AR Record&Replay: Situated Compositing of Video Content in Mobile Augmented Reality

Tobias Langlotz[#], Mathäus Zingerle[*], Raphael Grasset[#], Hannes Kaufmann[*], Gerhard Reitmayr[#]

[#]Graz University of Technology
Inffeldgasse 16, 8010 Graz, Austria
{langlotz, grasset, reitmayr}@icg.tugraz.at

[*]Vienna University of Technology
Favoritenstrasse 9-11, 1040 Vienna, Austria
{zingerle, kaufmann}@ims.tuwien.ac.at

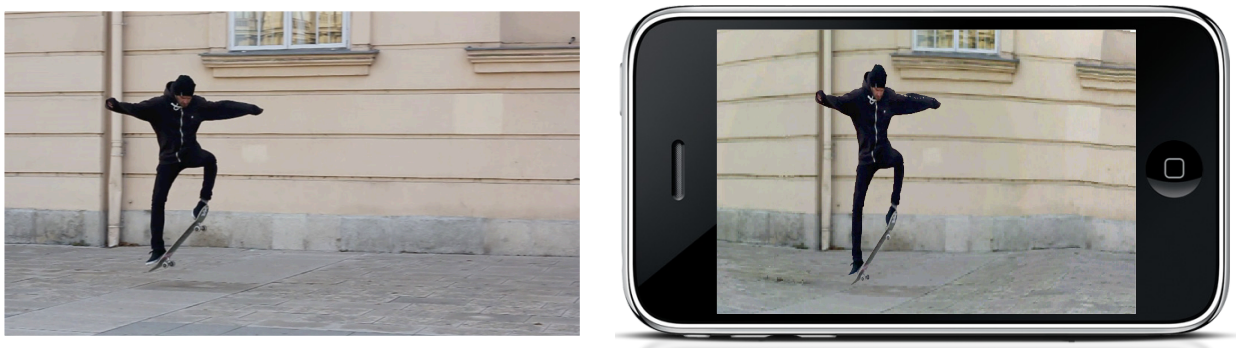

Figure 1. Illustration of situated video augmentations. (Left) Original video footage recorded using a mobile phone. (Right) Illustration of the augmented video application. The foreground video object – in this case the skateboarder – is augmented in the users view.

ABSTRACT

In this paper we present a novel approach to record and replay video content composited in-situ with a live view of the real environment. Our real-time technique works on mobile phones, and uses a panorama-based tracker to create visually seamless and spatially registered overlay of video content. We apply a temporal foreground-background segmentation of video footage and show how the segmented information can be precisely registered in real-time in the camera view of a mobile phone. We describe the user interface and the video post effects implemented in our prototype as well as our approach with a skateboard training application. Our technique can also be used with online video material and supports the creation of augmented situated documentaries.

Author Keywords

Augmented Video, Augmented Reality, Mobile phone

ACM Classification Keywords

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – Artificial, augmented, and virtual realities

OZCHI'12, November 26–30, 2012, Melbourne, Victoria, Australia.

INTRODUCTION

The availability of inexpensive mobile video recorders and the integration of high quality video recording capabilities into smartphones have tremendously increased the amount of videos being created and shared online. With more than 50 hours of video uploaded every minute on YouTube and billions of videos viewed each day[1], new ways to search, browse and experience video content are highly relevant.

Current user interfaces of online video tools mostly replicate the existing photo interfaces, however. Features as geo-tagging or browsing geo-referenced content in virtual globe application such as Google Earth[2] (or other map-based applications) have been mainly reproduced for video content.

More recently, efforts have been made to explore further the spatio-temporal aspect of videos. Applications such as Photo Tourism (Snavely et al., 2006) have inspired work of Ballan et al., 2010, allowing end-users to experience multi-viewpoint events recorded by multiple cameras. Their system allows a smooth transition between camera viewpoints and offers a flexible way to browse and create video montages captured from multiple perspectives.

[1] http://www.youtube.com/t/press_statistics

[2] http://www.earth.google.com

However, these systems limit themselves to produce and explore video content on desktop user interfaces (e.g. web, virtual globe) out of the real context. *Augmented Reality* (AR) technology can overcome this issue, providing a way to place geo-referenced video content on a live, spatially registered view of the real world.

For example, Höllerer et al., 1999 investigated situated documentaries and showed how to incorporate video information into a wearable AR system to realize complex narratives in an outdoor environment. Recent commercial AR browsers such as Layar[3] or Wikitude[4] are now integrating this feature, supporting video files or image sequences but with limited spatial registration due to the fact that the video is always screen aligned and registered using GPS and other sensors.

Video augmentation has also been explored for publishing media. RedBull[5] for example, presented an AR application that augmented pages of their Red Bulletin magazine with video material using *Natural Feature Tracking* (NFT). The application was running within a webpage as an Adobe Flash application, detected a magazine page and played the video content spatially overlaid on top of that page.

As these projects generally present the video on a 2D billboard type of representation, other works have been exploring how to provide more seamless mixing between video content and a live video view. MacIntyre et al. investigate within their *Three Angry Men* project the use of video information as an element for exploiting narratives in augmented reality (MacIntyre et al, 2003). They proposed a system where a user wearing a *Head Mounted Display* (HMD) can see overlay video actors virtually seated while discussing around a real table. The augmented video actors were prerecorded and foreground-background segmentation was applied to guarantee a seamless integration into the environment, created with their desktop authoring tool (MacIntyre et al, 2001, MacIntyre et al, 2002).

Whereas MacIntyre et al. used static camera recording of actors, the 3D Live (Prince et al, 2002) system extended this concept to 3D video. Prince et al. used a cylindrical multi-camera capture system, allowing capture and real-time replay of a 3D model of a person using a shape-from-silhouette approach. Their system was supporting remote viewing, by transmitting the 3D model via a network and displaying the generated 3D video onto an AR setup at a remote location as part of a teleconference system.

While these applications were proposed for indoor scenarios, Farrago[6], an application for mobile phones, proposed video mixing with 3D graphical content for outdoor environments. This tool records videos that can be edited afterwards by manually adjusting the position of the virtual 3D objects overlay on the video image, but requires the usage of 2D markers or face tracking. Once the video is re-rendered with the overlay, it can be shared with other users.

In this work, we investigate how we can offer a new user experience to a mobile user through compositing the user's view of the real world with prerecorded geo-referenced video content. Similar to MacIntyre et al., we are interested in extracting the salient information from the video (e.g. moving person or objects) and offer the possibilities to spatially navigate the video (by rotating the phone) mixed with the view of the real world. Differently to their work, we focused on mobile platforms in outdoor environments and also looked at offering simple ways to record and capture this type of video content with only a minimal input. We also have fewer restrictions during the recording, as we support rotational camera movements and do not rely on a green screen type of technology for recording the video augmentations.

In this paper, we present our interactive AR technique offering accurate spatial registration between recorded video content (e.g. person, motorized vehicles) and the real world with a seamless visual integration (e.g. extracted break dancer recorded the day before overlaid on one's camera video). Our system allows to replay geo-referenced video sequences, to re-enact a past captured event for a broad range of applications covering sports, history, cultural heritage or education. We support a variety of tools for the user to control video playback and apply video effects, hence delivering the first prototype of what can be a real-time AR video montage tool for mobile platforms (see Figure 1).

The presented system operates in three steps. The first step ("record") is the shooting of the video, including geo-tagging and uploading to a remote server for further processing (or social access).

In a second step, we extract the object of interest in the video frames, which we later augment in place. This pre-processing task can be performed remotely (server hosting the video) or can be done locally (desktop PC). We apply a segmentation only requiring that the user outlines the object of interest in the first frame. We also extract the background information of the video and assemble it into a panoramic representation of the background, which we later use for the precise registration of the video content into the environment.

The final step of the system is the "replay" mode with our system. This mode is enabled once a mobile user moves close to the position where a video sequence was shot. The system downloads previously created information. While the user explores the environment the video is registered using computer vision and augmented into the users view.

The proposed system contributes to the field of Augmented Reality demonstrating how to incorporate seamlessly video content into outdoor AR applications and allowing end users to participate in the content creation processes.

[3] http://www.layar.com

[4] http://www.wikitude.com

[5] http://www.redbullusa.com

[6] http://www.farragoapp.com

Figure 2. (Left) Manual initialization of the segmentation step. User sketches the foreground object (red) and outlines the background (blue). (Middle) Result of applying GrabCut segmentation to subsequent video frames: Segmented foreground object and size-optimized texture (green outline). (Right) Tracking the segment using Lukas-Kanade Tracker allows segmentation of later frames even in case the appearance changes.

SITUATED VIDEO COMPOSITING FOR AR

In the following we give a detailed description of our approach and describe the algorithms used for our AR video compositing technique.

Video shooting ("record")

The video capture is performed using standard video devices such as a smartphone or a digital camera (compressed, high definition). Our approach requires the camera location to be fixed while recording – but rotational movements of the camera are possible. The recorded video is then geo-tagged with the user's current GPS location for later use, and uploaded to a cloud-based server or transferred to a personal computer.

Offline video processing

In this step we process the video to extract relevant information. The main challenge here is to separate the object of interest in the video (foreground) from the remaining information such as the background or other moving objects that are not of interest. We later use the object of interest as an overlay but we also want to keep the background information as it is needed to register the video overlay into the new scene. This preprocessing can be done on a personal PC or on a cloud server hosting the uploaded video.

Foreground Segmentation. We start segmenting the video by applying a variation of the GraphCut algorithm, namely GrabCut, presented by Rother et al. (Rother et al, 2004). To initiate the algorithm, the user has to roughly sketch the object of interest (the foreground object) and mark some of the background pixels on the video image (in practice close to the foreground object, see left Figure 2). Then GrabCut delivers a segmentation of the video image into the foreground object and the background.

This approach was developed for static images and now needs to be applied to each frame as we are using temporal content. To avoid the cumbersome task of marking every individual video frame manually, we extended the method in a similar way to the approach presented by Mooser et al. (Mooser et al., 2007). The idea is to use the segmentation output of the GrabCut algorithm of the previous video frame to initialize the segmentation computation for the current frame.

As there is likely a movement of the object of interest between the two frames we cannot naively apply the result of the segmentation from the previous frame to the current one. We address this issue by estimating the position of the segmented foreground object in the current frame by computing the optical flow of pixels between the previous and the current frame using the Lucas-Kanade algorithm (Lucas and Kanade, 1981). This gives us an approximation of the foreground object's position in the current frame. We also dilate the estimated foreground object's footprint to compensate for tracking inaccuracies.

We compute the boundary of the estimated foreground object and select pixels within (pixels of the foreground object) and outside (background pixels) and use them as input for GrabCut. Applying this approach for each frame yields the foreground objects for all consecutive frames of the video. We apply a dilate and erosion operation on the segmented foreground objects to remove noisy border pixels and only keep the largest connected component as foreground object in case the segmentation computed more than one segment. Contrary to the approach of Mooser et al. (Mooser et al., 2007) we do not consider the edge saliency for tracking the object within the camera frames and also do not apply a banded graph cut-based segmentation. However, we also kept an option to manually initialize the GrabCut for specific frames in case the object of interest is not segmented properly.

The segmented foreground object is often only a fraction of the size of the full video frame (see Figure 2). To reduce the data we store the foreground object by only saving the bounding rectangle around it and its offset within the video frame.

Background Information. Once the foreground object is extracted we also need the background information, which we use for registering the object into the view of the user. We take the segmented frames and focus in the following only on the background pixels.

Due to the possibility that a user can rotate the camera while recording the video, the recorded frames hold different portions of the scene's background. Furthermore, the foreground object also occludes parts of the background, reducing the amount of visual features that are later available for vision-based registration. As we want to reconstruct as much background information as possible we do not only take into account the background information from one video frame but from all frames and integrate them into one panoramic image.

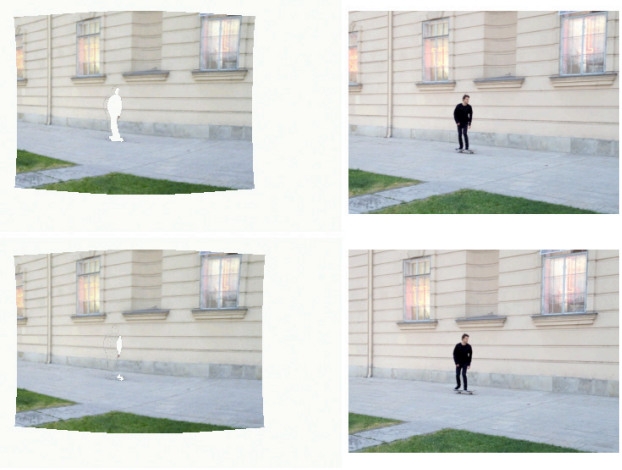

Figure 3. Creating a panoramic image containing only background information. (Top) Holes that are caused by occlusion are closed by adding the background pixels from later video frames (Bottom). The right side always shows the latest video frame that is used with the result of the segmentation as input for the panorama computation.

We create this panoramic image keeping the background pixels by using a modified version of the panoramic mapping and tracking approach presented by Wagner et al. (Wagner et al, 2010). In their original implementation, they use features in the incoming video frames to register the frames and stitch them into a panoramic image. They also demonstrated in their approach how to track the camera motion R_S of the recording camera while constructing the panoramic image. Similarly, we assume the camera movements are only of rotational nature.

We adapted their technique to handle alpha channels and to only map pixels into the panoramic image that are considered to be background pixels. The resulting holes in the panorama caused by the occluding foreground object that are not mapped, can be closed by also mapping later frames of the video as the foreground object moves within the camera frames revealing occluded background information in later frames (see Figure 3). We store the resulting panoramic image that contains the background information of the video as well as we also store the camera rotation R_S for each video frame.

All this information – the segmented video, the panoramic image holding the background information and the camera rotation for each frame, GPS geo-location – is packaged into a specific data structure and saved in a compressed file. This packaged dataset can be easily shared online and made available via a cloud repository.

Online video processing ("replay")

For this step we assume that the user's phone is also equipped with GPS similar to the recording device. We can query video augmentations located in the vicinity of the current user location: they can be retrieved from local storage or online from a cloud repository. Once the data is decompressed, we start registering the video into the current user's view.

For this purpose, we also use the technique from Wagner et al., 2010: We build a new panoramic image from the current camera feed and also track the current camera rotation R_T.

The use of the panorama-based tracking allows for a higher precision of the registration and the tracking, as we do not rely on noisy sensor values. As mentioned earlier this comes with the drawback of supporting only rotational movements. However, most users only perform rotational movements while using outdoor AR applications (Grubert et al, 2011) making this constraint acceptable in most scenarios.

While building the new panorama of the environment we try to match the loaded panorama holding the background pixels against this newly built panorama. The matching is performed using a point feature technique (in our case with PhonySIFT, Wagner et al., 2008). As soon as the overlapping area – the area holding image information that are in both panoramas – is big enough, the matching using PhonySIFT should succeed and provide the transformation T_{ST} describing the relative motion between the camera used to record the video (the source camera *s*) and the camera where the video information should be registered in (the target camera *t*).

By assuming that the user of the system is roughly at the same position where the video was recorded (identified via GPS) we can constrain the transformation T_{ST} to be purely rotational (see Figure 4).

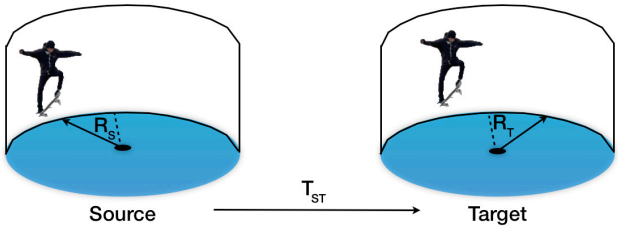

Figure 4. Illustration of the applied transformation between the source camera and the target camera used for replaying the augmented video

Using this transformation T_{ST} we can transform each pixel from the source panorama into the target panorama and vice versa. This allows us to play the video information by overlaying the current environment with the object of interest from the video frame. We therefore load the video frames and by applying for each video frame the combination of the transformation R_S (the orientation of the source camera computed in the offline video processing step), the transformation T_{ST} (the transformation between the source and the target camera gained from the registration) and the transformation R_T

(the orientation of the target camera computed using the panorama-based tracking) we can precisely augment the video content into the users view (see Figure 4).

Please note that using the panorama-based tracker we obtain an update of the transformation R_T at each frame. This allows us to rotate the target camera completely independent from the orientation of the source camera and to maintain the precise registration of the video in the current view.

PROTOTYPE

Video augmentations can be used for a wide range of applications covering areas such as entertainment or edutainment if integrated into outdoor AR applications such as browser systems.

We implemented our technique in a prototype of a mobile video editing AR application. Inspired by the current tools proposed in desktop video editing applications, we focused on some of their major features: video layers, video playback control and video effects.

In the following, we present an overview of the user interface and the post-effects implemented in our system. We also describe a case study of our technique.

User Interface

The interface of our prototype is inspired by the design of graphical user interfaces generally found in video editing tools. We created three different groups of functions distributed around the screen that can be operated with touch screen operations. The control groups of our implemented user interface are (illustrated in Figure 5): the video control group, the video layer group and the video effects group. In case the menu options in our interface are not used for a certain time (in our case 5 seconds) the menu disappears to give more space to the video, but will be redisplayed as soon as the user touches the screen.

Figure 5. Screenshot of our prototype showing a video augmentation together with the implemented interface with the control groups for video control (top and bottom), video layers (left side) and video effects (right side).

The functions in the video control group consist of the playback control found in most video players: play control buttons, time slider and speed buttons. In addition we added the possibility to slow down or increase the speed of the currently played video.

The video layer group provides access to the different video sequences accessible at this GPS location, organized in different depth layers. The end-user can control and activate/deactivate these different layers, which can be played independently of each other or simultaneously.

The last group of items, the video effects, trigger real-time video effects that can be applied to the different video sequences. Each effect can be switched on or off and the effects can be combined with each other.

Post effects and layers

Applying visual effects is an important part of video post-productions. Effects are used to highlight actions, and create views that are impossible in the real world, such as slow motion or highlighting of elements within the video. Normally these effects are applied to the video material in a rendering step that is carried out in an offline manner (Linz et al, 2010).

Because of the nature of our approach we are able to perform a wide variety of these video effects in real-time on a mobile device without the need of pre-rendering the video.

We explored space-time visual effects such as multi-exposure effects, open flash and flash-trail effects. Multi-exposure effects simulate the behavior of a multi exposure film where several images are visible at the same time. We can easily simulate this behavior for cameras with a fixed viewpoint by augmenting several frames of our videos at the same time. This results in having the subject appearing several times within the current view, such as in a multiple exposure image.

An extension of this effect is the Flash trail effect. This effect also allows seeing multiple instances of the same subject but the visibility depends on the time passed by (see Figure 6 right). This effect supports a better understanding of the motion in the recorded video. We implemented the Flash trail effect by blending in past frames of the augmented video with increasing amount of transparency. Thereby the strength of the transparency and the time between the frames can be freely adjusted.

Our approach allows us to play back more than one video at the same time by still allowing a seamless integration into the environment (see Figure 6 left). This allows it to compare actions that were performed at the same place but at a different time by integrating them into one view, thus bridging time constrains. Each video corresponds in our system to a video layer and the user can switch between these layers or play them simultaneously.

Other visual effects that can be enabled are different glow or drop-shadow variations that can be used to highlight the video object or in the case several video layers are playing at the same time the glow effect can be used to highlight a certain video layer.

All these presented effects and video layers do not require any preprocessing but are carried out on the device while

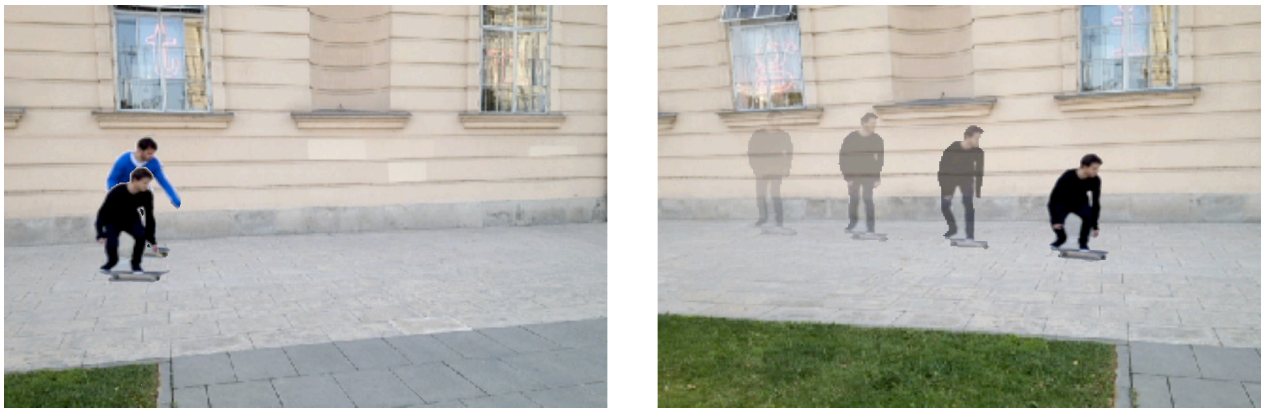

Figure 6. Examples of layers and an example for realized post effects as used in our skateboard tutor application as captured from an iPhone 4. (Left) Playing back two video augmentations layers allows the comparison of the riders' performance. (Right) Flash-trail effects visualize the path and the motion within the video.

playing back the video. Therefore, they can be combined or switched off on demand.

Skateboard tutor application

We demonstrated our system to end-users as part of a skateboard tutoring application. Skateboard videos are a good representative of dynamic real world content naturally evolving in our real world environment. The different ranges of maneuver performed with a skateboard (tricks) are largely bound to the environment and location through natural or artificial obstacles and ramps. Skateboarding videos make also use of a variety of camera shooting techniques (perspectives, movement, optics) due to the dynamic of the skateboarder evolving in the real environment. On the other hand, Tutorial/How-To videos represent a large part of YouTube videos (Sharma and Elidrisi, 2008) today, confirming the potential of this format for skills or competences learning. Skateboard tutorials (>30.000 hits on YouTube) serve therefore as a good application of our technique.

Our skateboard tutoring application allows recording skateboard tricks that can be shared with other users for demonstration and learning purposes. The application can be used to overlay the pre-recorded video content (extracted skateboarders) in place to replay and experience the tricks and actions performed by another user (or from a previous day) in the correct context (see Figure 7). It can support the learning process as online skateboard videos, generally recorded with fish-eye lens, can give a distorted perception of the skateboarder in the real environment.

Our test skateboard videos were recorded with normal digital cameras or smartphones and are processed using our approach of situated video compositing for AR. We also make use of the proposed post effects and layers. The layer approach allows recording skateboard maneuver on the fly, which can later be played back in parallel with other stored maneuver for comparison (e.g. speed, height of jumps). The flash-trail effect can be used to highlight the motion and the path of the rider.

We implemented the offline video processing using OpenCV[7]. The mobile skateboard tutoring application has been implemented on the iOS platform using the Studierstube ES framework (Schmalstieg and Wagner, 2008). We tested the application successfully on an Apple iPhone 3GS, iPhone 4S and an iPad2. Cross-devices, the application runs in real-time between 17fps (Apple iPhone 3GS) and 28 fps (Apple iPhone 4S/iPad2).

EVALUATION

We conducted a preliminary user study for gathering first user feedback on our technique as well as to identify flaws, improvements or to get additional ideas for the applicability of our technique.

Scenario and Setting

We evaluated our technique with the skateboard tutoring application as being a relevant use case scenario.

Producers of this kind of videos are usually also consumers, leveraging the possibility to collect feedback for both, the creation of video augmentations and experiencing video augmentations. We therefore invited skilled skateboarders (domain experts) who have experience in creating skateboard videos or tutorials already and who published their videos online via popular sharing platforms.

Our main objective with the user evaluation was to identify the usefulness and applicability of our approach as well as the usability of our created prototype.

In total we had 5 expert users with >7 years of skateboarding experience (all male, 25-28 years), all of them were involved in producing skateboard videos, some produced videos for marketing. All considered themselves as not tech-savvy. Two have minimal knowledge about augmented reality, none had any experience with any kind of AR application beforehand. One participant stated not to be very familiar with the

[7] http://sourceforge.net/projects/opencvlibrary/

Figure 7. Scenario as used during the user study. (Left) Skateboarder was recorded with a mobile phone while performing his actions. (Middle) Frame of the recorded video sequence. (Right) The same action as augmented within our skateboard tutoring application as captured from an iPhone 4.

usage of mobile devices as he restricted his usage of mobile phones to place calls or write messages.

Procedure

We gave all participants the chance to get hands-on experience with our prototype that we demonstrated on both an iPhone 3GS and an iPad2 (see Figure 8). After introducing the project, we gave them a short demonstration of the application, showing the different features. They were able to test the integrated effects as well as to try the video layers by playing two video layers that were augmented at the same time.

We selected 2 participants to create their own skateboard video that was later augmented, while all other users only had the chance to experience the augmented videos. After the participants finished trying out the prototype we asked them a series of questions as part of a semi-structured interview.

Figure 8. Evaluation of our prototype with domain experts using an Apple iPad2.

Results

In the interview all participants confirmed that our prototype was easy to use. Regarding the comfort factor with the device and interacting with the application, only one user didn't feel really comfortable (who was the participant not experienced with smartphones). We also asked them about the social aspect of using the application outdoor in a busy area, participants replied to be really comfortable on this aspect. All users commented that our system was easy to learn, and the current interface was well received.

Questioned about the usefulness, they all scored our application as really useful. They confirmed our hypotheses about the difficulty of perception and knowledge understanding with traditional online videos and the viability of our in-context video augmentation.

Three of the five participants said that they enjoyed the freedom of having control of the camera orientation during playback, as it was not relying on the recording camera orientation (fixed on traditional video recording technique). They highlighted the possibility of playing several videos/layers at the same time that are overlaid in parallel. They described it as a really useful for comparing their own runs with the tutorial video to detect differences. They also liked the flash-trail effect saying that this effect seems to be useful for studying "the line" a rider skates.

When asked about the general applicability and the usefulness of experiencing video augmentations in place the participants generally rated really positively regarding the usage in other application areas. However, two of them pointed out during the interview that the users have to visit the place, which makes more sense in certain specific cases. Both of them stated that they therefore generally see it more as a gadget as they could not extrapolate other convincing use-cases. As we presented them other possible use cases at the end of the interview (city guides, parades/events within the city) they answered that they see also potential in this kind of application but needed to experience it for a more reliable answer.

The last part of the interview focused on the visual quality of the technique, in term of spatial and visual integration. Three participants had the feeling that the scene and the rider were 3-dimensional and giving a sense of "authenticity", one perceived the rider to be 2-dimensional but the scene to be 3D, while the remaining two participants stated that it was all overlaid in 2D. They all commented that the movement of the augmented skateboarders within the scene was very realistic. Even after being explicitly asked they could not remember to have seen any drifting between the augmentation and the background. However, when asked about the seamless visual integration, we received more mixed answers.

They stated that sometimes the skateboarder did not have the same appearance, as the background was too dark or incorrectly lit. Two participants also noticed small segmentation errors (e.g. a wheel of the skateboard was disappearing in a couple of video frames).

The two participants that generated their own shooting video said that it was simple to use and the additional step acceptable for the generated outcome. When asked about constraints in the camera motion during shooting – limited to rotational movements of the camera - they said that it is likely to be acceptable in most cases. They explained that a huge majority of the people is making short videos with smartphone devices from a single point of view. One of the participant said: "The given constraints fit the medium, as I think the majority of the short online videos were shot in this [constrained] way". Our system delivers then all criteria generally used by laypersons recording skateboarding videos. Finally, during the open questions one participant proposed the possible use of our application as a mobile blue screen, which allows users to capture objects and scenes and assemble them together using the layer view.

DISCUSSION

Overall the evaluation using our skateboard experts showed that our approach has advantages over existing mobile video applications (shooting, video effects, playback). However the final outcome and the usefulness strongly depends on the use case. Even though all of our participants were not tech-savvy they had no problems to learn and handle our prototype application.

A major limitation with our prototype pointed out by the participants was the visual quality of the overlay. Even though the ratings were above average the users complained about the lack of visual coherence: The video augmentation looked different from the current environment. In our case this was mostly caused by cloudy weather during recording time resulting in low contrast actors, while it was mostly sunny during the playback of the video augmentations. This can be treated in future versions of our prototype by implementing an adaptive visual coherence. The basic idea is to compare the background panorama of the video with the current environment to adjust the video augmentation in terms of contrast and color.

Another problem was that the segmentation sometimes was not accurate enough, especially if applied to a well-structured background as required for vision-based registration. However, more sophisticated segmentation algorithms and better algorithms for tracking the segmented objects exist but require more expensive computation or GPU implementations and need to be investigated in the context of this work. Especially as the segmentation as used in our system has inverse requirements as the vision-based registration used in our approach: A less structured background achieves in general significantly better results in foreground-background segmentation, while it poses a hard problem when used to register the augmentation based on the background information.

Despite these drawbacks, our application showed that augmented video could be an interesting element especially as video content is often easier to create than 3D content, making our approach interesting for many applications.

Professional applications can benefit from video augmentations as realized in our approach. Augmented reality-based tourist guides could display more interactive content e.g. by capturing the guide for later replay. Furthermore authoring such content is less demanding than creating dynamic 3D content. This allows to easily create in-situ narratives similar to the concept of situated documentaries presented by Höllerer et al., 1999.

Many augmented reality applications can benefit from the simplicity of creating video augmentations using our approach, allowing laypersons to create content and share it with friends. This enables the creation of videos of certain events (e.g. parades, street artists etc.) and play-back in place at a different time.

More separation between the constraints used during shooting with the ones used for replaying can be explored further. Cinematography components such as camera type, camera movement, visual style of the image, location and its content are some examples of elements that can be altered, modified or "warped" between the record and replay. You can imagine to record a cyclist of the Tour de France with a rolling technique and replay it fixed in another location. Further, real-time montage with live video, online content and collaborative editing can leverage the full potentiality of mobile AR.

CONCLUSION

We presented an approach for in-situ compositing of video content in mobile augmented reality. We showed how to create and process video files for the use in mobile AR as well as how to register them precisely in the user's environment using a panorama-based tracking approach. Even though the approach is constrained to rotational movements of the cameras due to the usage of a panoramic representation of the environment, it could be applied to many existing outdoor AR applications, as this motion pattern is common for using AR browsers, as well as for shooting short videos.

We demonstrated the application with a skateboard tutoring prototype. Our prototype allows experiencing skateboard tricks and actions recorded by other people that are augmented in-place and displayed at interactive frame rates on mobile phones.

Future work will target better segmentation algorithms and an improved visual coherence between the overlay and the augmented environment. Porting the offline video processing to the mobile could be another future step.

Overall we hope that this work demonstrates possible usages of video footage in future mobile AR applications as well as it shows the advantages of interfaces that allow experiencing videos in place.

ACKNOWLEDGMENTS

We would like to thank all users participating in the experiments. We especially thank Holger Regenbrecht for his input and discussion. This work was partially supported by the Christian Doppler Laboratory for Handheld Augmented Reality.

REFERENCES

Ballan, L., Brostow, G.J., Puwein, J., and Pollefeys, M. Unstructured video-based rendering: Interactive Exploration of Casually Captured Videos. ACM SIGGRAPH 2010 papers on – SIGGRAPH'10, ACM Press (2010).

Grubert, J., Langlotz, T., and Grasset, R. Augmented Reality Browser Survey. Technical Report 2012, http://www.icg.tugraz.at/publications/augmented-reality-browser-survey, 2012.

Guven, S. and Feiner, S. Visualizing and navigating complex situated hypermedia in augmented and virtual reality. 2006 IEEE/ACM International Symposium on Mixed and Augmented Reality, IEEE (2006), 155-158.

Höllerer, T., Feiner, S., and Pavlik, J. Situated Documentaries: Embedding Multimedia Presentations in the Real World. In Proceedings of the 3rd IEEE International Symposium on Wearable Computers (ISWC'99), (1999), 79-86.

Linz, C., Lipski, C., Rogge, L., Theobalt, C., and Magnor, M. Space-time visual effects as a post-production process. Proceedings of the 1st international workshop on 3D video processing - 3DVP '10, ACM Press (2010).

Lucas, B.D. and Kanade, T. An iterative image registration technique with an application to stereo vision. IJCAI'81 Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2, (1981), 674-679.

Macintyre, B., Lohse, M., Bolter, J.D., and Moreno, E. Ghosts in the Machine : Integrating 2D Video Actors into a 3D AR System Georgia Institute of Technology. 2nd International Symposium on Mixed Reality, (2001).

MacIntyre, B., Lohse, M., Bolter, J.D., and Moreno, E. Integrating 2-D video actors into 3-D augmented-reality systems. Presence Teleoperators, (2002), 189-202.

MacIntyre, B., Bolter, J.D., Vaughn, J., et al. Three Angry Men: An Augmented-Reality Experiment In Point-Of-View Drama. IN PROCEEDINGS OF TIDSE 2003, (2003), 24 - 26.

Mooser, J., You, S., and Neumann, U. Real-Time Object Tracking for Augmented Reality Combining Graph Cuts and Optical Flow. 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, IEEE (2007), 1-8.

Prince, S., Cheok, A.D., Farbiz, F., et al. 3D Live: Real Time Captured Content for Mixed Reality. ISMAR'02 Proceedings of the 1st International Symposium on Mixed and Augmented Reality, (2002).

Rother, C., Kolmogorov, V., and Blake, A. „GrabCut": interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics 23, 3 (2004), 309.

Sharma, A.S., and Elidrisi, M. Classification of multi-media content (videos on YouTube) using tags and focal points. Unpublished manuscript, http://www-users.cs.umn.edu/~ankur/FinalReport_PR-1.pdf, (2008).

Schmalstieg, D. and Wagner, D. Mobile Phones as a Platform for Augmented Reality. Proceedings of the IEEE VR 2008 Workshop on Software Engineering and Architectures for Realtime Interactive Systems, (2008), 43-44.

Snavely, N., Seitz, S.M., and Szeliski, R. Photo tourism: Exploring photo collections in 3D. ACM Transactions on Graphics 25, 3 (2006).

Wagner, D., Mulloni, A., Langlotz, T., and Schmalstieg, D. Real-time panoramic mapping and tracking on mobile phones. 2010 IEEE Virtual Reality Conference (VR), (2010), 211-218.

Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T., and Schmalstieg, D. Pose tracking from natural features on mobile phones. 2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality, (2008), 125-134.