ABSTRACT
Mobile multimedia devices, such as smart phones and tablet PCs, are gaining increased popularity for the purpose of image and video recording. For many users, these ubiquitous devices are a good replacement for traditional and usually heavyweight cameras. However, while content creation is very simple on such mobile devices, browsing and searching still lacks of appropriate interaction means that are specifically designed for multi-touch-based interaction on small displays. In this work, we present our initial concepts and ideas to improve image and video browsing on smart phones and tablet PCs with touchscreen interaction.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Algorithms, Performance, Design, Experimentation

Keywords
Video Browsing, Image Browsing, Mobile devices, Touchscreens

1. INTRODUCTION
Due to their high availability, smart phones and tablet PCs are more and more used as a replacement for traditional and dedicated photo and video cameras. With the increased popularity of such devices and their ubiquitous use for recording purposes, the number of stored images and videos on mobile multimedia devices dramatically increases. However, as the volume of media files on such small devices increases, finding the proper image or video (e.g., to show it to a friend) becomes increasingly difficult. The reason for the problem is the fact that current mobile media browsing tools do not provide efficient features for searching and browsing. For example, the default image browsing app on an iPhone uses a fixed-size thumbnail list that can show only 24 images at a time, without any special features for search. The same app is used to browse videos, where videos are represented by a single key-frame. Efficient search within a video is another problem on such devices because most users are bound to video players with very limited navigation features.

In this work we present our initial concepts and ideas to improve exploratory image and video search on mobile multimedia devices. We focus not only on improved interaction models (e.g., multi-touch interaction) as well as advanced user interfaces (e.g., 3D visualizations), but also on the proper integration of available sensors (e.g., gyroscope and accelerometer). Moreover, we would like to investigate what kind of content analysis can be used on such devices for immediate content-based browsing right after content recording. Such a scenario is especially challenging because of the limited computing power, available on mobile devices.

2. RELATED WORK
Different approaches to improve image and video browsing on desktop PCs have been presented in the literature (see [16] for a recent review on video browsing). Many of these approaches use some kind of content analysis, similar to the approaches used in image and video retrieval [17]. In contrast to retrieval, in a typical browsing scenario the user doesn’t formulate a concrete query for what they are looking for. Browsing is more an interactive search process where the user has a rough idea of what she/he is looking for and tries to interactively find it. Recent research results show that 3D visualizations and similarity-based arrangements can improve visual search on mobile devices [13].

Proposed interfaces cover a wide range of concepts with and without variants of content analysis. Some examples are a Fisheye-like view [2], a 3D ring [14], different 3D globe configurations [1] and a hexagonal lattice [12]. In our current research we are using color analysis in order to enhance image browsing combined in combination with 2D and 3D visualizations.

In the domain of video browsing many concepts for improved interfaces on desktop PCs have been proposed. One example is the extension of seeker-bars through abstract content visualizations. In [16], for example, the dominant colors of frames are shown in an enhanced seeker-bar, making it easier to quickly jump to scenes that are similar (e.g., moderation scenes with similar background or repeating shots of ski jumpers).

HiStory [7] proposes a hierarchical approach and presents the user a storyboard-like arrangement of thumbnails of the video. The user can refine the view by choosing one of the thumbnails, representing a specific time range in the video. The view is then recreated with a new set of thumbnails, representing the time range of the original thumbnail in finer granularity. In a similar storyboard arrangement Comp2Watch [9] tries to further apply summarization techniques to give users a quick meaningful overview of the contents of the video. Another example of hierarchical image browsing uses a 3D ring for visualization as in [15] where each new hierarchy level is represented as a new 3D ring giving a user a better understanding of where in the hierarchy she/he currently is browsing.
Extensions of mobile video browsing with the help of touchscreen gestures are shown in [5] and [6]. Through the usage of swipe-gestures it is possible to navigate inside a video and also between different interrelated videos. Further utilizing touchscreen gestures with focus on very precise video browsing is ProPane [4]. The interface gives the user precise control on browsing speed. It can even be performed on a frame-by-frame style, which was especially attractive for professional users, as shown in their user study.

Extending video browsing to the usage of multiple mobile devices the authors of [11] show usage scenarios enabled by using multiple physical objects to interact with videos. They derive their approach from working with documents in the real world.

Providing the user a view of different similarity characteristics on a given query the Fork-, Cross- and RotorBrowsers of [3] are a prominent example for similarity based video browsing. In the center of the screen a user-chosen focal shot is displayed. In different arrangements (fork formation, star formation or cross formation) similarity threads of shots are displayed around this focal shot. A user can use the different threads to navigate through the content.

An approach that relies heavily on object tracking is presented in [10]. A user has the ability to navigate in a video by selecting and manipulating objects of the video with her/his finger. On example is the navigation in a soccer game, where dragging the ball to a certain position jumps to the respective scene in the video. Unfortunately the system only works with a connection to a server since the algorithms used for the object tracking are too demanding for mobile devices.

Finally, in [18] video browsing is performed by using the built-in accelerometer exclusively, making it especially attractive for single-handed use.

3. ON-THE-FLY CONTENT ANALYSIS

On-the-fly content analysis enables new usage scenarios for mobile devices especially in domain of video browsing and video editing. Although tablet-PCS and smart phones already use multi-core CPUs and multi-core GPUs, deep content analysis (e.g., methods like Bag-of-Visual-Words) is typically not feasible on mobile devices due to the high memory requirements and limited battery power. This is especially true for the scenario of on-the-fly content analysis that will be the focus of our research. Since there isn’t much prior work considering on-the-fly content analysis on mobile devices, we make first attempts in this field without claiming to provide optimal or final solutions. However, because of the current limitations of browsing tools on mobile devices, we expect that even rather simple content analysis methods can provide a significant improvement.

To better illustrate our first ideas, please consider the simple example of a mobile video browser that provides direct links to appearances of people to the user, based on face detection. Another example would be the usage by a small reporter team. In case of an unexpected but important event the team could use mobile devices instead of their currently unavailable professional equipment to record first footage of the event. After the footage is taken they could start filtering and editing the footage on-site with little to none delay, aided by an advanced browsing interface. The filtering process is enhanced by the usage of content analysis results that were already generated during the recording. In a short amount of time the team could send a pre-edited reportage video to the headquarters of the TV station.

4. IMAGE BROWSING DESIGNS

In this section we present our first prototypes to improve image browsing on mobile devices by combining color sorting with 2D and 3D visualizations. As we could show in [1] that such representations can improve search performance on a tablet-PC, we now want to extend our investigation to smaller devices. In a current user study we are evaluating them for small mobile devices like smart phones. All interfaces use the same color sorting algorithm: images are sorted by their dominant hue level obtained through a 24-bins HSV histogram. Additionally, very bright images are placed at the beginning and very dark images are placed at the end.

Two interfaces of our previous user study [1] (an image grid and a 3D globe) are joined with two new prototypes: The ImagePane and an improved version of the 3D ring interface proposed in [13].

![Figure 1: Image browsing on small touchscreen devices: a) Grid; b) 3D-Globe; c) 3D-Ring; d) ImagePane.](image)

To evaluate the performance of the three novel interfaces a traditional grid arrangement is included in the user test as a baseline (see Figure 1a). It is designed in the same way as the photo browser on iOS and similar devices. We add color sorting for ensuring fairness in comparing this interface to the other ones. The user can scroll the list upwards and downwards by using swipe and drag gestures.

In the interface shown in Figure 1b, images are distributed on the surface of a 3D globe according to the earlier mentioned color sorting algorithm [1]. The surface is organized in a grid-like arrangement. The sequence of images is applied in a one by one fashion, going north to south and west to east. A user can rotate and tilt the globe by using swipe and drag gestures. Zooming the view is possible by using pinch gestures.

Figure 1c shows the third interface, which is an advanced version of the ring interface presented in [13]. Images are arranged on the surface of a 3D ring. The ring can be rotated by swiping and dragging gestures. Swiping up results in a zoomed view of the front part of the ring. Swiping down returns the user to the default view. Double tapping results in a zoomed view of the back part of the ring, as if the user is standing in the middle of it. Additionally the ring automatically rotates so that the tapped area is positioned in the center of the screen.

An extension of the traditional grid is the ImagePane interface (see Figure 1d). All images in the collection are displayed as small thumbnails in the default view. To perform a zoom operation the user can double-tap on an area of the pane. The view then zooms to this selected area of images. In this view the user also has the opportunity to scroll the pane in any direction by applying drag and
swipe gestures. The user can leave this view again by performing another double-tap gesture.

5. VIDEO BROWSING DESIGNS

In this section we present our initial concepts for improved browsing and navigation within a single video. In future iterations we will integrate results of content analysis as well as extend our concepts for browsing many videos instead of a single video. These interfaces are currently targeted to be used on tablet-like devices.

Figure 2: Browsing a video using a 3D filmstrip.

Figure 2 shows an early version of our filmstrip interface. Each segment in the strip initially represents a 10 seconds segment of the video. To play the content of a certain segment a user has to tap the segment once. The video then starts playing back inside the region of the selected segment. The user can scroll the filmstrip in various ways. One way is a kind of fly-over view. For this, the user has to perform a drag gesture outside the filmstrip upwards or downwards on the screen. The filmstrip then moves towards or away from the user providing a quick overview of the contents of the whole video. It is also possible to change the tilting of the filmstrip. This is done by applying a drag gesture with two fingers up or down on the screen. Another possibility is the scrolling of the segments on the filmstrip. A drag gesture on the surface of the filmstrip to the left or to the right side results in a corresponding scroll action. The granularity of the segments can be adjusted by pinch-gestures. For example, to go from 10 second segments to 5 second segments the user has to perform a pinch-in gesture. On the other hand, a pinch-out gesture decreases granularity e.g. going from 10 second segments to 30 second segments. In a future iteration we will experiment with different segmentation methods (e.g., based on shot boundary detection).

Video browsing on tablets and smart phones is usually performed by holding the device in landscape orientation with both hands on the sides, to maximize usage of the available screen real-estate for the video content.

Figure 3: Sketch of the ThumbBrowser concept.

In such a configuration the thumbs become especially important. An intelligent video browser interface could consider this fact and optimize placement of browsing controls for the thumbs of the user. With traditional browsing interfaces the user has to release the comfortable two-handed holding position in order to be able to use the control buttons or the seeker-bar. A simple enhancement in this case would be the rearrangement of the controls to the sides (see Figure 3). Further, to reduce visual overhead, on-demand radial menus could be used to provide additional functions, e.g., for choosing between different content analysis result visualizations in contrast to buttons with fixed positions on the screen.

Bringing the idea of multiple windows of the same video to mobile devices is one of the ideas for the VideoPinch concept in Figure 4. A user starts with a single rectangle or frame resembling the whole video. By using a pinch gesture the user expands this single frame to multiple frames representing different time segments of the original video. She or he is therefore refining the search process, which can be repeated on all currently available frames. Although this concept isn’t completely new on traditional desktop computers, combining it with the abilities of modern touchscreen technology could increase usefulness and usability radically. The frames could be freely dragged around the workspace. This enables e.g. lining up two segments of a video side by side for direct comparison purposes.

Figure 4: Sketch of the VideoPinch concept.

Further, two buttons help to restructure the content: a reset button and a reorder button. The reset button deletes all frames except one, which resembles again the whole video. It should therefore be used to start over again. On the other hand the reorder button will re-align the current frames in the order they would appear in the original video. To play/pause a video frame the user can use a single tap. Double tapping on a frame opens the corresponding video frame in a full screen view. Finally a triple tap resets the frames to the first frame of their corresponding segment of the original video.

Figure 5: Sketch of the VideoRingX concept.

Extending the idea of the 3D ring image browser interface, the VideoRingX interface concept (Figure 5) displays screenshots of movie segments instead of images. The segments can be played/paused by single tap gestures. If users want to refine their browsing experience, they can tap and hold a segment. This will expand the segment in a horizontal filmstrip that can be scrolled up and down. Single tapping on the segments starts or pauses playback. Tapping and holding on a segment of the horizontal filmstrip further refines the search by expanding its time range into a new ring of segments.

6. RESEARCH METHOD

Due to the novel nature of the topic, an iterative research approach seems to be the best suited, where the phases Think/Design, Realize/Implement and Evaluate are forming a circulatory. The Think/Design-Phase will be performed by doing literature surveys of comparable prior work. The results of earlier evaluation phases will also be incorporated in future stages. The Realize/Implementation-Phase will result in the creation of artifacts, which will in our case primarily be prototypes.
In order to evaluate our prototypes for image and video browsing, we will perform several user studies for different use scenarios. One such scenario will be Known-Item-Search (KIS), where a user has some knowledge of a specific segment of a video (or of an image) and wants to quickly find it, but doesn't know where or how to look for it. Further research on alternative, efficient, partly automatic evaluation techniques will also be part of our work. For example, we want to extend the common evaluation methods by doing justice to the mobile nature of the devices. Most of the time mobile device interfaces are evaluated in a fixed lab-like environment, although they truly are designed to be used on the go. We therefore plan to evaluate our future prototypes also in some more realistic scenarios (e.g., in a walking setting). Participants should be tested using the prototypes during the performance of common tasks like walking down a hallway, having to pay attention not only on using the device but also to avoid running into other people. The same evaluation could also be repeated outside a building where the lighting conditions are quite challenging for LCD screens. A further interesting setting is usage during participating in public transportation.

Acknowledgments

This work was funded by the Federal Ministry for Transport, Innovation and Technology (bmvit) and the Austrian Science Fund (FWF): TRP 273-N15 and the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF), supported by Lakeside Labs GmbH, Klagenfurt, Austria.

7. REFERENCES


