https://doi.org/10.31256/Rq9Yg6l

Visually-based Prediction of Artist's Drawing

Chipp Jansen[1]
[1]*Dept of Engineering*
King's College London, UK
chipp.jansen@kcl.ac.uk

Elizabeth I Sklar[1,2]
[2]*Lincoln Institute for Agri-food Technology*
University of Lincoln, UK
esklar@lincoln.ac.uk

***Abstract*—This paper describes recent work in the development of a co-creative human-robot drawing system, which observes an artist's drawing process in real-time. Using the data gathered in a recent pilot study, a series of models were trained in order to recover the current state of the artist's drawing behaviour and pen attributes from a multi-camera multi-perspective set-up, aligned to a "ground truth" dataset obtained from a drawing tablet. Experiments, carried out with two computer vision models based on a CNN architecture, form a baseline for future, more sophisticated models.**

***Index Terms*—human-robot collaboration, co-creative drawing, computer vision, deep learning, sketch-based computing**

I. Introduction

Visual artists enjoy a large economy of creative digital tools to produce their work. However, as result of a recent study into co-creative artistic workflows [1] , we have found a desire for a more fluid transition between digital and analog media (e.g. pen and ink on paper), as artists often use physical media for initial idea exploration. Here, we investigate vision-based methods to understand artists' activity (e.g. are they currently drawing or not?) and output (e.g. predicting the pen position on the page) while drawing; and to understand which inputs (e.g. camera positions) are most useful for this modelling.

II. Background

Computer graphics and human-computer interaction have a rich literature on sketch-based computing and interaction via digital interfaces such as drawing tablets [2]. Neural network approaches to model sketching, such as the *sketch-rnn* model [3] (and the availability of large-scale drawn datasets, e.g. *QuickDraw!* [4]) have inspired many co-creative drawing systems [5]–[7]. Some of these co-creative systems respond to artists working with analog media and capture the drawing process for reflective post-processing [8], [9]. However, 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 [10], a feature which is currently lacking.

III. Models and Experiments

a) Set-up: We have developed a research prototype comprising multiple cameras that observe an artist's drawing surface: 3 RGB cameras (overhead, oblique right and left), 1 front facing depth camera (with RGB and infrared cameras integrating into a depth image). The artist draws on paper on top of a drawing tablet which records the position (x and y coordinates) and pressure of the drawing pen [11]

b) Data: In early 2020, we conducted a drawing data gathering study involving 13 professional and student illustrators who were prompted to engage in two drawing exercises: (i) observational drawing of a still-life; and (ii) drawing from imagination or memory. Here we utilise data from both drawing exercises for 5 participants, independently, to produce two types of datasets with corresponding models: *activity* and *pen_position*. The examples in each dataset are comprised of 6 temporally correlated images[1], which are resized to a fixed resolution (80×60 pixels) and labelled using the corresponding drawing tablet data as ground truth.

The *activity* dataset examples are labeled with a 3-class *pen_state* variable ("drawing", "hovering", "away") and two binary classes *is_drawing* and *is_present* based on the pen state. Each *activity* dataset had 2500 examples sampled at even intervals (200ms). The *pen_state* dataset had 3500 examples sampled only when the artist was drawing and are labeled with the normalised pen position: $(x, y) = ([0, 1], [0, 1])$.

c) Models: Each model takes 6 camera images as input (from individual sources or in combination). Each image is fed independently through a sequence of *Convolutional Neural Network (CNN)* layers, to be concatenated in a single layer that fully connected to output variables. There are three flavours of the *activity* model based on the variables: *pen_state*, *is_present* and *is_drawing*. The drawing model produces normalised pen position (as above). Models were built and trained using *Tensorflow*[2], with an 80/20 training/validation split on the datasets, using an ADAM optimiser with a learning rate of 0.01, for 30 epochs each.

d) Experiments: We experimented with 7 different combinations of input images (single individual image input and all images), on the three flavours of *activity* models and the *pen_state* model. Each model was trained and evaluated independently with a corresponding user-session dataset.

Research is supported through an EPSRC DTP Studentship "Collaborative Drawing Systems", Grant Reference EP/N509498/1

[1]4 RGB images for each camera, the infrared and depth image from the depth camera

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

IV. Results and Discussion

Figure 1 shows the results for the three flavours of the *pen_state* model, broken down by the different input image combinations, between which there are little differences. Overall, the predicting the *is_present* variable is the most accurate (mean 93.5%, std 2.9%, n=70), followed by the predicting the *is_drawing* variable (mean 73.3%, std 5.8%, n=70) and the ternary *pen_state* activity variable (mean 68.2%, std 6.9%, n=70). This suggests that the models are better able to predict whether the artist is present than drawing. This makes sense when one considers how visually close the drawing and hovering state are, and that the pen tip is often visually occluded from the image view.

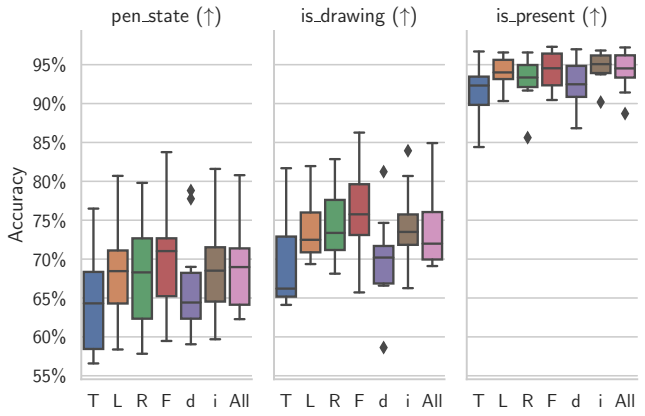

Fig. 1. Accuracy of *activity* models (*L-R*) *pen_state* (ternary), *is_drawing* (binary), is present (binary). Each error bar summarises 10 drawing sessions for the input images: (**T**)op, (**L**)eft, (**R**)ight, (**F**)ront (with (**d**)epth and (**i**)nfrared component), (**All**) six images combined as input.

Figure 2 shows the mean squared error (MSE) for the x and the y component of the *pen_position* model. Overall, the MSE for x (mean 0.0022, std 0.0064, n=70) was lower than y (mean 0.0032, std 0.0078, n=70). The combined MSE (mean 0.0054, std 0.0137, n=70), the model's training metric, was highest. There seems to be little difference amongst the RGB cameras (T, L, R, and F), while the individual depth (d) performs worse, and the individual infrared (i) has an outsized comparative variance. However, the combined images (All) yield a far better result than the individuals.

V. Summary and Future Work

We have shown that using a CNN architecture with camera images we can (1) predict activity and pen position across different artists and; (2) predict using all input sources and pairs of input sources. For predicting pen position, the combination of images performed better than individuals, where (surprisingly) they did not for predicting activity.

One limitation here is that each model is trained specifically for a user and drawing exercise. The rationale here would be that a co-creative system would train a model specific to an artist, and perhaps starting from a more general model. However, there are opportunities for future work in training

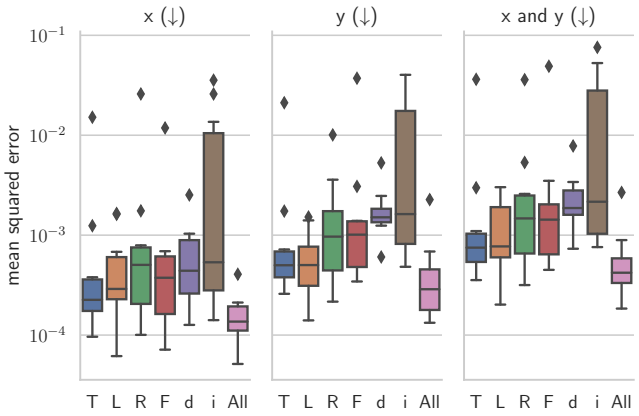

Fig. 2. Mean Squared Error (MSE) (log scale) *pen_position* models (*L-R*): x, y and combined $x + y$, Each error bar summarises 10 drawing sessions for the input images (same as Figure 1).

a generalised model as well as for *transfer learning* – how does one artist's model fair when evaluated on another artist's model? Or, how do models trained on observational drawing differ than that of the same artist drawing from imagination?

Further work in investigating advantages of different combinations of inputs are possible and would be more broadly applicable toward the human-robotic interaction research community, when considering the best view of a mobile camera for a robot when observing detailed human work, such as drawing, medical surgery or small electronics assembly.

In our work of human-robotic collaboration, our aim is for these models to contribute towards a framework for a co-creative drawing system, which is aware of the activity of the artist and what they are drawing based on visual input. Such a framework would benefit the co-creative computation community, and provide a basis to evaluate different co-creative approaches within the same context.

References

[1] C. Jansen and E. Sklar, "Exploring co-creative drawing workflows," *Frontiers in Robotics and AI*, vol. 8, p. 92, 2021.

[2] J. Jorge and F. Samavati, *Sketch-Based Interfaces and Modeling*. Springer Science & Business Media, Dec. 2010.

[3] D. Ha and D. Eck, "A Neural Representation of Sketch Drawings," *arXiv:1704.03477 [cs, stat]*, May 2017.

[4] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, and N. Fox-Gieg, "Quick, Draw!" https://quickdraw.withgoogle.com/, 2016.

[5] P. Karimi, M. L. Maher, N. Davis, and K. Grace, "Deep Learning in a Computational Model for Conceptual Shifts in a Co-Creative Design System," *arXiv:1906.10188 [cs, stat]*, Jun. 2019.

[6] J. E. Fan, M. Dinculescu, and D. Ha, "Collabdraw: An Environment for Collaborative Sketching with an Artificial Agent," in *C & C*, Jun. 2019, pp. 556–561.

[7] C. Oh, J. Song, J. Choi, S. Kim, S. Lee, and B. Suh, "I Lead, You Help but Only with Enough Details," in *CHI*, Apr. 2018, pp. 1–13.

[8] V. Cabannes, T. Kerdreux, L. Thiry, T. Campana, and C. Ferrandes, "Dialog on a canvas with a machine," *arXiv:1910.04386 [cs]*, Oct. 2019.

[9] P. Fernando, J. Weiler, S. Kuznetsov, and P. Turaga, "Tracking, Animating, and 3D Printing Elements of the Fine Arts Freehand Drawing Process," in *TEI*, 2018, pp. 555–561.

[10] C. Jansen and E. Sklar, "Co-creative Physical Drawing Systems," in *ICRA-X Robotic Art Program at ICRA*, 2019.

[11] ——, "Co-creative drawing with a robot," in *UK-RAS*, 2020.