Learning Sketch-based 3D Modelling from user's sketching gestures

Hugo Lopez-Tovar
University College London

Gabriel J. Brostow
University College London

ABSTRACT

To infer three-dimensional models from two-dimensional sketches, most of the existent research focuses on image similarity, requiring a minimum of drawing skills, which is often beyond regular user's capability. This paper proposes to learn the mapping from the user's sketch into the three dimensional model by considering the sketch as a set of gestures containing information that denotes the user's own style of sketching. The system learns from the specific user and provides a personalised inference for future sketches. A study is conducted with four representative users of different skills to consider the diversity on sketching styles over a set of three 3D primitive figures. The implemented system is validated by three representative users, demonstrating that the system learns the users' style independently of their skills.

Author Keywords

Sketch Recognition; Intelligent User Interfaces; User Adaptation; pen-based interfaces.

ACM Classification Keywords

H.5.2. User Interfaces: Interaction Styles

INTRODUCTION

Inferring a three-dimensional (3D) model from a two-dimensional (2D) sketch is a complicated task, given that there is no linear relation between them as the depth information from the three-dimensional figure has been lost during the projection to a plane.

Sketch-based modelling software is typically made with heuristics and established rules to use it. Artists must learn and obey such rules which may be difficult to follow without certain level of artistic skills. Moreover, some of these solutions move away from the mapping between 2D sketches and 3D models, providing tools to allow rotation and scaling of the sketch in the three dimensions. Our problem of interest focuses only in 2D sketches as the source, similar to those created with pen and paper.

Considering the information expressed in a sketch, we must take in mind that they reveal users' conception and do not portray reality. More over, the information expressed differs from novel and experts users [15]. Our assumption is that the specific information that users include in their sketches can be used to infer their own sketching style, denoting the sketched 3D model including its rotation and scaling, without considering visual similarity between the 2D sketch and the 3D rendered model.

Learning from the user's sketching gestures instead of considering the visual similarity between the sketch and the desired drawing is a neglected alternative, adaptive and personalised to any artistic skill. This method does not claim to generate better models than image similarity approaches, but suggests instead to acknowledge the difference on users' skills and to consider adapting to them. In this sense, our approach is not intended for a specific group of users (e.g. artists, designers) but for everyone. Moreover, the aim is not to produce complex models but to study on the learning of mapping between 2D sketches and 3D models.

For practical reasons, this paper limits the figures set to three 3D primitives: cuboid, cylinder and spheroid, providing the following advantages:

- It is expected that any user has visual experience with these figures in a variety of rotations and scaling poses.
- These three primitives cover a big range of characteristics from a wider set of figures by including both straight and curved lines and faces.
- These three primitives provide both visual similarities and differences between them. For example, cuboids and cylinders share some characteristics, such as right angles and straight lines, while spheroids and cylinders share curve lines. In the case of cuboids and spheroids, there are differences such as the first having right angles and straight lines, while the second has only curved lines.

HUMAN VISUAL PERCEPTION ON 2D AND 3D

For human sight sense, the difference between two-dimensional and three-dimensional images is given by the depth perception, which provides us with the ability as observers to discriminate the distance of objects and identify the three-dimensional shape of surfaces [5].

Stereo vision and movement through a scene give us a sense of depth which during our development is confirmed by touch sense. These experiences provide us with information which we turn into assumptions when we are unable to perform confirmation such as touching or moving around in order to analyse the three-dimensional environment exhaustively. This means that human visual perception compensates its limitations with experience. Something not seen before is fuzzy and we can do nothing but infer what is the most probably reality of what we see [2]. This is the case of 3D shapes that are

Sketch: Pen and Touch Recognition Workshop. IUI'14, Feb 24-27 2014, Haifa, Israel

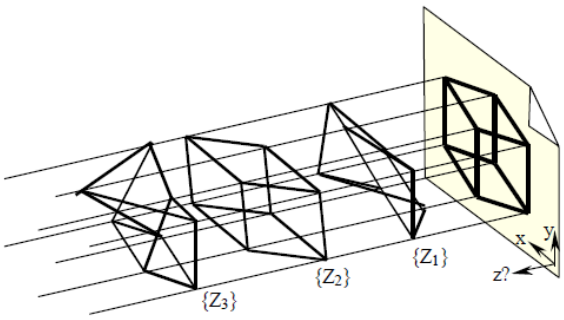


Figure 1: Thee-dimensional figures projecting into the same two-dimensional image. Figure from Lipson et al [8].

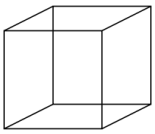


Figure 2: Necker cube, representing ambiguous projection of more than one cube pose.

hidden to sight in static images: how is the back of a figure that we can only see from the front?

Todd studied how observers were capable of perceiving metric structures in 3D shapes [13], obtaining as result constant failures and discrepancies increased as the viewing distance and orientation were varied. This means that even when we have experienced a simple primitive figure such as a cuboid, we can be confused when it is rotated, scaled and translated to a strange pose which we are not used to. More over, even when we see a known two-dimensional image as a projection of a three-dimensional shape, we tend to assume the most basic possible pose of this figure. The reality is that the loss of information during the projection, causes that multiple three-dimensional shapes share the same two-dimensional projection [8] as shown in figure 1. However, even when our brain tries to infer the more basic 3D model, we can be confused when there is more than one 3D object and pose to map to this projection, like the well known Necker Cube [Fig 2].

RELATED WORK

Sketch based interfaces provide a natural method to interact between the user and the computer based in strokes performed directly on a touch screen or a tablet, emulating the process performed by pencil and paper. Popular uses of this interface are to sketch two-dimensional drawings and three-dimensional models, known as Sketch Based Interface Modelling (SBIM). In the case of the three-dimensional modelling, the more natural but challenging method is to map from two-dimensional sketches into three-dimensional models.

There are complex tools which allow the user to explicitly provide much information to the system to obtain the required model, obtaining the best final results but requiring high artistic skills from the user. On the other hand, there are systems that provide a more simplistic interface to the user, expecting less input and inferring what the most probable model being drawn would be, sometimes obtaining fuzzy results, but being useful to a wider set of users. Frequently, these simplistic systems either provide tools to correct the results or expect the corrections to be made in an external tool.

Two interesting surveys [10] [3], include the use, research, and the most representative examples on this topic.

Sketch Acquisition

The modelling process starts with the sketch acquisition, usually as a set of two-dimensional points *(x,y)*, often including a value as the number of point *(x,y,n)* to define the order on how the sketch was drawn. This can be improved by replacing the number of point *n* by the time *(x,y,t)* to allow velocity analysis. Depending on the hardware capabilities, more advanced approaches include touch pressure and pen angle. The obtained raw data can be resampled and smoothed to overcome data capture constrains given by the used hardware (e.g. capture rate), then fitted to remove small errors and normalised to finally process the interpretation.

Resampling and Smoothing

Depending on the system, in case that the velocity of sketching is not required, it is recommended to resample to smooth the sketch strokes. This action ensures that a line is formed by equidistant points, which smooths the line. This means some closer points are removed and other points are inserted by interpolation to fill large gaps.

Fitting

To have a set of points representing a sketch could be considered as generic data with no much information. This set of points is usually fitted into a set of lines, process known as line segmentation. Depending on the needs, the segmentation may be limited only to straight lines [18] or include arcs and curve lines [17].

Feature Computation

Although there exist powerful feature methods based on image similarity, such as patch descriptors [19], some of them optimised for image matching and 3D reconstruction [16], these methods are measuring how similar the user sketch is to the final image, whether it is a reconstructed image or pulled from a database, which at the end is actually depending on the artistic skills of the user.

Long et al [9] tried to compare similitude between pen gestures instead of the sketch visualization using numerical calculated features (e.g. curviness, angles and density). However the study is limited to compare with human perception in the visual similitude field, and limited to simple strokes and not yet 2D geometric figures or 3D models.

A short study performed by Pastel et al [11] intended to investigate the information contained in simple gestures under the constrain of limited computation power as in PDA devices. They based their idea in basic information on a single slash, comparing it to a vector, as commonly used in mathematics fields, including in a single line both magnitude and direction. Although this idea seems correct, it is very difficult for

not high skilled users to perfectly draw a line in the desired direction and size.

Sketch Interpretation

Interpretation of the captured sketch can be done by mapping it to an existent model (i.e. image retrieval) or by reconstructing the model.

Image retrieval

Also known as Evocative Systems, these systems act as search engines for 3D models, accepting sketches as input. There are two main types: Iconic and Template retrieval systems.

The iconic systems [7], use evocative gestures, to define 3D primitive figures. For example, three linear strokes meeting at a point are interpreted as a box. The set of evocative gestures needs to be learned by the user, but once done, they are easy to use. Unfortunately the set of obtained figures is limited, in general to a mapping of one gesture to one model in an arbitrary pose.

Template retrieval systems [4] allow more complex figures, used as templates stored into a database. The sketch is interpreted as a set of objects and the more similar models are pulled from database. Their principal advantage is that the resulting models are complex, easily obtained, but with limited scope based on the content of the model database.

Model reconstruction

Freehand systems [1] have possibly, the most striking modelling results because they admit any kind of stroke from the user without limitation, interpreting smooth line drawings as 3D contours. However, they have two important disadvantages: the drawing skills required are very high and the use of these applications usually requires the user to learn a complicated set of gestures and commands in order to let the application know how the user wants to interpret or manipulate the given strokes.

3D Scaffolding and Sculpting is similar to analogue art. It intents to obtain complex figures based on 3D modification of initial simple 3D models, either by removing or adding volume to the figure [12, 6].

The related work here presented either makes use of image similarity methods, or provide tools for sketching in 3D rather than 2D, therefore not learning the mapping of 2D sketches to 3D models.

USER STUDY

A study has been carried out to have a sense of their different sketching styles for the three 3D primitive figures. A group of four representative users was selected, in a range of age from 14 to 35 years old. Two of them had not previous artistic preparation, one had amateur drawing experience and one with academic artistic background. A Lenovo x200 tablet with digitizer pen was used for the experiment.

Users were presented with 30 models in random order and asked to sketch each figure at a time by using three different markers: one for figurative lines, another for shadow lines and one more for geometric helpers [Fig. 3]. The use of markers for shadow and geometric helpers was optional.

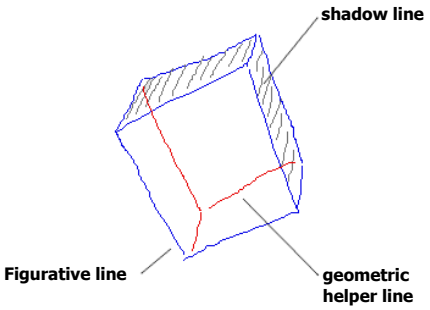


Figure 3: Thee line types for sketching during preliminary study.

User Study Results

From observation over the collected data, the following important facts were identified:

- The real size of the figure is not always important but the relative scale between each dimension is.
- It is not important how bright or dark the figures are, but how brighter or darker the faces in a single figure are in relation to the other faces.
- The same user may not be consistent on the use of shadows.
- The geometrical helper line was only used by one user. Therefore, it was decided not to include it in the experiment.

METHODOLOGY

Figure 4 shows the overall process of training, starting with a displayed model, which is sketched by the user, then the sketch is automatically segmented into lines and features are computed to train the system.

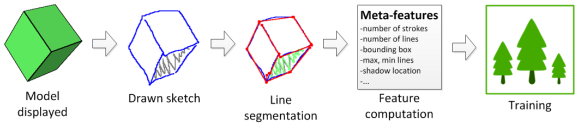


Figure 4: Overall training process.

Data Capture

The system is trained by asking the user to sketch the displayed 3D rendered model on screen, generated randomly, choosing from the three available primitives, with aleatory values from 0 to 1 (as a normalisation of 0-359 degrees) for each of the three dimensional rotations. In the same way, the three dimensional scaling factors were chosen from a random value between 0.5 and 1. The models were rendered as solid bodies with diffuse illumination to help the user realize its volume. Figures are displayed in green colour, which differs from the markers to sketch (blue and grey) to avoid user tendency to sketch color similarity.

Data structures

Sketches, Strokes and Points. The user's *sketch* is the two-dimensional drawing representing the asked figure and is conformed of one or more *strokes*. A stroke is the set of lines drawn without removing the pen from the screen. The stroke can be of any size and form, it can include both straight lines and curve lines, as well as corners and arcs. Each stroke is composed of a set of *points* in the two-dimensional plane.

Line Segmentation

Due to the fact that this project is not relaying on visual similarity, having the training data as a set of captured points by themselves does not say much. Instead, the points are grouped and connected as a set of lines which provide more valuable information on the gestures describing the user's sketching style for each primitive and its pose. This procedure is divided in two main steps: resampling and corner finding. The implemented segmentation algorithm is ShortStraw (aka iStraw) [17], a simple and effective corner finder for polylines. This algorithm is easy to implement and very fast, which allows it to be integrated into the system in order to provide immediate response to user sketches, permitting to keep a fluid training session for the user.

Feature Computation

Similar to [11], this project intents to demonstrate the intrinsic information contained in simplistic statistical features extracted from the segmented lines. At the same time, on the constrain of having a real time running application, the feature extraction must be kept simple to reduce computation time. Table 1 lists the computed features for each sketch. The first eight features are calculated from the detected lines on strokes drawn with the figurative marker, while the features starting with *Shdw_* in their name, are computed from detected lines on strokes done with the shadow marker. The name is self explanatory for most of them: there are features that count the number of lines, number of strokes, the total length of the segmented lines from the sketch, the width and height of the box bounding the sketch, the maximum, minimum and mean length of the lines. In the case of the shadow, there are four special computations: *Shdw_North*, *Shdw_South*, *Shdw_East* and *Shdw_West*, which contain binary values whether there is shadow in each of the four areas in the sketch (north, south, east and west). Once the user has provided all the requested sketches, the features are normalised to values from 0 to 1 considering all the training data.

Training

Although the nature of this research is to train the system for each user, the Random Decision Forests (RDF) include parameters to be set. In the aim of identifying those that maximise the accuracy and sensitivity while reducing the false positive rate, training is done over a data set of 120 sketches captured from a single user and cross-validated by Leaving One Out methodology to find the parameters for the Random Forests: *NTrees* and *R*.

The training process is set in two stages. First, primitive classification is performed by a RDF, and second, rotation and scaling regression is done by a set of six RDFs, i.e. one for each degree of freedom: rotation on X, Y and Z, and scaling on X, Y and Z.

For primitive classifications the *NTrees* parameter is set to 60, while *R* is set to 0.6. For rotation and scaling regression, the parameters are listed in table 2.

EVALUATION

To assess the results for classification, given as a discrete class label (Cuboid, Cylinder or Spheroid), Accuracy, Sensitivity and False Positive Rate (FPR) have been calculated. In the case of inferring the rotation and scaling degrees describing the pose of the 3D primitive, regression is measured by calculating the mean square error (MSE).

For Primitive classification, the Accuracy, Sensitivity and FPR obtained from the training are shown and compared in table 3. It is notable that the results show high accuracy, meaning that from the set of computed features, it is possible to differentiate information describing the user's style for each one of the primitives.

For Rotation and Scaling regression, table 4 shows the results.

Feature Vector
NumberLines
NumberStrokes
TotalLength
BoundingBoxWidth
BoundingBoxHeight
MaxLineLength
MinLineLength
MeanLineLength
Shdw_NumberLines
Shdw_NumberStrokes
Shdw_BoundingBoxWidth
Shdw_BoundingBoxHeight
Shdw_North
Shdw_South
Shdw_East
Shdw_West

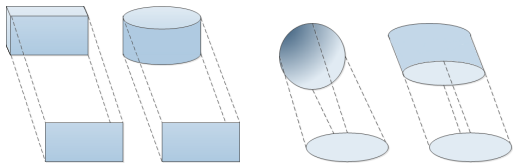
Table 1: Computed features for segmented lines from the user's sketch.

Parameter	NTrees	R
X Rotation	60	0.1
Y Rotation	70	0.1
Z Rotation	90	0.1
X Scaling	60	0.9
Y Scaling	90	0.3
Z Scaling	70	0.1

Table 2: Parameter setup to reduce MSE for Regression Random Decision Forest for each pose degree.

Measurement	**Accuracy**	**Sensitivity**	**FPR**
Primitive Classification	0.9917	0.9687	0.0000

Table 3: Accuracy, Sensitivity and False Positive Rate (FPR) for Primitive Classification.



(a) Example of Cuboid and Cylinder poses that have close similarity.

(b) Example of Cylinder and Spheroid poses that have close similarity.

Figure 5: Examples of primitive confusion when seen from frontal vies (bottom images) and different when translated, rotated and scaled (top images).

The rotation accuracy is affected by the uncertainty implied in the redundancy of rotation degree. For example, for a basic cube the visual representation repeats every 90°. The accuracy for scaling is between 85% and 90%.

Feature Relevance

To identify what set of features is more relevant to primitive classification, and to rotation and scaling regressions, the proposed method by [14] has been followed, which consists in generating an artificial variable constructed by randomly reordering one of the features and replacing it, keeping the rest of the features and labels in place and from there, calculate the Accuracy, Sensitivity and FPR for classification, and in the case of Regression, calculate the MSE and compare to our baseline.

For primitive classification, *Number of Strokes* is the most relevant feature by 39%, followed by the *Total Length* (sum of lengths from the lines detected on the sketch) by 8%, and then for both *Bounding box height and width* with 6%.

For rotation and scaling, the relevance for all the features is between 6.5% and 7.5%. Although the difference on the MSE between them is very low, we can identify the following features as more relevant: *Bounding Box Height*, *Total Length*, *Bounding Box Width* and *Line Maximum Length*.

Relevance of training data amount

Although training on a small number of sketches is enough for primitive classification, in practice it is suggested to extend the training to at least 20 poses for each primitive because there are some special cases in which specific poses could confuse the task of identifying the primitive [Fig. 5]. For rotation and scaling, the accuracy becomes stable after 60 training samples (20 for each primitive).

	MSE	% Error	% Accuracy
X Rotation	0.07518	0.27418	72.58%
Y Rotation	0.08845	0.29741	70.26%
Z Rotation	0.08073	0.28412	71.59%
X Scaling	0.01209	0.10996	89.00%
Y Scaling	0.00935	0.09669	90.33%
Z Scaling	0.02048	0.14309	85.69%

Table 4: Regression results for Rotation and Scaling.

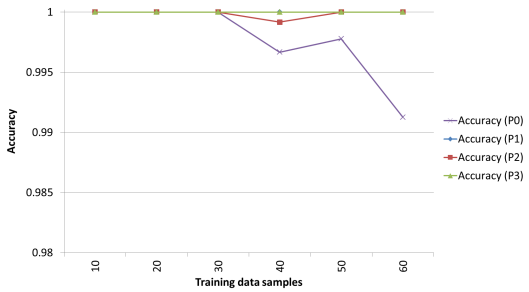


Figure 6: Accuracy comparison for Primitive Classification accuracy on baseline (P0) and three validation users (P1, P2, P3).

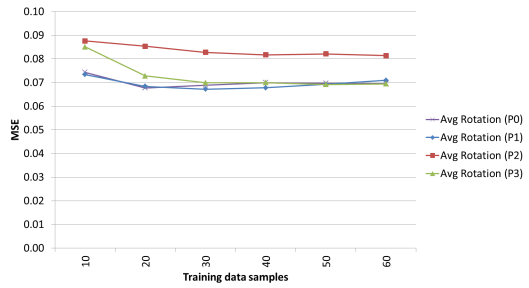


Figure 7: MSE comparison for Rotation Regression on baseline (P0) and three validation users (P1, P2, P3).

VALIDATION

The system has been validated over three representative users of different skills (two with no artistic background and one with artistic background).

Validation Results

The accuracy results on the primitive classification task for the three users (P1, P2 and P3) are compared against the accuracy obtained for the user participating in the deep study as baseline (P0), which are presented in figure 6. The results in this validation test are slightly better than the obtained on the deep study, which means these three users were more consistent in the sketching style across primitives. However, as the accuracy for all of them, including the baseline, is so high, the difference is not of much significance.

To simplify the comparison for Rotation and Scaling between the users, the three Rotation dimensions have been averaged to a single quantification of MSE for Rotation. The same has been simplified for Scaling. Figures 7 and 8 represent the MSE comparison for Rotation and Scaling Regression respectively, comparing the baseline user P0 and the three users (P1, P2, P3) for the Validation User Study. The results demonstrate that the model setup is extendible for users of different artistic skills, with similar results, and that to train on 20 sketches for each primitive is enough for these results.

For reference, examples of captured sketches are shown in figure 9.

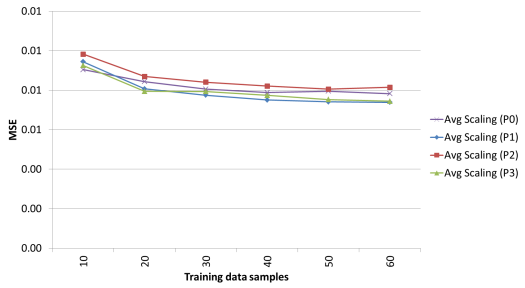


Figure 8: MSE comparison for Scaling Regression on baseline (P0) and three validation users (P1, P2, P3).

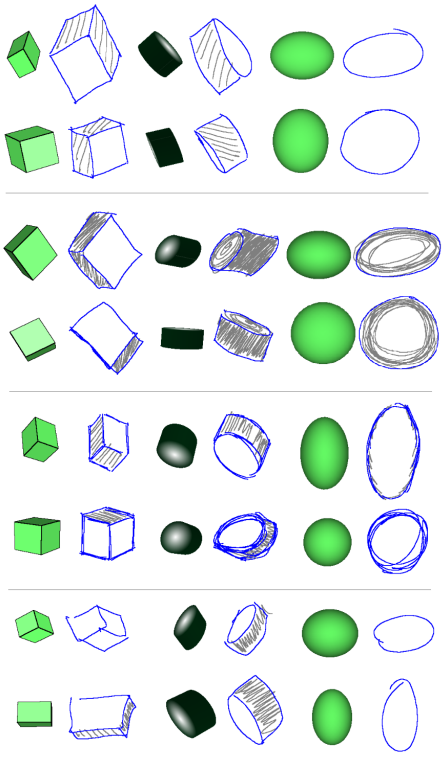


Figure 9: Examples of sketches by four different users. At top, sketches from the user participating in the deep study, the three users at bottom participating in validation study.)

DISCUSSION

There are some factors to take in mind when considering the results:

- The rotation degree range of 0-360°is redundant for 3D primitives. For example, for a basic cube the visual representation repeats every 90°. As this feature is not included in this model, the accuracy is decreased given the uncertainty it implies.

- Overlaying comparisons between a rendered 3D primitive and the user's sketch (e.g. figure 10) show that the sketches reveal users' conception and do not portray reality [15], illustrating how noisy the input sketch is.

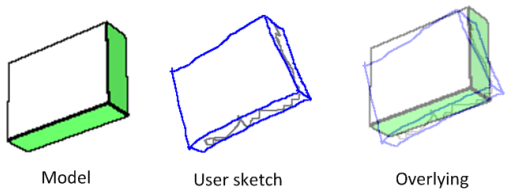


Figure 10: Overlaying comparison between the rendered model and the user's sketch.

- There is an intrinsic relationship between the three rotational dimensions and the three scaling dimensions which is not modelled in this project, and is relative to the above point. For example, if a cube is scaled up in the Y dimension and rotated 90°over the Z dimension, it will be visually similar to the a cube scaled up in the X dimension without rotation.

- There is strong evidence [13] that humans don't have accurate perception of 3D metric structure.

CONCLUSIONS

This paper has confirmed that relevant information exists on user gestures from sketching styles, even with simple features to map 2D sketches to 3D primitive models.

For Primitive classification, the results were very high, predicting the 3D primitive almost with full accuracy. This is attributed to the specific graphical characteristics each primitive has and makes it different to the others. In the case of Rotation regression, although the numerical results don't seem very satisfactory, they still provide a good approximation and maintain stability in the predictions, which also demonstrates intrinsic existence of valuable information on the user gestures to denote the rotation. The results on Scaling Regression are better than those in Rotation, maintaining stability as well.

This research has also demonstrated the following:

- Users have a tendency to keep the same sketching style to denote each primitive and its pose, independently of their sketching skills or academic background.

- This model captures the intrinsic style in the gestures, independently of the artistic skills of the user. Therefore, the model setup in this project is extendible for other users obtaining similar results.

- To train on 20 sketches for each primitive is enough to obtain good results, which takes from 10 to 30 minutes depending the user's speed.

Finally, it is important to remember that this method does not claim to generate better models than image similarity approaches, but suggests to acknowledge the difference on users' skills and to consider adapting to them.

Further Work

This research can be extended in several ways. To list only some: increasing the 3D model complexity beyond the three

used primitives, using more advanced features (e.g. line joints), and multi-object sketching recognition.

REFERENCES

1. Bae, S.-H., Balakrishnan, R., and Singh, K. Everybodylovessketch: 3d sketching for a broader audience. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, UIST '09, ACM (New York, NY, USA, 2009), 59–68.
2. Brookes, A. The adaptive nature of 3d perception. In *Proceedings of the second international conference on From animals to animats 2 : simulation of adaptive behavior: simulation of adaptive behavior*, MIT Press (Cambridge, MA, USA, 1993), 116–121.
3. Cook, M. T., and Agah, A. A survey of sketch-based 3-D modeling techniques. *Interact. Comput. 21*, 3 (July 2009), 201–211.
4. Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K., and Alexa, M. Sketch-based shape retrieval. *ACM Trans. Graph. (Proc. SIGGRAPH) 31*, 4 (2012), 31:1–31:10.
5. Glennerster, A. *The encyclopedia of Mind*. Sage Editorial, San Diego, USA, January 2013.
6. Igarashi, T., Matsuoka, S., and Tanaka, H. Teddy: a sketching interface for 3d freeform design. In *ACM SIGGRAPH 2007 courses*, SIGGRAPH '07, ACM (New York, NY, USA, 2007).
7. Jorge JA, Silva FN, C. D. Gides++. In *In Proceedings of the 12th annual Portuguese CG meeting* (2003).
8. Lipson, H., and Shpitalni, M. Correlation-based reconstruction of a 3d object from a single freehand sketch. In *ACM SIGGRAPH 2007 courses*, ACM (2007), 44.
9. Long, Jr., A. C., Landay, J. A., Rowe, L. A., and Michiels, J. Visual similarity of pen gestures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '00, ACM (New York, NY, USA, 2000), 360–367.
10. Olsen, L., Samavati, F. F., Sousa, M. C., and Jorge, J. A. Sketch-based modeling: A survey. *Computers & Graphics 33*, 1 (Feb. 2009), 85–103.
11. Pastel, R., and Skalsky, N. Demonstrating information in simple gestures. In *Proceedings of the 9th international conference on Intelligent user interfaces*, IUI '04, ACM (New York, NY, USA, 2004), 360–361.
12. Schmidt, R., Isenberg, T., Singh, K., and Wyvill, B. Sketching, scaffolding, and inking: A visual history for interactive 3d modeling.
13. Todd, J. T. The visual perception of 3d shape from multiples cues: Are observers capable of perceiving metric structure? *Perception and Psychophysics*, 65 (2003), 31 – 47.
14. Tuv, E., Borisov, A., Runger, G., and Torkkola, K. Feature selection with ensembles, artificial variables, and redundancy elimination. *J. Mach. Learn. Res. 10* (Dec. 2009), 1341–1366.
15. Tversky, B. What does drawing reveal about thinking? In *IN*, Citeseer (1999).
16. Winder, S. A. J., and Brown, M. Learning local image descriptors. In *In CVPR* (2007), 1–8.
17. Xiong, Y., and LaViola, Jr., J. J. Revisiting shortstraw: improving corner finding in sketch-based interfaces. In *Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling*, SBIM '09, ACM (New York, NY, USA, 2009), 101–108.
18. Ying, S., Lin, L., and Zhang, Y. Consistent line simplification based on constraint points. In *21st International Cartographic Conference*, International Cartographic Association (Durban, South Africa, Aug. 2003).
19. Zitnick, C. L. Binary coherent edge descriptors. In *Proceedings of the 11th European conference on Computer vision: Part II*, ECCV'10, Springer-Verlag (Berlin, Heidelberg, 2010), 170–182.