next, an SCF trained on the entire dataset is tested on CGD2011. The model achieved a score of 67.14% for this challenging scheme. These results are given in Table 2. The three columns correspond to the three testing schemes. Per pixel classification results for this case are given in the third row of Figure 6. The confusion matrix is depicted in the Figure 7. Evidently, most of the error is caused by two hand shapes that are very similar. Disregarding these shapes further increases the recognition rate to 92.54%.

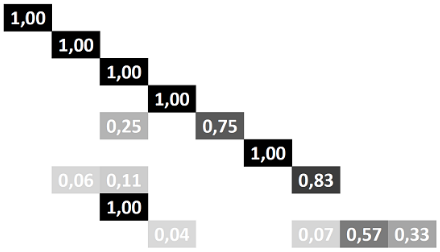


Figure 7. Confusion matrix for the third type of experiment. The SCF is trained using the external dataset collected, which is then tested on CGD2011 samples. The overall success rate is 67.14%. However, most of the error is caused by the last two gestures, that are extremely similar. The success rate on the rest of the gestures reaches 92.54%.

Half vs. Half	Leave–one–out	External Training
94.74%	74.3%	67.14%

Table 2. Success rates of different types of experiments.

4. Discussions and Conclusion

In this work, SCF is proposed as an effective solution to hand shape classification, which is demonstrated on a large ASL dataset and CGD2011. In comparison to recent work on the same dataset [12], SCF performs significantly better: 97.8% versus 69% when using half of the trainings set for training and the rest for validation, and 84.3% versus 47% in the case of leave–one–subject–out.

For CGD2011, we introduced a method that automatically finds and segments the hands from each video sample with a precision rate of 87.88% when there are multiple gestures, and 94.35% when there is a single gesture in the sample. We conducted several different types of experiments. The hand shape classification success rate is estimated to be 94.74% on a single lexicon of CGD2011. The success rate is 74.3% for the leave–one–subject–out scheme using four subjects for training and one subject for testing. We also collected a new dataset consisting of 4500 images by performing the same gestures in the lexicon. We trained the system on this dataset and tested on CGD2011, and achieved a score of 67.14%. These results show that SCFs are especially accurate if the training set contains samples similar to the one being tested. Also, SCFs have promising generalization capabilities, as demonstrated by the fact that they can effectively generalize to unseen subjects.

SCF runs at 30 frames per second on the CPU in real–time, and is capable of running on the GPU for further increase in speed. Furthermore, in contrast to the skeleton classification method introduced in [7], SCF can be trained using real depth images and require a smaller training set.

References

[1] V. Athitsos and S. Sclaroff. Estimating 3D hand pose from a cluttered image. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages II–432–9, 2003. 1

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

[3] T. de Campos and D. Murray. Regression-based Hand Pose Estimation from Multiple Cameras. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, pages 782–789, 2006. 1

[4] M. de La Gorce, D. J. Fleet, and N. Paragios. Model-Based 3D Hand Pose Estimation from Monocular Video. *IEEE transactions on pattern analysis and machine intelligence*, pages 1–14, Feb. 2011. 1

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

[6] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient Regression of General-Activity Human Poses from Depth Images. In *in Proceedings Thirteenth IEEE International Conference on Computer Vision. ICCV 2011*, volume 2011, pages 415–422. IEEE Comput. Soc, 2011. 1

[7] C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun. Real-time hand pose estimation using depth sensors. In *Proceedings Thirteenth IEEE International Conference on Computer Vision Workshops. ICCV 2011*, pages 1228–1234. IEEE Comput. Soc, 2011. 1, 2, 3, 6

[8] A. Lopez-Mendez, M. Alcoverro, M. Pardas, and J. R. Casas. Real-time upper body tracking with online initialization using a range sensor. In *in Proceedings Thirteenth IEEE International Conference on Computer Vision Workshops. ICCV 2011*, volume 2011, pages 391–398. IEEE Comput. Soc, 2011. 1

[9] S. Malassiotis and M. Strintzis. Real-time hand posture recognition using range data. *Image and Vision Computing*, 26(7):1027–1037, July 2008. 1

[10] Z. Mo and U. Neumann. Real-time Hand Pose Recognition Using Low-Resolution Depth Images. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, pages 1499–1505, 2006. 1

[11] I. Oikonomidis, N. Kyriazis, and A. Argyros. Markerless and efficient 26-DOF hand pose recovery. In *Proceedings of the 10th Asian conference on Computer vision-Volume Part III*, pages 744–757. Springer, 2011. 1

[12] N. Pugeault and R. Bowden. Spelling It Out: Real Time ASL Fingerspelling Recognition. In *in Proceedings of the 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision, in conjunction with ICCV 2011*, volume 2011. IEEE Comput. Soc, 2011. 2, 4, 5

[13] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff. 3D hand pose reconstruction using specialized mappings. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2000, pages 378–385. IEEE Comput. Soc, 2001. 1

[14] 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 single depth images. In *CVPR*, 2011. 1, 2, 3

[15] V. K. Singh and R. Nevatia. Action recognition in cluttered dynamic scenes using pose-specific part models. In *in Proceedings Thirteenth IEEE International Conference on Computer Vision. ICCV 2011*, volume 2011, pages 113–120. IEEE Comput. Soc, 2011. 1

[16] B. Stenger, P. Mendonça, and R. Cipolla. Model-based 3D tracking of an articulated hand. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages II–310–II–315. IEEE Comput. Soc, 2001. 1

[17] D. Uebersax, J. Gall, M. Van den Bergh, and L. Van Gool. Real-time sign language letter and word recognition from depth data. In *in Proceedings Thirteenth IEEE International Conference on Computer Vision. ICCV 2011*, volume 2011, pages 383–390. IEEE Comput. Soc, 2011. 2

[18] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys. Accurate 3D pose estimation from a single depth image. In *in Proceedings Thirteenth IEEE International Conference on Computer Vision. ICCV 2011*, volume 2011, pages 731–738. IEEE Comput. Soc, 2011. 1