STIVi: Turning Perspective Sketching Videos into Interactive Tutorials Supplemental Material

Category: Research

1 Implementation Details

1.1 Video processing

Removing the mouse pointer. Our video processing algorithm is inspired by related work on the analysis of timelapse paintings [1,3], which found that under the assumption of a static canvas, pixels that get covered in paint keep their intensity across subsequent frames. In contrast, pixels that are occluded by the drawing tool (mouse pointer in our case) exhibit rapid variations of intensity across frames. Following this observation, we filter out the mouse pointer by replacing the intensity of each pixel in each frame by the median intensity of that pixel in neighboring frames (we use a window of 11 frames centered on the frame of interest).

Extracting strokes. We extract pen strokes as they appear on canvas by computing the difference between each frame and its subsequent frame. To prevent false positives due to video compression artifacts, we only keep high-difference pixels that form short linear structures in the difference image, which we define as groups of pixels whose bounding box covers less than half of the frame, and whose bounding box diagonal is well covered by the pixels in the group (we rely on a user-provided threshold to define the minimum ratio of pixels covering the diagonal, which we adjust depending on the video quality and the drawing speed of the instructor).

Extracting planes. We detect planes in the drawing by searching for patterns of four intersecting lines made of two pairs of parallel lines converging to different vanishing points. However, some of these patterns might be due to lines that occlude each other in the drawing yet do not intersect over the intended 3D object (Figure 1). In the absence of a complete 3D reconstruction of the drawing, we heuristically filter out a significant portion of such false positives by only keeping planes that are not contained within a larger plane, and whose area exceeds a threshold (fixed to 2,000pxl in our implementation).

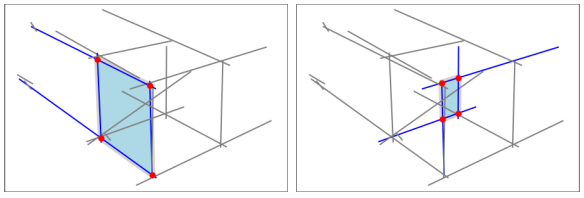

Figure 1: Extracting planes formed by pairs of parallel lines. A true positive case (left), and a false positive caused by occlusion (right).

Linking textual and visual elements. Figure 2 illustrates the interface we implemented to put the speech and sketch data in correspondence, where we display the visual and textual elements along a common timeline, we highlight the selected visual element over the drawing, and we represent its similarity with respect to textual elements as edges of varying opacity.

1.2 User interface

Rotating views. Our interface provides students the ability to visualize how parts of a drawing would look like under different viewpoints. For some scenarios, our solution does not permit to simultaneously translate the two vanishing points, e.g., to rotate all lines of a group around a common axis (Figure 3b). To deal with such situations, STIVi divides lines into two subgroups (left- vs. right-converging lines) and associates each subgroup with a different interaction widget (e.g., either the left or the right). It also allows students to simply click on the video window to display the vanishing lines associated with the group of highlighted elements (Figure 3a) and inspect how they converge, e.g., by dragging the pen to pan the scene.

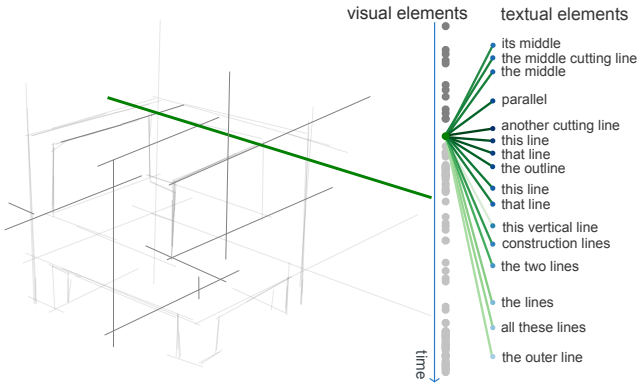

Figure 2: Visualization of visual-textual similarity. For the current visual element (highlighted in green in the drawing), we display similarity scores with all textual elements as edges of varying opacity. The rest of the extracted drawing is displayed in gray, with faint lines representing elements that have not been drawn yet.

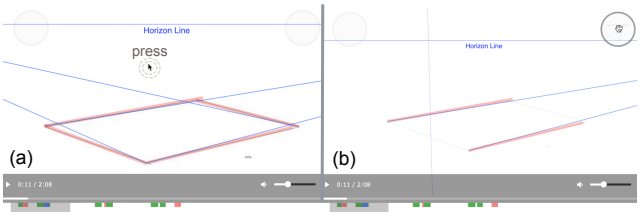

Figure 3: Assessing line convergence. (a) Pressing down the pen to display the vanishing lines over the sketch. Here, the perspective of the red lines converging to the left vanishing point is not accurate. (b) Interacting with the right circular widget makes left-converging lines disappear because our partial reconstruction approach does not allow for a consistent rotation around a common vertical axis in this case.

2 User Study

2.1 Methodological details

Background of participants. Among our 12 participants, one participant (P3) stated having a good grasp of perspective drawing, and five participants reported being able to follow some perspective rules. The remaining six participants had little to no experience with perspective. Ten participants had followed at least a small number of video tutorials in the past, while three of them were regular consumers of video tutorials across a range of topics such as

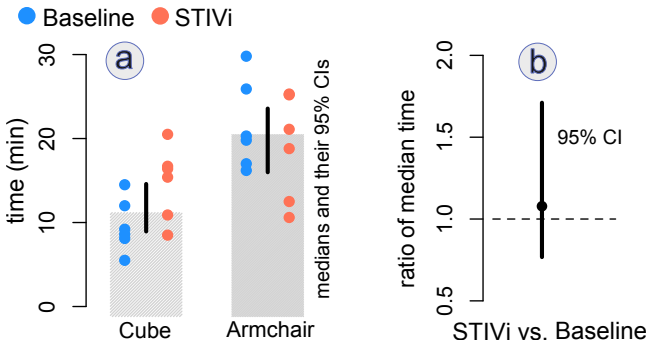

Figure 4: Additional results from the user study. (a) Time spent with each video tutorial and system configuration. We show CIs for medians, since our analysis is based on log-normal distributions. (b) Comparison of time (median) spent with STIVi versus time spent with BASELINE. A ratio equal to 1 indicates no difference.

knitting, drawing, photo editing, origami, and programming.

Apparatus. The participants interacted with a Wacom Cintiq 16 pen display (1920 × 1080 FHD resolution), which was connected to a Dell portable computer running Windows 10. The user interface was split into two windows (Google Chrome 111). The first window contained the video interface. The second window contained the canvas and basic drawing tools: brush sizes, a color palette, erasing and panning tools. It also provided functions for saving the workspace and activating the recording of interaction events (controlled by the experimenter). The above configuration allowed participants to conveniently size the drawing canvas and the video tutorial side by side.

Data collection. Participants were encouraged to think aloud during the tasks. Sessions were audio- and screen-recorded. We also recorded system events, which allowed us to observe how participants used the system functionality while completing the tasks.

2.2 Additional results

Time spent on video tutorials. We analyzed the time participants spent interacting with the video tutorials. This analysis was not planned and is largely exploratory, so we decided to not include it in the main paper. However, it provides additional intuition about the system use.

Participants spent on average 12.2 min (Median = 11.5 min, 95% CI [8.9, 14.6]) on the CUBE and 20.2 min (Median = 20.1 min, 95% CI [16.0, 23.6]) on the ARMCHAIR. We did not find any clear difference between the median time spent with STIVi and the median time spent with the BASELINE (see confidence interval in Figure 4b). However, we observe that participants spent considerably more time with the BASELINE when drawing the ARMCHAIR than when drawing the CUBE (see Figure 4a).

To further investigate this effect, we ran a mixed-design ANOVA, treating the SYSTEM (BASELINE vs. STIVi) as a repeated-measures factor and the GROUP of participants (starting with BASELINE vs. starting with STIVi) as a between-participants factor. Since we considered log-normal time distributions [2], we log-transformed all time values before analysis and based our inference on median values. Our results show no clear main effects for GROUP ($F_{1,10} = 3.33$, $p = .098$) and SYSTEM ($F_{1,10} = 2.23$, $p = .17$). In contrast, we find a strong interaction effect SYSTEM × GROUP: $F_{1,10} = 33.9$, $p = .00017$, partial $\eta^2 = .77$, 95% CI [.48, 1.]. Specifically, we observe that the increase of the time spent on the more complex model is less pronounced for STIVi than for the BASELINE. This result may suggest that exposure to STIVi's functionality has affected the way participants used the BASELINE. A possible explanation is that this group of participants spent additional time drawing in closer interaction with the video when using the BASELINE after STIVi. However, since we did not control for possible confounds, this result requires additional investigation.

Task evolution. Figure 5 present temporal patterns for all the 12 participants. Observe that logs for P4 and P12 when they used the BASELINE are not available due to technical issues.

REFERENCES

[1] C. Amati and G. J. Brostow. Modeling 2.5D Plants from Ink Paintings. In M. Alexa and E. Y.-L. Do, eds., *Eurographics Workshop on Sketch-Based Interfaces and Modeling*. The Eurographics Association, 2010. doi: 10.2312/SBM/SBM10/041-048

[2] E. Limpert, W. A. Stahel, and M. Abbt. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*, 51(5):341–352, 05 2001. doi: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

[3] J. Tan, M. Dvorožňák, D. Sýkora, and Y. Gingold. Decomposing time-lapse paintings into layers. *ACM Transactions on Graphics*, 34(4), 2015.

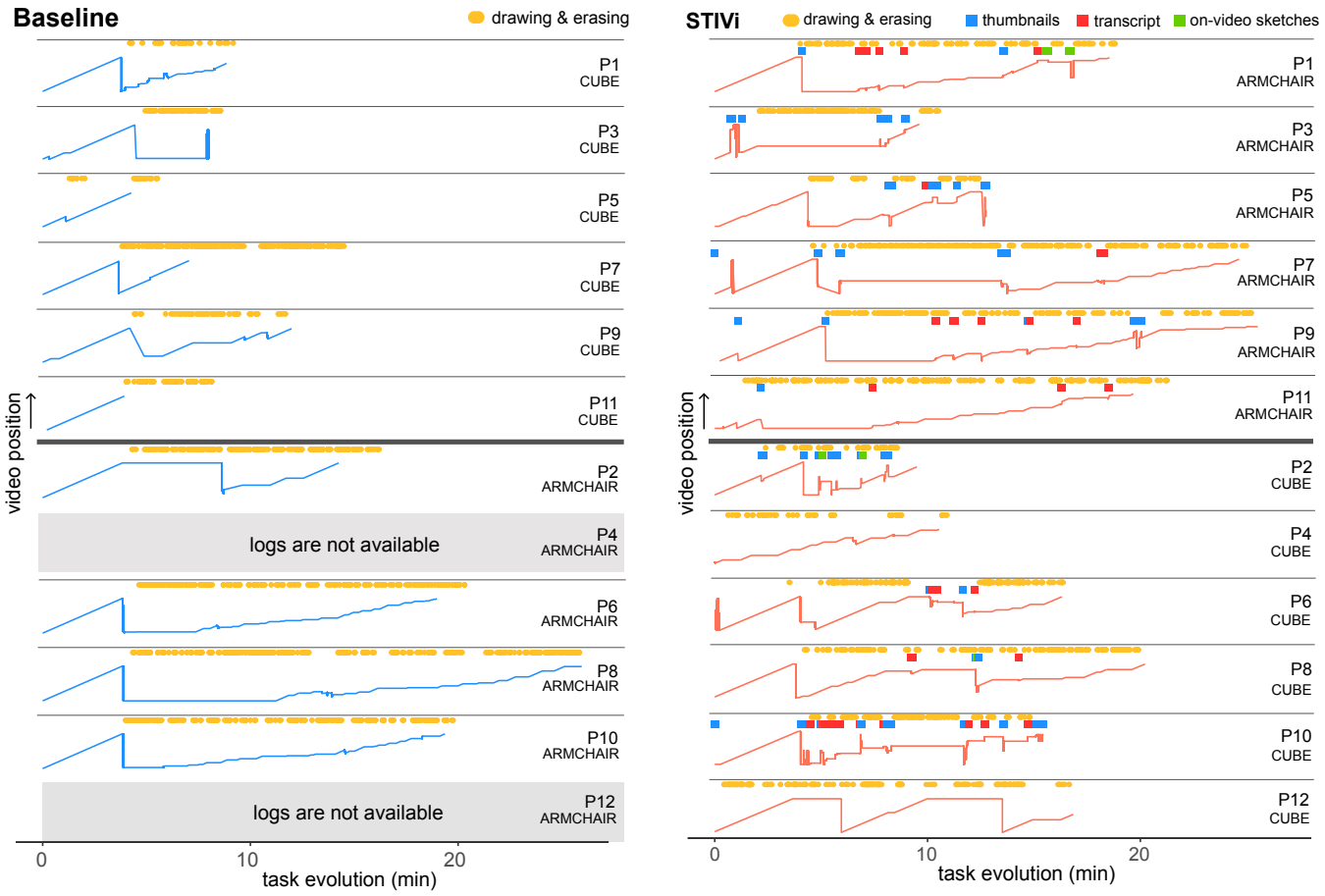

Figure 5: Evolution of the task for the 12 participants while interacting with BASELINE (left) and STIVi (right). The line trajectories show the time position on the video as a function of the time the participant spends on the task. The gray area corresponds to the initial phase of watching the video tutorial. For STIVi, we highlight user interaction events as yellow dots (drawing events) and colored squares (navigation events through thumbnails, transcript hyperlinks, and on-video sketches).