Gesture-based Object Recognition using Histograms of Guiding Strokes

Amir Sadeghipour[1]
sadeghipour@uni-bielefeld.de
Louis-Philippe Morency[2]
morency@ict.usc.edu
Stefan Kopp[1]
skopp@techfak.uni-bielefeld.de

[1] Cognitive Interaction Technology - Center of Excellence,
Bielefeld University,
Bielefeld, Germany

[2] Institute for Creative Technologies,
University of Southern California,
Los Angeles, CA, USA

Humans perform iconic gestures to refer to entities through embodying their shapes. For instance, people often gesture the outline of an object (e.g. a circle for a ball) when referring to it during communication. In this paper, we present a new gesture descriptor, called Histograms of Guiding Strokes (HoGS), well-suited for automatic recognition of iconic gesture depicting 3D objects.

Iconic Gesture Dataset We created a dataset, called 3D Iconic Gesture dataset (in short, 3DIG [1]), which contains 29 subjects (20 males and 9 females) performing iconic gestures to refer to 20 different 3D objects. Using MS Kinect™, we captured 1739 gesture performances (∼87 gestures per object), each in three formats: color video, depth video and motion of the tracked skeleton (as 3D positions of 20 joints in 30 fps).

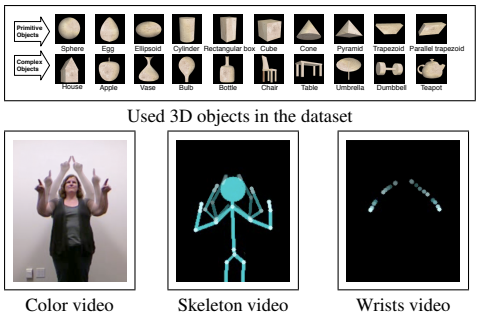


Figure 1: *top:* 3D object in the dataset, *bottom:* sample captured videos.

Analyzing the recorded gesture performances shows that people use very different techniques when performing gestures. Thus, the main challenge addressed by our HoGS descriptor is to tolerate the variations among the gesture performances, while enabling robust discriminative classification. This means that the HoGS needs to be invariant to the intra-class variabilities, while it should still represent the features which best discriminate between different classes. In the following, the mostly observed intra-class variations while analyzing the color videos of the 3DIG dataset:

- The wrists' movement direction and velocity
- The size of the used spatial space for a gesture performance
- Degree of simplification: Ignoring/considering the objects' details
- Repetition: Repeating some parts of a gestural wrist movement
- Position: The location of the used spatial space for a gesture.
- The ordering of referred components of an object.
- Handedness: The contributed hand(s).

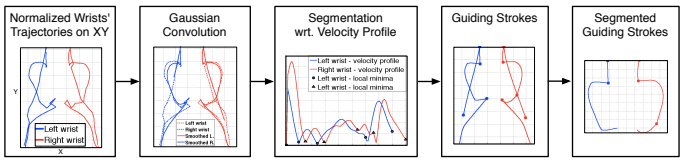


Figure 2: The processing pipeline to extract segmented guiding strokes.

Extracting Guiding Strokes To cope with the intra-class and inter-class variations, we employ the idea of decomposing the movement trajectories into *guiding strokes* [1] with respect to their spatial and kinematic features. For this aim, first we smooth the normalized 3D movement trajectories of both wrists. Then, as illustrated in Figure 2 the trajectories are split based on their velocity profiles. Finally, we segment the guiding strokes, by extracting the ones which contain the semantic content of a gesture and not the preparing and retraction movement parts. This is done automatically, through thresholding with respect to the spatiotemporal features of each guiding stroke (i.e. velocity, position and ordering). As a result, a gesture is represented as a set of guiding strokes. Each guiding stroke is a segment of the trajectory which can be parametrized. Such a parameterized representation of gestures makes it possible to extract gesture features from the attributes that reduces the intra-class variation (see Table 1).

[1]The dataset is available online at http://projects.ict.usc.edu/3dig

Computing HoGS We propose Histograms of Guiding Strokes (HoGS) as descriptors for iconic gesture recognition. To this end, we compute the histograms of the relevant attributes of the guiding strokes (see Table 1). This is done for the guiding strokes laying on each projection plane (*xy*, *xz* or *yz*) separately.

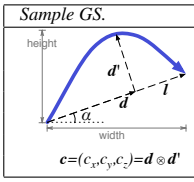
Sample GS.	*Attribue*	*Domain of values*
	Length (normalized)	$(0,1]$
	Width (normalized)	$(0,1]$
	Height (normalized)	$(0,1]$
	Curvature $= \lvert d\rvert/\lvert l\rvert$	$[0,1]$
	Curvature side $\equiv sign(c_z)$	$\{L,R\}$
	Skewness $\propto \lvert d'\rvert/\lvert l\rvert$	$[-0.5,0.5]$
	Orientation $= \alpha$	$[0,\pi)$

Table 1: The extracted attributes of guiding strokes for HoGS.

Results For the classification of iconic gestures in 3DIG dataset, we applied 1-to-1 Support Vector Machines (SVM) using HoGS as descriptor. In order to validate our approach, we designed different experiments to test and compare it with respect to three aspects. First, we compared the performance of the proposed HoGS descriptor to the descriptors applied in sketch recognition applications. Second, the performance of the SVM-based classification is evaluated against common gesture recognition approaches working with instantaneous features. Third, in an online study we asked people to recognize the captured gesture based on different video types (see Figure 1, bottom row). We applied the resulted human judgment performance as a scoring ground-truth for this task.

Classifier	*Descriptor*	*F_1 of all gestures*	*F_1 of drawing gestures*
SVM	**HoGS**	**0.61**	**0.77**
SVM	pen gesture features [2]	0.53	0.57
Bag of Trees	HoGS	0.48	0.53
NN+DTW	Instantaneous	0.44	0.40
HMM	Instantaneous	0.33	0.44
Humans	**Color video**	**0.74**	**0.76**
Humans	Skeleton video	0.38	0.44
Humans	Wrists video	0.34	0.43

Table 2: The classification results of 20 objects (i.e. chance level is 0.05)

Our approach (SVM with HoGS features) outperforms the other alternatives and compares favorably to human judgment performance. Instantaneous features, which are used in common gesture recognition approaches, do not achieve any accurate result in this dataset. The gestures performed through drawing technique are better classified in all conditions. This difference can be observed even in humans' judgment, yet very slightly when watching color videos.

[1] Stefan Kopp and Ipke Wachsmuth. Synthesizing multimodal utterances for conversational agents: Research articles. *Comput. Animat. Virtual Worlds*, 15(1):39–52, 2004.

[2] D.J.M. Willems and R. Niels. Definitions for features used in online pen gesture recognition. Technical report, NICI, Radboud University Nijmegen, 2008.