PaperVideo: Interacting with Videos On Multiple Paper-like Displays

Roman Lissermann[1], Simon Olberding[2], Benjamin Petry[1], Max Mühlhäuser[1], Jürgen Steimle[2]

[1]Technische Universität Darmstadt
64289 Darmstadt, Germany
{lissermann,benjaminpetry,max}@tk.informatik.tu-darmstadt.de

[2]MIT Media Lab
Cambridge, MA, USA
{simon_o, steimle}@mit.edu

ABSTRACT

Sifting and sense-making of video collections are important tasks in many professions. In contrast to sense-making of paper documents, where physical structuring of many documents has proven to be key to effective work, interaction with video is still restricted to the traditional "one video at a time" paradigm. This paper investigates how interaction with video can benefit from paper-like displays that allow for working with multiple videos simultaneously in physical space. We present a corresponding approach and system called PaperVideo, including novel interaction concepts for both video and audio. These include spatial techniques for temporal navigation, arranging, grouping and linking of videos, as well as for managing video contents and simultaneous audio playback on multiple displays. An evaluation with users provides insights into how paper-based navigation with videos improves active video work.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interface

General Terms

Design, Human Factor

Keywords

Tangible user interface, electronic paper, flexible display, thin-film display, multiple displays, video, pile.

1. INTRODUCTION

People from a variety of professional backgrounds are confronted daily with large amounts of video footage that they must sift trough and make sense of: TV news editors have to deal with approximately 30 hours of video material offered per news agency and day (such as Reuters). A Hollywood movie director must distill hundreds of hours of footage into a blockbuster movie. Analysts and researchers must make sense of information that is contained within many videos, such as CCTV recordings or recordings of scientific experiments. The Youtube era extends these tasks of sifting and making sense out of many videos to the general population, for hobby and scholarly activities. These examples show that *active video work* with large amounts of video material (as opposed to passive watching of a single video) is a daily routine of many people. As such it is obvious that better usability for active video work is a research topic of primary importance.

In contrast to active video work, sifting and sense-making of *paper-based* information is a well-researched field. Research shows that the key is using not only one, but multiple documents or sheets of paper simultaneously, in order to manipulate and organize information in physical space. Amongst others, this has proven to effectively support comparison, generating an overview and better orientation [11,28]. All these activities are also of paramount importance when working with videos.

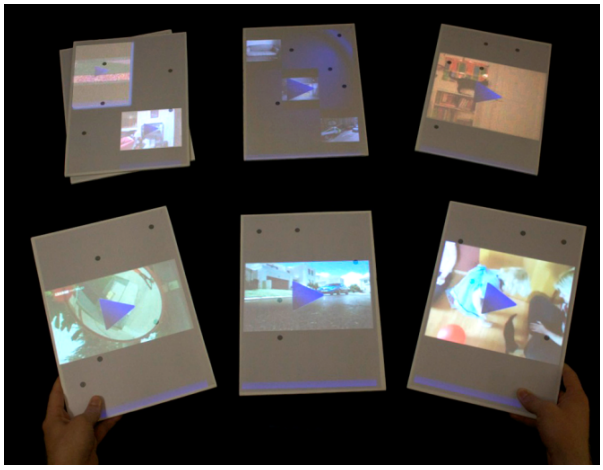

Figure 1: PaperVideo: A system for interacting with multiple videos in physical space

Compared to these paper-based practices, today's user interfaces for active video work are characterized by two main shortcomings: (1) While standard navigation techniques for videos have their obvious benefits, they lack the effectiveness of physical interaction [11,13,28] for spatially structuring videos. (2) The traditional "one video at a time" paradigm does not leverage the whole spectrum of human perception. While humans are able to focus only on limited information, they are able to grasp a much higher amount of information in the periphery, which is helpful for getting an overview and structuring tasks.

We argue that these affordances of working with multiple documents in physical space can be beneficially transferred to the domain of video. We advocate a paradigm for videos which consists of using many videos simultaneously, similarly to how we lay out multiple printed documents on our desk. In this paper, we investigate how interactions known from physical documents can be transferred to the world of videos and fitted to a computing device that offers multiple electronic displays, which are very similar to paper. Note that we do not aim to replace existing practices of watching a movie for entertainment purposes. We aim at supporting three widespread scenarios that all involve actively working with multiple videos:

1. *Systematic analysis* of videos, e.g. for learning or research, similar to active reading of text documents [1]. Thereby the main activities are sorting, comparing and (re)structuring.

2. *Playful exploration* of video contents, e.g. at installations in museums or shops, by providing an intuitive way of interaction and
3. *Lightweight video editing,* where people select and combine video clips, e.g. to create a personal excerpt.

In this paper we present PaperVideo, a coherent system that allows users to play back and navigate through videos and collections of videos with multiple physical displays (Fig. 1). It enables users to create an overview of multiple videos as well as structure and organize video contents by leveraging physical arrangements. We contribute a set of interaction techniques for video content that takes advantage of the characteristics of dynamic displays. These techniques go beyond established physical interactions such as arranging and piling of paper. Furthermore, we introduce interaction techniques supporting the management of contents on multiple displays. Addressing the possibility of viewing multiple videos in parallel, we finally contribute several novel sound concepts that allow the user to mentally grasp multiple audio sources of videos simultaneously. Results from a qualitative user study show that the system effectively supports active video work.

To technically realize displays that can be manipulated like paper, the system makes use of a projection-tracking setup. Passive cardboards are tracked in real-time. Visual content is automatically projected onto them, correcting for perspective distortions. We envision such systems to be installed at offices, schools, libraries, museums, and stores. Moreover, given the rapid advances in mobile devices, future tablet devices are very likely to be much thinner and more lightweight than nowadays. This will eventually render projection-tracking obsolete and also allow for mobile use cases of our system.

The remainder of the paper is organized as follows. First, we discuss related work then describe the underlying requirements of our design. Next, we present the system design, starting with interactions for spatial arrangements of video, followed by interactions for managing multiple physical displays and ending with sound concepts for multiple videos. Later we present the implementation of our system. Finally, we discuss the results of a user study.

2. RELATED WORK

Prior research on video navigation investigated interaction on PCs [3,10,18,26] and mobile devices [7,8,33]. Manske introduced a concept for browsing a video on a large display by visualizing the video as a 3D content tree [16]. Schoeffmann has concentrated on hierarchical video browsing, with 3D graphics support, and provided an intuitive way of content navigation within a video [25]. Nevertheless, all of these systems provide only one single screen. In contrast, CThru [9] combined multimedia content (images, videos and text) with a storytelling educational video on an interactive tabletop and duplicated the view onto a wall size display. None of these works have focused on digital spatial management and spatial cues while handling video content.

A large body of empirical work [11,28,34] shows that physical space management is important for overview and organization of information and that physical interactions with paper have different qualities than physically-inspired interactions on touch screens. The usage of multiple physical displays has been investigated for several purposes. Rekimoto [20] presented a system where multiple transparent tiles were transformed into interactive controls by placing them onto a flat panel display. Siftables [17] demonstrated the technical feasibility of a system with tiny, wirelessly interconnected color displays, introducing multi-display interactions for gaming and educational purposes. Other work presented examples of how several tiny bezel-less screens can be used for interactive board games [22] and studied gestures for linking multiple displays [5]. These works inspired us to provide dynamic visual contents on multiple displays.

Several systems support tangible interaction with video contents: Video Mosaic offers a tangible interface for editing video [13]. A snippet of normal paper can be used as a physical token that represents a video. By holding the snippet in front of a camera, the video is played back on a PC screen. A similar approach, using RFID tokens, is presented in [29]. However, these systems do not display the video on the paper snippet, but on a nearby screen. This creates an indirection that is overcome in our work. Video mosaic targets video editing whereas our focus is on sifting, exploration and sense-making of video collections. Finally, Tangible Video Editor [36] presented a set of small active displays that can each host a video snippet. By physically arranging displays in a linear sequence, the temporal sequence of clips can be edited. The full video can then be displayed on a computer screen. This work influenced our approach of physically arranging video displays. In contrast, our work supports large video displays, a wider range of activities and introduces novel tangible interactions.

Ongoing advances in OLED display technology allow for displaying full color video on very thin and lightweight displays. Interaction with lightweight and thin displays has been a focus of various research projects since the seminal DigitalDesk [35] introduced first user interfaces such as a projected virtual calculator. Within this stream of research, PaperWindows [6] is very influential work. It was the first to present a user interface that is distributed over a set of very thin and lightweight paper-like displays. PaperWindows further contributed a set of interaction techniques for basic windowing tasks, however, it did not address interaction with videos. PaperLens [30] demonstrated how the space above the tabletop can be used for interactions with single paper-like displays. Our conceptual model (cf. section 4) was inspired by previous works [6,31] that have investigated gestures and an interaction vocabulary of paper-like displays. We improve upon these works by addressing paper-based interaction with both visual and audio contents, by introducing a set of interactions for managing multiple displays and by providing first empirical insights into how people use systems with multiple paper-like displays.

Our system setup allows users to play back multiple videos simultaneously. Multiple parallel audio outputs located in space produce the well-known cocktail party effect [2]. Our sound concepts are inspired by Audiosteamer [23] and Audio Hallway [24]. These works presented interaction techniques for browsing multiple audio sources by creating an audio-only environment and virtually placing the audio sources around the user's head. By simply turning the head, the user could select his sound of interest. In our work, we refine this research for audiovisual contents and tangible interaction while introducing further spatial audio concepts.

3. DESIGN REQUIREMENTS

3.1 Video-based Activities

With PaperVideo we aim at supporting three main scenarios of active video work, which benefit from multiple video displays:

1. Systematic video analysis: *Active reading* [1] is a well-studied domain. Active reading involves intensely engaging with

documents, for instance by following references, annotating, and comparing documents. People often work with multiple documents simultaneously and by effectively arranging them in a physical space to support their reading. Analogously, we propose systematic video analysis as a way of actively working with video material: people explore a set of videos, prioritize the content, study related content and compare and (re)structure the content. As outlined in the introduction, these activities are of crucial importance in a wide range of professions.

Different areas like film (post)production by novice/professional users or analysts of a huge amount of multiperspective camera recordings from a catastrophe are in need of prioritizing, comparing and (re)structuring video snippets. Based on the mentioned needs of actively working with videos, we are convinced that also in the case of videos, the use of space provides effective support for such highly creative and dynamic activities.

2. Playful video exploration: Multiple displays can be beneficial for exploring collections of videos, for instance at installations in places like stores, museums, or exhibition booths. We envision videos spread on booths or tables, where visitors can stop by and playfully explore new topics or products individually or collaboratively. The focus of such systems is not only on functionality, but also on high user experience as well as ease of access. Further, it enables users to serendipitously discover content. The system should be intuitively usable to allow a playful exploration and a positive experience to people of all ages and professional backgrounds.

3. Video editing: Simple video editing is a common part of using videos as a consumer. For instance, people trim video snippets or they order and align several snippets in a personal excerpt. PaperVideo supports such simple editing tasks. They are conceptually similar to highlighting or excerpting passages in active reading of text documents, which serve for better understanding and condensing the contents. This is opposed to advanced video editing that focuses on production of videos, including specific functions like time-stamping or synchronizing footage from multiple cameras over time, for sorting the recordings, annotating and augmenting video snippets or position in a video with additional information.

All of these scenarios have a set of functionality in common: users require functionality for quickly getting an overview of a single or multiple videos, for prioritizing content, finding related content, comparing content, and (re)structuring content. The system should support quick temporal navigation, cross-video use for overview, comparison and linking, as well as flexible means for prioritizing, grouping, and structuring.

3.2 Technical Requirements

Today's PCs and mobile devices usually have only one or two displays. Our system should support a significantly higher number of displays to support physical interactions that are known from the world of paper. These displays should be very thin and lightweight such that they can be easily moved and arranged in physical structures, such as piles. The displays should provide color output, high resolution and high update rate to play back video. Each display should also support direct input for navigation purposes as well as sound output. To allow for physical interactions that span multiple displays, each display should have knowledge about its relative position in space with respect to its neighbors.

While large tabletop displays would allow for laying out multiple videos in space, interaction on tabletops is inherently limited compared to using multiple physical displays: First, while tabletop interfaces mimic basic interactions with physical objects, the resulting interaction styles have been shown to be fundamentally different [34]. In particular, while people make ample use of both hands in physical setups, they mostly restrict interaction to only one hand at a time on tabletops. Physicality also offers a number of advantages such as cues for implicitly assessing the quantity of objects. It is also more difficult for users to arrange objects in a way that can be ergonomically read or viewed on a tabletop than with physical displays (this aspect was called micro-mobility [12] in the literature). Finally, tabletops require a static, immobile setup. In contrast, several small physical displays can be used in nomadic setups. While currently available technology does not yet allow us to realize our system without a static setup, nomadic uses can be supported in the near future.

To technically realize a system with multiple paper-like displays, our prototype uses a tracking-projection setup. Details about our prototype are provided in the implementation section below. The setup is well-suited for installations at fixed places, e.g. at work, in museums, stores and schools. The tracking-projection approach allows us to realize paper-like displays already today, even though currently available tablets are still too heavy and too thick for our interactions. This is very likely to change in the near future when thin-film OLED technology will further slims down the form factor of tablet devices. Our proposed interaction techniques can be fully transferred to such devices and will then allow for use in mobile settings in addition to the settings that can be supported currently.

4. INPUT DESIGN SPACE OF MULTIPLE DISPLAYS

For a systematic design of our interaction techniques, we investigated the design space of how input is performed with multiple location-aware displays. This allowed us to identify several interaction primitives, grouped along three different basic forms of input (see Fig. 2):

Spatial location input: Moving the display in space is translated into input. Thereby the absolute position of a display in physical space is captured.

Display proximity input: Changes in the relative positioning among two or more displays are translated into input. Interaction primitives include piling of displays and using one display as a pointer for selecting content on another display.

Gestures with and input on the display: The user can perform physical gestures with one display or with a set of displays, e.g. by shaking. Moreover, the user can directly interact with contents on a display using direct touch or pen input (our prototype currently supports only pen input).

5. INTERACTING WITH VIDEOS IN PHYSICAL SPACE

In this section, we present interaction techniques that support a set of base activities for individual videos and collections of videos. These techniques leverage the manipulation and arrangement of one or several displays in physical space.

5.1 Temporal Navigation

Temporal navigation within a video is one of the most basic functionalities. It is required to get an overview of the video as well as for quickly accessing specific passages. Similarly to

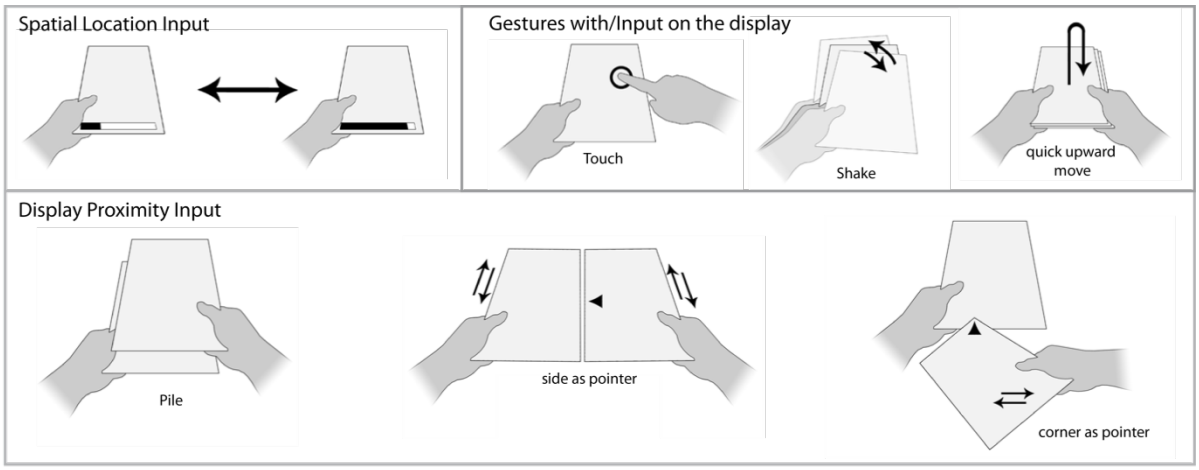

Figure 2: Input design space of multiple location aware displays

existing user interfaces for desktop computers and mobile devices, our design allows users to start, pause, and skim a video by directly interacting with widgets on the display using a stylus (see Fig. 3a, 3b). In contrast to most existing interfaces, it is possible to play multiple videos simultaneously.

Space is a strong cue for encoding information, for instance sequences [11]. This motivated us to design a technique in which the physical workspace encodes temporal positions. The timeline of the video is virtually spread out in physical space, extending from left to right within the user's arm reach (Fig. 3c). Each spatial position is mapped to a temporal position within the video. By moving a display through space (simultaneously moving several displays is also possible), the user navigates through the video. This technique allows for quickly skimming as well as for jumping back and forth between several passages of a video. To avoid interfering with free arrangements of multiple displays (see next subsection), temporal navigation is activated when the user lifts the display up to a higher level above the table surface, the "temporal layer".

5.2 Arranging

Similar to arranging objects in the real world, the thin and lightweight displays can be freely arranged on the table surface. To state only a couple examples, two videos can be compared by placing them side-by-side while multiple videos can be ordered in a spatially-encoded sequence. Videos can be prioritized by placing them closer or more distant to the user. Such arrangements enable powerful ways of organizing information in space [11].

5.3 See-through Pile

A large body of research shows the relevance of piling for managing information [14,15]. Users can place multiple displays on top of one another to form a pile of videos. Our pile is more advanced in comparison to piles of ordinary physical objects. Since the system is aware of which displays are occluded, the content of the *entire* pile is visualized on the topmost display, resulting in an "x-ray style" view (see Fig. 4). All of the content on the topmost display is fully interactive. So the user can view, play or skim any video in the pile easily.

Piling or unpiling does not interrupt playback; the video continues to play inside or outside a pile.

5.4 Accessing Related Videos

Many videos are organized in collections, in which they are linked to related videos. This is the case for influential video platforms such as YouTube, iTunes U [37], and OpenCourseWare [38]. We present a spatial technique for navigating video relations using multiple displays. By bringing an empty display near to a display with a video (see Fig. 5), the mode for selecting related videos is entered. A list of related videos visualizes on the video. By moving the empty display or the video display up or down, a related video can be selected from the list. While doing so, a preview of the currently selected video is shown on the empty display. By slightly removing one display from the other, the list shows categories or groups of videos instead of individual videos, allowing for a selection at a higher level. By moving one display apart, the related video (or group of videos) is eventually selected and displayed on the previously empty display (see Fig. 5).

There are two main advantages in this gesture while working with multiple displays. First, the original video is not replaced by the related one, as in most current solutions, but remains visible. Hence a spatial overview can be easily generated by leaving a trace of "where we came from". Second, multiple related videos can be opened by using multiple displays. These videos can then be spatially arranged and also viewed in parallel, if desired.

5.5 Linking Videos

The user can create his own hyperlinks between any two videos. This is done by taking two displays with different videos and bumping them against each other (see Fig. 6). From now on, the videos appear in the respective lists of linked videos.

5.6 Lightweight Video Editing

When people actively work with text documents, they highlight passages that are of high interest, write excerpts, and create text collages by copying and pasting relevant passages into a new document. In contrast, video documents are usually consumed as-is, without personalizing them. We propose a lightweight interaction technique for cutting videos. We do not aim for professional video editing, but on providing a simple interaction technique. This can be used for focusing on specific passages of a video and for composing a "video excerpt".

For cutting out a section of an existing video, an additional empty display is needed. By placing one corner of the empty display

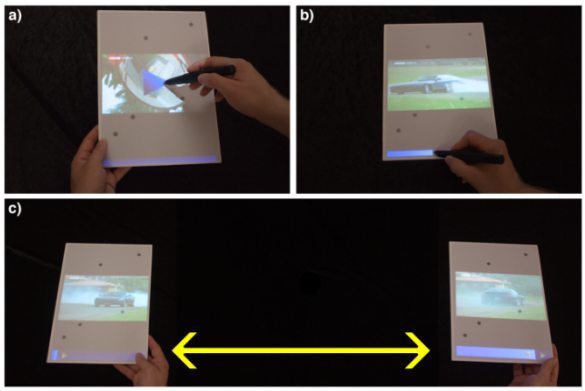

Figure 3: a) Playback, b) skimming, c) temporal navigation in space

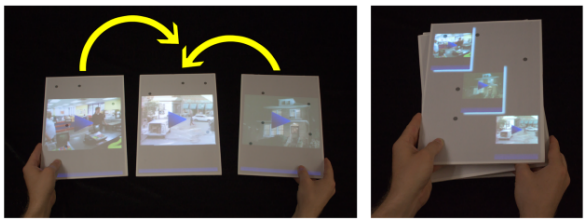

Figure 4: Physical piling of videos. The topmost display allows for interacting with all videos.

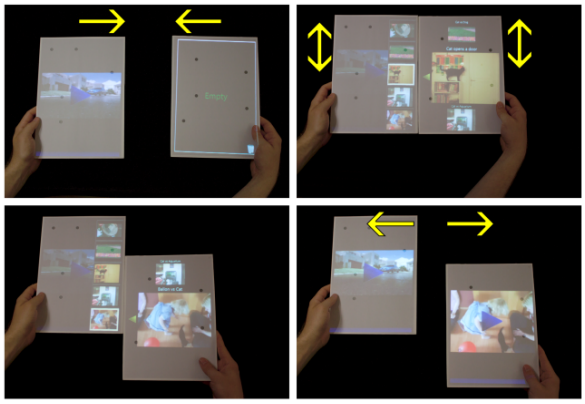

Figure 5: Accessing related videos by bringing displays side by side, selecting a related video and moving the displays apart

onto the timeline of the video, the corners are used as a video cutting tool. The start and end positions of the cut are selected with the upper-left and upper-right corners, respectively. While selecting, the start and end frames are visualized on the previously empty display and the entire passage is highlighted in the video timeline (see Fig. 7). By moving the display apart, the cut is executed and the newly created video snippet is made available on this display. From now on, the user can interact with this snippet as with any ordinary video. The next section discusses how several physical video snippets can be combined to one video.

6. MANAGING MULTIPLE DISPLAYS: VIRTUALIZING AND MATERIALIZING CONTENTS

Paper documents have a static mapping between contents and the physical carrier medium. One page of content is permanently

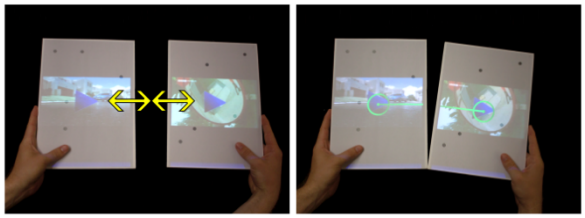

Figure 6: Linking two videos by bumping them against each other

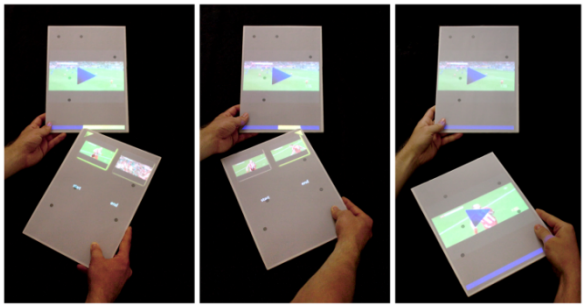

Figure 7: Playful video cutting with multiple displays

bound to one page of the carrier medium: this is a one-to-one mapping. Most computing devices have only one screen. Here, in contrast, the content is dynamic. Potentially, an infinite number of content can be displayed on one physical screen: here we have a many-to-one mapping. Both types of content mappings are well-understood.

Given the expectation that future displays will be low-priced and lightweight, we imagine that users will have a number of displays which combined are much smaller than the number of sheets of paper that we typically use with printed documents today. Hence we are not limited to tight one-to-one mapping of content to displays. This would also not be desirable, as it would limit the display's capability of dynamically changing its content.

Therefore systems that offer many paper-like displays have a many-to-many mapping. Such systems mimic the physical interactions of paper, however with a smaller number of carrier media. Previous work has shown how contents can be easily transferred from one display to another [6,21]. However, it is not clear how the handling and association of content on many displays should work. How can content be temporarily disassociated from displays to generate free carriers for displaying additional contents? How can such "virtualized" contents be "materialized" again and bound to physical carriers? We present interaction techniques that allow the user to combine contents onto one single display, distribute content over multiple displays and to clear and to restore content.

6.1 Combine and Distribute Content

By combining videos that are currently bound to a physical display, physical displays can be freed from contents. Thereafter they can serve as physical carriers for additional videos. To combine one or several videos, the user creates a pile out of the respective displays. Quickly moving the entire pile upwards combines all videos into the topmost display. The remaining displays inside the pile become empty. The metaphor of this interaction is to push all videos up, which are caught by the topmost display.

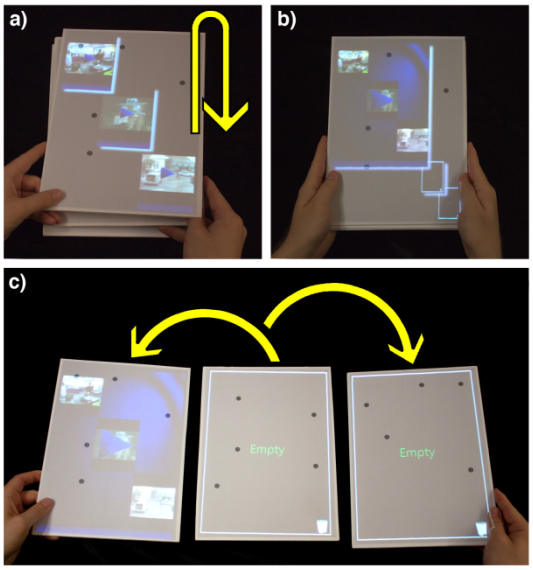

Figure 8: Virtualizing a physical pile (a), onto one single display (b). This clears the remaining displays (c).

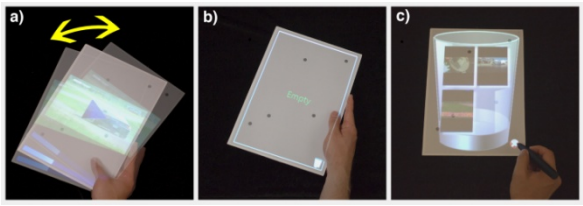

Figure 9: Clear content by shaking the display (a and b). Restore content from the recycle bin (c).

The reverse direction, distributing videos from a video collection, is done by placing one or several empty displays underneath a display containing a video collection. By quickly moving the pile downwards, the videos from the collection are distributed onto the empty displays in the pile (see Fig. 8).

6.2 Clear and Restore Content

Contents can also be virtualized by clearing a display. Clearing is performed by shaking the display, as if one shook contents off. Cleared contents are available in the recycle bin, which can be accessed by tapping on an icon that is available on all empty displays (see Fig. 9).

7. SOUND CONCEPTS FOR PARALLEL VIDEOS

Since multiple videos can be laid out in physical space and played back simultaneously, multiple audio sources can be active at the same time. This produces the well-known cocktail party effect [2], which might make it difficult to perceive information conveyed on the audio channel.

The standard case with multiple display devices is that each display has a built-in speaker to generate the audio of the video that is displayed on this device. This makes sure that both visual contents and audio track of one video are located at the same position in space. The sound is perceived in space at a position relative to the user's position and head orientation. Moving the sound source away from the user reduces its volume slightly. We call this sound concept Real-World Behavior (see Fig. 10a) and implemented it as our baseline.

In this section we introduce three additional sound concepts that will allow users to more effectively mentally grasp (focus) on one or multiple sound sources that are located in space, reducing the cocktail party effect.

7.1 Distance-based Focusing

In work with paper documents, it is a well-established practice to focus on documents by placing them directly in front of the user. Documents are brought out of focus by placing them farther away, but still within an arm's reach [28]. Inspired by this behavior, we propose a sound concept for focusing on sound sources based on their distance.

Moving a display closer to or more distant from the user increases or decreases its volume (see Fig. 10b). In contrast to the Real-World Behavior, where distance has only a barely noticeable effect on the volume, the Euclidean distance between the user and the display is mapped inverse exponentially to the volume. As a result, volume can be finely adjusted, somewhat similar to a slider of an audio mixer. Placing the display an arm's length away is distant enough to reduce the volume to zero.

7.2 Orientation-based Focusing

When people focus their attention on a person or on object, they usually look at it. We propose a concept that leverages head orientation of the user for focusing on sound sources.

A virtual line originates from the user's head in the head's orientation. The volume of each display is mapped inverse exponentially to its relative distance from the virtual line. Hence sound originating from displays that are directly within the user's orientation has the highest volume and is located in the center. Sound originating from displays to the left or right side of the line has a lower volume and is located to the left or right of the user. Sound from displays at the extreme outer sides is muted (see Fig. 10d). By reorienting his head, the user can easily and quickly change his focus to different videos.

7.3 Pick-up-based Focusing

It is common practice to pick up an object to focus on it. Inspired by this behavior, we introduce a further sound concept. Sound from displays that are lying down on the table is set to mute. By lifting up one or multiple displays, the sound of these displays is played back (see Fig. 10c). While picked up, sound sources expose a Real-World Behavior, being correctly located in space.

8. TECHNICAL REALIZATION

Our prototype system realizes paper-like displays by tracking passive cardboards in real-time and projecting contents onto them. An overview of the system is shown in Fig. 11. Our system consists of an optical tracking system with 6 infrared cameras, two full HD projectors mounted on the ceiling, and a set of cardboards, each augmented with infrared retro-reflective markers. The high-resolution projection frustum measures approximately 200 * 120 * 40 cm^3. The information which is provided by the tracking system (position, orientation of the cardboards), is used to warp the projected images onto the cardboards in real-time. In our software toolkit, we simulate the environment by constructing a Direct3D world model. In an initial calibration step, the two Direct3D cameras are set to the positions and orientations of the two projectors, thus the camera "sees" the multiple cardboards and renders their contents from the correct perspective. The projectors display the camera views, which are generated by Direct3D, while the world model is continuously

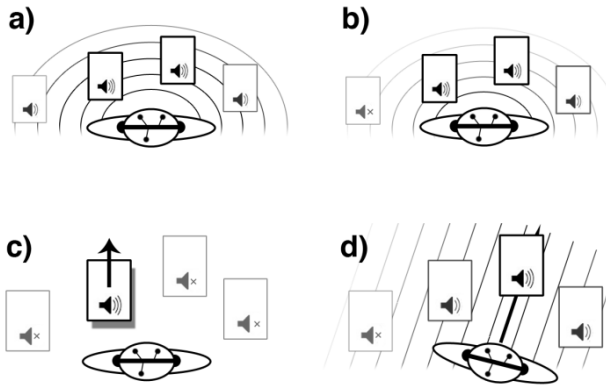

Figure 10: Sound concepts a) Real-World Behavior b) Distance-based Focusing c)Pick-up based Focusing d)Orientation-based Focusing

updated by the tracker data. For recognizing the different gestures, we implemented a gesture recognizer that analyzes positional information of each of the displays.

The application is implemented in C# and WPF. Each display owns its own WPF window that is screen captured and rendered onto the display on demand. Our prototype currently supports seven different cardboard displays, but could be easily modified to support more.

Stylus input is realized using an Anoto Digital Pen ADP-301 and the Letras software framework [4]. Each cardboard is therefore augmented with the Anoto pattern.

We added a layer of Anoto pattern on each cardboard. By using Letras Framework the pen could send pen coordinates, which we then converted to mouse events in the WPF application. We refrained from supporting touch input for the following reasons. Pressure sensitive or capacitive touch foils either require tethering or too bulky of electronic components. While optical touch tracking would be a suitable approach with a single display or a small number of displays, it is too unreliable with the large number of displays supported by our system and the corresponding large number of markers which is required for tracking of displays.

For generating a 3D perception of sound, we used the OpenAL Framework [39]. The user was equipped with headphones which where augmented with markers to track the user's head position and orientation so that the sound sources could be positioned accurately in space.

This system setup is suitable for stationary installations, e.g. in schools, libraries, museums, or stores. In the near future, an alternative system setup can be created which consists of multiple thin-film OLED displays that are connected wirelessly instead of the stationary tracking-projection system. This will make the system easily portable and will allow to it to be used in at various places.

9. EVALUATION

PaperVideo introduces a novel way of tangible interaction while actively working with videos. It allows for a broad range of new styles of working with videos. Rather than focusing onto single variables (like time efficiency) and thereby limiting our view to a subset of scenarios, it is of primary importance to understand the broad range of new styles of working with videos that are enabled by our system. In particular, an evaluation must provide first

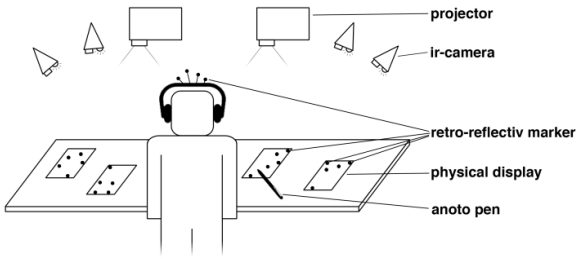

Figure 11: Technical setup

insights into how users treat and use multiple displays simultaneously and how this affects the interaction with videos. Hence, answering the 'why' and 'how' is very important; this requires a qualitative research methodology, rather than a quantitative one, and a detailed analysis of a small, but focused sample. The in-depth analysis of users' practices and mental models which is presented in this section provides the foundation for future quantitative analyses of specific questions that are beyond the scope of this paper.

9.1 Study Design

Six experts participated in single-user sessions, each of approximately 3 hours length. All of them are video power users, spending a large amount of time on watching videos (avg=13 hours a week, SD=4.2), and having extensive recording and video authoring experience. To ensure a wide range of viewpoints, we recruited participants with different professional backgrounds: medicine, computer science, philosophy, cultural science and product design. All participants were male. Their median age was 24. All of them were familiar with modern computing devices, all owning a smart phone and two of them owning an iPad.

Each session was organized as follows. After 10 minutes of guided introduction to the system and its basic interaction techniques, the participant was given ample time to get familiar with the devices. Then the participant was asked to perform the following tasks using a think-aloud protocol:

1. The first task consisted of navigating within one video on a single display. The participant was asked to give a short oral summary about two different scenes of the video, one situated near the beginning, the other near the end of the video. He was free to decide which interaction technique to use (pen or timeline in space).
2. The participant was given all 7 displays. His task was to explore a collection of 6 related videos, to group them in two meaningful groups, to explain his grouping, and to regroup all videos following another criterion.
3. Before the third task, we introduced the interaction techniques for combining and distributing content. The task consisted of exploring a collection of 14 related videos and grouping it into two given topics, followed by giving an overview of all videos in the collection. This task was performed twice with different contents: once using all 7 displays, once using only 3 displays (the order was counterbalanced).
4. Before the next task we introduced the lightweight video editing technique to the participant. He had to provide a summarized video of soccer goals by creating a video excerpt which contained the goal sequences from 3 soccer videos.

5. Before the last task, the participant was introduced to the four sound concepts. Then he had to perform four sub-tasks, each with a different sound concept. The order was randomized. In each sub-task he was given a set of 5 videos on 5 displays and had to explore and decide which of the videos he liked.

These tasks allowed us to study a range of phenomena: how users leverage space to interact with multiple videos on multiple displays (tasks 1–3), how they combine and distribute content over different amount of displays (task 3), how they perform lightweight video editing (task 4), as well how they were able to manipulate and focus on multiple parallel sound sources (task 5).

9.2 Data Gathering and Analysis

As methodologies, we used semi-structured interviews (at the end of each task and after the whole session) and observation. The entire session was videotaped. Interviews and observations were transcribed and analyzed using an open coding approach [32].

10. RESULTS AND DISCUSSION

10.1 Task Support

The system itself was positively received by all participants. For instance P5 mentioned: "*I have the video with the page really in my hands (…) that's great*." All participants explicitly mentioned that they could easily gain an overview of the videos and could easily structure them by having multiple displays that could be rearranged in space. To more quickly get an overview of a video collection, one participant (P1) watched two videos in parallel on two adjacent displays. This allowed him to decide whether he liked the video and sort the video into the appropriate category.

Four participants emphasized that they would use the system for working purposes, such as learning, ordering and organizing videos as well as video editing. However, these participants were skeptical in using the system to just watch videos, such as watching a theater movie or YouTube video. P4 stated that the system is *"good for teaching video navigation to PC novices"*. P5 envisioned that the system be used at an exhibition booth in order to attract people via promotional videos.

10.2 Functional Zones

In the tasks where participants could use all seven displays, table space was used similarly to how it is reported in the literature about traditional paper documents [19,27]. In the video recordings, we clearly identified two different functional zones: The area situated directly in front of the participants, in the center of attention, can be characterized as the working area. Less important videos were moved to the periphery of attention, at the outer, more distant zones of the table surface. Following Scott et al., we call this the storage area.

Most intense interaction with video displays (such as playback, seeking, accessing related videos, clearing contents, video cutting) was done in the working area. Only one single participant navigated videos in the storage space. In contrast, all participants moved displays that were currently not needed (whether they were filled with content or be they empty) out of their attention into the storage space. These displays were placed on a free spot or piled onto other displays. 3 participants loosely arranged displays in the storage area, without piling. P6 explained: *"I have not piled because I'm not a person who orders things directly"*. In contrast, other participants preferred less cluttered arrangements and piled the displays as soon as they placed them in the storage area.

In contrast to the above findings for seven displays, participants behaved differently in tasks where they disposed only of three displays. 2 participants kept all displays inside the working area. Four participants placed only one single display in the storage area, which contained a virtualized pile. These findings show that the number of available displays directly influences the spatial practices of how the system is used. While with only three displays the system is used very much like an enhanced computing device with only very limited spatial interaction, already a relatively small number of additional displays is sufficient for unleashing the power of paper-based interactions.

10.3 Functional Roles of Displays

As long as the number of videos did not exceed the number of available displays, all participants realized a fixed one-to-one mapping of videos to displays. The display has thereby one single functional role: being a physical carrier of the video content.

In cases where the number of videos exceeded the number of displays, we observed two general strategies of participants to cope with the fewer number of displays. One group (3 of 6) can be characterized as "materializers". These participants preferred having as many videos as possible available in tangible form. They filled as many empty displays with content as possible. Only once all displays were full, they combined videos onto one carrier display to get free displays. In turn all of these displays were filled before virtualizing again. In contrast, the other group (3 of 6) can be characterized as "virtualizers". They filled just one (or at most two empty displays) and directly grouped the videos onto a virtual pile.

Participants assigned stable functional roles to physical displays. We identified three different roles of display usage that all participants assigned to displays in tasks 2-4: (a) information source, (b) working display, (c) information container. The information source was a display that contained all the related videos. The working display was used to iteratively open, watch and assess a related video before moving it into an information container. An information container display was used to group and store several videos that represented a topic. Both "materializers" and "virtualizers" attributed these roles to displays, with the only difference being that virtualizers had only one working display, whereas materializers had several.

Despite the possibility of using many displays, P6 and P4, both "virtualizers", used only 3 or 4 of them. P6 stated: *"Oh yes that goes well with 3, with 3 I have a better overview of the displays."* P4 even felt three displays to be more efficient than with more of them and was amazed how easily the tasks could be performed. In contrast, P1 and P5, both "materializers" found it inconvenient to work with only 3 displays.

We conclude from the results that depending on their strategy, users prefer more or less working displays. The system should provide enough displays to leave the choice to the user.

10.4 Dynamic vs. Static Content

In this section we focus on how dynamic content on paper displays was perceived in contrast to traditional static content on paper. Five participants found that dynamic content was easy and intuitive to use. P1 stated: *"It is better that videos can be detached from the medium whereas content on paper is bound to paper."* Furthermore the *"x-ray view"* (P5) through the pile was found to be beneficial to paper by all participants: *"I can see through the pile and still interact with it."* (P2). P5 mentioned: *"It saves me from flipping through pages. That's convenient; I can directly continue to work."*

All participants positively perceived that contents of the displays within a pile change dynamically when a pile is virtualized. P6

commented: *"This allows me to work with less displays."* However, comments about how specifically display contents should change in this case revealed two different mental models of the participants. This is best made clear by an example given by P5: The participant wanted to add a video to an existing virtual pile of videos. To do so he placed the display with the single video on top of the virtual pile and then performed the up gesture for virtualizing the pile. This automatically "pushed" all contents to the topmost display, which now contained the virtual pile. The remaining displays were empty. Thereby, the former working display changed its role to an information container and vice versa. This participant had a mental model that focused on the roles of the *physical* displays. He disliked their changing roles, which he described as "computer logic" (P5). In contrast the remaining 5 participants had a mental model that focused rather on the dynamic contents. They did not even notice that the physical carrier medium changed its role.

From these results we conclude that dynamic content on physical displays is appreciated and does not need high rethinking or high cognitive effort in contrast to the known behavior of traditional paper. However, to account for the mental model that focuses on the physical carrier medium, systems should equally support interactions that keep the roles of the physical displays steady. For instance, videos can be added to an existing pile by dragging and dropping them from the working display onto the information container.

10.5 Physical vs. Touch Input

We were interested to find out which interactions should be delegated to physical input (manipulating displays in space) and which ones to surface-based input (direct touch or pen input).

We observed that interactions like arranging and piling videos in space as well as combining and distributing contents were performed intuitively by physical input. In contrast, all participants intuitively used surface-based input for playback and skimming in videos. The interaction for temporal navigation in space was rarely used. Participants stated that skimming with the pen requires less effort and less space and moreover is more precise than moving the display in space. Furthermore, two participants (P2, P4) proposed a function for reordering videos within a pile by dragging & dropping their small representations on the topmost display. P3 suggested that the list of related videos be accessed by touch and that related videos be opened by dragging & dropping them from the list onto an empty display. However, P2 appreciated the physical gesture and mentioned browsing relations, by bringing displays near to each other, is intuitive and easy to use.

From these results we conclude that interactions that naturally anchor information in physical space would rather be done using physical interactions. For instance, if users arrange displays or pile them they expect the physical arrangements to change. On the other hand, interactions that have no spatial anchor and are without spatial consequences are performed using surface-based input. This particularly concerns interactions that apply to only one single display, such as temporal navigation within a video. Future work should explore spatial temporal navigation with multiple displays that are linked to only one video. By moving and arranging displays in space, the user could sneak-peak into different temporal locations of the video simultaneously and easily compare contents within one video. We assume that in this case, users prefer physical over touch input.

10.6 Sound Concepts for Multiple Videos

In this section, we evaluate and contrast the four sound concepts.

10.6.1 Real World Behavior

The evaluation showed that the sound concept which realizes real world behavior is not suitable for viewing multiple videos on multiple displays at the same time. Three participants watched the videos one after another. P5 who tried to view the videos in parallel stated: *"It feels better when only one video is playing."* All other sound concepts where judged to be better than this one.

10.6.2 Distance-based focusing

Our observations and comments from the users clearly showed that the distance-based focusing concept is much better suited for watching multiple videos. With this technique, five of six participants watched videos in parallel. Four of six participants rated this to be the best of all concepts, since it allowed for the most flexible sound manipulation with a very intuitive mapping. For instance, P2 stated that *"It is easy to manipulate the volume"*. The remaining participant did not watch videos in parallel with any of the concepts.

10.6.3 Orientation-based focusing

Three of the five participants who watched videos in parallel criticized this concept because of too much noise coming from displays at the outer sides. Two other participants mentioned that this concept is good for only focusing on audio without visual feedback, but not both combined.

10.6.4 Pick-up-based focusing

The pick-up-based focusing had the advantage that many videos can be played back in parallel without generating sound disturbance. Two participants, P1 and P5, started all the videos right at the start: *"I do not miss anything, I can still see everything*." One participant (P6) mentioned: *"I can focus on one video more clearly*." However, two participants (P1, P4) stressed that they *"have just two hands"* so that they can only hold and listen to two videos at a time. Moreover, one of them (P4) feared that holding the display for a long period could be tiring. P4 proposed as an improvement that *"shortly picking up a video could toggle between active and deactivated sound"*.

We conclude that with our sound concepts, in contrast to the real-world behavior, it is possible to watch videos in parallel and be able to explicitly and easily change the sound focus. The results show that distance-based focusing, preferred for its high flexibility and intuitiveness, was the best technique. Pick-up-based focusing also has its strengths in situations where users focus only on one or two videos at a time from a set of many videos that are simultaneously played back. A video installation at an exhibition booth is one example.

11. CONCLUSION

In this paper we proposed a novel paradigm for spatial interaction with video. We introduced a set of interaction techniques and spatial sound concepts that support playback, flexible navigation and spatial organization of videos on multiple physical displays. Results from a qualitative evaluation shed light on how people use multiple interactive displays simultaneously and how this affects the interaction with video contents. The results show that users can flexibly organize and structure videos in physical space while generating a good overview of multiple videos. They thereby flexibly attribute three different functional roles to paper-like displays: information source, working display and information container. The study further showed that advanced spatial sound concepts effectively support a user in simultaneously viewing

multiple videos which contain audio tracks. Finally, we have characterized different mental models and strategies of users ("materializers" vs. "virtualizers") to cope with a restricted number of displays.

In future work, we plan to explore the use of the system in a collaborative environment..

12. REFERENCES

[1] Adler, M.J. and Van Doren, C.L. *How to read a book*. Simon and Schuster, 1972.

[2] Arons, B. A review of the cocktail party effect. *Journal of the American Voice IO Society 12*, 7 (1992).

[3] Glass, J., Hazen, T.J., Cyphers, S., Malioutov, I., Huynh, D., and Barzilay, R. Recent Progress in the MIT Spoken Lecture Processing Project. *Artificial Intelligence*, (2007).

[4] Heinrichs, F., Steimle, J., Schreiber, D., and Mühlhäuser, M. Letras: an architecture and framework for ubiquitous pen-and-paper interaction. *Proc. EICS '10*, (2010).

[5] Hinckley, K. Synchronous gestures for multiple persons and computers. *Proc. UIST 03*, (2003).

[6] Holman, D., Vertegaal, R., Altosaar, M., Kl, C., Troje, N., and Johns, D. PaperWindows : Interaction Techniques for Digital Paper. *Interfaces*, (2005).

[7] Huber, J., Steimle, J., and Mühlhäuser, M. Toward more efficient user interfaces for mobile video browsing. *Proc. MM '10*, (2010).

[8] Hürst, W., Götz, G., and Welte, M. Interactive video browsing on mobile devices. *Proc. MM '07*, (2007).

[9] Jiang, H., Viel, A., Bajaj, M., Lue, R.A., and Shen, C. CThru: exploration in a video-centered information space for educational purposes. *Proc. CHI '09*, (2009).

[10] Karrer, T., Weiss, M., Lee, E., and Borchers, J. DRAGON : A Direct Manipulation Interface for Frame-Accurate In-Scene Video Navigation. *Evaluation*, (2008).

[11] Kirsh, D. The intelligent use of space. *Artificial Intelligence 73*, (1995)

[12] Luff, P. and Heath, C. Mobility in collaboration. *Proc. CSCW '98*, (1998).

[13] Mackay, W.E. and Pagani, D.S. Video Mosaic: Laying Out Time in a Physical Space. *Proc. MM '94*, (1994).

[14] Malone, T.W. How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems 1*, (1983).

[15] Mander, R., Salomon, G., and Wong, Y.Y. A "pile" metaphor for supporting casual organization of information. *Proc. CHI '92*, ACM Press (1992).

[16] Manske, K. and Mühlhäuser, M. OBVI: Hierarchical 3D Video-Browsing. *Proceedings of ACM Multimedia*, (1998).

[17] Merrill, D., Kalanithi, J., and Maes, P. Siftables: towards sensor network user interfaces. *Proc. TEI '07*, (2007).

[18] Mertens, R. and Vornberger, O. Hypermedia Navigation Concepts for Lecture Recordings Navigation in Hypermedia. *World Conference on ELearning in Corporate Government Healthcare Higher Education*, (2004), 2480-2487.

[19] O'Hara, K. and Sellen, A. A comparison of reading paper and on-line documents. *Proc. CHI '97*, ACM Press (1997).

[20] Rekimoto, J., Ullmer, B., and Oba, H. DataTiles: a modular platform for mixed physical and graphical interactions. *Proc. CHI '01*, ACM Press (2001).

[21] Rekimoto, J. Pick-and-drop: a direct manipulation technique for multiple computer environments. *Symposium on User Interface Software and Technology*, (1997).

[22] Rooke, M. and Vertegaal, R. Physics on display: tangible graphics on hexagonal bezel-less screens. *TEI '10,* (2010).

[23] Schmandt, C. and Mullins, A. AudioStreamer: Exploting Simultaneity for Listening. *Proceedings of CHI 95*, (1995).

[24] Schmandt, C. Audio hallway: a virtual acoustic environment for browsing. *Proc. UIST 98*, (1998).

[25] Schoeffmann, K. and Fabro, M. del. Hierarchical video browsing with a 3D carousel. *Proc. MM '11*, (2011).

[26] Schoeffmann, K., Hopfgartner, F., Marques, O., Boeszoermenyi, L., and Jose, J. Video browsing interfaces and applications: a review. *SPIE Reviews 1*, 1 (2010).

[27] Scott, S.D., Sheelagh, M., Carpendale, T., and Inkpen, K.M. Territoriality in collaborative tabletop workspaces. *Proc. CSCW 04 04pages*, (2004).

[28] Sellen, A.J. and Harper, R.H.R. *The Myth of the Paperless Office*. The MIT Press, 2001.

[29] Sokoler, T. and Edeholt, H. Physically embodied video snippets supporting collaborative exploration of video material during design sessions. *Proc. NordiCHI '02*, (2002).

[30] Spindler, M., Stellmach, S., and Dachselt, R. PaperLens: advanced magic lens interaction above the tabletop. *Proc. ITS '09*, ACM (2009).

[31] Spindler, M., Tominski, C., Schumann, H., and Dachselt, R. *Tangible views for information visualization*. ACM Press, New York, New York, USA, 2010.

[32] Strauss, A. and Corbin, J. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage, 2008.

[33] Sun, Q. and Hürst, W. Video Browsing on Handheld Devices - Interface Designs for the Next Generation of Mobile Video Players. *IEEE Multimedia 15*, 3 (2008).

[34] Terrenghi, L., Kirk, D., Sellen, A., and Izadi, S. Affordances for manipulation of physical versus digital media on interactive surfaces. *Proc. CHI 07 8*, September (2007).

[35] Wellner, P. The DigitalDesk calculator: tangible manipulation on a desk top display. *Communications of the ACM 36*, 7 (1993).

[36] Zigelbaum, J., Horn, M.S., and Jacob, R.J.K. The Tangible Video Editor : Collaborative Video Editing with Active Tokens. *Design*, (2007).

[37] iTunes U. http://www.apple.com/education/mobile-learning/.

[38] MIT OpenCourseWare. http://ocw.mit.edu/.

[39] OpenAL. http://connect.creativelabs.com/openal/default.aspx.