Comparing pose recognition algorithms and introducing a new approach

Gabriel Peixoto de Carvalho
Center for Mathematics
Computing and Cognition
Federal University of ABC
Santo André – SP - Brazil
Email: gabriel.carvalho@ufabc.edu.br

Fernando Teubl Ferreira
Center for Mathematics
Computing and Cognition
Federal University of ABC
Santo André – SP - Brazil
Email: fernando.teubl@ufabc.edu.br

André Luiz Brandão
Center for Mathematics
Computing and Cognition
Federal University of ABC
Santo André – SP - Brazil
Email: brandao@daad-alumni.de

Abstract—**Gesture-based interaction with 3D objects is commonly done using color and depth (RGB-D) information. This kind of scheme requires specialized software and hardware (camera and depth sensor) which increases both financial and computational cost. Most of the public does not have access to such kind of technology due to these increased costs. Our ongoing study will be an approach for interaction with 3D objects with color (RGB) information and hand poses recognition. The interaction target is those made by one person with a single webcam. In this study, we intend to evaluate the use of commonly available algorithms to implement a pose acquisition scheme and present a comparative study of classification algorithms to achieve a low-cost and real-time interaction. Preliminary results demonstrate satisfactory performance with the hand pose acquisition scheme and classification algorithms, pointing to a real-time interaction capability.**

I. Introduction

Since 2010, with the launch of the Kinect, interaction with games and virtual environments became feasible for consumer devices. The Kinect sensor is composed of two main parts: a depth sensor and a RGB camera [1]. The depth information made body gesture recognition more precise than with color information only. On the other hand, depth information requires specialized hardware sensors and algorithms which are not commonly available, and add an extra computational cost for interaction with hand poses. The devices that work with color and depth information (RGB-D) are in general more expensive than webcams. Thus, we propose an approach for 3D interaction with hand poses images obtained from common webcams, and with a lower overall monetary cost. To achieve such goal, this study presents an approach for a hand tracking scheme and a comparative study of hand pose classification methods.

Pisharady and Saerbeck [2] present a methodical and systematic review of the last 15 years in gesture recognition research. The authors introduce most popular algorithms, techniques and databases used in the reviewed works. This information was essential to identify which are the successful techniques and current challenges faced in the field.

Lowe [3] published in 1999, an algorithm for detecting and describe local features in images, which was reported to be invariant to scale, affine transformations and rotation. This algorithm is named by its characteristics Scale-Invariant Feature Transform (SIFT). Bay et al. [4] published a similar algorithm to SIFT, which extract image descriptor features, this algorithm is similar to SIFT but claims to be faster. This algorithm is named Speeded up robust features (SURF). These two algorithms have been extensively used in studies of the Computer Vision area. Dalal and Triggs [5] present another algorithm to extract image descriptor features, which is used for object detection and outputs interesting information of the object contour and shape.

Gosh and Ari [6] present a comparative study of features sets (LCS and block features) with Support Vector Machine (SVM) classifier. Badi [7] also present a comparative study of feature sets (contour and complex moments) with artificial neural network classification algorithm. Bastos et al's [8] study also served as an inspiration to our present study. Bastos et al present also a comparative study of feature sets (HOG and Zernike moments) with neural network classification. They used images of the Brazilian sign language to evaluate the performance of the feature sets.

To evaluate our approach, we present a feasibility study of a hand pose capture, tracking and segmenting scheme using the Camshift algorithm to track and histogram back-projection to segment the hand pose. We also present a comparative study of classification techniques for hand pose recognition. SURF, SIFT, and HOG are compared for the performance as hand pose image feature descriptors. Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP) algorithms are compared by the pose classification performance. Thus, this approach is divided into two main parts: (1) pose acquisition and (2) pose classification.

In the pose acquisition part, the Camshift algorithm is adopted for its speed and accuracy in tracking the desired object. Segmentation is done by the histogram back-projection, since we are dealing with skin color. In the pose classification part, the feature extraction algorithms were adopted based on the reports of the survey Pisharady and Saerbeck [2]. SIFT and SURF feature descriptors were chosen based on these algorithms robustness for scale, rotation and affine transformation. HOG algorithm was chosen because it outputs information of contour and shape, which we believe is interesting for hand

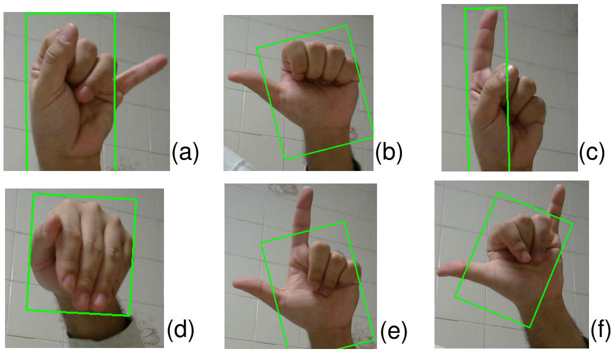

Fig. 1. Proposed system class samples: rotate left (a), rotate right (b), rotate up (c), rotate down (d), zoom in (e) and zoom out (f).

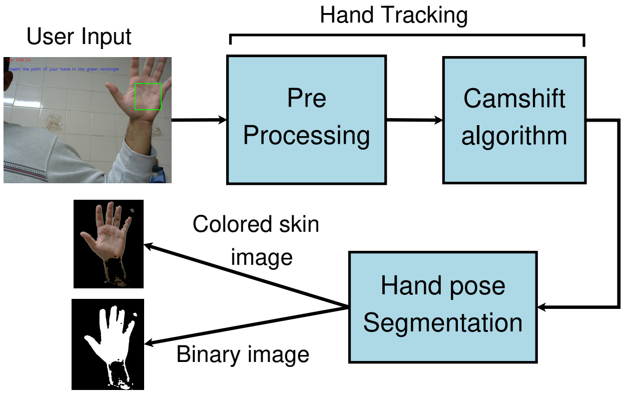

Fig. 2. Hand Pose acquisition system block diagram

pose description. The classification algorithms were chosen based on the reports Gosh and Ari [6], Badi [7] and Bastos et al's [8]. SVM and MLP were chosen by their performance with multi-class problems and robustness for no linearity in the problems we are presenting, hand pose image classification.

Our contribution is a low computational and monetary cost approach for interacting with 3D objects with hand pose. With our approach, we intend to simplify the access of 3D interaction without the demand of expensive tools. This study can be used to visualize any 3D model dynamically and with low cost. For instance, it also can be applied in museums to visualize virtual sculptures or in hospitals for 3D visualization.

This work is divided in five sections. In Section II we detail our approach for a hand pose interaction with 3D objects. In Section III we describe how the experiments were conducted to access our approach. In Section IV we present some of the results we obtained in our experiments. In Section V we describe a discussion and possible improvements that can be experimented with our approach.

II. The proposed approach

As mentioned in Section I, we divided our proposed approach into two parts, pose acquisition, and pose classification. We define 6 basic pose commands for this approach: rotate left, rotate right, rotate up, rotate down, zoom in and zoom out. Figure 1 presents a sample of each mentioned command. In the next Subsections, we will explore in details both parts of our approach. Subsection II-A will present the hand pose acquisition scheme and the Subsection II-B will present the hand pose classification part of our approach, which presents a comparative study of classification algorithms and feature sets.

A. Hand Pose acquisition

In out proposed scheme, the user triggers the hand pose input. Figure 2 shows the stages of the hand pose acquisition. The user must position the hand in the defined location. The location is represented by a Green square in the input of the block diagram (see Figure 2). Then the captured image is fed to the tracking block which tracks the hand by the skin color space.

The Camshift [9] algorithm needs some pre-processing stages to obtain the color distribution of the desired tracking region. The first pre-processing stage in our approach is to convert our image to HSV color space for a better illumination invariance. Next, we compute the color histogram of the region. We use a 2-D histogram (the normal procedure involves calculating 1-D histogram) of the H and S components (channels) of HSV for better tracking accuracy. In the last pre-processing stage our approach calculates the back-projection of the histogram which creates an image with the probability information of each pixel that could belong to our region of interest (RoI). In our case, the RoI is positioned on the hand skin color. In Figure 2, we can observe a summary of the described hand tracking process.

The Camshift algorithm receives the back-projection information and then it outputs position, size, and the most probable location of the tracked object. Camshift is an interactive algorithm, thus its accuracy is proportional on the times it is executed.

The hand region is segmented from the background using the 2-D histogram back-projection information. The back-projection information can be used to calculate the probability of each pixel being the pixel of the desired color region. This information is used to generate a map and with thresholding operation, the hand region can be segmented from the background. Figure 2 shows the segmented skin and binary images as an output of the scheme.

All the algorithms used in this scheme were chosen based on their accuracy, computational cost and time to implement. We want to use algorithms which have optimized implementations because this will reduce the developing and testing time. More robust algorithms may have a better performance but add an additional layer of complexity in the development which is not desired, at this time, for our approach.

B. Hand Pose Classification

Hand pose classification part of our approach is a hand pose image classification scheme. This part consists of feature extraction and classification algorithm. Figure 3 presents a summary of the steps in this scheme.

For feature sets, we use SIFT and SURF for their robust capacity of describing objects and HOG for its shape description capability. However, SIFT and SURF algorithms do not output same length vectors for each image and this is a problem since SVM and MLP only accept fixed size inputs. Thus, to normalize SURF and SIFT feature vectors we used the Bag of Words method (BoW) [10].

The BoW method treats the image as a document, and the image features are the "words". The first stage consists in extracting the features and then these features are converted to codewords. The set of all features compose the codebook.

To convert the features to codewords, the k-means [11] algorithm to cluster similar features with different sizes, then the number of codewords in the codebook is the number of clusters in the k-means algorithm. At the end, each image can be represented by a histogram of codewords. Figure 3 presents the BoW block for the SIFT and SURF feature sets.

The HOG feature always outputs the same number of features for every image, thus there is no need to normalize the feature vector, thus, it does not enter in the BoW block.

III. Experiments

To evaluate the feasibility of our approach we use OpenCV to implement the hand pose acquisition system and feature extraction, and Scikit-Learn to implement the classification algorithms.

To evaluate the classification of hand poses we use the dataset from Marcel and Bernier [12]. This dataset of images is composed of 4872 training images divided into 6 classes and 372 more complex testing images. Figure 4 shows samples of classes from this dataset.

We divide the 4872 samples in the dataset in 3639 samples for training and 1233 samples for testing, which are randomly selected. We also used the 372 samples from the complex dataset to test the algorithms.

The image quality and the number of classes in this dataset match our case study. This dataset also has complex backgrounds, scale and rotation variations in hand poses which are interesting characteristics to evaluate the classification algorithm and thus the feasibility of our approach. In order to save time in collecting a dataset and evaluate the classification algorithms, and give the similarities of our proposed scheme, this dataset is sufficient for our comparative study.

To evaluate the performance of the classification algorithms, we explore several parameters for the dataset images and for the algorithms. For the images, we explore parameters such as, image dimensions and color space. For the algorithms we explore parameters such as, C and Gamma for SVM and hidden layer size and number of neurons and neuron activation function for the MLP.

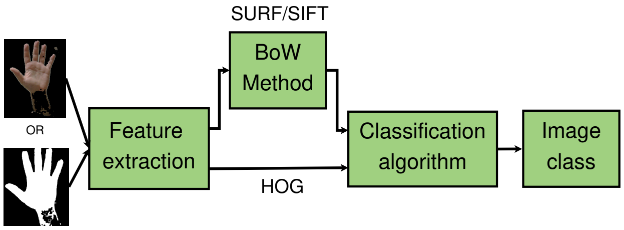

Fig. 3. Hand pose classification system block diagram

The metrics that we used to evaluate our approach were established by Pisharady and Saerbeck [2] in their survey work. For the gesture acquisition part we use metrics such as, background and illumination invariance, and user independence. For the classification algorithms we use metric of accuracy and inference time.

We do not combine two features in this evaluation, only one type of feature is extracted and used for training and testing. Combining two features would add another unnecessary layer of complexity because to make an ensemble of features first we need to evaluate them individually.

The tests were conducted using a 720p with 30 fps capture webcam (Logitech C270) connected to a computer (Thinkpad T430) with a core i5 3320M CPU, 8 GB of RAM and Linux operating system.

IV. Preliminary results

The hand pose acquisition scheme performed satisfactorily in different illumination condition and backgrounds. Each frame was processed in 0.01s giving us a theoretical 92 fps, which is sufficient for a real-time execution.

The Camshift algorithm does not output a fixed size region, its size varies from one frame to another. To output a fixed size pose image, which is best for the feature extraction and classification parts, we used the points given by Camshift algorithm to draw a rectangle bounding them. The image dimensions were obtained empirically. Figure 2 shows the results obtained by this process as an output of the scheme.

For the classification of hand poses, we obtained the best results with HOG feature sets with 97% of accuracy for both MLP and SVM algorithms. The tests with the complex set achieved an accuracy of 79% and 81.4% for SVM and MLP respectively. Regarding the MLP parameters, we experimented with one, two and three hidden layers of variable size and the activation function.

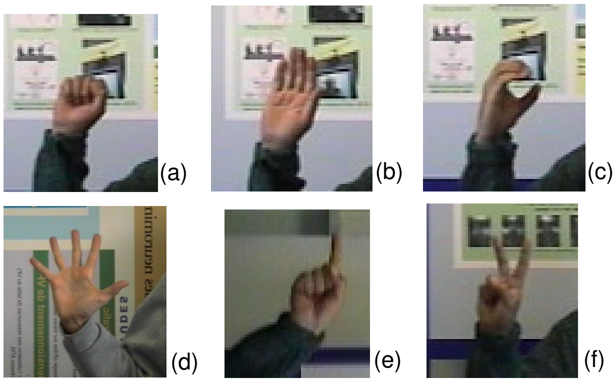

Fig. 4. Sample of the gesture classes in the Sebastian Marcel Hand Posture Database: A(a), B(b), C(c), Five(d), Point(e), V(f)

The best results were obtained using one hidden layer with 200 neurons and hyperbolic tangent activation function. The SVM parameters were 1000 for C and 0.001 for Gamma with radial basis function kernel. The inference time of the MLP is approximately 10 times faster than the SVM inference time.

The Tables I and II presents the confusion tables for the MLP and SVM, respectively. Confusion tables are a good tool to evaluate the classification models, because they show information regarding which class is more difficult for the algorithm to correctly classify.

In Table I shows that the most difficult classes to classify were C, V, and Point. For the SVM classifier on Table II, the most confusing classes were also C, V, and Point. The confusion tables show the performance of both classifiers is similar, however, testing with the complex set show that MLP has a better generalization capability.

TABLE I
MLP CONFUSION TABLE

Actual \ Predicted	A	B	C	Five	Point	V
A	118	0	0	0	0	5
B	0	147	0	0	0	0
C	0	0	119	4	1	2
Five	1	0	2	160	0	2
Point	1	0	2	3	295	5
V	0	3	0	2	2	359

TABLE II
SVM CONFUSION TABLE

Actual \ Predicted	A	B	C	Five	Point	V
A	119	0	0	1	0	3
B	0	147	0	0	0	0
C	0	0	117	4	2	3
Five	0	0	2	159	0	4
Point	0	0	2	3	298	3
V	0	0	1	1	3	361

V. DISCUSSION AND FUTURE WORKS

The use of SURF and SIFT feature sets add one more layer of complexity for the approach, however, there is no performance gain to justify this increased complexity. HOG features are simpler and faster than SIFT and SURF to process and extract. SIFT and SURF features perform very good in object recognition, however for our case study they do not perform satisfactorily. This is due to the fact that, hand pose is essentially the same object with slight difference in shape.

The hand pose acquisition system is simple but effective in different environmental conditions, such as illumination and background. The inclusion of a classification technique such as MLP will not affect the overall real-time execution of the algorithm because of its short inference time.

We intend to collect or own hand pose dataset and integrate both parts in one system, based on the results reported in this study. We also intend to evaluate more feature sets, like Hu [13] invariant moments and feature set ensemble (combining two features).

We can improve the hand detection using the object detection framework proposed by Viola and Jones [14]. With a proper detection scheme we will remove the necessity of user input to start the algorithm. For better tracking we can explore the use of SURF and SIFT features to track the hand throughout the frames. With a more robust hand pose acquisition and classification schemes, we can expand the number of classes to incorporate more commands to this approach.

ACKNOWLEDGMENT

This research has been supported by CAPES (Brazilian Coordination for Improvement of Higher Level Personnel).

REFERENCES

[1] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 1297–1304.

[2] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Computer Vision and Image Understanding*, vol. 141, pp. 152 – 165, 2015, pose amp; Gesture.

[3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ser. ICCV '99. IEEE Computer Society, 1999, pp. 1150–.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008, similarity Matching in Computer Vision and Multimedia.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, ser. CVPR '05. IEEE Computer Society, 2005, pp. 886–893.

[6] D. K. Ghosh and S. Ari, "Static hand gesture recognition using mixture of features and svm classifier," in *2015 Fifth International Conference on Communication Systems and Network Technologies*, April 2015, pp. 1094–1099.

[7] H. Badi, "Recent methods in vision-based hand gesture recognition," *International Journal of Data Science and Analytics*, vol. 1, no. 2, pp. 77–87, 2016.

[8] I. L. O. Bastos, M. F. Angelo, and A. C. Loula, "Recognition of static gestures applied to brazilian sign language (libras)," Conference on Graphics, Patterns and Images, 28. (SIBGRAPI). IEEE Computer Societys Conference Publishing Services, 2015.

[9] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision (WACV'98)*, ser. WACV '98. IEEE Computer Society, 1998, pp. 214–.

[10] G. Qiu, "Indexing chromatic and achromatic patterns for content-based colour image retrieval," *Pattern Recognition*, vol. 35, no. 8, pp. 1675 – 1686, 2002, colour Imaging.

[11] H. Steinhaus, "Sur la division des corp materiels en parties," *Bull. Acad. Polon. Sci*, vol. 1, pp. 801–804, 1956.

[12] S. Marcel and O. Bernier, "Hand posture recognition in a body-face centered space," in *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, ser. GW '99. Springer-Verlag, 1999, pp. 97–100.

[13] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, February 1962.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I–511–I–518 vol.1.