Evaluating Video Visualizations of Human Behavior

Mario Romero[1], Alice Vialard[2], John Peponis[2], John Stasko[1], Gregory Abowd[1]

[1] School of Interactive Computing
Georgia Institute of Technology
{mromero, abowd, staskso}@cc.gatech.edu

[2] School of Architecture
Georgia Institute of Technology
avialard3@gatech.edu, john.peponis@ca.gatech.edu

ABSTRACT

Previously, we presented *Viz-A-Vis*, a **VI**suali**Z**ation of **A**ctivity through computer **VIS**ion [17]. Viz-A-Vis visualizes behavior as aggregate motion over observation space. In this paper, we present two complementary user studies of Viz-A-Vis measuring its performance and discovery affordances. First, we present a controlled user study aimed at comparatively measuring behavioral analysis preference and performance for observation and search tasks. Second, we describe a study with architects measuring discovery affordances and potential impacts on their work practices. We conclude: 1) Viz-A-Vis significantly reduced search time; and 2) it increased the number and quality of insightful discoveries.

Author Keywords

Information Visualization, Video, Behavior, User Studies.

ACM Classification Keywords

H.5.2 **[Information Interfaces and Presentation]**: User Interfaces-Graphical user interfaces. Evaluation / Methods.

General Terms: Human Factors.

INTRODUCTION

Many disciplines spend considerable resources studying behavior. Methods range from qualitative pen-and-paper observation to automatic video content analysis. We present a semi-automated method where a network of overhead cameras captures behavior. The images are processed and visualized for rapid search and visual pattern analysis. Overhead video has the temporal and spatial resolution to potentially open new insights into everyday behavior by objectively revealing its invisible spatiotemporal structures. If analyzed thoroughly, it may function as a window into how people relate to each other and how they appropriate natural spaces and the objects within. Overhead video has potential for new analytical applications in multiple domains. For example, it may capture and evaluate the long-term effects of behavioral therapy in especial classrooms. It may track developmental progress in a baby's nursery. It may provide objective, long-term, and continuous physical therapy reports in natural places beyond the doctor's office. It may trace factory operations to increase industrial productivity. It may uncover subtle customer behaviors to boost retail space marketability. In this paper, we explore two domains: Behavioral Analysis and Architecture. Behavioral analysts track the topography (physicality and context) and the function (goal) of target human behaviors [8]. Architects analyze the relationship between the environment and people's behavior to evaluate designs and gain lessons for theory [16].

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Previously, we developed *Viz-A-Vis*, a **VI**suali**Z**ation of **A**ctivity through computer **VIS**ion [17]. Viz-A-Vis captures behavior using overhead cameras, it processes the video with simple and robust computer vision, and it visualizes behavior as aggregate motion over the places of observation. Video Figure 1 demonstrates Viz-A-Vis.

Here, we evaluated Viz-A-Vis through two complementary user studies. A *performance* study measured its low-level usability and a *discovery* study measured its impact on high-level analysis. The performance study compared task-based user preference and performance against two systems. It determined that the tool is superior for some of the most critical tasks of behavior analysis. More importantly, it set a foundation that simplified the discovery study, where we did not test low-level usability. The discovery study reports Viz-A-Vis's clear positive impact on the practices of a group of architects, including increased opportunities for the discovery of actionable insights.

Additionally, we briefly discuss our lessons learned in evaluation design. While measuring performance in the laboratory is a bounded effort, the typical field study of a system's impact is not. We argue that our two-part evaluation may approximate the findings of a field study.

This paper's sections present related work in visualizations and evaluations, Viz-A-Vis's system architecture, the performance study, the discovery study, a discussion on the evaluation design, and its conclusions and future work.

RELATED WORK

Video Visualizations of Behavior

The first image sequences visualizing action and behavior are the beautifully pioneering photographs of Muybridge and Marey from the 1880s [10]. The first 3D space-time representation of a video cube (VC) is the 1970s work on motion by Ullman [20]. Fels et al. were the first to describe interactive cutting planes for visually filtering a VC [7].

Daniel and Chen present one of the first abstract visualizations of behavior in video [5]. They visualize

motion in a translucent space-time cube by mapping greater motion to greater opaqueness, thus enabling an operator to see through inactive regions. Ivanov et al. present a visualization of the history of living spaces [9]. The authors provide 2D visualizations of motion sensor and raw video data. Through motion detection they visualize contextual paths and provide detail through strategic camera views. Botchen et al. present a 2D time lapse video visualization with highlighted abstractions of target objects and activities [1]. We propose similar goals and techniques to these papers, except our video has a near one-to-one correspondence with architectural space that naturally supports space-centric queries.

TotalRecall visualizes long-term video from real environments [13]. The main difference from Viz-A-Vis is that TotalRecall visualizes video in a 2D representation that introduces ambiguity between time and space. It slides frames like cards spread out from a deck. The visual effect is that each 2D location in the visualization is an ambiguous combination of multiple spatiotemporal coordinates.

Due to our image-to-space correspondence, we were inspired by GeoTime [11], which vertically maps temporal data as linear paths above a 2D geography. However, unlike GeoTime, Viz-A-Vis visualizes dense 2D layers of activity over 3D space. While the visualization is more challenging, the result is a more thorough view of activity across the entire space for each time frame.

Evaluations of Video Visualization Systems

While the number of video visualization systems is considerable, there are alarmingly few rigorous evaluations. Daniel and Chen's work has a follow-up publication that describes a rigorous study validating very specific usability claims of visual signatures [4]. Chen et al. argue that video analysis without human input is impossible for unbounded sequences and that a human must be in the loop of decision making. The role of video visualization is to fill in the gap between vast data sets that humans cannot practically search linearly and automation that is not computationally tractable. By placing the human in a critical role, the authors recognize the intrinsic need of user studies for video visualizations. In their study, the authors use computer graphics to carefully synthesize a clean video for evaluation that only models translations of one sphere. While this study rigorously answers questions about users' ability to interpret the visual signatures of the synthetic video, its level of artificiality fails to answer the ecologically-valid questions raised by the complexities of real data and tasks. In both of our studies we provided participants with real data and ecologically valid tasks.

Wang et al. developed a spatially contextual video representation that was based on requirements gathering and on understanding current security operator tasks [21]. They conclude with an informal user study based on tasks and usage patterns. In a follow-up, Wang et al. present a rigorous user study comparing performance through path

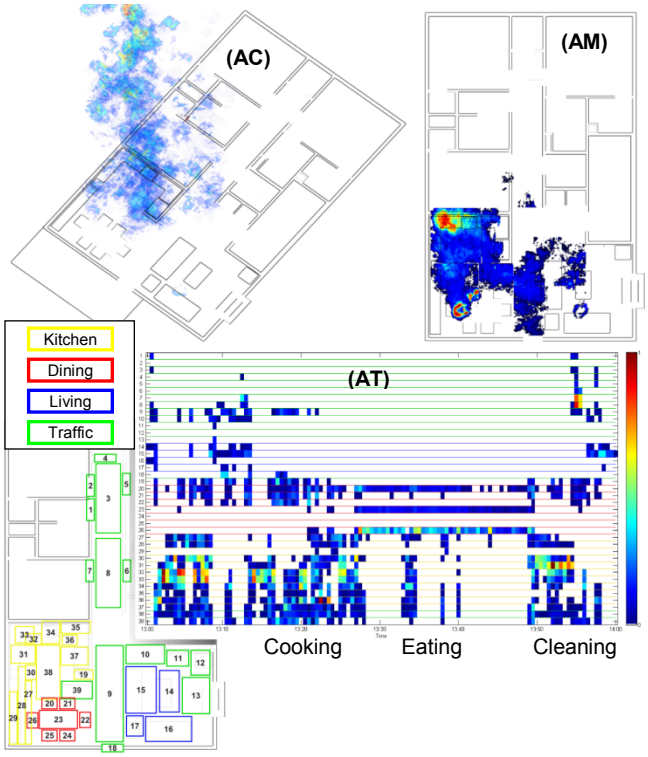
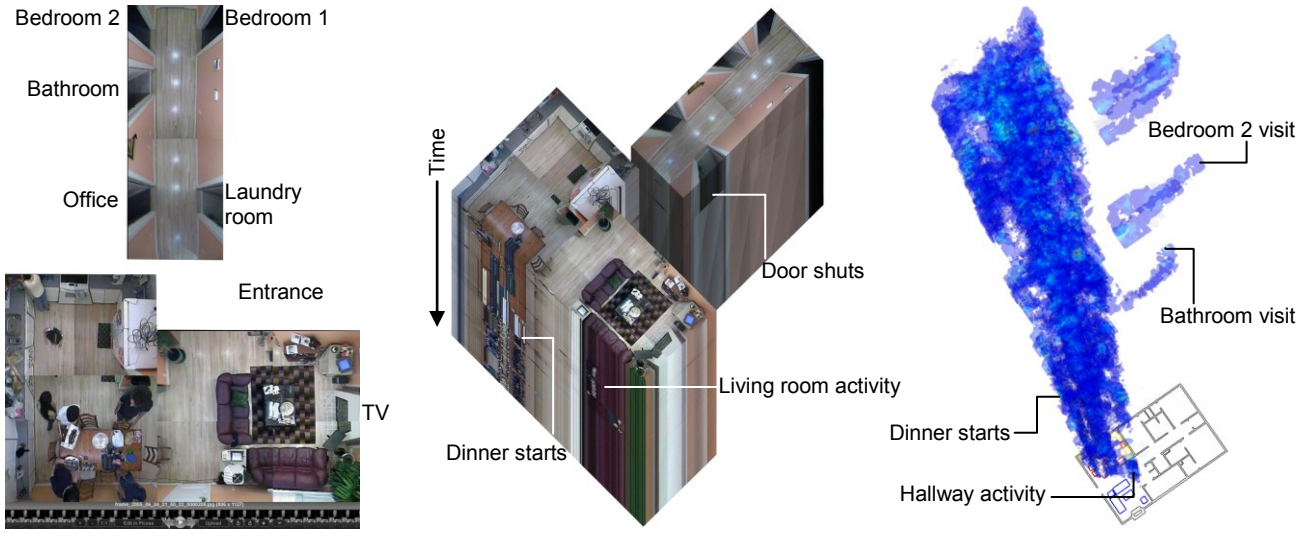


Figure 1. Viz-A-Vis visualizing two people cooking and eating on the activity cube (AC), map (AM), and table (AT).

reconstruction tasks [22]. They compare two contextualized video design factors and two levels of knowledge in participants. We gathered our requirements and tasks from interviews with domain experts, both in Behavioral Analysis and Architecture, and from the domain literature [8, 16]. Also, we trained participants until they self-reported proficiency in 3D navigation and filtering. Finally, we compared three experimental conditions.

Our performance study measured the user's preference and performance through time-to-task completion, precision, recall, coverage, and exit surveys. Numerous authors have proposed similar methods for evaluating information visualizations [2, 3]. In particular, Plaisant categorizes the types of evaluations based on the tasks, users, and goals [15]. Our performance study is an instance of a laboratory experiment comparing three tools: 1) the commonplace – a video player (VP); 2) the state-of-the-art – a video cube (VC); and 3) our experimental prototype and the central element of Viz-A-Vis – the *activity cube* (AC). Plaisant characterizes the fundamental problem of matching tasks, tools, users, and relevant high-level goals. Furthermore, her recognition that discovery requires real expertise, needs, context, and prolonged exposure to occur is central to the design of the two studies. There are a number of discovery-focused field studies that perform the costly evaluations we approximate [6, 18, 19]. In the performance study, participants execute predetermined tasks with correct answers. In the discovery study, users pose and answer novel questions creating a discovery loop. The discovery study goes deeper into questions of analytic insight.



(a) Video Player (VP) (b) Video Cube (VC) (c) Activity Cube (AC)

Figure 2. Three experimental conditions visualizing behavior from overhead video mapped onto architectural space.

VIZ-A-VIS SYSTEM ARCHITECTURE

Viz-A-Vis stands for **VI**sual**Z**ation of **A**ctivity through computer **VIS**ion [17]. Primarily, it is a capture-and-access system for the analysis of human behavior. In this paper's instantiation, the capture module is a synchronized network of overhead cameras that provide full coverage over the kitchen, dining room, living room, and hallways of the Aware Home [12]. Each frame results from un-warping, scaling, translating, rotating, stitching, and cropping overhead images to match pixels with locations. The result resembles a single orthographic projection (see Figure 2a). Ten wide-angle cameras collect 24-bit color, 320-by-240-pixel JPEG files at 2 Hertz. We continually captured over 200 hours in the life of a married couple and their guests.

Viz-A-Vis combines 2D and 3D histograms and heat maps of aggregate motion (see Figure 1). We compute motion through frame differencing and we aggregate it over regions and periods of interest. The main overview structure is the *activity cube* (AC). AC is a volumetric geographic information system (GIS), where the geography is the home's floor plan. Time maps to the vertical axis. The heat maps doubly-encode aggregate motion into color and translucency – the more opaque and red the area, the more active. Users navigate and filter AC by rotating, zooming, and translating the cube and by placing cuts along the sub-volumes of interest, respectively. The cube includes an index to the original frames that allows the user to reify hypotheses about its visible patterns of behavior.

The *activity map* (AM) is a 2D aggregate summary of the activity cube. The user defines a period of aggregation from pre-computed aggregates, from a few seconds to a few hours. Also, the user can zoom and translate the map.

The *activity table* (AT) is a 2D array of aggregate motion across space (rows) and time (columns). In this version of Viz-A-Vis, the system presents manually pre-define regions of interest and aggregates motion over the regions across a window of time. A future version of Viz-A-Vis will allow dynamic region definition by the user and automatic region definition by the system. A cell on the table holds the value of the spatiotemporal aggregate of motion, which maps to a 2D color histogram equivalent to the heat maps (greater motion maps to red). The user can zoom and filter AT and index original frame sequences in the video. We implemented the backend of Viz-A-Vis in C++ and Matlab and the frontend in Ruby as a plug-in for Google Sketchup.

CONTROLLED LABORATORY PERFORMANCE STUDY

Preference and Performance Study Design

Our research question is: what are the task-based user preference and performance operating the Activity Cube (AC) compared to a video player (VP) and a video cube (VC) as measured by exit surveys, time-to-task-completion, precision, recall, and coverage? To answer it, we designed a counterbalanced-order, within-subject user study. We intentionally simplified this user study by evaluating only the activity cube and not the table (AT) or map (AM). First, it is a natural progression to go from a video player, which uses time to view time, to a video cube, which uses space to view time, to an activity cube. The activity cube also uses space to view time, but its view of activity goes deeper into the cube at a loss of detail. Second, training users to understand and operate the activity table and map would have tripled the resources necessary for this study, without much further insight into Viz-A-Vis's usability. Finally, the activity table and map are not as natural progressions from the video player and cube as the activity cube is.

We recruited 24 participants (18 male, 22.9 average age) with normal vision from two classes, HCI and CogSci, where they received 1% extra credit on the final grade as compensation. Given the within-subject design, we measure 24 data points per condition-task pair. Through an initial survey, we determined that most participants were

computer scientists and considered themselves experts at interfaces (some at 3D navigation), good at programming, and experienced with data analysis and visualizations. On the other hand, most participants had never analyzed behavior and had no experience with Picasa or Sketchup.

Condition 1, VP, provides standard video playback functionality. We use Google Picasa Image Viewer to browse the raw JPEG frames (see Figure 2a). Condition 2, VC, provides a 3D structure of frames across time with interactive cutting surfaces to remove occluding volumes and standard 3D navigation tools (see Figure 2b). Condition 3, AC, provides the same 3D structure and interaction model, except it visualizes a stack of translucent heat maps of aggregate motion (see Figure 2c).

We evaluated the three conditions in counterbalanced order for each participant during three one-hour sessions on separate days. For each condition, participants trained until they self-determined proficiency. Training times varied across conditions. On average, the training required for VP was 3 minutes, for VC, 18 minutes, and for AC, 23 minutes. We placed an upper time limit on tasks and most participants completed them before reaching the limit.

We conducted this study in a usability laboratory. The computer had two 19-inch monitors, a 2.4 GHz Intel Core 2 CPU, 4 GB of RAM, and a necessary NVIDIA GeForce GTX 280 GPU for the visualizations to flow without lag.

We collected a dataset ripe with target events for this study during a four-hour dinner party. Eight friends in their 30s prepared food, had dinner, cleaned up, and played a board game (see Figure 6). There were 3 married couples, 2 single males, 7 Latin Americans, and 1 American. The first author and his wife hosted. All signed consents and were aware of the recording. We stated our goal: "to visualize natural human behavior." We asked them to act naturally, which they did within a few minutes. We purposefully included two activities into the soirée: a raclette and a game of Cranium™. A raclette is an electric grill surrounded by raw ingredients at the table and people cook their own meal. Cranium is a board game where two teams compete by performing a number of tasks, some very physical (acting, sculpting, and drawing) and some not (spelling backwards).

We carefully split the data into three scenes and showed a different scene during each experimental condition to avoid data learning effects. The scenes contained equivalent targets for each condition. We always presented the scenes in chronological order, regardless of the condition. Scene one, presented on the first session of participation, includes arriving, preparing dinner, setting the table, and starting the raclette. Scene two includes ending the raclette, cleaning up, and preparing and eating dessert. Scene three includes ending dessert, cleaning, moving to the living room, and starting Cranium. All scenes include bathroom visits.

While behavioral analysis tasks routinely include high-level statistical comparisons, for instance, they also include low-level tasks. Our study focuses on nine typical low-level, evidence-gathering tasks of behavior observation:

- **Interacting** is operating the application's low-level controls (clicking & dragging, filtering, navigating).
- **Overviewing** is verbalizing a shallow narrative of behavior and its context across an entire dataset.
- **Describing** is verbalizing the details and context of the behavior of all subjects during a target event.
- **Tracking** is following the location and describing the actions of one target subject during a target event.
- **Searching** is spatiotemporally locating sporadic and brief target behaviors and events.
- **Counting** is enumerating the repetitions of recurrent and brief target behaviors and events.
- **Finding transitions** is locating the periods where the entire group switches between activities.
- **Short-bounding** is finding the tight spatiotemporal boundaries of activities lasting a few seconds.
- **Long-bounding** is finding the tight spatiotemporal boundaries of activities lasting minutes or hours.

For all conditions, we carefully presented each task through a script that clearly defined it, provided examples, set a time limit, and asked participants if they had any questions. We also invited participants to formulate a strategy before starting the task in order to model expert users.

For all datasets, participants overviewed, described, tracked, short-bounded, and found transitions by choosing their own targets. We tasked participants to search for bathroom visits in all datasets. We asked participants to count raclette reaches, ice scream spoonfuls, and game board reaches in the first, second, and third dataset, respectively. Finally, we tasked participants with long-bounding dinner, dessert, and game play. For this task, we asked participant to define the boundaries. Through pilots, we determined, for example, that dinner starts for some participants when all are at the table and, for others, when someone starts eating. We needed a concrete a-priori definition to consistently measure performance, yet we wanted to observe the process of defining very concrete boundaries and let the users experience it as well.

We measured preference through an exit survey. We asked participants to rank the three conditions based on how well they support each task. We also asked them to design a hypothetical analysis system for an airport where the goal is to understand typical behavior and learn to discover outlier behavior. The design had to be based on at least one of the conditions and at most be any combination of the three.

We measured performance through time-to-task-completion (TTC), precision, recall, and coverage. *TTC* is a bounded period between the start and end of a task, including repetition until user satisfaction. *Precision* is the percentage of correct targets in the set of retrieved items. *Recall* is the percentage of retrieved targets from the set of possible targets. *Coverage* is the percentage of the dataset reviewed.

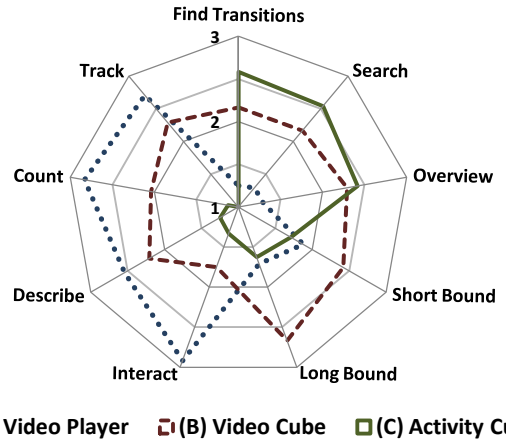


Figure 3. Task-based user preference: VP, VC, and AC.

Not every task lends itself to all measures of performance. Interacting, overviewing, describing, tracking, and finding transitions present subjective and variable definitions of the quality of the results, thus precision and recall do not apply.

For analysis, we summarized the data as mean ± standard error. We conducted a one-way repeated measures analysis of variance (ANOVA). We used the Tukey test to conduct pair-wise comparisons between conditions and considered differences at $p < 0.05$ to be statistically significant.

Preference and Performance Study Results

We present the preference results first and we use these to frame the performance results. The target users of Viz-A-Vis are expert analysts and their preferences are paramount to the success of the tool. Figure 3 presents a radar plot that visualizes the average of the 24 participants' ranking of the three conditions across the nine tasks. The radial scale of the graph goes from 1 to 3, where 3 is preferred. We sorted the plot clockwise in decreasing preference for AC. The first observation is that there is a clear complement between VP and AC, except for bounding, where VC is preferred. Second, participants preferred some conditions for certain tasks: VP – tracking, counting, describing, and interacting; VC – long and short bounding; AC – finding transitions, searching, and overviewing. We expected most of these results (tracking, counting, describing, and interacting). It is clear that VP is not only simpler to use, but actually required for its detailed and controlled video traversal. We expected AC to outperform in the other tasks, including bounding. Though there was the extra cost of performing cuts, participants preferred VC for bounding because it unambiguously visualized activity boundaries. With AC, users were not sure they could clearly interpret boundaries.

Given the design of the study and its metrics of performance, it is possible to compute performance for counting, long-bounding, and searching only. Since people simply could not count with AC, because long-bounding was difficult, and in the interest of space, we only present the statistical analysis of searching performance.

We analyze the results visualized in Figure 4. With

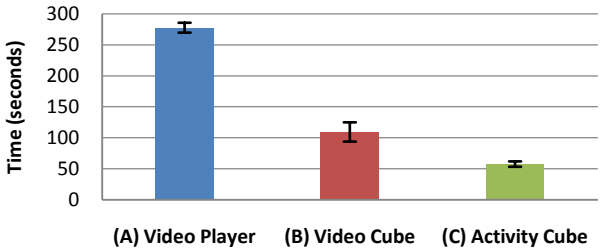


Figure 4. Search time-to-task completion: VP, VC, and AC.

statistical significance ($p<0.01$), AC's average time to *search* completion (57 seconds) outperformed the video player (278 seconds) by nearly 5-to-1 and the video cube (110 seconds) by nearly 2-to-1, while maintaining precision and recall at 100% and increasing coverage from VP's 78% to 100%. This is particularly relevant since, according to our interviews with professional behaviorists, their most time-consuming task is searching. Furthermore, since we limited task execution time to five minutes, we restricted the improvement factor. Without restriction, we extrapolate the improvement to be 8-to-1. Moreover, as 3D navigation and interpretation proficiency increase and as sequences lengthen, the improvement factor may grow further.

As a dramatic example, one participant, an extreme outlier, searched in 2 seconds. He orbited AC to its side, detected the relevant patterns, pointed to each target event, and said: "there!" His two-second search of random targets buried in over 7000 frames measured 100% precision, recall, and coverage, while managing to avoid two subtle false positives – an improvement factor of 139!

Finally, we present the results of the hypothetical design question and relevant comments. First, AC was the only condition unanimously chosen, always in complement with VP or VC, though. Users cited overviewing and searching for outlier behavior as the primary tasks of AC. Also, participants stated that AC could help learning the shape of normal patterns, thus outliers would easily stand out. Participants highlighted the importance of privacy in designing behavior capture systems but noted the tradeoff between security and privacy, for instance. They volunteered a number of applications: tracking disabilities in the home, monitoring child development with a baby-cam, observing social behavior for gender studies, tracking behavioral changes in children with autism in classrooms, performing ubiquitous computing and augmented reality studies, where physically observable behavior is part of systems' experience, and studying body language in group dynamics. To finish, all twenty-four participants expressed admiration and found value in both 3D models:

> *Wow! I get the illusion that I see the scene from different perspectives. It feels like I'm moving the [capture] camera! Although I know it's not true, I feel I can see faces better when I look from the side [of VC].*
>
> *The activity clouds show where the action is and the type of action by the amount of activity. You lose who is doing what, but you get to see longer periods of time and where things happened.*

DESIGN STUDIO & FOCUS GROUP DISCOVERY STUDY

Discovery Study Design

Our research question is: can the visualization of behavior raise opportunities for discovery and change work practices and outcomes for a domain-specific application? To answer it, we devised a two-group design studio and focus group with architects. The control group used current methods to inform their design and the experimental group augmented those practices with Viz-A-Vis. They viewed the activity cube (AC), the activity table (AT), the activity map (AM), and indexed original frames with the video player (VP).

Environmental psychologists are architects who formulate design choices through the systematic study of the relationships between space and behavior. Their data gathering and analysis methods are arduous. For example, architects gather flow and occupancy by observing and manually counting or by interviewing and surveying.

During a design studio, we observed two groups consisting of five and six doctoral architecture students each. Their task was to renovate the interior public spaces of the Aware Home given a number of constraints and requirements as stipulated in writing and verbally by fictional clients. Each architect worked individually, but shared the work space, the delivery of the requirements, and the clients' answers to the questions posed by other architects in the same group.

The study had two sessions on separate days for each group. The first session was a five-hour design studio. The second session was a two-hour focus group. The design and the focus group sessions took place, respectively, in the dining room and the living room of the Aware Home.

The design studios consisted of the delivery of the design program, the fictional clients' requirements statement, questions from the architects, sketching, a second round of questions, refinements, and the architects' presentation of their designs. For the experimental group the presentation of the requirements and current patterns integrated Viz-A-Vis visualizations. The client requirements included supporting a mutual sense of presence during parallel activities and providing space for entertaining friends, shelving books, watching movies, and listening to music.

From the start, both groups were aware of the general goal of the study: "to understand your current design practices and to determine the efficacy of a software tool aimed at supporting part of those practices." The control group was aware of the existence of the tool and they knew they would not see it until the focus group, where we showed them a number of episodes from daily living in the home and asked them to relate the visualizations back to their original design. We also motivated them to project how they could use the visual data in future designs.

We started the experimental group with a presentation and discussion of the system. We visualized a number of episodes from the everyday life of the fictional client occupying the home during a period of nine days and asked the participants to input queries into the system, for example, "what does typical cooking look like?" Figure 1 shows the result of this query with some context around it. Notice in AM the regions of highest activity around the kitchen and in AT, the period of dispersed activity, cooking, focused activity, eating, and dispersed again, cleaning.

Participants asked questions that would then be answered with visualizations. We displayed the results of the queries to all participants and let them verbally guide the interactive views, allowing them to interpret the data. To sidestep training the participants and to exert a uniform impact, we delivered the queries through a dedicated technician instead of hands-on participant interaction. We were not testing the controls of the interface in this study. Rather, we tested whether participants could interpret and utilize the visualization to support their design task.

The experimental group had equal time limits to complete their design and shared the same deliverables. We presented to the experimental group the results of the individual queries mid-way through their design and we collected their deliverables at the end. On a separate day, we conducted a focus group with emphasis on what worked, what did not work, what influenced their design, what was missing from the tool, and how they could use it to inform future designs.

We observed, recorded, and transcribed the design studios and the focus groups. We collected questions, comments, suggestions, and critiques, as well as the presentations of their design in visual, verbal, and textual media.

It is important to expose a potentially confounding factor in our study. The first author played four roles during the design studio and one more role during the focus group. First, he created the system. We did not hide this fact in order to motivate the participants by providing them with the real opportunity to have impact on the tool. Second, the first author and his wife played the fictional clients. They lived in the home and recorded the nine days of activity visualized during the study. We modeled the fictional clients' behavior closely based on the real life behavior of the couple. Third, he was part of the team observing the architects during their practices. The observation included taking notes, photographs, and video recording. It did not include questions during the design studio. During the presentation of the designs of the architects, the author played both the role of the client and the role of the observer when asking questions. Fourth, for the experimental group, the author played the role of the technician. He collected the queries, asked enough questions to eliminate any ambiguity, executed the queries, and presented the results being careful not to interpret them. Finally, the first author also moderated the focus groups.

To mitigate the impact of these factors, we took a number of steps. First, the study included five observers, three of whom are professional architects. Second, we carefully modeled and practiced playing the clients in order to deliver exactly the same descriptions and return equivalent answers

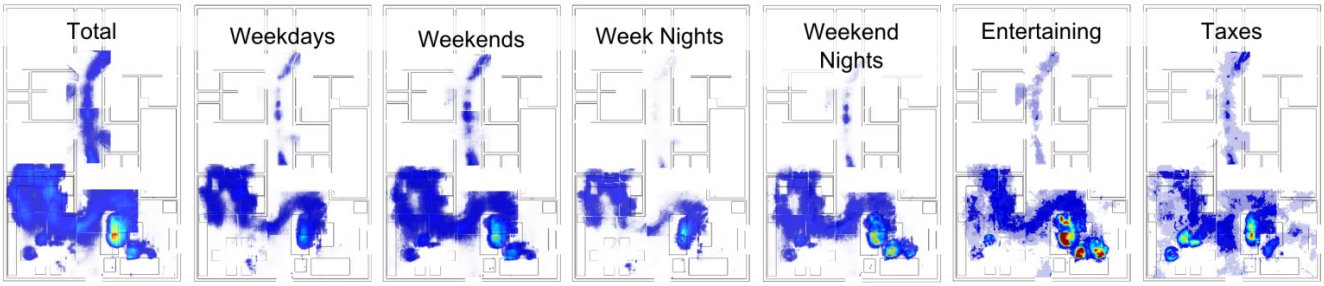


Figure 5. Activity maps showing aggregate motion heat maps presented as examples of everyday living episodes.

to similar questions. Third, we carefully controlled the technician's role. His task was only to deliver the results of the query. We avoided including behavioral interpretations of the results. Fourth, we rapidly established an amicable environment where we constantly encouraged criticism.

The two groups' previous design experience differed. While the control group had a 10-year design experience on average, the experimental group only had a 5-year experience. Participants were randomly divided into the groups based on their availability. While this difference is significant, we were less concerned about its impact because it was the control group with more experience. The experimental group, if anything, was at a disadvantage.

We defined the same task and schedule for both groups. They were in charge of renovating the kitchen, dining room, living room, foyer, media closet, coat closet, south end of the main corridor, and balcony. Both groups had 30 minutes for initial data gathering, 120 minutes for initial sketches, 15 minutes for further data gathering, 60 minutes for final sketches and presentation material, and 5 minutes per architect for the presentation of the final design. The total running time for the control design studio was 4 hours and 20 minutes and for the experimental design studio, it was 4 hours and 50 minutes. The extra time of the experimental group was due to the additional architect and the 25-minute presentation of the system at the beginning of the session. For the data gathering sessions, we balanced the time of showing query results with the time of clients delivering their verbal accounts of their lifestyle. We kept it in the same time limits of 30 and 15 minutes each.

We observed the practices and evaluated the product of design employing a technique called *architectural moves*, which analyzes the design's impact on the elements, features, and programs in the layout. A program is the set of intended uses of a space together with the architectural affordances. During the design studio, we observed participant questions, comments, critiques, descriptions, and presentations of their designs.

During the focus groups, we collected the architects' reflective evaluation based on any new information provided by Viz-A-Vis, their interpretations and use, if any, of the visualizations, their critiques of the technology, and proposed future improvements and applications. We used focused coding for the analysis of the results [14].

Discovery Study Results

We present the results through five themes: 1) discovery of patterns of behavior; 2) architectural moves; 3) creation of a new spatiotemporal ontology of behavior; 4) creation of behavioral design sketches; and 5) comments and critiques.

The first theme is *discovery of patterns of behavior*. Figure 5 shows activity maps summarizing behavior across multiple days and events. We presented these samples of daily living to the experimental group at the start of their design studio and to the control group during their focus group. Both groups discovered a number of behavioral patterns, some of which the clients were not aware of. In the interest of space, we present the most striking pattern.

During the control focus group discussion, one of the architects (A) remarked: "[the clients] seem to be introverted." The moderator, who was also one of the clients and who did not believe it, replied: "What do you mean?" A: "Well, [the clients] always stay away from the windows [pointing at Figure 5]. When I'm at my house, I like to have coffee by the window and watch the world outside." After an extended discussion and analysis of the evidence, we concluded that the clients were not introverted. They were avoiding Atlanta's 10th Street, which is crowded, polluted, noisy, and public. Living in this home was different from their regular home, which bordered on the Chattahoochee River National Park. There, the clients would spend many hours by the windows. But at the Aware Home, the clients avoided the outside at all cost. The most striking aspect of this discovery is that the clients were not aware of this behavior – it simply happened. It required an extended discussion grounded on the objective evidence provided by the visualization to arrive at this conclusion.

The second theme, *architectural moves*, refers to the designed changes in form or function of the architectural space. We synthesized five types of architectural moves from the eleven designs: 1) the inclusion of the balcony into the indoor space; 2) the creation of a foyer, an entrance; 3) the establishment of visual links between the public spaces; 4) the bounding of spaces with half walls or furniture; and 5) the creation of a space solely dedicated to media consumption. Between the control and the experimental group, we only observed a significant difference in the fifth architectural move, the creation of media spaces. All the other four architectural moves had roughly the same number of instantiations for both groups.

Figure 6. Architects used this activity table (AT) of a dinner party to create a spatiotemporal ontology of behavior.

While none of the more experienced architects in the control group created a dedicated media space, four out of the six architects in the experimental group created it. The experimental group discovered and used a behavioral pattern of extreme media consumption in the "Taxes" visualization on the right-hand side of Figure 5. The activity map depicts the clients preparing their tax returns. It shows activity in the living room and in the dining room. After an inquiry from the architects, the wife explained that they started filing their taxes electronically in the living room. When they attempted to electronically submit the return, the "free" service charged a $50 fee. The clients moved into the dining room to redo their returns on paper. The unexpected behavior, visible in the map, is that the clients sat on the far side of the table. After prompting from the architects, the wife answered that they were watching Spiderman 2 and they needed to sit on the far left to continue to view the television on top of the fireplace on the far right.

The third theme is the *creation of a new spatiotemporal ontology of behavior*. Figure 6 presents an activity table with the data presented in the controlled performance study. The clients invited a group of six friends to a dinner party. We presented the table to the control group during their focus group. We did not label the activities, yet the architects were able to describe the sequence of events. They found the table very insightful. It allowed them to quickly understand "density of occupancy" and "patterns of flow." As they discussed the image, they started creating a new vocabulary to describe behaviors in relation to space and time. One of the architects stood up and, in front of the screen, stated: "these bathroom visits are punctual over space and time, these periods of transition are distributed over space and punctual over time, this socializing is punctual over space and distributed over time, and this preparing dinner is distributed over space and time." She outlined four broad behavioral categories for the use of space and time and created a spatiotemporal ontology of behavior and a new vocabulary to describe it. After further discussion, the group's senior architect agreed that this new ontology was worth exploring in Architectural Theory. He imagined an example where a museum curator would be interested in distinguishing between three patterns of patron behavior: 1) "translation," going from one place to another in the museum; 2) "vibration," staying in one place but moving a lot, like during a conversation; and 3) "contemplation," remaining relatively static and contemplating a compositional space within the museum where several exhibition pieces create one visual statement.

The fourth theme is the *creation of behavioral sketches*. The architects in the control group did not create any type of sketches that outlined behavior over space. During their focus group discussion, we established that it is not part of their practice. On the other hand, two of the six architects in the experimental group created sketches that depicted a model of behavior the architects had extracted from the visualizations (see Figure 7). When we queried them, they explained that they analyzed the relationship between behavior and space, abstracted some patterns from the old space, and instantiated the abstractions into their designs. In other words, they ran a thought simulation partly motivated by the data driven visualizations. The sketches on figure 7 depict lines of communication from particular points in the floor plan. On the left, yellow depicts inward communication and red depicts outward communication.

The fifth theme compiles the most relevant *comments and critiques* from the architects. First, the architects found the activity table and the activity maps more useful to their analysis than the activity cube. They had a hard time visualizing a summary of activity from the 3D structure. We did not expect it, but in retrospect it is clear. Architects are not as interested in the sequence of events as behaviorists are. Architects focus on the event-based relationship between space and time. The 2D

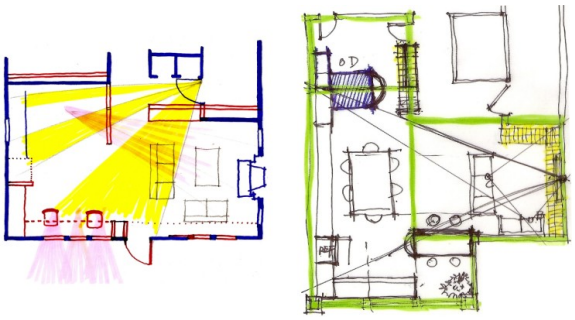


Figure 7. Behavioral sketches for architectural design.

representations provided these summaries clearly and succinctly. One caveat is that they did not have to search for the temporal windows of aggregation. The interaction with the activity cube is crucial for these searches.

Second, architects stated that visualizing household activity was not the most justified use of the powerful tools provided by Viz-A-Vis. A house, in their opinion, is a place of relative simplicity, where at least one dweller understands the overall pattern of occupancy. They stated that more complex environments, where no single individual understands the overall patterns of activity, would dramatically highlight the virtues of our system and would undoubtedly have an impact on the theory, the practice, and the product of Architecture. They volunteered a number of complex spaces for using the tool at its potential: nurses' desks, hospitals, plazas, museums, ground and air terminals, and public transportation lines.

Third, they stated that it would be of great benefit to their practice to include identity in the visualization. Viz-A-Vis is intentionally simple; it only visualizes motion. Individuals are indistinguishable in this visualization. A reification step, indexing original frames, is necessary to understand individual behavior. Blob tracking is considerably more complex and less reliable than motion aggregation and we purposefully avoided it for this stage of our research, but the point is well taken.

DISCUSSION

We did not set out to contribute to evaluation methodologies of visualizations. Nevertheless, we learned a number of lessons that we consider can be applied to future evaluations and we report these here.

The ultimate goal of visualizations is to promote discoveries that support actionable insights. This is difficult to evaluate. It typically requires long-term field studies that determine current practices and products and the visualization's impact on both. The studies occur in the workplace with relevant datasets. Participants typically require extensive training and monetary incentives. Field deployments consume thousands of human-hours [18, 19].

Through our performance and discovery studies, we informally approximate the results of a field study at a fraction of the cost. In the lab, we test the low-level usability performance of the system compared to the state-of-the-art and the commonplace: a three-condition, within-subject, counterbalanced-order study. The performance study consumed approximately 80 hours for testing and 140 hours for analysis. Testing includes one participant and one researcher, thus its total cost is roughly 300 human-hours.

Prior to the discovery study, we answer the foundational questions: "can participants use and understand the visualization and can they be more effective or more efficient than with regular and advanced tools?" In the discovery study, we focus on high-level and domain-specific questions of insightful discovery. We tested two groups: a control and an experimental group. To optimize the time of engagement with highly-skilled domain experts, we do not train participants on the low-level operations of the visualization. The performance study established this usability. Rather, we present the interpretative affordances of the visualization. Next, we use our own datasets and, together with a domain expert, we designed a work exercise aimed at closely mimicking real practices. Familiarity with the datasets facilitates rapidly answering search queries. The key is to avoid interpreting the query or the results. The technician must clarify the query beyond ambiguity and must present the results without any interpretation. In a sense, this approach is a Wizard-of-Oz intelligent interface.

Through this approach, we compress what typically takes many days of regular work per person into a five-hour period with each group. By parallelizing domain expert participation, we not only optimize time, we standardize within-group conditions. Every participant in the group consumes exactly the same sequence of queries and contributes and benefits from the group discussion interpreting the information. It is important to stress that participants work independently to maintain plurality. Each participant uses the same information differently for individual work goals. On the other hand, the efficiency tradeoff comes at a cost. We can't claim to have five statistically independent data points. The designs are not fully independent because the query results informing the designers are consumed by all the members of one group.

Excluding study design, the discovery study took 18 hours and consumed 90 researcher-hours and 99 participant-hours. The analysis consumed approximately 72 researcher-hours. The approximate cost is 261 human-hours.

CONCLUSIONS AND FUTURE WORK

We presented a two-part evaluation of the preference, performance, discovery, and impact of an information visualization of human behavior in everyday environments called Viz-A-Vis. We presented the complementary results of these evaluations with respect to our visualization. From the performance study, we highlight that our system greatly out-performed the other conditions for the critical task of searching for target events. The performance study also clearly establishes system usability, a necessary condition for acknowledging the discovery study's results. From the discovery study, we emphasize that we cost-effectively

provided multiple and conclusive evidence of the visualization's support for the discovery of actionable insights in the real practices of domain experts. Finally, we discussed the principles we learned for the general design of cost-effective evaluations of the visualization's power to raise opportunities for insightful discovery.

Our future work includes three venues. First, we are testing blob tracking algorithms and identity visualizations. That was the most unanimous unfulfilled requirement from participants in both studies. Second, we are recruiting domain experts from different fields and collecting data from significantly more complex spaces. As stated by the architects, the virtues of the visualization should become more apparent as the full complexity of the observation environment escapes human understanding. Finally, we are planning to run performance and discovery evaluations alongside long-term field deployments in order to compare the quality of the results with the justifiability of the costs.

ACKNOWLEDGEMENTS

This work was funded in part by the NSF Expeditions Award 1029679. We thank the contributions of architect Julie Zook to the discovery study. We thank our friends who donated their frank behaviors in the Aware Home during our data capture. We thank the 33 participants of our two user studies. Mostly, the first author thanks his wife, Dr. Natalia Landázuri, for her relentless support of this research, including donating over 200 hours of her private life to the data capture. Finally, Dr. Mario Romero dedicates this work to the memory of his mother.

REFERENCES

1. Botchen, R.P., Schick, F., and Ertl, T., Action-Based Multifield Video Visualization, in *Visualization and Computer Graphics, IEEE Transactions on*, 2008. 14(4): p. 885-899.
2. Card, S., Mackinlay, J. and Shneiderman, B., *Readings in Information Visualization: Using Vision to Think*. The Morgan Kaufmann Series in Interactive Technologies. 1999, San Francisco, Calif. Morgan Kaufmann.
3. Chen, C. and Yu, Y., Empirical Studies of Information Visualization: a Meta-Analysis. *International Journal of Human Computer Studies*, 2000. 53(5): p. 851-866.
4. Chen, M., Botchen, R., Hashim, R., Weiskopf, D., Ertl, T., Thornton, I., Visual Signatures in Video Visualization, in *Visualization and Computer Graphics, IEEE Transactions on*, 2006. 12(5): p. 1093-1100.
5. Daniel, G. and Chen, M., Video Visualization, in *Proceedings of the 14th IEEE Visualization 2003 (VIS'03)*. 2003, IEEE Computer Society.
6. Fayyad, U., Grinstein, G., and Wierse, A., *Information Visualization in Data Mining and Knowledge Discovery*. 2002, San Francisco, CA: Morgan Kaufmann.
7. Fels, S., Lee, E., and Mase, E., Techniques for Interactive Video Cubism. *Proceedings of ACM Multimedia*. 2000.
8. Grant, L. and Evans, A.N., *Principles of Behavior Analysis*. First ed. 1994, New York: HarperCollins College Publishers.
9. Ivanov, Y., Wren, C., Sorokin, A., Kaur, I., Visualizing the History of Living Spaces, in *Visualization and Computer Graphics, IEEE Transactions on,* 2007. 13(6): p. 1153-1160.
10. Jaschko, S., Space-Time Correlations Focused in Film Objects and Interactive Video, in *Future Cinema: The Cinematic Imaginary after Film*. 2003, MIT Press.
11. Kapler, T. and Wright, W., GeoTime Information Visualization, in *Information Visualization, INFOVIS 2004*. 2004. Austin, Texas.
12. Kidd, C., et al., The Aware Home: a Living Laboratory for Ubiquitous Computing Research. *Proceedings of the Second International Workshop on Cooperative Buildings*—CoBuild - 1999.
13. Kubat, R., DeCamp, P., Roy, B., and Roy, D., TotalRecall: Visualization and Semi-Automatic Annotation of Very Large Audio-Visual Corpora, in *Ninth International Conference on Multimodal Interfaces (ICMI 2007)*. 2007.
14. Lofland, J. and Lofland, L., *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. 3rd ed. 1995: Wadsworth Publishing Company.
15. Plaisant, C., The Challenge of Information Visualization Evaluation, in *Proceedings of the working conference on Advanced Visual Interfaces*. 2004, ACM: Italy.
16. Proshansky, H., *Environmental Psychology: People and Their Physical Settings*. 2nd ed. 1976: Holt McDougal.
17. Romero, M., Summet, J., Stasko, J., Abowd, G., Viz-A-Vis: Toward Visualizing Video through Computer Vision, in *Visualization and Computer Graphics, IEEE Transactions on,* 2008. 14(6): p. 1261-1268.
18. Saraiya, P., North, C., and Duca, K., An Evaluation of Microarray Visualization Tools for Biological Insight, in *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*. 2004, IEEE: Austin, TX.
19. Shneiderman, B. and Plaisant, C., Strategies for Evaluating Information Visualization Tools: Multi-Dimensional In-depth Long-term Case Studies, in *Proceedings of the 2006 AVI workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*. 2006, ACM: Venice, Italy.
20. Ullman, S., *The Interpretation of Visual Motion*. 1979, Cambridge, MA: MIT Press.
21. Wang, Y., et al., Contextualized Videos: Combining Videos with Environment Models to Support Situational Understanding, in *Visualization and Computer Graphics, IEEE Transactions on,* 2007. 13: p. 1568-1575.
22. Wang, Y., et al., Effects of Video Placement and Spatial Context Presentation on Path Reconstruction Tasks with Contextualized Videos, in *Visualization and Computer Graphics, IEEE Transactions on*, 2008. 14(6): p. 1755-1762.