MosaicUI: Interactive media navigation using grid-based video

Arjen Veenhuizen
TNO
Brassersplein 2
Delft, The Netherlands
+31 8886 61168
arjen.veenhuizen@tno.nl

Ray van Brandenburg
TNO
Brassersplein 2
Delft, The Netherlands
+31 8886 63609
ray.vanbrandenburg@tno.nl

Omar Niamut
TNO
Brassersplein 2
Delft, The Netherlands
+31 8886 67218
omar.niamut@tno.nl

ABSTRACT

Intuitively navigating through a large number of video sources can be a difficult task. The problem of locating a video asset of interest, whilst keeping an overview of all the available content arises in video search and video dashboard application domains, utilizing the new possibilities provided by recent advances in second screen devices and connected TVs. This paper presents initial results of an attempt to solve this problem by creating real-time mosaic-like video streams from a large number of independent video sources. The developed framework, called MosaicUI, is the result of cooperation between the European FP7 project FascinatE and the Dutch Service Innovation & ICT project Metadata Extraction Services. Using the combination of an intuitive user interface with a media processing and presentation platform for segmented video, a new method for user interaction with multiple video sources is achieved. In this paper, the MosaicUI framework is described and three possible use cases are discussed that demonstrate the real world applicability of the framework, both within the consumer market as well as for professional applications. A proof-of-concept of the framework is available for demonstration.

Categories and Subject Descriptors

H.5.1 [Information Interfaces And Presentation]: Multimedia Information Systems – *video navigation, video search, immersive media, interactive media, metadata extraction*

General Terms

Experimentation, Verification.

Keywords

Natural user interfaces, second screen, TV, mosaic, video search

1. INTRODUCTION

With an increasing number of video content available to viewers on multiple screens, it becomes more difficult to find videos of interest. Furthermore, to keep videos accessible to all viewers on all their devices, semantic cue or metadata-based access has become a necessity. Several content-based video retrieval systems or video search engines, that enables a user to explore large video archives quickly and with high precision, have already been developed. Janse et al [1], were one of the first to present a study on the relationship between visualization of content information, the structure of this information and the effective traversal and navigation of digital video material. Currently, the MediaMill Semantic Video Search Engine [2] is one of the highest-ranking video search systems, both for concept detection and interactive search. Most of these systems include a video browsing or navigation interface for user interaction. For example, both the Investigator's Dashboard and the Surveillance Dashboard prototypes, developed in the MultimediaN project [3], use video browsing as a method to assist users in their video-related tasks.

TV viewers face a similar task when trying to find content to their taste in an ever-expanding offer of TV channels and on-demand content. Barkhuus et al. [4] describes how techniques have changed the experience and planning of TV watching. The active choosing of what content is to be watched resembles other types of media consumption such as reading, listening to music, or going to the cinema. However, interactivity in TV deployments has not yet seen significant advances, as one of the key issues that need to be resolved is how to interact with and navigate through the content. Most of the solutions available so far have been based on advanced remote controls and hierarchical menus. Natural user interfaces and second-screen applications are seen as promising candidates for a next step in TV interaction, adding a new dimension and introducing new possibilities to navigating video content.

In this paper, we describe the concept of MosaicUI as a framework for interactive video browsing and navigation using, for example, a tablet or smartphone. A number of video sources are combined into one large grid which is displayed on a screen. The combination of high resolution, multi-layer, spatially and temporally segmented video with a state of the art metadata search engine and associated media content creates a multitude of new interactive video applications, which enable the user to interactively navigate videos, selecting the one of interest and exploiting the new possibilities in media navigation introduced by the latest advances in media interaction. Video selection can be performed in numerous natural ways, e.g. motion tracking or gestures.

MosaicUI finds its origin in results from two ongoing projects. The Metadata Extraction Service (MES) [5] project aims at near real-time extraction and indexing of multi-modal metadata from live and stored multimedia content (video, audio, text, images, etc.). Metadata extraction includes, but is not limited to, video concept detection, extraction of text embedded within an image or video (OCR) and speech recognition. Source multimedia content is stored using a long term storage array, and by combining this archive with a metadata index which can be queried, one is able to search for content in a multimodal manner.

Within the EU FP7 project FascinatE [6], a capture, production and delivery system capable of supporting pan/tilt/zoom (PTZ) interaction with immersive media is being developed. End-users can interactively view and navigate around an ultra-high resolution video panorama showing a live event, with the accompanying audio automatically changing to match the selected view. The output is adapted to their particular kind of device, covering anything from a mobile handset to an immersive

panoramic display. The FascinatE delivery network uses spatial segmentation and tiled streaming to enable interaction on mobile devices.

In the remainder of this paper, the MosaicUI framework is explained in further detail and its application domains are discussed.

2. MOSAICUI FRAMEWORK

The potential synergy between the MES and Fascinate Project has been explored by developing the MosaicUI framework. The capabilities of the two initially unrelated platforms are combined to form a new platform enabling a number of new use cases. For example, one could create a grid of live TV shows and display it on a HDTV, while the end user interactively browses and selects his program of interest on a second screen (e.g. a tablet). Alternatively, a number of CCTV feeds could be joined together into a video grid in real time. Another example would be to take the MES metadata search engine into the equation, enabling interactive high resolution multimedia content search and navigation, potentially using a second screen. Figure 1 shows the basic concept of a MosaicUI video grid.

Figure 1. A mosaic of multiple video sources.

In order to achieve the desired level of user interaction and allow for each of these use cases, a number of high-level requirements have been identified. Besides the ability to support a wide variety of media sources, such as live TV, internet videos, local videos, images and segmented content, a controller interface is required to navigate the content. Most importantly, a system is required which is able to combine the different sources together to a grid-like live video feed in real-time, which can then be streamed to a play-out device. Based on these use cases, two, largely equal, architectures have been developed, as shown in Figure 2 and 3. These architectures differ from one another in that the first architecture uses client side media processing, while the second architecture uses server side media processing. This allows the MosaicUI framework to be used both on devices that have a limited amount of processing power available as well as on more powerful devices, creating a versatile platform which can be utilized on numerous types of devices and which introduces new dimensions to the previously explored possibilities of mosaic based video content navigation.

2.1 The MosaicUI Functional Components

From a functional point of view, five key components and one optional component have been defined that together constitute the MosaicUI architecture. First, one or multiple media sources are required. In the MES system, this source can be any form of media, e.g. a live broadcast, a local audio file, a YouTube video or CCTV capture stream. Second, this media source must be indexed in some way so that the user is able to "find" and retrieve that specific media source. Indexing could be for example a form of metadata which describes the geographic location of a CCTV capture stream, a transcript of the closed caption of a broadcast item or the tags associated to a specific YouTube video. A metadata database is required to store and index this information. This enables one to actually find the media that is of interest to the user. It is envisioned that this metadata index could be a multi modal platform and could be based on existing (metadata) databases (e.g. YouTube). The current implementation is able to use both the MES database and a set of TV channel listings. Third, a user interface is required to control the play-out of the mosaic and to allow the user to adapt the mosaic to his liking. A controller facilitating this interaction (scrolling, zooming, fast forwarding/rewinding the content, but also changing the selection of the sources being shown in the mosaic) can be implemented by e.g. a touch-based, gesture recognition or motion tracking platform. The current implementation features a touch-based controller application on a tablet. Fourth, a combiner is required which places the requested media sources together in a grid to form a single mosaic media stream with which the user can interact. The role of this combiner is to stitch together multiple media sources into one high resolution stream in real-time. The most difficult aspect of the combiner to control is its ability to react as fast as possible to changing user requests, in order to get an intuitive user experience. The current implementation of the combiner is largely based on the FascinatE tiled streaming platform described in [7]. Fifth, a play-out function is required. The play-out interface actually displays the mosaic video stream with which the user can interact. This combined media content could for example be shown on an HDTV, beamer, smartphone or tablet. Sixth, an optional stream server can provide the means to stream the combined media content to a client. This is required in case the server-side combiner architecture is used. As stated in the previous paragraphs, an important distinction in the potential use cases is the fact that some of them require that the combination of media sources is performed locally (e.g. client side), while others require server side combination of media sources. From a functional point of view these two use cases seem to be identical, but from an architectural point of view, a clear distinction can be identified. Figure 2 shows the MosaicUI architecture with a client-side combiner, while Figure 3 shows a server-side combiner.

In the current framework, the media sources, the metadata index and play-out are provided by the MES project while parts of the combiner and controller are based on work performed in the EU FP7 project FascinatE [6].

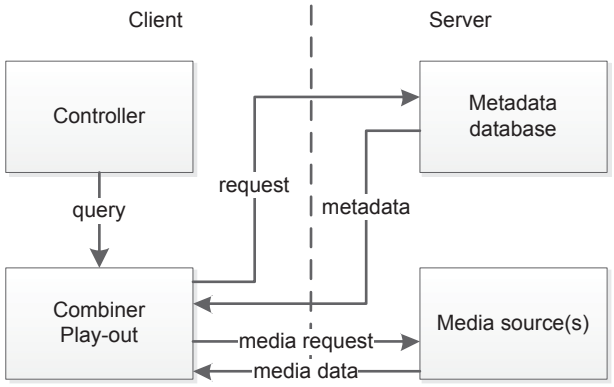

Figure 2. Client side media source combination.

2.2 Creating a mosaic video stream

The basic data flow in the MosaicUI framework can be described as follows. First the user requests a particular set of video sources using the controller. This request could for example be a fixed set of video streams or a search query to find a specific type of video. This query is sent from the controller to the combiner, which is either local or remote. Next, the query is relayed as a request from the combiner to the metadata database in order to look up the metadata describing the relevant media fragments and their location. This information is sent back to the combiner, which then requests the media streams from the identified sources.

Upon reception of the first frames of the media streams, the combiner starts the decoding process, running multiple parallel decoders. After a frame has been decoded, the combiner places it in the grid. In case a server-side combiner is used, the resulting grid frame is sent to an encoding process and streamed to the client. In case a client-side combiner is used, the resulting grid frame is sent to a video buffer for output on the display. It should be noted that in order for the recombination process to start, it is not necessary for the combiner to wait until the first frames of all video sources are available. The combiner can just add sources to the grid as they become available. Since the combiner needs simultaneous access to all video sources, the necessary bandwidth can become problematic. However, since the resolution of the resulting mosaic video will in most cases not be much larger than the full resolution of a single conventional video source, it is sufficient for the combiner to access a low-resolution version of each media source. Sources that are available in multiple resolutions, such as is often the case with adaptive bitrate content in the form of e.g. MPEG DASH or Apple HLS, are especially useful in this regard. A further method for limiting the bandwidth requirements at the combiner is by using FascinatE tiled streaming technology [7], which provides an efficient and scalable delivery mechanism for streaming parts of a high-resolution video – the video grid in this case.

3. APPLICATION DOMAINS

The MosaicUI architecture is suited for use in a wide variety of application domains. This section will examine the possibilities for MosaicUI in three of those domains: as a novel method for browsing TV channels; as an intuitive user interface for searching video and as a method for viewing multiple CCTV streams on a mobile device.

3.1 TV browsing with a second-screen

One of the more obvious applications of MosaicUI technology is as a method for interacting with, and navigating between, TV channels. In recent years several new types of TV user interfaces

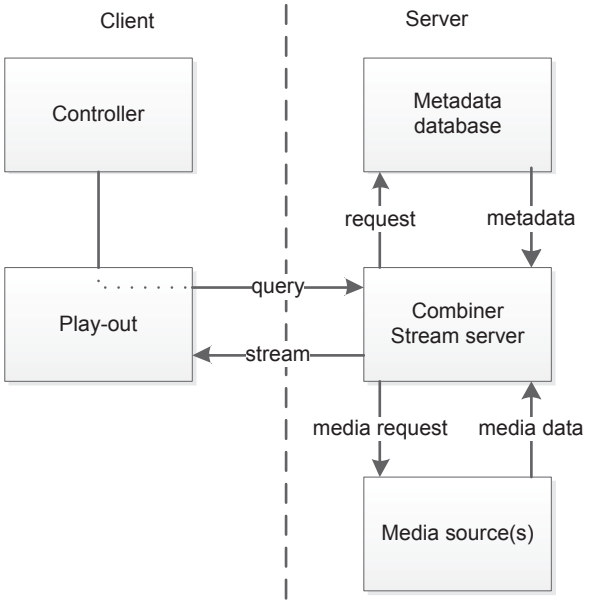

Figure 3. Server side media source combination.

have been proposed both in the academic community as well as in commercial applications: from touch-based interfaces to gesture recognition and voice recognition. The underlying principles of navigating between TV channels have not changed however. Instead of pressing a channel up or down button on a remote control, one might make an analogous gesture to a TV. And instead of pressing a specific channel number on a remote control, one might now shout the name of the channel to the TV. However, the essence of navigating between TV channels is still either linear (the channel up/down button) or directed (pressing '7' on your remote control). The only major innovation in this regard has been the introduction of the EPG, which allowed users to see an overview of the available programming before making a selection. MosaicUI can be seen as an evolution of this EPG. By giving users a visual overview of all available content, in this case TV channels, they can make a selection in a more intuitive way than by reading program titles from an EPG.

In figure 4, one can see an example of a second screen device being used to navigate through TV content. In this case the second screen device receives a mosaic video that incorporates the live video streams of all the available TV channels. By clicking on a particular channel, the second screen device instructs e.g. the TV to switch to the selected channel. Alternatively, the second screen device itself switches to a high resolution version of the selected channel, allowing the user to watch the particular channel on the second screen device instead of on the TV. Depending on the application, the selection of the TV channels that are included in a mosaic can be generic, such as a default mosaic including the 25 most popular channels, or personalized. In the personalized case, the selection of channels in the mosaic could for example be based on the user's viewing history or on a pre-selected list of channels. It is also possible to further configure the mosaic by for example including metadata information or channel logos either in the mosaic stream itself, or in the second screen application being used to display the mosaic. The second screen TV channel application might also be used to allow a personalized picture-in-picture stream. For example, a user might want to watch multiple channels simultaneously. In this case, he selects the desired channels from the mosaic (see the visual overlay in Figure 4), presses a button, and the second screen application sends the newly created video stream, consisting of the selected videos, to

the TV, for example through Apple Airplay. As discussed in the previous section, this newly configured mosaic stream can either be generated by a network-side process or as part of the second screen application.

Figure 4. Interactively selecting videos in a mosaic grid on a second screen device.

3.2 Interactive video search

Another possible application for MosaicUI is as a new method of displaying video search results. In most current video portals that allow for video search, e.g. YouTube and Vimeo, the results of a particular search query are displayed as text accompanied by either a static or animated thumbnail while navigating the results utilizes, again, a linear interface. This mostly text-based and static display of results can make it difficult for a user to assess which of the listed videos best matches with what he was looking for. The MosaicUI framework can be used to visually present the results of a particular search query, e.g. by including the 25 most relevant video results in the mosaic video. Upon seeing the results in their actual video form, a user would then be able to more quickly assess the results and either change his search query or select a particular video from the mosaic grid. This system could be extended with a function that works in a way similar to Google Instant, which updates your results continuously while you type. In a MosaicUI system, this would mean that the videos included in the mosaic would continuously be updated while the user refines his search query. Certain videos in the mosaic might be swapped out for others, while the position of those that are in the mosaic might change depending on their relevance. The most relevant videos might for example be placed in the middle of the mosaic. Also, the size of each tile in the grid could vary, depending on for example the relevance of that search result, content type or user specific criteria. Such an interactive video search system could support browsing video along multiple threads, as described in [8].

3.3 Video surveillance

One of the advantages of the MosaicUI framework is that it allows for simultaneous playback of multiple videos on devices that normally are not capable of doing so. Examples of such devices are smartphones and tablets that do not have the processing power for media decoding and are limited by a single hardware decoder. One application where this kind of functionality is useful is in the area of video surveillance. In this case, MosaicUI can be used to present security officials, working in the field, an immediate overview of a particular area or location by placing all relevant camera feeds into a grid and sending the resulting mosaic to e.g. a smartphone. By clicking on a particular video, the security official can quickly get a higher resolution version of a particular camera feed.

4. CONCLUSION AND FUTURE WORK

In this paper, a novel platform for navigating through multiple video sources has been presented. Three potential use cases, navigating between TV channels, intuitive video search and interactive video surveillance have been implemented, discussed and demonstrated. There are, however, a multitude of other applications and media sources which could be implemented by or connected to this platform. As part of the future work, we will look at the aforementioned combination of, and integration with, different (types of) media sources and experiment with different forms of interaction. Currently, a tablet has been functioning as user interface. In the future we will investigate the use of motion tracking and gesture recognition platforms (e.g. the Kinect) for interacting with the mosaic video. Connecting the platform to a platform like YouTube or Vimeo will be investigated as well. Preliminary results show that although technically possible, the respective APIs do not allow for a large number of concurrent connections. It is concluded that new intuitive user interfaces like multi-touch tablets, combined with the ability to create grid-like compositions of multiple video sources (optionally augmented with indexed metadata) creates a number of exciting new possibilities. In contrast to earlier research in the field of grid based media navigation, the utilization of new user interfaces greatly extends the flexibility and possibilities in media navigation. Application of these new possibilities proves not to be limited to over the top concepts, but to real world user centered cases as well with immediate usage in end-user environments and security environments, to name a few.

5. ACKNOWLEDGMENTS

Part of the research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248138.

6. REFERENCES

[1] Janse, M.D., Das, D.A.D., Tang, H.K. & Paassen, R.L.F. van (1997). Visual tables of contents: structure and navigation of digital video material. IPO Annual Progress Report, 32, 41-50. http://www.tue.nl/publicatie/ep/p/d/144348/?no_cache=1

[2] Snoek, C. G. M.; van de Sande, Koen E. A.; de Rooij, O.; Huurnink B.; Gavves, E.; Odijk, D.; de Rijke, M.; Gevers, T.; Worring, M.; Koelma, D.C.; Smeulders, A. W. M. "The MediaMill TRECVID 2010 semantic video search engine". In Proceedings of the 8th TRECVID Workshop. Gaithersburg, USA, November 2010.

[3] MultimediaN Golden Demos http://www.multimedian.nl/en/demo.php. Visited March 2, 2012.

[4] Barkhuus, L.; Browns, B. "Unpacking the Television: User Practices around a Changing Technology". In ACM Transactions of Computer-Human Interaction 16, 3, September 2009.

[5] SII Metadata Extraction Services. http://www.si-i.nl/project.php?id=82. Visited March 2, 2012.

[6] FascinatE. http://www.fascinate-project.eu/. Visited March 2, 2012.

[7] van Brandenburg, R.; Niamut, O.; Prins, M.; Stokking, H.; , "Spatial segmentation for immersive media delivery," in Proceedings of 15th International Conference on Intelligence in Next Generation Networks (ICIN), 4-7 October, 2011.

[8] de Rooij, O.; Worring, M. "Browsing Video Along Multiple Threads," In IEEE Transactions on Multimedia, Vol. 12, No. 2, February 2010.