Predicting Artist Drawing Activity via Multi-Camera Inputs for Co-Creative Drawing*

Chipp Jansen[1][0000−0003−2951−6396] and Elizabeth Sklar[2][0000−0002−6383−9407]

[1] Centre for Robotics Research, King's College London, London, UK
chipp.jansen@kcl.ac.uk
[2] Lincoln Institute for Agri-food Technology, University of Lincoln, Lincoln, UK
esklar@lincoln.ac.uk

Abstract. This paper presents the results of computer vision experiments in the perception of an artist drawing with analog media (pen and paper), with the aim to contribute towards a human-robot co-creative drawing system. Using data gathered from user studies with artists and illustrators, two types of CNN models were designed and evaluated. Both models use multi-camera images of the drawing surface as input. One models predicts an artist's activity (e.g. are they drawing or not?). The other model predicts the position of the pen on the canvas. Results of different combination of input sources are presented. The overall mean accuracy is 95% (std: 7%) for predicting when the artist is present and 68% (std: 15%) for predicting when the artist is drawing. The model predicts the pen's position on the drawing canvas with a mean squared error (in normalised units) of 0.0034 (std: 0.0099). These results contribute towards the development of an autonomous robotic system which is aware of an artist at work via camera based input. In addition, this benefits the artist with a more fluid physical to digital workflow for creative content creation.

Keywords: human-robot collaboration · co-creative drawing · computer vision · deep learning · convolutional neural networks · sketch-based computing

1 Introduction

Visual artists enjoy a large economy of creative digital tools to produce their work. In our recent study into co-creative artistic workflows [10], we found artists often use physical analog media (e.g. pen and ink on paper) for initial idea exploration and desire for a more fluid transition from analog to digital media. In addition, when considering collaboration with an Artificial Intelligence (AI), we found that artists preferred an inspirational or co-creative AI role to that of a didactic one.

* This research is supported through an EPSRC (UK) DTP Studentship "Collaborative Drawing Systems", Grant Reference EP/N509498/1.

In this paper, we investigate vision-based methods to build models of artists' **activity** (e.g. are they currently *drawing*—pen touching the page—or *not*—pen hovering above the page while the artist is thinking about what to draw next) and **output** (e.g. predicting the pen position on the page to understand what is being drawn on the canvas) while drawing.

Understanding the pen movements would then allow for a vision-based system to digitally recreate a drawing without being tethered to a drawing tablet or to rely on a scanner set-up. A camera-based system would allow an artist freer physical range in the studio, as well as a more diverse set of mediums to draw upon—an important point of feedback gathered previously [10].

Ultimately, we see these vision-based methods as models that would be components of a co-creative drawing system enhanced with visual-based awareness of the artist. Since the drawing process is a 3-D activity, despite having 2-D outputs, we evaluate which image inputs (e.g. camera positions for observing the artist) are most useful for these models through the experiments presented in this paper. We believe the results of these experiments would not only be useful for the creative computing community, but also the greater human-robotic interaction community, as they describe predicting fine human motor control (e.g. drawing) at a personal robot scale.

2 Background

Artist's drawing behaviour with physical media has been studied in psychology, from manual annotation of video frames of an artist's hand motion [18] to using techniques such as *saliency analysis*, or analysing the movement of an individual's eye fixation to understand where the their attention lies [17, 16]. Sensor fusion has also been used to study the painting process through combining axis-aligned cameras and acoustic sensors attached to a canvas to record the contact of a paint brush onto the canvas surface [6]. Within the computer graphics and human-computer interaction literature, there is a rich tradition of sketch-based computing and interaction via digital interfaces such as drawing tablets [15, 12].

Co-creative drawing systems aim to be a drawing partner, such as the *Drawing Apprentice* [4], where an improvising drawing agent analyzes the user's input and responds with its own artistic contributions upon a shared digital canvas. Neural network approaches to sketching, such as the *sketch-rnn* model [7] (and the availability of large-scale drawn datasets, e.g. *QuickDraw!* [11]) have inspired a class of deep learning driven co-creative drawing systems [13, 5, 14]. In all of these systems, the medium is digital drawing, which is immediate for the creative agent to observe the state of the artist, where they are drawing and for the agent to interact with the canvas. However, there are a few recent examples of physical co-creative work with robotic systems, such as *D.O.U.G* [2], which involves an industrial robot to mimic what the artist is drawing and in turn the artist can respond; and the the *ArtTherapyRobot* [3] which uses a Baxter[3] robot

[3] https://robots.ieee.org/robots/baxter/

to conduct research into socially assisted robotics for art therapy. Instead of a robot, projected interfaces serve as a platform to physical co-creative drawing as well, such as the *DialogCanvasMachine* [1].

Most of these physical co-creative drawing examples feature a bespoke system created to facilitate the co-creation with a specific individual artist or artist-group, as opposed to being research into more general physical co-creative drawing. In addition, while some of the examples capture the artist's drawing process for reflective post-processing [1, 6], none of these systems build a real-time model of what the artist is currently drawing or their behaviour. In addition, artists and illustrators still use physical media as part of their workflow and desire a more fluid way of capturing their drawings [8], a feature which is currently lacking in contemporary sketch-based computing research.

3 Research Set-up

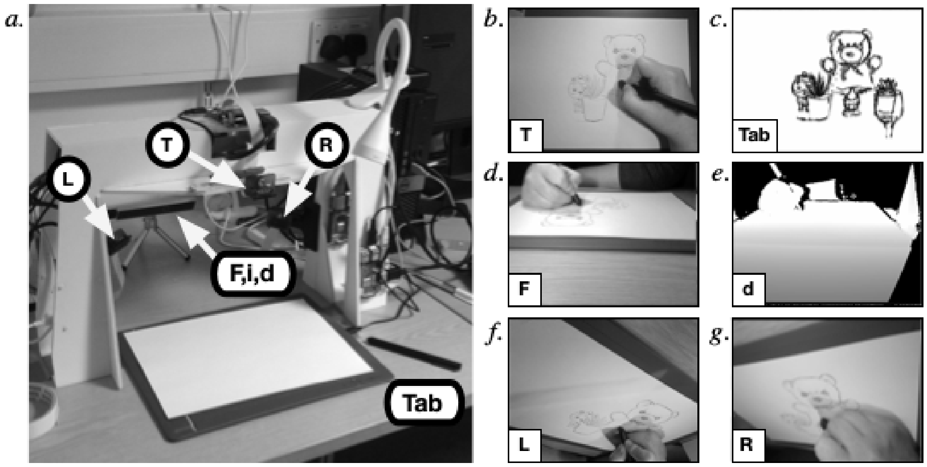

Fig. 1. *(a.)* Prototype hardware setup with components: top, right and left cameras **T**, **R** and **L**; front camera **F** with infrared **i** depth **d** components; and drawing tablet **Tab**. *(b-g.)* input from the components: top camera, drawing tablet, front camera rgb, front camera depth, left camera and right camera respectively (front infrared camera component is not shown).

We have developed a co-creative drawing system research prototype [9], shown in Figure 1, comprising multiple cameras that observe an artist's drawing surface. There are 3 RGB cameras [4](an overhead top down camera (**T**), and side oblique right (**R**) and left (**L**) cameras). There is also a front facing depth

[4] Raspberry PI Camera Module V2 https://www.raspberrypi.org/products/camera-module-v2/

camera [5] (with separate RGB (**F**) and infrared sensors (**i**) integrating into a depth image (**d**)). All of the cameras record at 25 frames per second to produce images at 1280 × 960 resolution (for **T**, **R**, **L**) and 640 × 480 resolution (for **F**, **d**, **i**). In addition, the artist draws on paper on top of a drawing tablet (**Tab**), which records the position (x and y coordinates) and pressure of the drawing pen at 200 vector points per second at a discrete 0.01mm resolution. This set-up allows us to gather drawing data that correlates camera images with a drawn vector representation from the tablet.

4 Drawing Data Gathering Study

In early 2020, we conducted a drawing data gathering study ($n = 13$) involving full-time drawing practitioners (professionals and students) to test our prototype system and to collect the drawing dataset used in the models presented in this paper. Participants were instructed to undertake two separate drawing exercises: draw from *observation* of a still-life, and draw freely from *imagination*. For each exercise, the participants were asked to draw for at least 10 minutes, but no more than 30 minutes (with a time reminder every 10 minutes). In total, our research prototype recorded 26 drawing exercises. However, due to technical issues, our prototype was only able to record from all input sources for both drawing exercises from 7 participants.

In this paper, we utilised data from these 7 participants to produce two types of datasets with corresponding models: *activity* and *pen_position*. The examples in each dataset comprise 6 temporally correlated input images (**T**,**R**,**L**,**d**,**i**), which are individually resized (using nearest-neighbor) to a smaller and more computationally tractable resolution (80 × 60 pixels). Each example is labelled using the corresponding drawing tablet data as ground truth. From each of the 7 participants' two drawing sessions, 14 *activity* and 14 *pen_position* datasets were produced. Every dataset had 3500 examples, which we split into 80% training ($n = 2800$) and 20% testing sets ($n = 700$).

Each *activity* dataset was randomly sampled from the entire drawing session, prioritising examples that had the lowest temporal difference amongst the 6 image frames. Categorising artist's "activity" while drawing is a multi-faceted and deeply complex phenomenon. For the purposes of these experiments, we take advantage of the drawing tablet, which senses the proximity of the pen once it is within 2-3cm. The pen senses a pressure level as an integer value ($[0, 2047]$), which is a relative measure of the pressure of the pen's tip upon the drawing surface. A pressure level of 0 indicates that the pen is "hovering" above the page. A pressure level > 0 indicates the pen is "drawing". Otherwise, when no points are being recorded, the pen (and thus the artist) is "away". We use these pen states to label the *activity* dataset examples with a 3-class *pen_state* variable ("drawing", "hovering", "away"). While these are a natural classification from the pen; however, from the perspective of developing a controller for a co-creative system, it is useful for the AI to be able to discern two things: first,

[5] Intel Depth Camera SR305 https://www.intelrealsense.com/depth-camera-sr305/

Table 1. Distribution of *pen_state* classes for each participant exercises dataset. Rows are labelled by the participant id (**1** to **7**) and the drawing exercise: observation (**obs**) and imagination (**img**).

	training				test			
	total	drawing	hovering	away	total	drawing	hovering	away
1 img	2800	1985 (71%)	593 (21%)	222 (8%)	700	468 (67%)	181 (26%)	51 (7%)
obs	2800	1654 (59%)	721 (26%)	425 (15%)	700	427 (61%)	164 (23%)	109 (16%)
2 img	2800	1071 (38%)	1181 (42%)	548 (20%)	700	261 (37%)	317 (45%)	122 (17%)
obs	2800	1389 (50%)	1277 (46%)	134 (5%)	700	380 (54%)	281 (40%)	39 (6%)
3 img	2800	873 (31%)	1372 (49%)	555 (20%)	700	220 (31%)	357 (51%)	123 (18%)
obs	2800	596 (21%)	957 (34%)	1247 (45%)	700	142 (20%)	232 (33%)	326 (47%)
4 img	2800	921 (33%)	1236 (44%)	643 (23%)	700	227 (32%)	314 (45%)	159 (23%)
obs	2800	1295 (46%)	1342 (48%)	163 (6%)	700	328 (47%)	329 (47%)	43 (6%)
5 img	2800	2001 (71%)	473 (17%)	326 (12%)	700	501 (72%)	108 (15%)	91 (13%)
obs	2800	2113 (75%)	594 (21%)	93 (3%)	700	552 (79%)	133 (19%)	15 (2%)
6 img	2800	1443 (52%)	1201 (43%)	156 (6%)	700	363 (52%)	294 (42%)	43 (6%)
obs	2800	1009 (36%)	1419 (51%)	372 (13%)	700	253 (36%)	330 (47%)	117 (17%)
7 img	2800	1388 (50%)	1263 (45%)	149 (5%)	700	342 (49%)	315 (45%)	43 (6%)
obs	2800	1532 (55%)	1136 (41%)	132 (5%)	700	393 (56%)	279 (40%)	28 (4%)

is the artist present; and second, is the artist drawing. We derive two further binary classes: *is_drawing* ($activity == drawing$) and *is_present* ($activity == drawing \lor activity == hovering$) based on the pen state.

Table 1 shows the distribution of examples for each of the three *pen_state* classes. Due to the random sampling regime, the distribution for each participant-drawing exercise *activity* dataset varied. For the experiments presented here, the balance of examples across classes was not adjusted; future work will consider re-balancing (e.g. boosting) some classes to improve prediction accuracy.

Finally, each *pen_position* dataset was randomly sampled *only when the artist was drawing* and are labeled with the normalised pen position: $(x, y) = ([0, 1], [0, 1])$.

5 Visual Based Models

We produced two types of models, based on the previously described datasets: *activity* and *pen_position*. Each model takes the 6 camera images as input (from individual sources or in combination of multiple sources). Each image is fed independently through a sequence of *Convolutional Neural Network (CNN)* layers, to be concatenated in a single layer that is fully connected to output variables. The concatenation layer is then connected via a single hidden layer to the output. There are three different variations of the *activity* model, each based on the multi-class output variables: *pen_state* (3 classes), *is_present* (2 classes) and *is_drawing* (2 classes). There is a single *pen_position* model, which produces pen positions (x and y), normalised to the width and height of the drawing tablet.

Models were built and trained using *Tensorflow*[6], with a split of 80:20 on the training to the test data subsets, using an ADAM optimiser with a learning rate of 0.01, for 30 epochs each. The *activity* models were trained to optimise a cross-entropy loss for multi-class variables (Boolean variables were treated as multi-class to maintain consistency in the experimental methods) with an accuracy metric evaluated on the test dataset. The *pen_position* models were trained to minimise the combined *Mean Squared Error (MSE)* loss for the normalised x and y output variables.

6 Experiments and Results

We experimented with 22 different combinations of input images (6 single individual image input, 15 pairs of images and the set of all images) on the three flavours of *activity* models and the *pen_position* model. Each model was trained and evaluated independently with a corresponding user-session dataset to explore 308 variations per model type.

Figure 2 shows the accuracy results for the three *activity* models: *pen_state*, *is_drawing*, *is_present* for specific image combinations. Each bar is a summary of the 14 participant-exercises datasets.

Overall (all sessions and combinations together), the accuracy for the *is_present* binary model was higher (mean 95.7%, std 6.7%, n=308) than the *is_drawing* (mean 68.3%, std 15.1%, n=308) model. Accuracy for the 3-class *pen_state* (mean 68.5%, std 16.0%, n=308) model had a wide variation amongst the different input combinations, with the Front camera (**F**) having the best performance. The Right camera (**R**) had noticeably worse performance, as shown by the spread in the *is_present* model. All of the participants in the selected datasets were right-handed and their hand often occludes the pen tip in the Right camera view, which may explain this variation. The Front infrared (**i**) camera also performed poorly within the *is_present* model, although the RGB component of the same camera (**F**) produced a high mean accuracy for the *pen_state* model.

Figure 3 shows the MSE of the x and y components, and the combined x *and* y training metric for the *pen_position* model for specific image combinations. The MSE is in terms of normalised x, y positions of the pen with respect to the width (29.7cm) and height (21.6cm) of the drawing tablet.

Overall, the MSE for x (mean 0.001298, std 0.004564, n=308) was lower than y (mean 0.002054, std 0.005789, n=308). The combined (x *and* y) MSE (mean 0.003352, std 0.009936, n=308) was highest. For the *pen_position* model, there seems to be little difference amongst the individual RGB cameras (**T**, **L**, **R**, **F**), while the individual depth (**d**) performs worse, and the individual infrared (**i**) has an out-sized comparative variance. In addition, the pair-wise images also seem to have little difference amongst themselves. However, models that use all the input images (**All**) yielded a far better result than the individual image sources, and had the best mean MSE overall.

[6] https://www.tensorflow.org/

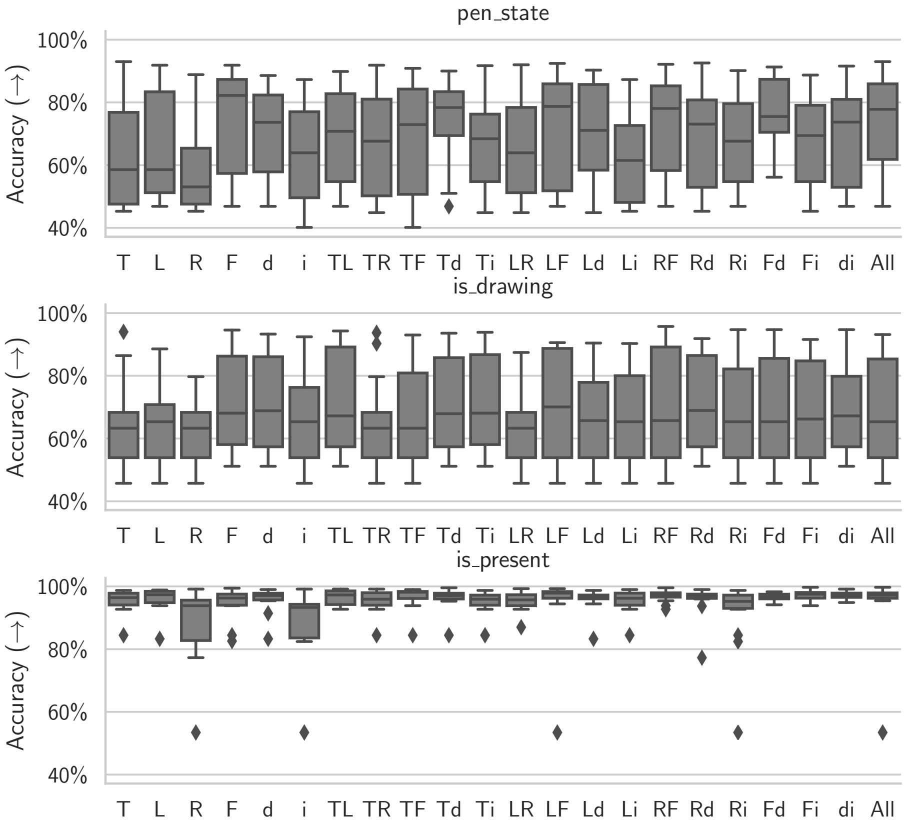

Fig. 2. Accuracy of predicting the activity of the artist: (*top to bottom*): *pen_state*, *is_drawing*, *is_present*. Accuracy values fall between 40-100%, higher is better (↑).

7 Discussion and Limitations

Sensing when the artist is present visually, using the *is_present* model, is by far the most successful model from our experimentation aside from relying solely on the same-handed oblique side camera (i.e. Right camera (**R**) for a right-handed artist). Sensing when artist is drawing, using the *is_drawing* model, proves to be more difficult. This may be due to the slight visual differences between the pen touching the canvas and that of the pen hovering just above the canvas, especially at the lower image resolution of 80 × 60. In addition, the wide variation in the balance for the different *pen_state* classes as shown in Table 1 may be a reason for the results for the *activity* models having a wide variation.

Basing the artist's activity on the pen state also has limitations. For example, an artist who draws with grand arm motions will, at moments, lift their pen beyond the 2-3cm bounds of pen proximity for the drawing tablet, thus recording "hover" as "away" activity. Or, when the artist sets their pen down

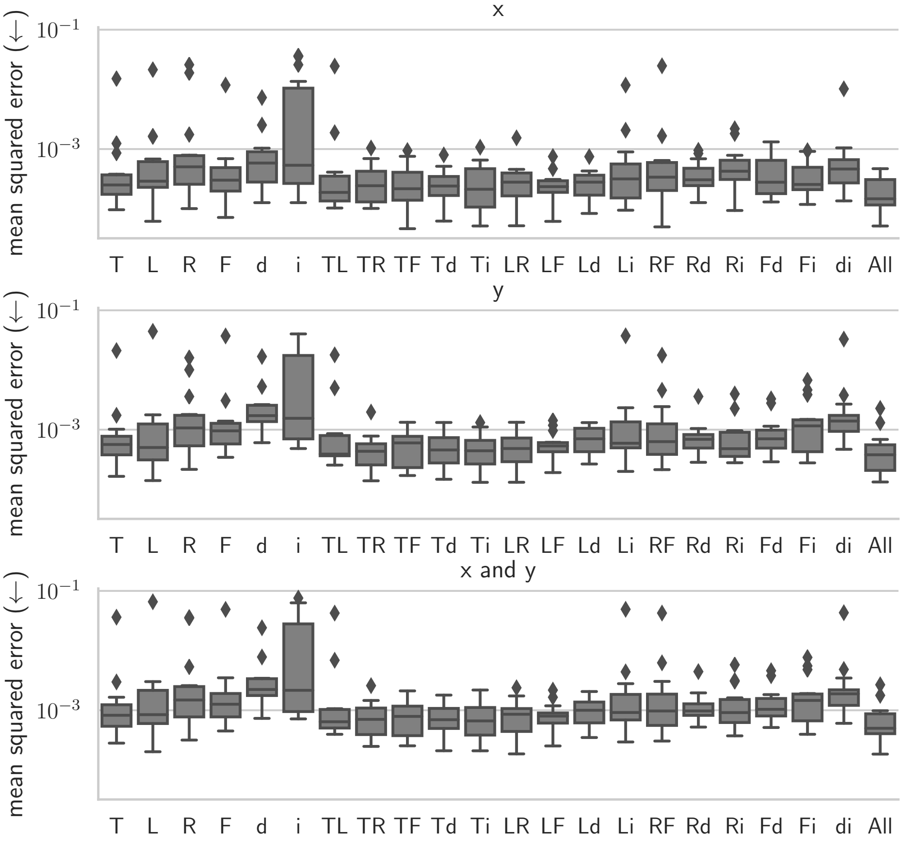

Fig. 3. Mean Squared Error (MSE) (log scale) of the pen attribute predictions for the *pen_position* models (*top to bottom*): *x*, *y* and combined *x and y*. Each error bar summaries the 14 drawing sessions for the specific images combination, lower is better (↓).

upon the tablet to take a break, this will be recorded as a continuous stream of "hover" points, but the artist is in fact "away". These limitations reinforce the advantage of having a vision based system which adds additional context to recording an artist's activity. These labels could be further refined, through manual annotation of the artist's states from the camera images.

Predicting the pen's position had a clearer result with the combined image inputs model (**All**) having the lowest error. While the MSE for the *pen_position* model is low, initial attempts to use a model with visual-driven drawing did not produce coherent results (Figure 4). This might be due the variation in predicted points being too high at the camera frame rate (i.e. 25 frames per second as opposed to the 200 point per second produced by the drawing tablet).

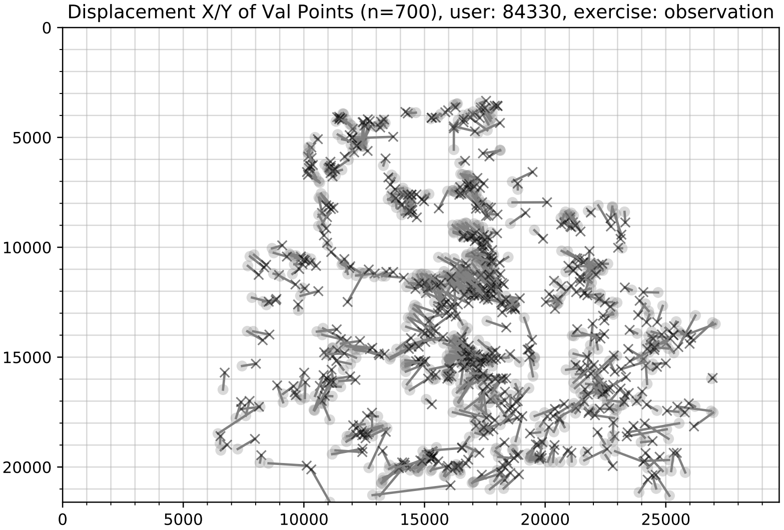

Fig. 4. Example rendering using the *pen_position* model using test points ($n = 700$) from an observational drawing session. Actual points (light grey dots) are connected to predicted points (black X's).

8 Summary and Future Work

We have demonstrated that using vision-based input from a multi-camera system with a trained CNN can predict the activity and output of an artist drawing with physical media—being able to predict that an artist is present and drawing within a relatively localised area on the canvas.

While these models were trained and evaluated on individual drawing session datasets, possible future work in *transfer learning* is possible to evaluate one artist's model on another artist's drawing data. Our current rationale for training only on an individual session is to work towards a system which is bespoke and custom to a particular artist's drawing style. However, another avenue of work would be to train a more general purpose model that later adapts to a specific artist's style.

Next steps in our research is to integrate these models into a framework for co-creative drawing, and to evaluate this framework with various co-creative drawing agents in an artist's studio setting.

References

1. Cabannes, V., Kerdreux, T., Thiry, L., Campana, T., Ferrandes, C.: Dialog on a canvas with a machine. arXiv:1910.04386 [cs] (Oct 2019)
2. Chung, S.: Drawing Operations (DOUG). https://sougwen.com/project/drawing-operations (2015)

3. Cooney, M., Berck, P.: Designing a Robot Which Paints With a Human: Visual Metaphors to Convey Contingency and Artistry. In: ICRA-X Robots Art Program at IEEE International Conference on Robotics and Automation (ICRA). p. 2. Montreal QC, Canada (May 2019)
4. Davis, N., Hsiao, C.P., Singh, K.Y., Magerko, B.: Co-Creative Drawing Agent with Object Recognition. In: Aritificial Intelligence in Interactive Digital Entertainment. p. 8. Burlingame, California, USA (2016)
5. Fan, J.E., Dinculescu, M., Ha, D.: Collabdraw: An Environment for Collaborative Sketching with an Artificial Agent. In: Proceedings of the 2019 on Creativity and Cognition. pp. 556–561. C&C '19, Association for Computing Machinery, San Diego, CA, USA (Jun 2019). https://doi.org/10.1145/3325480.3326578
6. Fernando, P., Weiler, J., Kuznetsov, S., Turaga, P.: Tracking, Animating, and 3D Printing Elements of the Fine Arts Freehand Drawing Process. In: Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction - TEI '18. pp. 555–561. ACM Press, Stockholm, Sweden (2018). https://doi.org/10.1145/3173225.3173307
7. Ha, D., Eck, D.: A Neural Representation of Sketch Drawings. arXiv:1704.03477 [cs, stat] (May 2017)
8. Jansen, C., Sklar, E.: Co-creative Physical Drawing Systems. In: ICRA-X Robots Art Program at IEEE International Conference on Robotics and Automation (ICRA). p. 2. Montreal QC, Canada (May 2019)
9. Jansen, C., Sklar, E.: Towards a HRI system for co-creative drawing. In: ACM/IEEE International Conference on Human-Robot Interaction (HRI), Workshop on on Exploring Creative Content in Social Robotics (2020)
10. Jansen, C., Sklar, E.: Exploring co-creative drawing workflows. Frontiers in Robotics and AI **8**, 92 (2021)
11. Jongejan, J., Rowley, H., Kawashima, T., Kim, J., Fox-Gieg, N.: The Quick, Draw! - A.I. Experiment. https://quickdraw.withgoogle.com/ (2016)
12. Jorge, J., Samavati, F.: Sketch-Based Interfaces and Modeling. Springer Science & Business Media (Dec 2010)
13. Karimi, P., Maher, M.L., Davis, N., Grace, K.: Deep Learning in a Computational Model for Conceptual Shifts in a Co-Creative Design System. arXiv:1906.10188 [cs, stat] (Jun 2019)
14. Oh, C., Song, J., Choi, J., Kim, S., Lee, S., Suh, B.: I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–13. CHI '18, Association for Computing Machinery, Montreal QC, Canada (Apr 2018). https://doi.org/10.1145/3173574.3174223
15. Olsen, L., Samavati, F.F., Sousa, M.C., Jorge, J.A.: Sketch-based modeling: A survey. Computers & Graphics **33**(1), 85–103 (Feb 2009). https://doi.org/10.1016/j.cag.2008.09.013
16. Sarvadevabhatla, R.K., Suresh, S., Babu, R.V.: Object category understanding via eye fixations on freehand sketches. IEEE Transactions on Image Processing **26**(5), 2508–2518 (May 2017). https://doi.org/10.1109/TIP.2017.2675539
17. Tchalenko, J., Nam, S.H., Ladanga, M., Miall, R.C.: The gaze-shift strategy in drawing. Psychology of Aesthetics, Creativity, and the Arts **8**(3), 330–339 (Aug 2014). https://doi.org/10.1037/a0036132
18. Van Sommers, P.: Drawing and Cognition: Descriptive and Experimental Studies of Graphic Production Processes. Cambridge University Press, Cambridge [Cambridgeshire] ; New York (1984)