Markerless Hand Pose Recognition Through Shape Distributions

Fabio Luiz Marinho de Oliveira, Helena de Almeida Maia,
Liliane Rodrigues de Almeida, Rodrigo Luis de Souza da Silva
Computer Science Department
Universidade Federal de Juiz de Fora
Email: fabio@ice.ufjf.br, helena.maia@ice.ufjf.br, liliane.rodrigues@ice.ufjf.br, rodrigoluis@ice.ufjf.br

***Abstract*—In Human-Machine Interface literature there are many works aiming to recognize hand gestures, since they provide a more natural way to interact with virtual environments and to transmit information. Some of the visual-based approaches about it extracts features from simple cameras, found in most multimedia devices. Instead of building 3D models, we propose a method for hand pose recognition by computing the shape distribution over a 2D image from a single viewpoint. A database was built with hand images of the 24 static gestures of the American Sign Language (ASL).**

***Keywords*—ASL, hand recognition, fingerspelling, shape distribution.**

I. Introduction

As the popularity and quality of digital image acquisition devices rise, we can see an increasing number of applications exploiting non-conventional interactions with these devices. Such is the case with hand pose estimation using a single camera. The problem of hand pose estimation is relevant in applications on human-machine interface, such as robot learning by demonstration [1], sign language recognition [2], and animation synthesis [3]. The hand pose recognition is still a challenge, specially with a vision-based approach, considering that self occlusions, background clutter, and variety in hand types are very common.

This paper presents a markerless hand gesture recognition method based on a monocular camera, by measuring the shape similarity. The hand pictures of the database have a correspondent shape distribution to be compared with the incoming frame. For evaluation purposes, we take 5 very distinct hand poses and also discuss a case where 24 poses from the American Sign Language (ASL) are used.

II. Related Works

There are several approaches on literature to recognize hand gestures, summarized on the survey [4]. They can be distinguished by three main aspects: hand model, hand detection and gesture description and classification.

To model the hand, some works search for the joints in order to build a hand skeleton, so it is easier to reconstruct a 3D model [5] [6]. These approaches are generally able to recognize dynamic gestures by considering the temporal component [7]. Image-based approaches often do not try to rebuild the hand skeleton. Instead, they recognize sign language using datasets with static gestures [2] [8] [9]. In this work, we propose a image-based model and a new dataset with several variations such as different actors, positions and illumination.

Most static approaches are based on depth-images, due to the challenge of detecting a hand in RGB images with a cluttered background. For instance, Pugeault and Bowden [2] use OpenNI+NITE framework with Microsoft Kinect to detect the hand. To solve this problem in RGB images, Wang and Wang [9] propose a SIFT-based detector. During a training stage, SIFT features are extracted in a controlled environment, so the hand is detected without background noise. This way, during the test stage, the system is able to recognize the hand in a real environment. Our work is based on RGB images from a monocular camera. For the sake of simplicity, we use a dark background for hand images in both training and test stages, keeping focused on describing the hand shape.

In [2], the hand is described by the response of a bank of Gabor filters averaged across overlapping Gaussian basis functions. The gesture is classified using a multi-class random forest. In [9], Adaboost is combined with SIFT descriptors of hand images to represent a gesture as a category. We propose a way of describing the hand pose through shape distribution, a method presented in [10]. They reduced the problem of shape matching to one-dimensional function comparison, originally for 3D models. They sampled points over the mesh, applied a function to measure the angle, the area, or the distance between them, and built a distribution of these measurements. These distributions provide a signature for the 3D model that can be used in the classification process. Building a 3D model for hand from images is a complex task and computationally expensive. Moreover, regular monocular cameras are way more common and accessible than 3D scanners, depth cameras or even monocular camera arrangements that are capable of producing 3D models. So, in this work, we adapt their idea for 2D images and evaluate it specifically in the hand pose recognition scenario.

III. Proposed Method

Our proposed method for hand pose recognition consists of, for each hand image, computing a called *shape distribution* that can be used as a feature vector for further classification. This process is detailed in the following subsections.

A. Shape Distribution

The shape distribution computation is composed of 3 main steps. The first one is the segmentation of the hand. The second

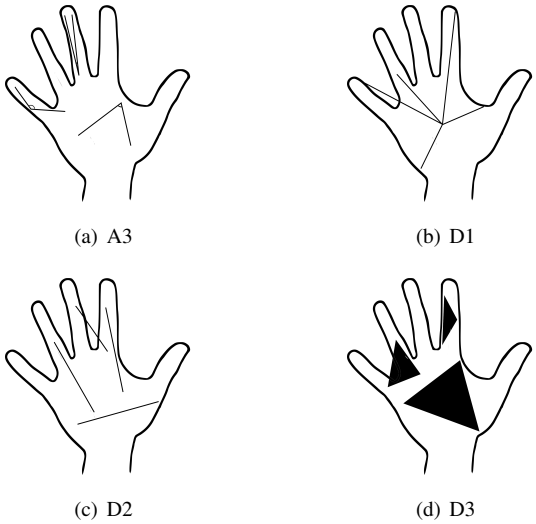
one is the sampling of the segmented hand. The third one is building a piecewise linear function, or the so called shape distribution.

The segmentation is necessary to obtain which pixel coordinates are to be considered inside the hand. To avoid the common challenges related to segmenting objects in images, we use a dark, plain background for the images in the dataset. This way, a simple greyscale thresholding procedure is enough to segment the hand from the background.

The sampling is done by selecting random points from the segmented hand and taking a measurement between these points. The measures, and how we refer to them, similarly to the ones presented in the original paper by [10], are:

- Angle between 3 random points (A3);
- Distance between the center of the image and a random point (D1);
- Distance between 2 random points (D2);
- Square root of the area between 3 random points (D3);

Figure 1 shows simple examples of measurements taken.



(a) A3 (b) D1

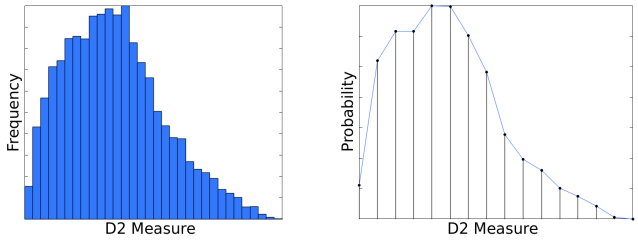
(c) D2 (d) D3

Fig. 1: Measures used to obtain samples. Endpoints of black lines are the randomly selected points from the sampling procedure. Note that line segments between sample points don't have to be entirely within the hand.

One important aspect of these measures is that they are all invariant to rotation, that is, regardless of the orientation of the hand in the image, the measurement values do not change between the points and their rotated counterparts. The number of samples collected from the images is denoted as S and it is a parameter of the method, which is further explored in Section IV.

After sampling the image, the samples are accumulated into a histogram, with bins as uniform subdivisions of the interval between 0 and the maximum measurement value obtained, both inclusive. For each sample, the measure taken increases the value of its pertaining bin. The number of bins, denoted as B, is also a parameter of the method.

From the histogram, we build a piecewise linear function, with V vertices, such that $V \leq B$ and each vertex is equally spaced. The value of the function on each vertex is equal to the the value of the corresponding bin in the histogram. Function values for points other than the vertices are calculated via a linear interpolation of the values of the two nearest vertices. Figure 2 shows an example histogram and its corresponding function.



(a) Histogram. (b) Distribution.

Fig. 2: An example histogram and distribution of D2 measures for the letter C sign. (a) The histogram has 32 bins where the horizontal axis represents the measure values and vertical axis the number of occurrences of such measure values binned into each interval. (b) Its corresponding piecewise linear function has 16 vertices and provides an interpolation of values between vertices.

Once a function is built, it undergoes a normalization process that makes the complete interval coincide with $[0, 1]$ and the area below the function become equal to 1, thus making it a probability density function (PDF) over the interval $[0, 1]$. Being f and g two image distributions, we use the following metrics to compare them:

- χ^2: $D(f,g) = \int \frac{(f-g)^2}{f+g}$
- Bhattacharyya: $D(f,g) = 1 - \int \sqrt{fg}$
- Minkowski L_1: $D(f,g) = \int |f-g|$
- Minkowski L_2: $D(f,g) = \left(\int |f-g|^2\right)^{1/2}$
- Minkowski L_∞: $D(f,g) = max\{|f-g|\}$,

The choice of metrics follows the original paper by Osada [10]. χ^2 and Bhattacharyya are used to estimate the overlap of two statistical distributions. The Minkowski metrics are commonly used to compare feature vectors (descriptors). As it is possible to see, we have an array of parameters to balance between quality and execution speed. It is desirable that the whole process achieves real-time computing capability, if it is to be used as a human interface tool. And that also takes into account the classifying process, which is subject of the next subsection.

B. *Classifier*

For the classification, we build a nearest neighbour scheme. From the input distributions, we obtain the centroids of the letter classes/clusters. The centroid of a class is an "artificial" distribution, in the sense that its vertices are the mean value of the corresponding vertices in all of the distributions of that class. To determine the class of a new input distribution d,

we evaluate the distance from d to each one of the centroids, and the one with the least distance, represents the class where d belongs. The distance function can be any of the aforementioned metrics.

C. Database

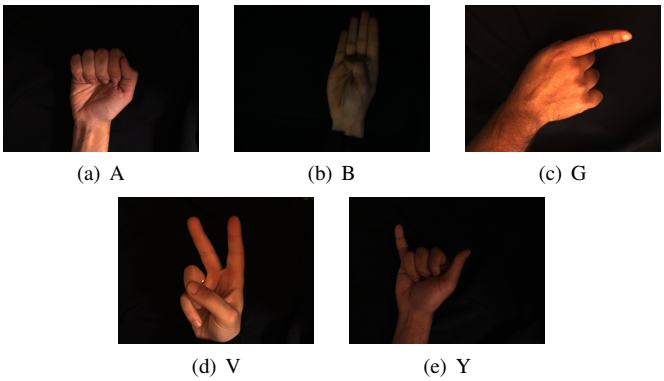
We have constructed a database of 701 images of hands posing as letters from the American Sign Language (ASL), excluding the letters *j* and *z*, which require movement to be properly spelled. Figure 3 shows some of the images from the database. The 24 different signs were performed by 7 people with slight variations on pose, camera angle, and illumination conditions. Additionally, 8 more people performed a restricted set of 5 letter signs which resemble common day-to-day gestures, like showing a palm meaning "stop", or "halt", pointing with the index finger, and so on. These signs represent letters A, B, G, V and Y, and are highly distinctive. Figure 4 shows the selected 5 signs for additional experiments. This restricted set is composed of 270 out of the 701 images, and also has more than one image for each pose, producing the same variations as the full set.

Fig. 3: Example images from the database with different person and varying illumination and orientation of the hands.

For all the images in the database, the distributions are precomputed, in order to speed up the recognition process and avoid recomputation upon every execution. The classifiers are also precomputed for the same reason. This way, the pose recognition achieves real-time computation speed.

IV. Experiments

We conduct two main experiments. The first one uses the 24 signs and 7 people from the database, as described previously. Out of the 551 images, 414 (5 people) were used for classifier calibration/training, and the remaining 137 (2 people) for testing. The second one uses the subset of 5 signs (258 images) and 15 people in total, as also previously described. For this experiment, 10 people (180 images) are used for training, and 5 for testing (78 images). As for the parameters exploration, Table I shows all the parameter values with which we experiment.



(a) A (b) B (c) G (d) V (e) Y

Fig. 4: Example images from the subset of 5 signs of the database.

TABLE I: Parameter Exploration

Parameter	Values
Samples	512 1024 2048 4096 8192 65536
Bins	128 256 512 1024
Vertices	16 32 64 128
Measures	A3 D1 D2 D3

Our best results achieved for the ASL and restricted set experiments are shown in Tables II and III, respectively.

The low accuracy values can be mainly attributed to the high number of classes and very high inter-class similarity. Not only some signs are very similar, but during segmentation and the construction of the shape distribution, some distinctive features might be lost. The area between three random points seems to be the most distinctive metric in this case.

TABLE II: Best results for full 24 letter signs from ASL experiment

Samples	Bins	Vertices	Measure	Distance	Accuracy
65536	128	16	D3	Bhatt	0.2594
8192	256	128	D3	L_1	0.2540
65536	128	128	D3	L_2	0.2540
65536	128	64	D3	L_1	0.2486
8192	128	64	D3	L_1	0.2324

TABLE III: Best results for restricted set of 5 signs experiment

Samples	Bins	Vertices	Measure	Distance	Accuracy
8192	1024	128	D3	L_1	0.6666
8192	512	128	D3	L_1	0.6410
8192	1024	128	D3	Bhatt	0.6282
65536	128	64	D3	$L_{\inf}$	0.6282
8192	512	16	D3	L_2	0.6153

As these results suggest, the reduction in the number of classes contributes to an increase in accuracy, since we have a sparser, and possibly linearly separable, disposition of the distributions in the classification space. Another factor that might be responsible for the higher accuracy rates is the fact that all the signs chosen for this experiment have a very distinct silhouette. More often than not, the segmentation of two similar hand signs yields the exact same binary image, imposing a barrier for the method to distinguish between these signs. This phenomenon can be presumed as not occurring in this second experiment, and the accuracy rate can be more reliably correlated with other aspects of the process. Once again, the area between three random points is the most distinctive metric.

On top of these main experiments, we also group some of the distributions as to have more than one representation of the same image in the classification process. This way, it is expected that the centroids are a more reliable representative of the class, since outliers have a lesser impact on the centroid coordinates. Contrary to our expectations, the results are consistently worse than the ones shown previously, not going much above 20%. This could happen because the enlarged clusters might have bigger intersections, leading to a non-linearly separable classification problem, since all classes are "forcefully" labelled *a priori*.

V. Application

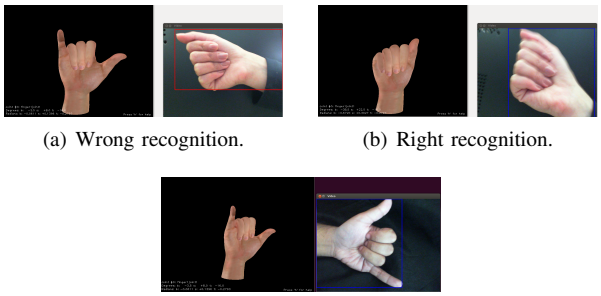
We also propose an application for the hand recognition method as a human-computer interface, whether for entertainment or teaching purposes. We use LibHand to provide a 3D hand model, which mimics the hand pose detected by the classifier. As with the database, all the hand poses are previously known, so the model is not able to reproduce poses other than the ones in the database. Figure 5 shows some poses and views of the 3D hand model.

(a) Letter C sign. (b) Letter W sign. (c) Letter Y sign.

Fig. 5: Example poses reproduced by LibHand's 3D model.

Note that, not all frames in the video need to be classified. Some frames represent transitional or repeated gestures. To reduce this computational effort, we select key frames by the movement intensity. This way, only the first frame after a sequence of transitional frames will be classified. In Fig. 6, the frame in which the classification occurs, has a blue bounding box around the hand. Other frames have a red bounding box around the hand, indicating transition between gestures or no movement at all since the last classified frame. The pose of the 3D hand model doesn't change between classifications.

As we've shown in Section IV, the system failed in some recognitions. However, since it is a interative system, a slight change in position can improve the result.



(a) Wrong recognition. (b) Right recognition.

(c) Right recognition.

Fig. 6: Example of detected poses. On the left side, the LibHand 3D models. On the right side, the input camera feed. (a) shows an attempt to recognize the letter A gone wrong. (b) Producing a slight variation in hand pose, the system is able to recognize the sign. (c) Another succesful recognition, this time for the letter Y.

VI. Conclusion and Future Works

We propose an adaptation of the work by [10] for a 2D application of hand pose recognition. We also employ a classification scheme for the obtained shape distributions. Unsatisfactory accuracy rates and weak discrimination of distributions raises the question whether the adaptation of the 3D case is appropriate.

Further exploration of the method is still required in order to ascertain about its validity. This could be done by exploring the recognition of different objects, not only hands, given that even in the work by Osada [10] objects within the same category are often confused. Another improvement could be made by employing a SVM classifier, in order to drop the assumption of linearly separable distributions. As it stands, the descriptor's invariance to illumination and rotation could prove itself useful in other object recognition scenarios.

Future works may include other improvements on object detection/segmentation and use of user feedback in the interative system. To overcome the limitation regarding the letters J and Z, it would be necessary to extend our approach to sequences of images, since these letters are spelled with movement.

References

[1] J. Romero, H. Kjellström, and D. Kragic, "Monocular real-time 3d articulated hand pose estimation." *9th IEEE-RAS International Conference on Humanoid Robots*, 2009.

[2] N. Pugeault and R. Bowden, "Spelling it out: Real-time asl fingerspelling recognition." in *ICCV Workshops*. IEEE, 2011, pp. 1114–1119.

[3] S. Jörg, J. Hodgins, and A. Safonova, "Data-driven finger motion synthesis for gesturing characters," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 189:1–189:7, Nov. 2012. [Online]. Available: http://doi.acm.org/10.1145/2366145.2366208

[4] S. Mitra and T. Acharya, "Gesture recognition: A survey," *Trans. Sys. Man Cyber Part C*, vol. 37, no. 3, pp. 311–324, May 2007. [Online]. Available: http://dx.doi.org/10.1109/TSMCC.2007.893280

[5] S. U. Lee and I. Cohen, "3d hand reconstruction from a monocular view," *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, 2004.

[6] A. State, F. Coleca, E. Barth, and T. Martinetz, "Hand tracking with an extended self-organizing map." in *WSOM*, ser. Advances in Intelligent Systems and Computing, P. A. Estévez, J. C. Príncipe, and P. Zegers, Eds., vol. 198. Springer, 2012, pp. 115–124.

[7] H. Liang, J. Yuan, D. Thalmann, and Z. Zhang, "Model-based hand pose estimation via spatial-temporal hand parsing and 3d fingertip localization," *Vis. Comput.*, vol. 29, no. 6-8, pp. 837–848, Jun. 2013. [Online]. Available: http://dx.doi.org/10.1007/s00371-013-0822-4

[8] N. Dardas, Q. Chen, N. D. Georganas, and E. M. Petriu, "Hand gesture recognition using bag-of-features and multi-class support vector machine," in *Haptic Audio-Visual Environments and Games (HAVE), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1–5.

[9] C.-C. Wang and K.-C. Wang, "Hand posture recognition using adaboost with sift for human robot interaction," in *Recent progress in robotics: viable robotic service to human*. Springer, 2008, pp. 317–329.

[10] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," *ACM Transactions on Graphics*, vol. 21, pp. 807–832, 2002.