EagleView: A Video Analysis Tool for Visualising and Querying Spatial Interactions of People and Devices

Frederik Brudy[1], Suppachai Suwanwatcharachat[1], Wenyu Zhang[1], Steven Houben[2], Nicolai Marquardt[1]
[1]University College London, London, UK; [2]Lancaster University, Lancaster, UK
f.brudy@cs.ucl.ac.uk, s.houben@lancaster.ac.uk,
{s.suwanwatcharachat.16, wenyu.zhang.16, n.marquardt}@ucl.ac.uk

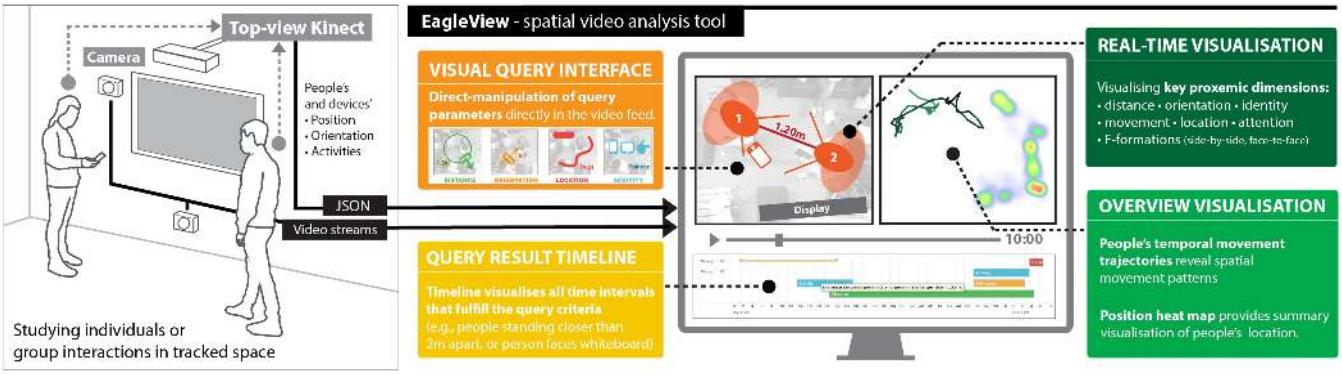

Figure 1. EagleView workflow with querying and visualisation techniques facilitating spatial interaction analysis.

ABSTRACT

To study and understand group collaborations involving multiple handheld devices and large interactive displays, researchers frequently analyse video recordings of interaction studies to interpret people's interactions with each other and/or devices. Advances in ubicomp technologies allow researchers to record spatial information through sensors in addition to video material. However, the volume of video data and high number of coding parameters involved in such an interaction analysis makes this a time-consuming and labour-intensive process. We designed *EagleView*, which provides analysts with real-time visualisations during playback of videos and an accompanying data-stream of tracked interactions. *Real-time visualisations* take into account key proxemic dimensions, such as distance and orientation. *Overview visualisations* show people's position and movement over longer periods of time. *EagleView* also allows the user to query people's interactions with an easy-to-use visual interface. Results are highlighted on the video player's timeline, enabling quick review of relevant instances. Our evaluation with expert users showed that *EagleView* is easy to learn and use, and the visualisations allow analysts to gain insights into collaborative activities.

Author Keywords

Video analysis; cross-device interaction analysis; spatial interaction; group collaboration; interaction analysis; information visualisation

INTRODUCTION

To understand and evaluate interactive systems, researchers often use video analysis of individual or collaborative interactions of people and the devices they use (e.g. [1,11,12,26]). The analysis of the recorded video data is a tedious and labour-intensive task, requiring researchers to review the raw video iteratively and identify relevant tags and codes in the footage [13]. While various commercial and research tools exist to support the viewing and tagging of videos (e.g. [5,25,34]), these mostly focus on facilitating the navigation, transcription, and annotation of videos, not the actual interaction analysis. However, spatial characteristics of group interactions are important for such analysis, including where people stand, how close they are to each other, which devices they use, and so on. These features need to be manually observed and annotated by the researcher, as current tools do not support the analysis of spatial characteristics well.

Recent ubiquitous computing (ubicomp) environments enables recording of proxemic and spatial interaction data, such as people's location and orientation, as well as the devices they interact with (for example with the open-source Prox-

ISS '18, November 25–28, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-5694-7/18/11...$15.00
https://doi.org/10.1145/3279778.3279795

imityToolkit [19] or EagleSense [32] platforms) and their activities (e.g. holding phone, standing, sitting). While recording this additional data during studies allows for deeper insights into the interactions, the analysis still needs to be done manually as there are currently only very few tools that allow for automated insights into participants' interaction with their devices, each other, and their surroundings (e.g. [21,28,33]). Inductive video analysis has become the main pathway to gain insights of recorded interactions [13]. However, thorough video coding remains a challenging task, in particular once multiple people and multiple devices are involved: rather than being able to focus on *one* user's interaction with *one* system, an analyst has to study many relations between multiple users and devices. Using sensor-data can help to find moments of interest in video data, but more detailed manual analysis is still needed, as relying solely on sensor data can lead to false conclusions.

To better support spatial interaction analysis, we contribute *EagleView* (Figure 1), a novel video analysis tool that allows expert users (researchers, conducting analysis in multi-user, multi-device scenarios) to review and analyse multiple videos and an accompanying spatial tracking data-stream by providing a querying interface, real-time visualisations, and multiple overview visualisations of the interactions through a web-based interface. *EagleView* allows users to create new queries on the videos and tracking data through an easy-to-use direct-manipulation visual-query interface. Examples of such queries are "*when are participants closer than 1 metre*", "*when is a person facing the screen*", or "*when are people pointing at the large display"*. We evaluated both the visualisations and the querying interface of *EagleView* through two user studies with expert users (HCI researchers with several years of experience in video analysis of group interactions and/or collaborations). Our studies show that *EagleView* was easy to learn and use, and that the querying tool enabled analysts to quickly select interaction scenarios of interest.

In summary, the contributions of this paper are 1) the design and techniques of the *EagleView* query tool as well as the real-time and overview visualisations; 2) the insights gained from expert users through two user studies, exploring two different aspects of *EagleView* (real-time and aggregated overview visualisations and the querying tool). We share *EagleView* with the HCI research community as open-source software at https://github.com/frederikbrudy/eagleview.

RELATED WORK

Our research builds on the foundations of previous work on (i) interaction analysis; (ii) systems supporting interaction analysis; and (iii) video analysis and visualisation tools.

Interaction Analysis

Interaction analysis [13] describes the empirical lens through which researchers analyse how people interact with each other and their surroundings. Increasingly, social science and psychology theories are used as lenses for conducting interaction analysis. For example, Hall's theory of *proxemics* [8] describes how people physically engage and communicate with other people and the devices in their surroundings through their use of distance (intimate, personal, social, and public), orientation, and posture. Kendon's *F-formations* describe how multiple people use and share a physical space through their distance and relative body orientation to indicate when and how they interact as a group (in a circular, vis-à-vis, L-, or side-by-side arrangement). Both concepts have been used in ubiquitous computing as a lens for interaction analysis, for example to discover patterns of collaboration in a tourist information centre [22] or museum [4]; to support large display interactions (e.g. [14,29]); and to enable cross-device interactions in multi-device scenarios (e.g. [20]). The current state of the art in interaction analysis is through iterative coding of observations in video recordings. Such video recordings of studies or experiments support repeated observations to gain an in-depth understanding of a scenario [13]. However, the analysis of those recordings requires many hours of reviewing video data.

Systems supporting interaction analysis

To facilitate the analysis of interactions using video recordings a multitude of tools have been developed, which allow users to review and annotate videos [3], and several commercial (e.g. [25,34]) as well as research tools (e.g. [2,5,7,28]) aim to support video analysis.

Advances in ubicomp systems allow users to record more information besides the video data, such as proxemic and spatial interaction data through sensors. For example, the ProximityToolkit [19] uses infrared cameras and visual markers attached to record people's and devices' identity, location, and orientation. More recently, markerless top-down tracking systems have been introduced, tracking people's identity, position and orientation (e.g. [18,20]), posture (e.g. [10]), as well as their activity (e.g. [9,31,32]). These systems show that there is an increasing supply of systems not only tracking people and their devices, but also recognising what they are doing (e.g. holding a paper, pointing, using a tablet, or smartphone). As an example, in this paper we record the tracking data from EagleSense [32], which uses a Kinect v2 camera mounted to the ceiling to capture the space underneath it, recording people's position, orientation, and their activity (e.g. *standing*, *holding phone*).

Several tools aiming to support the analysis of these multi-stream recordings have been developed. For example, VACA [2] allows a synchronised playback of video data with accompanying sensor data. The sensor data can then be used as an additional cue for finding the relevant parts of the video. Similarly, VCode and VData [7] enable synchronised playback and annotation using sensor data and video recordings. ChronoViz [5] enables synchronised playback of multiple video and data streams, allowing analysts to add annotations in form of tags and textual descriptions.

Figure 2. *EagleView's* user interface, showing the visualization panel with real-time visualizations, playback control, and different angled videos. On the right the analyst can change preferences and switch to a view, showing aggregated data.

While previous work proposed to use crowd workers to annotate [30] or enable natural language querying [16] on video material, this is cost-intensive and might compromise privacy. EXCITE [21], on the other hand, enabled researchers to conduct search-queries on the recorded video+sensor data of interactions in ubiquitous computing environments, using a descriptive query language. This allowed researchers to gain insights that were not as easily attainable before.

Our work builds on this prior work of tools that support interaction analysis, in that we allow analysts to review video data, combined with tracking information about people's position, orientation, and activity. *EagleView* further allows analysts to visually create queries on the video+sensor data, and then navigate between search results for further review. Similar to EXCITE [21], *EagleView* allows analysts to analyse group interactions involving the use of mobile as well as fixed devices, enabling analysts to focus on high-level analysis of the interactions, rather than focussing on finding low-level evidence for a hypotheses.

Visualisation tools for video analysis

Other specialised tools support researchers in visualising video recordings of multi-device and/or multi-person interaction scenarios. For example, VICPAM [24] shows users' activities over time and the duration of each activity on an overview timeline. VisTACO [28] focussed on analysing spatial interactions around a tabletop display and GIAnT [33] enabled users to analyse and visualise interactions of people with a large, interactive wall display.

We build on and extend previous work with our real-time and overview visualisations. Similarly to slit-tears [27], *EagleView* helps the researcher by summarising a longer period of time in a static *overview visualisation* in addition to *real-time visualisations* during playback. In particular, we use the five proxemic dimensions [6] (distance, orientation, movement, identity, location) to create visualisations about individual people and objects. We further leverage the notion of F-formations [15,20] of people and devices, enabling analysts to identify critical moments of interactions.

SCENARIO DESCRIPTION

Throughout this paper we will refer to the following scenario about the analysis of a multi-device multi-user interaction in a museum. This scenario helps us to better situate our tools and techniques that we introduce shortly within the context they will be used.

Mary is developing a new application for a museum, which allows visitors to explore details about each exhibit through an interactive display next to each item. While visitors roam the museum and interact with various touch displays, they can also navigate the collection on their own smartphone through the museum's app and website. After deploying the system, Mary wants to learn more about how people approach the exhibit's displays and use them together with their personal devices. She installs a Kinect camera in the ceiling above the interactive exhibits, as well as 3 cameras to record the interactions from a side perspective. She records an entire day in the museum through EagleSense [32]*, which uses a ceiling-mounted Kinect camera to track visitors in a gallery and their activity. At the end of the day she has 10 hours of footage from each camera, totalling 40 hours of video material she needs to analyse. The museum was well attended on that day, but not everyone approached or interacted with the exhibits and/or their smartphone.*

EAGLEVIEW OVERVIEW

EagleView is a web-based video analysis tool that allows users to explore people's spatial interactions using visual analysis of recorded video data together with automatically tracked spatial measures (*location* and *distance* of people and devices; their *orientation*; *movement*; *identity*; *activity*; recorded with EagleSense [32] or similar tools, e.g. [19]). The user interface consists of the following elements as shown in Figure 2: all *parallel video playbacks* on the left (Figure 2b); *real-time visualisation* interface in the middle (Figure 2c, 2d); *overview visualisations* on the right (Figure 2e, 2f, 2g); *video timeline* with manual annotations and tags (Figure 2h) and *query results timeline* (Figure 2i) at the bottom. Through a tabbed interface at the top (Figure 2a), a user can switch from the *Visualisation* interface to the *Query Creator* interface (Figure 11, introduced later in the paper).

Getting started. To begin the analysis, a person first selects one or multiple video files from the recorded study and the accompanying spatial tracking data file (in our case recorded by the EagleSense tracking framework [32], but potentially provided by other tracking systems). In this *configuration step*, the user can also create objects that are fixed in the environment (e.g. displays or tables) by drawing them on a still image of the top-down video recording and adding a descriptive label. Further, an offset can be configured for each video to synchronise start times. All video playbacks are displayed in the top half of the interface (Figure 2b), are time-synchronised, and can be started and stopped using the video controls. A progress bar shows the playback progress and allowsanalysts to go forward or back in the video. The first view a user is presented with shows the *real-time and overview visualisations* (Figure 2c and 2d). *EagleView* includes features similar to current video analysis tools: users can add annotations to a timeline either by clicking on pre-defined tags (configurable in the *configuration step*) or by entering comments in a text area, and each of the annotations is then shown in a timeline (Figure 2h). By clicking an annotation, analysts can jump to that moment in the video.

Beyond conventional video analysis through reviewing and tagging, *EagleView* allows analysts to review the video by creating search-queries (Figure 11) based on spatial features, and gain insights into user interactions through real-time and overview visualisations. We will describe both key functions in more detail after the technical details. Figure 3 shows the stages of interaction analysis supported by *EagleView*.

TECHNICAL IMPLEMENTATION

EagleView is built using modern web technology (HTML5, JavaScript, CSS) and runs entirely on a client's device (tested in Chrome 67). We use CreateJS[1] for easy HTML5 Canvas manipulation and Vis.js[2] for the timeline component.

To visualise the spatial properties and interactions, *EagleView* consumes spatial tracking data recorded with the top-down tracking system EagleSense [32], through its API. Specifically we record an array of time instances, each including a `timestamp` and people's location, orientation, and whether they are sitting, standing, and if they are holding a paper, a tablet, or a phone (as an array of `skeletons`, with properties `{id, activity, activity_tracked, head{x, y, z, orientation}}`) in a JSON file. As *EagleView* is impartial to the tracking technology used, other input sources can be used if the recorded data is in the same format. Further, *EagleView* displays one or multiple video recordings. In Figure 2, two top-down recordings (an RGB video as well

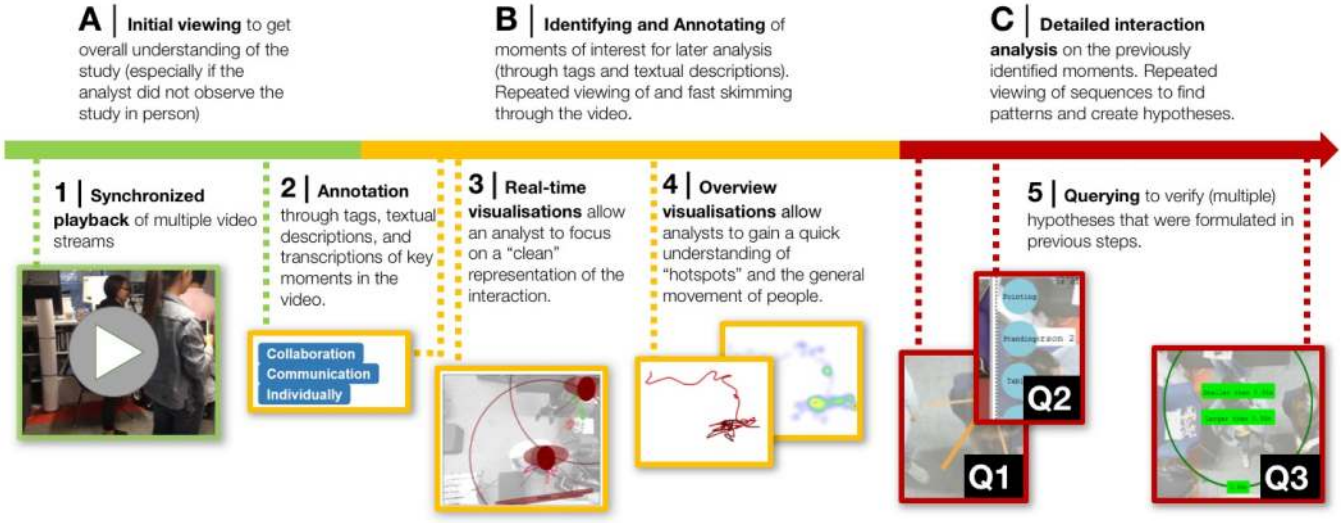

Figure 3. Timeline with an overview of how the different components of *EagleView* support the video analysis workflow. The modularity and flexibility of *EagleView's* tools allows analysts to use our visualisations and/or querying functions in any possible order, so that they best fit to the analysing researcher's workflow.

[1] https://www.createjs.com/

[2] http://visjs.org/

as the depth video, captured with the Kinect v2 used by EagleSense [32]), as well as two different side-views are shown. The main video view displays the top-down RGB video as a semi-transparent video (if toggled on). The videos' playback is synchronised based on their timestamp.

Limits. EagleView is running entirely on a user's local machine and web browser, and no data is sent to any remote server. As a result, loading time of a dataset is kept to a minimum of only a few seconds and any video format can be used that is supported by the user's browser. However, since computation entirely happens client-side, a more resourceful computer is required. In our experience, a limit of 3-5 videos playing in parallel is easily achievable on any current laptop.

For our studies (reported on later), we manually cleaned the JSON data after recording the tracking information from EagleSense [32], to remove any tracking errors (e.g. removing false activities or sudden jumps of location). We did this because we were interested in how, in an ideal case, spatial data can be analysed. We envision that these tracking artefacts of third-party systems will become rarer with better tracking systems in the future and therefore continued to use data without these artefacts. We limited the data saved from the EagleSense API to 4 frames per second. In our experience this was a reasonable trade-off between clean-up time required and still having detailed data.

REAL-TIME AND OVERVIEW VISUALISATIONS

First, we focus on how researchers can analyse recorded interactions through two different types of visualisations: *real-time and aggregated overview visualisations* (Figure 3B).

Real-Time Visualisation

The real-time visualisations are shown in the middle column on the tab *Visualisations* (Figure 2) during video playback and visualise the data around the current playback time. They show the spatial properties recorded from the EagleSense API (people's location, orientation, and activity) as well as fixed objects (such as wall displays, whiteboards, or tabletop displays as defined by the user in the configuration phase). The video, recorded from the top-down camera, can be optionally displayed in the background.

People's **locations** are displayed as two ovals, representing head and shoulders (Figure 4). Their **viewing direction** is indicated by two lines, marking their field of view. (The angle is configurable by the user.) If any activity is recorded for them, the respective **activity** is visualised through an icon in their field of view (e.g. phone, tablet, paper). Each person's **identity** (as ID number) is displayed alongside their location. **Fixed objects** are drawn as outlines in their locations on the top-down view with their descriptor.

In addition, *EagleView* visualises the following information in the *real-time visualisation*:

Distance: As shown in Figure 4, analysts can choose to show a circle around a person to indicate proxemic distances (e.g. personal or social zones [8]). In addition, distances between different entities (e.g. two people or a person and an object) can be shown as a line between them. The distance, as well as the textual description of the proxemic zones [8], is shown underneath the line.

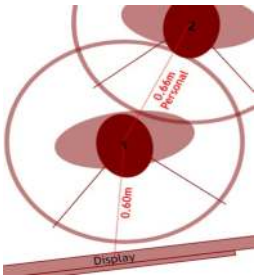

Figure 4. The distance circle (set to 0.5m) and lines between P1 and P2 as well as the display are visible.

Movement trajectories: people's past and future movement trajectories can be displayed as coloured lines (Figure 5), fading to more transparency the further in the past or future the respective locations are. The length of time used for these trajectories is configurable by the user, as well as the colour for past and future movements.

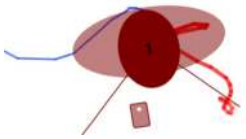

Figure 5. The person's movement of the previous 10 (blue) and next 15 seconds (red) is shown.

Zones: During the setup phase, analysts can define rectangular zones of interest. When a tracked person enters a zone, the zone will be highlighted (bold border and opaquer colour; Figure 6). This allows users to quickly skim the video and easily spot when a person enters a particular area.

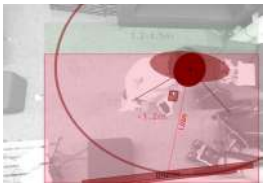

Figure 6: Defined zones to observe (red/green rectangles are defined areas).

Attention grouping (Figure 7 and Figure 8): *EagleView* highlights tracked users in the same colour if i) they directly face a fixed object; ii) their attention is focused on each other; or iii) if two people stand next to each other and they are facing the same fixed object.

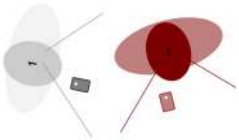

Figure 7: Person on the right is highlighted (e.g., when facing an observed object, or another person).

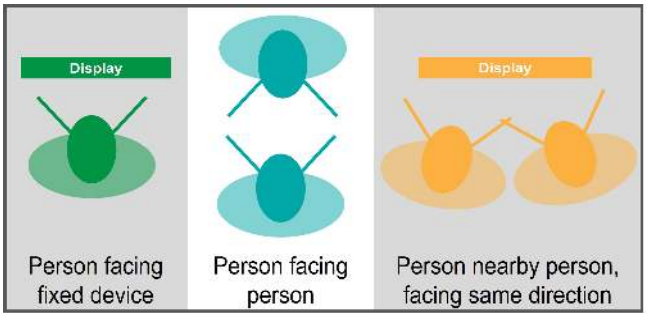

Figure 8. *EagleView* supports different conditions for attention grouping: a person facing a fixed object, two people facing each other, or when two people face the same fixed object.

In the scenario: *Mary needs to analyse the video recordings of her deployment in the museum. She loads the video footage of the four cameras, as well as the recorded tracking data into EagleView. During the configuration she marks the interactive screens' locations by drawing rectangles on a still video frame and saving each as a fixed entity. She switches to the "Visualisation" tab and plays the videos while watching the real-time visualisation in the centre.*

She activates the movement trajectories to know where the visitors are about to move so she can appropriately switch her attention to the according sideview camera on the left. The attention grouping highlights the tracked entities in the same colour whenever one of the three conditions is met (Figure 8). As a result, Mary can quickly notice instances where visitors are looking at an interactive screen or are talking to each other (as they change colour whenever a condition is met). When she finds an interesting interaction, she uses her pre-defined tags (configured during the configuration phase) to quickly tag those moments for later review. She can switch the background image of the top-down camera in the real-time visualisation on and off if she needs more clarity or wants to check for tracking errors.

Aggregated overview visualisation

Aggregated overview visualisations show a summary of people's location over time (Figure 2e,f,g). They are shown in the tab *Summary View* in the right column. Analysts can choose the time interval for which they want to show aggregated overviews through two sliders at the top. Each tracked user is shown in an individual visualisation. We implemented two visualisations: *heatmap*, showing where users were most active (2f) and *movement trajectories*, showing a user's movement path (2g).

In the scenario: *Mary wants to know what path one particular active visitor took through the exhibition. She switches to the overview visualisation of the movement trajectories of this visitor. She narrows the time visualised down to the duration of his visit, allowing her to get an overall picture of that visitor's movement throughout the museum. She realises that he frequently walked around each interactive display but moved less around non-interactive exhibits.*

Through the heatmap visualisation, she can confirm her observation: the visitor has spent most of his time around the displays. She compares his heatmap with those of other visitors, by overlaying them on top of each other in the real-time visualisation view. She realises that all visitors spent most of their time around interactive exhibits and very little around non-interactive ones. Something worth investigating!

Evaluating Real-time and Overview Visualisations

In this first study, we validate *EagleView* through *demonstration* and provide an external validation of its usability and usefulness to support interaction analysis [17], through a scenario-based *usage evaluation* with experts, following our previously described scenario (following strategies from [17]). The focus of this first study is on evaluating the real-time and overview visualisations for interaction analysis.

Participants

We recruited seven researchers in the HCI domain (2 female, 5 male) between 24 and 34 years (M=28, SD=3.9) from the UK and Canada. Participants were all active researchers (two post-doctoral researchers) or students (four PhD students, one MSc student) and all had previous experience conducting research involving group collaborations. They had between 2 and 10 years of experience in conducting HCI research (M=5.4, SD=2.6) and all but one had conducted video analysis. We conducted the study either in person or remotely (Skype). We incentivised participation in the study with gift vouchers. The in-person study lasted for 60 minutes and remote participation lasted for 90 minutes (due to longer technical setup and accounting for technical issues).

Study Design

We used two sets of videos for the study scenario for participants to analyse. Both sets were re-enacted interactions of multiple users with each other, a public display, and their own handheld devices, recorded through a top-view Kinect v2 and EagleSense [32]. We captured the data from the EagleSense API, as well as the raw RGB and depth video. Additional cameras recorded the interaction from the side. The first set of videos was used for training purposes and getting used to the interface of the querying tool; the second was used for the evaluation task.

Procedure

Participants were invited to our lab. After an introduction to the study, and after giving informed consent, participants answered a pre-study questionnaire with basic demographic data and prior experience in HCI and video analysis. Questionnaires were administered on paper for local participants and via an online questionnaire for remote participants. We used the answers to the questionnaire as the basis for a semi-structured interview to gain insights into prior video analysis experience.

Participants then received training about how to use the tool. We allowed them to freely explore the tool, answering any questions they had. We provided guidance along the way to make sure that they explored every aspect of the tool and were familiar with all its functionality. After answering any questions, the second set of data was provided, and a general description of the scene depicted was given to set the context. Participants were asked to answer eight questions, while thinking aloud. No time limit was given. After task completion, participants filled in a post-study questionnaire, addressing the usefulness as well as the usability of each visualisation component (both rated on a 5-point Likert scale). Afterwards, a semi-structured interview was conducted to elicit (i) insights into difficulties during the tasks; (ii) usefulness of certain tool aspects related to the task given; (iii) incorporation of the tool into their own workflow; and (iv) ideas for other features or changes.

The interaction of participants with the system was video- and audio-recorded for later analysis, either over their shoulder or via screen-recording (remote participants).

Prior experience and focus points in prior experience

Our participants reported that they had used video analysis for a varied set of study tasks (e.g. in public settings indoors and outdoors, in controlled lab environments, as well as in classroom experiments), and the analysed videos lasted anything from a few minutes to several hours. The focus points of their analysis could be that of a pre-existing or adapted coding scheme, entirely open coding, or a mix of both. Participants reported using different commercial or research tools (e.g. ChronoViz [5], NVIVO [25], or ATLAS.ti [34]) as well as simple playback in a video player on one screen with a spreadsheet to record information on a second screen.

Regarding their general approach, most participants reported that they first watched the entire video (sometimes at a higher speed) to get an overall feel of the interaction. This was particularly used when the researcher did not observe the actual interaction (e.g. in a public long-time deployment, or when the study was conducted by a different researcher). During this initial screening, participants reported that they often already note down events of interest relating to their search objectives (e.g. through tagging start- and end-point). This allowed them to later go back to these sequences and analyse them in detail. During this second step they then visually searched for instances of their pre-defined objectives or analysed the previously marked instances in more detail. They used tags and annotations to mark the video. Since a sequence of a video could contain multiple search objectives, researchers might watch the same sequence multiple times to completely identify all the details in the video.

After tagging and annotating the video in this way, researchers reported that they often exported the data to analyse it further with statistical tools or through qualitative methods. For example, they might compare the number of occurrences of each interaction or try to find usage patterns by using event marker visualizations.

Results

We report the results of our evaluation on the dimension of usability and usefulness.

General feedback

We observed that participants learned how to use the system well within the given time. Figure 9 shows an overview of participants' answer from the post-study questionnaire. Participants were comfortable in using the system (Q2. Median=2; inter-quartile-range=1) and found it easy to use (Q3. Md=1; iqr=1). Most participants felt that that the system was not complex (Q5. Md=4; iqr=1) and not cumbersome to use (Q7. Md=5; iqr=1).

Learnability

The study was setup in a way that participants learned the usage of *EagleView* on-the-go while they were conducting video analysis with the first set of video data. They explored the functionality of the tool and were guided by the study facilitator to ensure that every aspect of the system had been explored. All but one participant disagreed or were neutral to the question of whether they needed to learn a lot (Q6. Md=4; iqr=2) and all expected that most people would learn how to use it quickly (Q1. Md=2; iqr=1).

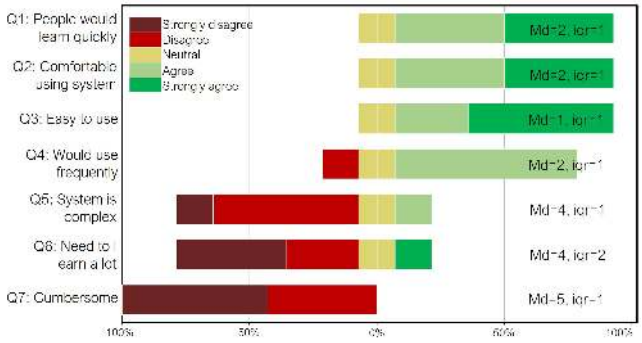

Figure 9. Results of post-study questionnaire about *EagleView's* usability. Answers are on a 5-point-Likert-scale. N=7.

An example of a feature that was not readily understood is *EagleView's* 'attention grouping' (Figure 7 and Figure 8), which shows one or multiple tracked users in the same colour if they have a shared point of attention (directly facing a fixed object; focussing on each other; two people standing next to each other and facing the same fixed object). The third condition at times confused participants, e.g. in instances where two people, who have no interaction with each other, were looking at the same (public) display. P5 suggested the usage of individual colours for each person, so that "*if two people are both looking at the display then one person is blue, one person is red, then the display will show red and blue, so we know that both are looking at the display. (Otherwise) it's hard to tell them apart. Especially with two or three people*".

Use in real-world practices

Although one participant disagreed, most participants could well imagine using it frequently (Q4. Md=2; iqr=1). However, most of the participants stated that they would not rely solely on the visualisations but would want to use them in conjunction with the video data to gain more detailed insights into the users' interactions. In particular, participants who currently use a video player with a spreadsheet or notes application on the side could imagine themselves using the tool for their analysis.

On the other hand, some participants used the real-time visualizations as a simplified version of the raw video. It provided them with a cleaner view of the information they were really interested in (people's position, orientation, and interactions). For example, P7 turned off the video background during the evaluation task and solely relied on the visualisations as it "*gives the most objective view of the relationship[s] that are happening*" (P7).

Usefulness

Responses to our post-study questionnaire showed that the different visualisation components were perceived as useful

for video analysis. For example, P7 described how he would use *EagleView* in his own research to easier identify when people look at a screen alone or together. This information was not available with other video analysis tools. He further said that he liked how "y*ou get to see when people use devices, when people are talking to each other, when people meet each other, which I wouldn't [be] able to get from other tools*" (P7). P2 highlighted that the overview visualisations gave insights that were not available by watching video playback, for example the "*heat map […] is very difficult to get out of a transcription because how would you do that*". He also commented on the use of movement traces stating that "*with the traces, you can do predictions that you couldn't do [otherwise]*".

Overall, we found that *EagleView* well supported analysts to quickly gain an *overview understanding* of the interactions recorded and gain more *detailed insights* through the real-time visualisations during the initial viewing of the video material. After initial viewing, researchers then often conduct detailed analysis of key interaction scenarios. To better support these, *EagleView* allows analysts to create search-queries on the key proxemic dimensions [6].

EAGLEVIEW SEARCH-QUERIES

Once a researcher has gained a (basic) understanding about the interactions in the recorded sequence, they can analyse the recorded interaction sequence by means of spatial queries based on people's distance, orientation, location, and activity within the captured tracking data (Figure 3C).

Search-queries

Search-queries can be created through a graphical interface on a still-frame whenever the video is paused. Such a query constrains the search within the recorded spatial tracking data to only the events that fulfil the criteria defined in the query, and allows researchers to find all relevant instances in the (possibly large) video dataset that meet these conditions.

New queries can be created by selecting one of the properties (Figure 11.1) and then adjusting the parameters on a graphical interface overlaid on top of the video (Figure 11.2 and 10.3). Once a query is created, the spatial tracking data will be parsed for matching conditions of the query and the results are highlighted in an additional timeline (Figure 11 bottom). A click on each result jumps to the position in the video.

Query creation

Our query creation tool is inspired by EXCITE's [21] idea of allowing analysts to search video for interaction events. However, rather than using a declarative language to describe a query, *EagleView* uses a graphical user interface for query creation and setting of parameters. A search-query is a combination of **property** (distance, orientation, location, activity), **entity** (person and/or object), and **parameters** (specifying a value for the property being searched for). Query creation is a three-step process (Figure 11): First the analysts selects a property (Figure 11.1). They then select the relevant entities (e.g. person or large screen) by clicking on the overlaid items on the still frame (Figure 11.2). Lastly, the search parameters can be adjusted (Figure 11.3 a-d) for a single entity (e.g. a person's orientation, location, or activity) or between two entities (e.g. specifying the relevant distance threshold between two people or between a person and a fixed object). Matched results for each query show up in the event timeline. Each query is shown in its own timeline.

For example, in Figure 11a, an analyst creates a query to search for instances where the distance between person 2 and 3 is smaller than 100cm to find all instances when the two people are standing in close proximity. The analyst then selects the orientation property (Figure 11a) and changes the relevant opening and orientation angle, to find all events when the two people in the video face each other. Last, the analyst specifies a query for all instances when any of the two people stand in front of the large display (Figure 11c).

In the scenario: *Mary now wants to understand further why the visitors spent most of their time around the interactive exhibits. She creates a "Location" query by marking the area around the interactive displays. The search results indicate all instances where a user is in front of an interactive display. Through this search-query, the 10 hours of video are narrowed down to 120 results of 30-90 seconds – only 120 minutes in total. Mary can now review those instances in more detail by clicking on the search results in the timeline. Through the review, she finds that visitors frequently get their smartphone out around the interactive exhibits to select the accompanying audio-guide and listen to the narration.*

Compound Queries

Analysts can also create *compound queries*, which are comprised of two or more queries. The results are then filtered to only include instances where all queries match (AND logic connection). Compound queries show as a combined section in the event timeline. For instance, in our example the researcher is interested in finding all instances of so called L-shape F-formations (a sociological lens for analysing pair interactions based on their spatial characteristics [15]). For finding these formations, the analyst creates a compound query: *finding all instances where the orientation angle between two people is around 90 degrees (by visually adjusting to a wider tolerance angle; Query1), AND distance is below 2 meters (Query2)*. Once completed, the new compound query highlights these F-formation instances in the timeline.

In the scenario: *Mary now wants to know whether people who are visiting the museum in a group also use their smartphone in some similar way to single visitors. She therefore needs to find all instances where two people are in front of an interactive display, while they are using their smartphones. She creates a three-part compound query: first she creates a "Location" query like in the previous example to filter for location matches around interactive exhibits. She then adds a "Distance" query to filter for instances where two people are closer than 100cm. Last, she adds an "Activity" query with its search parameter set to "Smartphone" to*

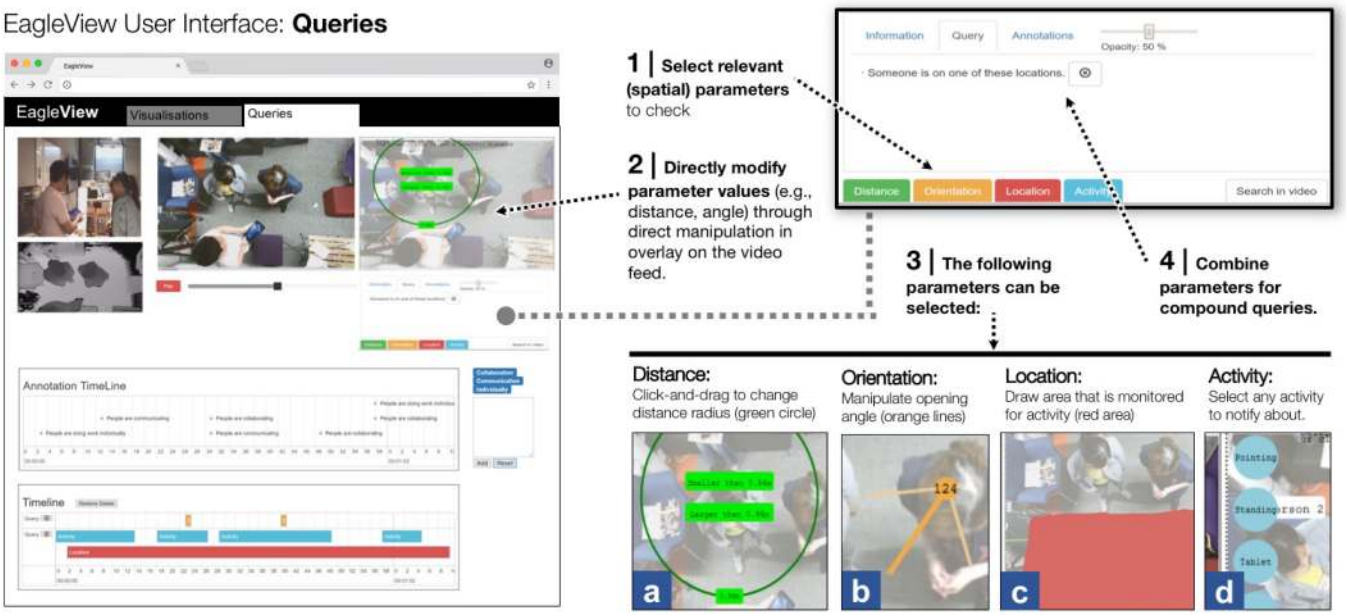

Figure 10. Queries are created via a GUI laid on top of a still video frame (steps 1-3). Queries can be combined to make up more specialised compound queries (4).

filter for moment when both visitors are using their smartphone. By reviewing the 70 search results, she finds that group visitors less frequently listen to the audio guide and rely on reading the museum guide.

Evaluating querying interface

We ran a separate, second user study (with a new set of expert user participants to the first one) with the study focussing on the *query-creation part* of the tool.

Participants

We recruited four expert users (3 female, 1 male) aged between 23-33 years (M=28, SD=3.8), all HCI researchers who had prior experience (3-12 years) in conducting video analysis for interaction design. The study lasted for ~60 minutes and participants received a £10 GBP voucher.

Study Design and Procedure

We used the same dataset as the first study, comprising two staged interaction scenarios.

Participants were invited to our lab. After giving informed consent, they answered a pre-study questionnaire about basic demographic data and prior experience in HCI and video analysis. We used the answers to those questions to conduct a short interview about prior experience. Participants received training for the tool in a similar way to our first study. They were then given a second, separate set of videos, received a general description of the scenario they could see in the video, and were asked to imagine that they wanted to conduct analysis on this video data. After answering any questions about the scenario and system, they received a set of five question and were asked to answer them during the analysis of the video, while thinking aloud. No time limit was given. After task completion, participants answered a post-study questionnaire and we conducted a short semi-structured interview to follow up on our observations and to gather further insights into how they envisioned *EagleView* might support their work. The session was video- and audio-recorded for later analysis.

Results

Current Practices of Participants

Participants reported to having conducted video analysis for different purposes and focus. They reported that they generally first skim through the entire video to get an overall understanding and quickly mark sequences of interest. Participants reported that this was the most time consuming and "*boring*" (P4) task. They then review each of the highlighted sequences repeatedly, to annotate them and conduct further tagging or analysis on the video data (much like the participants in our first study). For example, P4 stated that "*I will watch [each clip] 7-8 times slower to get an idea of [the movement]. […] I will try to draw a diagram [that] gives me the sequence of movement*". Often "*tags to annotate a period of time [were used, with] different colours which help differentiate them*" (P3).

Findings About Use of Query Tool

In the following, we focus on reporting qualitative insights and observations. We refrain from reporting quantitative results because of the small number of participants.

Generally, participants agreed that the querying tool of *EagleView* has a good usability and "*is rather easy [to use]. With just one or two queries you can know what specific activity one is doing, the distance between him with others, and also his attention*" (P2).

The task given required participants to create two single-property queries and three compound queries. We observed that all participants successfully created the queries to find the instances of the single-property events (using *location* for the first and *activity* for the second question) and P4 said "*it is very straightforward to do so*". P2 added that "*the query makes the analysis less time-consuming. […] You still have to do a little bit analysis by yourself, but you probably want to do that anyway, because you do not want to trust the system to do the analysis fully*".

However, for the last three questions, participants had to create compound queries to find events that included more than one property (e.g. activity of one person while facing a fixed object). We found that participants fell back to either manually searching for the specified interaction or that they created partial queries to narrow down the data to a few instances that contained at least one of the relevant properties (i.e. activity) and then manually check for other property (i.e. orientation). P1 noted that "*if I use the activity query it takes me less time, but I still need to see if the guy is looking at somewhere else, so I need the orientation. But I think [manually searching and compound query creation] will take me [a] similar time*". This might have to do with the length of the overall video in our study (<3 minutes); in a longer scenario analysts' behaviour might have differed. Participants stated that they would expect the querying tool to be particularly useful *"in a complicated research analysis situation […where] people would be using the query rather than manually [searching for interactions]"* (P3) and that "*it might be more useful if [they] need to analyse longer video[s]*" (P2) or "*want […] to keep track [of] complicated things*" (P4).

DISCUSSION

EagleView supports researchers to explore their video material in a way which was not easily possible before, and supports them throughout their analysis process (Figure 3). Researchers do not necessarily need to know what conditions they are looking for but can follow an iterative approach of switching between using the visualisations for serendipitous discovery and the **search-queries** for a more fine-grained analysis. The **overview visualisations** enable users to see movement over a longer period of time, acting as a *summary* of what is going on. It allows analysts to discover trends, directly linked to the videos. For example, the movement traces can be used to easily spot familiar movement patterns, such as *audience funnels* [23]. The **real-time visualisations** enable researchers to view a *clean* representation of the data, while having the ability to switch to the video feed. Our study showed that the visualisation-only playback of the data stream allows researchers to use this sensor data as an *additional video feed*, joining it with the camera feed whenever more detail is desired. As a result, analysts can get focus points for their analysis and can quickly gain an overall understanding of the material.

F-formations [15,20] are a powerful lens for interaction analysis. While *EagleView* allows analysts to easily create queries to search for these, they currently need to be created manually every time they are used. *Query-templates* would allow for quick searches, filtering for the most frequent interactions. Further, search-queries currently can only be used for parallel conditions. Creating *temporal queries*, that highlights a sequence of interactions, would enable researchers to easier analyse temporal aspects, such as with the *audience funnel* framework [23]. We are currently planning to use *EagleView* for the analysis of a real-life interaction study. This will allow us to gather further insights into which templates would be useful outside of a fictional scenario (like during our user study).

The current implementation of *EagleView* uses top-down tracking information. Although tracking-system-agnostic, this is a limitation, as a top-down camera cannot be installed in every location. Further, the current implementation of viewing direction only acts on 2D information. If a person is e.g. looking *over the top* of a display, this still counts as an interaction. The researcher then has to use an additional side-view camera to gain further insights.

Our two studies evaluated the use of visualisations and showed that the querying function enabled analysts to quickly select interaction scenarios of interest. While both our studies were conducted with domain experts (i.e. researchers who have actively conducted video analysis), we acknowledge that these studies do not show how such a tool would perform in everyday or long-term use. Further, from our study participants we learnt that interaction analysis on the video material is only the first step and a more detailed quantitative (time, duration, and frequency) and qualitative (on verbal transcriptions and other annotations) analysis is often needed. Currently, this has to be done on the exported JSON data of tags and annotations. We see further potential for future extensions of *EagleView*. For example, our current real-time visualisations of F-formations could be extended with dashboard-like summary views of the different formations recorded. Furthermore, it would be interesting to consider visualisations and query constructs for more fine-grained gestures that people perform.

CONCLUSION

EagleView is a novel tool directly supporting researchers performing interaction analysis through video coding and integrates a querying tool as well as real-time and overview visualisations, making it easier to find relevant sequences in the videos to interpret. We invite the ISS and HCI research community to use (and extend or re-appropriate) our tool: *EagleView* is available as open-source software at https://github.com/frederikbrudy/eagleview.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their feedback and all of our study participants. This work has been supported by Microsoft Research through its PhD Scholarship Programme.

REFERENCES

1. Frederik Brudy, Joshua Kevin Budiman, Steven Houben, and Nicolai Marquardt. 2018. Investigating the Role of an Overview Device in Multi-Device Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 300:1–300:13. https://doi.org/10.1145/3173574.3173874

2. Brandon Burr. 2006. VACA: a tool for qualitative video analysis. In *CHI'06 extended abstracts on Human factors in computing systems*, 622–627. Retrieved from https://doi.org/10.1145/1125451.1125580

3. Stamatia Dasiopoulou, Eirini Giannakidou, Georgios Litos, Polyxeni Malasioti, and Yiannis Kompatsiaris. 2011. A Survey of Semantic Image and Video Annotation Tools. In *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, Georgios Paliouras, Constantine D. Spyropoulos and George Tsatsaronis (eds.). Springer Berlin Heidelberg, 196–239. https://doi.org/10.1007/978-3-642-20795-2_8

4. Pryce Davis, Michael Horn, Florian Block, Brenda Phillips, E. Margaret Evans, Judy Diamond, and Chia Shen. 2015. "Whoa! We're going deep in the trees!": Patterns of collaboration around an interactive information visualization exhibit. *International Journal of Computer-Supported Collaborative Learning* 10, 1: 53–76. https://doi.org/10.1007/s11412-015-9209-z

5. Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D. Hollan. 2011. ChronoViz: a system for supporting navigation of time-coded data. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 299–304. Retrieved May 11, 2017 from https://doi.org/10.1145/1979742.1979706

6. Saul Greenberg, Nicolai Marquardt, Till Ballendat, Rob Diaz-Marino, and Miaosen Wang. 2011. Proxemic interactions: the new ubicomp? *interactions 18*, 42–50.

7. Joey Hagedorn, Joshua Hailpern, and Karrie G. Karahalios. 2008. VCode and VData: illustrating a new framework for supporting the video annotation workflow. In *Proceedings of the working conference on Advanced visual interfaces*, 317–321. Retrieved from https://doi.org/10.1145/1385569.1385622

8. Edward Twitchell Hall. 1969. *The hidden dimension*. Anchor Books New York.

9. Gang Hu, Derek Reilly, Mohammed Alnusayri, Ben Swinden, and Qigang Gao. 2014. DT-DT: Top-down Human Activity Analysis for Interactive Surface Applications. 167–176. https://doi.org/10.1145/2669485.2669501

10. N. Hu, G. Englebienne, and B. Kröse. 2013. Posture recognition with a top-view camera. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2152–2157. https://doi.org/10.1109/IROS.2013.6696657

11. Petra Isenberg, Danyel Fisher, Paul Sharoda A., Meredith Ringel Morris, Kori Inkpen, and Mary Czerwinski. 2012. Co-located Collaborative Visual Analytics Around a Tabletop Display. *IEEE Transactions on Visualization and Computer Graphics* 18, 5: 689–702. http://dx.doi.org/10.1109/TVCG.2011.287

12. Mikkel R. Jakobsen and Kasper Hornbæk. 2014. Up close and personal: Collaborative work on a high-resolution multitouch wall display. *ACM Transactions on Computer-Human Interaction* 21, 2: 1–34. https://doi.org/10.1145/2576099

13. Brigitte Jordan and Austin Henderson. 1995. Interaction Analysis: Foundations and Practice. *Journal of the Learning Sciences* 4, 1: 39–103. https://doi.org/10.1207/s15327809jls0401_2

14. Wendy Ju, Brian A. Lee, and Scott R. Klemmer. 2008. Range: exploring implicit interaction through electronic whiteboard design. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 17–26. Retrieved from https://doi.org/10.1145/1460563.1460569

15. Adam Kendon. 2010. Spacing and Orientation in Co-present Interaction. In *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer, Berlin, Heidelberg, 1–15. https://doi.org/10.1007/978-3-642-12397-9_1

16. Walter S. Lasecki, Mitchell Gordon, Danai Koutra, Malte F. Jung, Steven P. Dow, and Jeffrey P. Bigham. 2014. Glance: rapidly coding behavioral video with the crowd. 551–562. https://doi.org/10.1145/2642918.2647367

17. David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation Strategies for HCI Toolkit Research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 36:1–36:17. https://doi.org/10.1145/3173574.3173610

18. S. C. Lin, A. S. Liu, T. W. Hsu, and L. C. Fu. 2015. Representative Body Points on Top-View Depth Sequences for Daily Activity Recognition. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2968–2973. https://doi.org/10.1109/SMC.2015.516

19. Nicolai Marquardt, Robert Diaz-Marino, Sebastian Boring, and Saul Greenberg. 2011. The proximity toolkit: prototyping proxemic interactions in ubiquitous computing ecologies. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 315–326. Retrieved from https://doi.org/10.1145/2047196.2047238

20. Nicolai Marquardt, Ken Hinckley, and Saul Greenberg. 2012. Cross-device interaction via micro-mobility and f-formations. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, 13–22. Retrieved from https://doi.org/10.1145/2380116.2380121

21. Nicolai Marquardt, Frederico Schardong, and Anthony Tang. 2015. EXCITE: EXploring Collaborative Interaction in Tracked Environments. In *Human-Computer Interaction*, 89–97. Retrieved from https://doi.org/10.1007/978-3-319-22668-2_8

22. Paul Marshall, Yvonne Rogers, and Nadia Pantidi. 2011. Using F-formations to analyse spatial patterns of interaction in physical environments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 445–454. Retrieved from https://doi.org/10.1145/1958824.1958893

23. Daniel Michelis and Jörg Müller. 2011. The audience funnel: Observations of gesture based interaction with multiple large displays in a city center. *Intl. Journal of Human–Computer Interaction* 27, 6: 562–579. https://doi.org/10.1080/10447318.2011.555299

24. Roshanak Zilouchian Moghaddam and Brian Bailey. 2011. VICPAM: a visualization tool for examining interaction data in multiple display environments. In *Symposium on Human Interface*, 278–287. Retrieved from https://doi.org/10.1007/978-3-642-21793-7_32

25. Nvivo. 2017. Nvivo, www.qsrinternational.com. *Nvivo*. Retrieved November 25, 2016 from http://www.qsrinternational.com/

26. Stacey D. Scott, M. Sheelagh T. Carpendale, and Kori M. Inkpen. 2004. Territoriality in collaborative tabletop workspaces. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, 294–303. Retrieved from https://doi.org/10.1145/1031607.1031655

27. Anthony Tang, Saul Greenberg, and Sidney Fels. 2008. Exploring Video Streams Using Slit-tear Visualizations. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (AVI '08), 191–198. https://doi.org/10.1145/1385569.1385601

28. Anthony Tang, Michel Pahud, Sheelagh Carpendale, and Bill Buxton. 2010. VisTACO: visualizing tabletop collaboration. In *ACM International Conference on Interactive Tabletops and Surfaces*, 29–38. Retrieved from https://doi.org/10.1145/1936652.1936659

29. Daniel Vogel and Ravin Balakrishnan. 2004. Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*, 137–146. Retrieved from https://doi.org/10.1145/1029632.1029656

30. Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision* 101, 1: 184–204. https://doi.org/10.1007/s11263-012-0564-1

31. Andrew D. Wilson and Hrvoje Benko. 2010. Combining Multiple Depth Cameras and Projectors for Interactions on, Above and Between Surfaces. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology* (UIST '10), 273–282. https://doi.org/10.1145/1866029.1866073

32. Chi-Jui Wu, Steven Houben, and Nicolai Marquardt. 2017. EagleSense: Tracking People and Devices in Interactive Spaces using Real-Time Top-View Depth-Sensing. 3929–3942. https://doi.org/10.1145/3025453.3025562

33. Ulrich von Zadow and Raimund Dachselt. 2017. GI-AnT: Visualizing Group Interaction at Large Wall Displays. 2639–2647. https://doi.org/10.1145/3025453.3026006

34. ATLAS.ti: The Qualitative Data Analysis & Research Software. *atlas.ti*. Retrieved July 6, 2018 from https://atlasti.com/