Affordances of Input Modalities for Visual Data Exploration in Immersive Environments

Sriram Karthik Badam*  
University of Maryland  
College Park, MD, USA

Arjun Srinivasan†  
Georgia Institute of Technology  
Atlanta, GA, USA

Niklas Elmqvist‡  
University of Maryland  
College Park, MD, USA

John Stasko§  
Georgia Institute of Technology  
Atlanta, GA, USA

Abstract

There has been a consistent push towards exploring novel input, display, and feedback technologies for sensemaking of data. However, most visual analytic systems in the wild that go beyond a traditional desktop just utilize commercial large displays with direct touch, since they require the least effort to adapt from the desktop/mouse setting. There is a plethora of device technologies that are yet to be tapped. Through this paper, we want to bring attention to available modalities for input for visual exploration within immersive display environments. These can be environments that contain multiple wall and floor displays, or projections to create immersion. We first isolate the low-level interaction tasks performed by a user based on a usage scenario for visual exploration. We focus on egocentric visual exploration in these environments, and introduce input modalities that enable interactions directly between a human body and objects of interest on the interface without a mediator in the middle (e.g., a handheld smartphone). Based on this, we identify affordances of different input modalities—touch, speech, proxemics, gestures, gaze, and wearable—in terms of the interaction tasks from the envisioned scenario. Finally, we discuss how modalities can be combined to complement each other and leverage their advantages in the immersive environment. By doing so, this paper provides guidelines for new system developers to utilize the best input technologies for their immersive analytics applications.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction styles;

1 INTRODUCTION

Developing an immersive analytics system is not straightforward. Beyond the monetary challenges, these systems need to be assembled to some extent, if not entirely, by coupling multiple technologies together. For instance, CAVE systems1 can provide floor and wall displays within a room to create immersion, but to interact with them in the 3D space, motion capture platforms2 are often used. However, the selected input technology may not provide the required freedom of expression in terms of interaction for visual exploration [25]. Case in point, gestural and proxemic interactions for visual analytic tasks have so far been only observed to be effective for simple interactions (e.g., changing zoom levels and switching between specific level of details in visualizations) [2, 11, 13]. Each input modality has specific affordances in terms of the interactions for visual analytic operations, and therefore, these affordances should be closely considered when developing new immersive visual analytics systems.

In this paper, we identify affordances of input technologies for supporting visual exploration. We focus on input modalities that enable egocentric exploration through interactions that are directly between the human body and objects of interest in the immersive environment, and not through an intermediate device (e.g., a handheld smartphone). To do so, we first discuss a potential usage scenario involving visual exploration in immersive room that contains wall displays. We extract specific low-level interaction tasks based on the scenario. Following this, we present the input modalities of interest—direct touch, speech, proxemics, gestures, gaze, and wearable input (seen in Figure 1) and discuss their affordances in terms of supporting the interaction tasks. The affordances are discussed based on the freedom of expression of each input modality. For instance, touch input is very expressive and enables users to easily focus on specific visualizations in an interface, navigate through direct manipulation, and easily pick items of interest in a visualization for further exploration. However, creating visualizations using touch input requires interface options to capture the user’s specification.3 This freedom of expression based categorization led to Figure 2, which highlights the affordances for interaction tasks on a four-point scale for each modality. Based on this, it is apparent that speech and touch inputs can be coupled to best support all the interaction tasks. However, direct touch is only feasible when the user is close to a display. Other interesting and potentially useful couplings include, speech + mid-air gestures, proxemics + wearable input, and speech + gaze input, within the immersive environment. While the affordances are explained through a specific scenario, the principles can be applied generally to other immersive environments created by AR/VR headsets or even futuristic room-scale holographic displays.

1CAVE: http://www.visbox.com/products/cave/

2VICON: https://www.vicon.com/products/camera-systems

3Previous interactive design tools [12, 20] helped convey the visualization specification in traditional desktop/mouse settings.

Figure 1: Input modalities for visual data exploration covered in this paper.
2 Envisioned Environment and Usage Scenario

The choice of target device(s) and the immersive environment created by them play a pivotal role in supporting immersive analytics. Considering a single environment with all possible input modalities and displays (touch-screen displays, projections, AR, VR, etc.) together is beyond the scope of one paper. To limit our scope to practical use cases, we envision the target setting for discussions in this paper to be a common thinking space such as board rooms with interactive surfaces or displays distributed in the room.

We imagine a visual interface in this immersive environment that presents a dataset through multiple visualizations. For the purpose of generality, we assume a simple visual design that exposes the data items within the dataset directly on the visual interface by utilizing the important visual variables (i.e., location, size, shape, and color). The individual visualizations use granular designs where each point in the visualization has a data context; for instance, a line chart with each point on the line representing an attribute from the dataset. This design is ecological since appropriate data transformations including aggregation and sampling can be used to create these point-based designs. Since each point in the visualization has a data context, the regions in space also have a context within the original dataset (e.g., corresponding to data attributes).

Usage Scenario. As an example of how such a setting may be used in practice, consider Sarah, an executive officer for a book publishing house. Sarah needs to make a presentation on the annual performance for the previous fiscal year to inform decisions about product lines and types of products (books) the company should focus on for the upcoming year. Sarah has the data for the previous fiscal year with details about individual sales, product type, region of sale, etc. To brainstorm about her upcoming presentation and explore this data freely, Sarah loads the dataset in an open space board room in her office with an interactive wall display and a comfortable couch. Below, we describe a usage scenario highlighting Sarah’s experience in this setting. We discuss the scenario in terms of tasks to give a general overview. We discuss the possible input modalities and their affordances, and give examples of how various modalities could be used to support these tasks in the next section.

Settling herself comfortably on the couch, Sarah begins her exploration. She first creates a bar chart for the overall sales over quarters (create). To see the distribution of sales across product types, Sarah modifies the bar chart to a stacked bar chart showing sales by book genre (reconfigure data mapping). Noticing “fantasy literature” as the highest selling book type across quarters, she decides to explore it further. She highlights with color and annotates the bars for fantasy literature sales with the percentage of sales they contribute to in each quarter (reconfigure graphical property). Sarah adds a new map visualization to the wall showing distribution of sales for fantasy literature books around USA (create, focus). She notices that southern California and Florida have most sales. Intrigued by these hotspots, she looks more closely and notices that the hotspots are around Los Angeles, California and Orlando, Florida (pick, find, filter). Knowing that both locations are close to Walt Disney parks and Universal studios, she assumes that the higher sales attribute to an effect of park visits, as children get intrigued by park experiences and buy books to read new stories and engage with their favorite fictional characters.

She walks around thinking some more about what she could explore next and returns to sit on the couch, casually glancing through the visualizations she generated. Looking at the stacked bar chart (focus, navigate), Sarah notices that art books are not among the highest selling book types and wonders why since it seems logical to have them as an extension to fantasy literature. To identify the reasons, Sarah creates a scatterplot matrix showing the investment and profit for books across all genres (create). Next, Sarah compares the fantasy and arts genres in terms of investment-profit ratio (reconfigure data mapping). She notices that art books have a higher profit margin but have not been invested in by their company. On the other hand, fantasy literature books have a moderate investment-profit ratio and a large investment. She starts preparing her presentation based on these findings.

3 Tasks and Affordances

In the usage scenario, there are certain interaction tasks that are used by Sarah to visually explore the data in her enhanced office room. These tasks resemble the interactions specified in popular task taxonomies [25]. To identify the affordances of input modalities for supporting these tasks, we first define them more generally to extend them to a general taxonomy [25]. We then describe the user actions with each input modality to perform the task, along with the freedom of expression supported by the input modality for the particular task. We also discuss the aspects of fatigue, distance of action, and role of input modality in general for justifying the affordances. As mentioned earlier, we are interested in input modalities that create egocentric exploration opportunities in the immersive environments, where the interactions happen between a human body and objects of interest without a mediator in the middle (e.g., a handheld smartphone) [19].

3.1 Interaction Tasks

The usage scenario covers eight interaction tasks.

- **Create** a new visualization on the interface.
- **Focus** on a specific visualization within the interface.
- **Pick** individual data items within a visualization.
- **Find** data items satisfying a predicate logic (e.g., \( a < 5 \)).
- **Filter** selected data items by removing others.
- **Navigate** a visual with pan and zoom around focii.
- **Reconfigure graphical properties** in a visualization.
- **Reconfigure data mappings** driving a visualization.

These low-level interactions are connected to the high-level tasks defined by Yi et al. [25], and extend them to our scenario. For instance, pick and find are two ways to select content in a visualization. However, in contrast to Yi et al. [25], our list differentiates them since the cognitive effort for performing these two types of selections can be different from a user’s perspective—forming predicates can be more complex than picking individual data items. Similarly, our reconfiguration tasks resemble the encode, reconfigure, and abstract/elaborate tasks from Yi et al. [25], but differentiate based on the graphical properties vs. data mappings. We admit...
that our interaction task taxonomy is by no means exhaustive or tai-
lored towards immersive analytics, but we believe it offers a starting
point to identify the affordances.

3.2 Affordances of Input Modalities

We are interested in six input modalities for visual exploration in
our scenario. The capabilities of the input modalities in terms of
the task affordances are broadly categorized in Figure 2. Here we
introduce each modality and its input technology followed by its
affordances for the tasks in terms of the freedom of expression.

3.2.1 Touch

Direct touch input with the displays in the environment can help
directly interact with the visualizations. This input can be enabled,
for instance, through capacitive sensors embedded within a display.
Naturally, this input is only possible when very close to the tar-
get display. Touch interaction has three main actions: tap specific
points in space, drag/move fingers in space, and gesture with single
or multiple fingers. Due to these degrees of expression, touch is
one of the most expressive modalities. Touch input can effectively
support (1) focusing on specific visualizations in the environment
by choosing them directly, (2) pick individual items in a visualiza-
tion by tapping them, (3) navigating within a visualization by drag
movements to zoom and pan, and also perform multi-focus naviga-
tion through multi-touch.

Other interactions can be performed with touch input, but need
further interface support. To create new visualizations, specific in-
terface features to specify the data attributes of interest and visual
mappings by touch need to be present. To find data items based
on predicates, additional interface widgets such as dynamic slid-
ers and option menus are required. Touch gestures—e.g., a remove
gesture—needs to defined to filter out uninteresting items in visu-
alizations. Finally, reconfiguration interactions involving graphi-
cal items on a visualization—changing mark shapes or arrange-
ment within a visualization—can be done directly through touch ac-
tions, but additional interface widgets are required to change color
schemes, data mappings (e.g., replacing attributes), and data trans-
formations (e.g., aggregation parameters).

3.2.2 Speech

Prior work has shown that users of visualization systems may be
able to express their questions and intents more freely using natu-
ral language [1,7], allowing them to perform a range of tasks [23].
In the envisioned scenario, speech input can allow the user to issue
queries and interact with a visualization without being too close to
the display or in its field of view. Speech queries can have a high
freedom of expression, only bounded by the user’s ability to express
a query in natural language. Therefore they can effectively help the
user, (1) create new visualizations by expressing which attributes
and/or representations, or potentially, even the analytic intent, (2)
find data items satisfying some predicate logic, and (3) filter the
visualizations by expressing the intent to remove uninteresting items.
With sufficient expressive modalities, speech input can be used to
reconfigure data mappings by specifying

3.2.3 Proxemics

Proxemic interaction refers to utilizing the spatial relationships [9]
between people and artifacts in an environment including posi-
tion, movement, distance, orientation, and identity. These vari-
cables can be obtained by using motion capture cameras within the
environment. Interfaces built on these interactions automatically
adapt to these user/device attributes [14]; for instance, a proxemic
video player will start playing a movie when the user sits on the
couch [3]. Due to the variety of proxemic dimensions, this input
modality, representing spatial interaction, can be quite useful for
visual exploration tasks. Within our usage scenario, proxemic in-
put is best suited for interface-level operations such as focusing on
a new visualization in the interface [2]. With some assumptions,
proxemics can also be used for navigating within a visualization
(pan and zoom) based on the attention of the user (e.g., orientation
of their body) [13]. Furthermore, previous research has also consid-
ered manipulating the level of detail in visualizations—reconfigure
data mappings—by using interaction zones in front of a display.
For instance, when far from a display, a particular amount of visual
content is shown [2,11]. However, the remaining interactions in
our scenario, are quite challenging to perform with just proxemic
input. Finding data items that satisfy a predicate involves design
considerations in terms of how to translate a user’s field of view
into a logical operation. Remaining interactions—create visualiza-
tions, pick individual data items, filtering, and reconfigure graphi-
cal properties—are hard to map naturally to proxemic dimensions
making them unsuitable for this modality.

3.2.4 Mid-Air Gestures

Gestural input, especially mid-air, can help explicitly navigate the
content shown in visualizations (e.g., zoom and pan) [2]. A long
list of gestures designed for an interface may increase the freedom
of expression of the user, making it possible to perform many in-
teraction tasks. However, this is not a feasible choice as it com-
plicates the user training and potentially confuses the users during
the exploration. Hence we discuss the affordances based on two
main types of gesture designs: (1) pointing gestures to convey in-
terest in a specific unit in the interface, and (2) gestures utilizing
hand movements based on the expected changes within a visualiza-
tion [15]. Given these choices, gestures—similar to touch input—
can effectively support (1) specifying the focus of the user within
the entire interface and (2) picking specific items in the visualiza-
tions through pointing actions. With sufficient disambiguation and
some assumptions, it is possible to filter visualizations and navigate
within a visualization through gestural actions. Furthermore, with
some interface options, simple predicates can be constructed for
find interactions based on hand movements along axes of data at-
tributes. However, this input is not suitable for reconfiguring graphi-
cal properties or data mappings as there will be complex design
considerations in developing gestures that freely support them.

3.2.5 Gaze

Gaze input possible through eye tracking technologies offers poten-
tial opportunities to adapt visual interfaces based on the visual
attention of the user. However, this also adds a challenge: using
the visual channel for input as well as output can lead to unwanted
changes to the content of the interface. For this reason, gaze in-
put is typically based on the visual focus of the user’s eyes as well
as the duration of focus. Compared to other modalities, gaze sup-
ports fewer interactions. It is ideal for tracking the focus of the user
within the entire interface. It can also be useful to pick specific
items in a visualization and navigate (pan and zoom) towards the
items in user’s focus; however, the duration of gaze needs to be ef-
effectively configured to make the interaction seamless. Remaining
interactions—create visualizations, find items based on predicates,
filter, and reconfigure—are not suitable with gaze since the com-
plexity of these interactions is too high to be easily supported by
the freedom of expression of gaze input.

3.2.6 Wearable input device with a display
Wearables with a display attached to them such as smartwatches
and on-body input devices [10] enhance an immersive environment
with secondary interactive displays that are attached to the user’s
body to offer more freedom of expression in interactions. At
the same time, wearable input modality falls into egocentric interaction
by leveraging proprioception and eyes-free interaction for body-
centric operations [22] (which is not possible with handheld de-
vices). This modality also enables remote interaction from a dis-
tance from the immersive displays. With sufficient interface op-
tions on the wearable device, the focus of the user in terms of the
visualization in the interface can be easily conveyed. For instance,
a smartwatch can be used to perform swipe input to pick a visual-
ization on a large display, without even looking at the watch. To
support other interactions, the on-body/wearable interface needs to
be designed effectively. For instance, to create new visualizations
there should be options on a smartwatch interface to specify the
attributes to visualize. Find, filter, navigate, and reconfigure inter-
actions similarly require interface elements on the watch to convey
the specific intent. Finally, picking individual data items through a
wearable input device is not suitable since a responsive version of
the large display interface should be placed on the small wearable
display to pick the items; this has considerable limitations due to
the mismatch in display size.

4 DISCUSSION
Each input modality has specific expectations in terms of the sys-
tem behaviors. In some cases, this is for reducing the amount of
effort by being proactive; for instance, to automatically react to
user’s gaze rather than prompting them to gaze. Coupling the in-
put modalities can also further enhance the visual exploration by
overcoming individual tradeoffs. Here we discuss these aspects.

4.1 Combining Modalities: Why? When? How?
As stated earlier, Figure 2 discusses affordances of various input
modalities individually. However, prior work in the broader HCI
community [17, 18], and even recent work within the visualization
community [2, 24] has shown that it is natural and effective to use
multimodal input. Coupling input modalities can create more de-
sign opportunities, where advantages of one modality can be used
to overcome the challenges of another (e.g., the ambiguity in speech
can be corrected by precise selections offered via touch).

Coupling modalities is not straightforward, however. Users may
prefer different modalities for different tasks and the usage patterns
of modalities may vary as well (e.g., sequential vs. simultaneous
multimodal interaction). One of our primary goals with Figure 2
was to present an overview of strengths and weaknesses of various
modalities with respect to specific tasks in a scenario such as the
one presented earlier. A direction for future work is to consider
how two or more of the discussed modalities, that are suited based
on Figure 2, can be combined to facilitate a full spectrum of tasks
and foster a more fluid interaction and visual analysis experience
for the users [5].

Exploring preferences for input modalities based on tasks, under-
standing how people combine modalities, and elucidating if pref-
ferences change for different combinations of modalities (e.g., is
speech always the most commonly used input modality? do peo-
ples use it less when combined with wearable technology as com-
pared to touch?) are all open questions. For example, contrary to
common assumption, prior work has shown that when using speech
and gesture, people rarely perform “put-that-there” [4] style inter-
duction using both modalities simultaneously [16]. We hope discus-
sions and ideas presented in this article help future systems consider
which modalities can be mixed to foster an enhanced visual analy-
sis workflow, and encourage them to identify the most effective
input combinations for their specific visual exploration tasks.

4.2 System Behavior: Proactive vs. Reactive
Most existing work in the visualization community exploring post-
WIMP input has considered reactive system behavior and how sys-
tems can respond to “intentional” user interactions. However, in a
setting such as the one presented in the usage scenario earlier, some
input modalities are potentially more suited for proactive system
behavior by default whereas others may be more suited for reactive
behavior. For instance, proactively adjusting the size of the visual-
ization in focus based on the user’s proximity to the display can
help preserve the focus and allow the user to look at the same set of
data from different distances (e.g. while sitting or moving around
in the scenario discussed earlier). Reactively doing so demands
additional effort from the user, which is not ideal for longitudinal
analytics sessions. On the other hand, proactively listening to user
utterances and adapting to them would lead to an interruptive and
frustrating user experience.

With some input modalities, proactive behavior can be more
suited for certain tasks and less for others. For instance, proac-
tively updating the “focus” visualization or resizing a visualization
based on a user’s location (proxemics) may be more effective than
navigating the visualization content (zoom/pan within a visualiza-
tion). This was identified in previous work to combine proxemic
(implicit/proactive) and gestural interactions (explicit/reactive) [2]
into a hybrid interaction model for visual exploration. Additionally,
in a setting with multiple modalities, depending on the sequence of
use of the modalities, proactive behavior in response to one modali-
ity may interfere with the input/output for another. Preserving the
user’s workflow and identifying when to activate/deactivate proac-
tive behavior is also an open challenge when mixing modalities for
visual exploration. For instance, proxemics may be well suited for
identifying shifts in the focus (or target) visualization when there
are multiple visualizations to choose from. When combin-
ing speech and proxemics, systems would need to identify when
to apply proactive behavior and shift focus versus when not to do
so based on the user’s movement and potential changes in global
coherence in the dialog between the user and the system [8].

ACKNOWLEDGMENTS
This work was supported by the U.S. National Science Foundation
award #1539534. Any opinions, findings, and conclusions or rec-
ommendations expressed in this material are those of the authors
and do not necessarily reflect the views of the funding agency.

REFERENCES
and A. Johnson. “Show me data”: Observational study of a conversational
designing for a proximity and orientation-aware environment. In Pro-
cceedings of the ACM Conference on Interactive Tabletops and Surfaces,
[4] R. A. Bolt. “Put-that-there”: Voice and gesture at the graphics in-
terface. In Proceedings of the 7th Annual Conference on Computer


