Intelligent Film Assistant for Personalized Video Creation on Mobile Devices

Dominik Schörkhuber
Interactive Media Systems Group, Vienna University of Technology, Vienna, Austria
dominik.schoerkhuber@tuwien.ac.at

Florian Seitner
emotion3D GmbH, Vienna, Austria
fse@emotion3d.tv

Benedikt Salzbrunn
University of Applied Sciences Technikum Wien, Vienna, Austria
salzbrunn@technikum-wien.at

Margrit Gelautz
Interactive Media Systems Group, Vienna University of Technology, Vienna, Austria
margrit.gelautz@tuwien.ac.at

Georg Braun
emotion3D GmbH, Vienna, Austria
gbr@emotion3d.tv

ABSTRACT
We describe the development of an intelligent film assistant system that supports the creation of professional video content on mobile devices such as smartphones by amateur users. Cinematographic expert knowledge on scene composition and camera motion is provided to the user in the form of storyboards that are tailored to specific use cases. We give an overview of the project concept and some selected components including algorithms that are required for video stabilization and shot classification. The goal is to translate the scene characteristics detected by the vision algorithms into real-time feedback to the user during video recording and to support post-production. A major focus of the project is the incorporation of human-centered design principles with usability studies and expert interviews into the whole research process.

CCS CONCEPTS
• Human-centered computing → Mobile phones; Human computer interaction (HCI); • Computing methodologies → Image and video acquisition;

1 INTRODUCTION AND PROBLEM STATEMENT
Today’s smartphones are frequently equipped with high-quality cameras and provide a multitude of assistant technologies to support users during the acquisition of still images and videos. These assistant technologies range from simple image adjustment (e.g. automatic white-balancing, color correction, contrast adjustment, etc.) to techniques based on more advanced image analysis such as closed-eye detection, face detection and smile detection for improved focusing and triggering during capture. While these technologies enable even amateur users to create photos of an exceptional image quality, professional video creation - due to its higher complexity - still remains an exclusive domain for film professionals. The capturing of videos and the creation of video content require a high level of technical expertise (e.g. regarding the planning of a storyboard or the cinematographic execution), and errors during capture often cannot be corrected after shooting. In addition, post-production steps for video content such as cutting are typically more complex and time-consuming than manipulation of still images. Despite the large availability of smartphones, this combination of technical and artistic complexity and higher time-effort for performing post-production steps on large video data currently limits the number of users that can create professional video content.

The motivation behind our work is the development of an intelligent film assistant that supports amateur users during the capturing of video clips with smartphones and in the creation of professional videos. We implement a template-based storyboard concept that enables users to systematically capture content for specific genres or objectives. Based on the used template (e.g. a 3-minutes image film), the system suggests a sequence of video clips and provides cinematographic directions to the user on how individual video clips shall be recorded. Throughout the capturing of individual video clips, the user is supported by various assistant tools that aim at reducing common mistakes and improve the quality of the captured video clip. Based on the captured video clips and the underlying template, our system is capable of performing an automatic video creation that handles tasks such as cutting and synchronization of video clips with music. The overall system enables amateur users...
to create professional video content and opens up new possibilities for private as well as business use cases that can benefit from high-quality video content.

2 RELATED WORK
In the following, we review some related work in the context of cinematographic assistance for video creation and video stabilization. Furthermore, several commercial products that are relevant to our project are outlined.

2.1 Cinematographic Assistance
In a cinematographically inspired approach, Hasan et al. have designed a novel camera motion histogram descriptor (CAMHID) [7] for videos, which relies on singular value decomposition of motion vector fields and support vector machines. Their primary goal is to classify videos into different shot types (stationary, tracking, focus-in, focus-out, establishing, chaotic). Likewise, Bhattacharya et al. [1] have produced a classification system based on mapping known inter-frame transformations into a vector space and classify the result into aerial, bird-eye, crane, dolly, establishing, pan, tilt, and zoom with a support vector machine. It should be noted that shot classes are often defined in the literature depending on specific use cases and needs of directors, rather than referring to a standard definition. While Hasan et al. [7] used mainly videos from Hollywood movies and videos shot by themselves, Bhattacharya et al. [1] also incorporated unconstrained amateur videos from public sources into their test data.

The techniques presented in [7] and [1] are not assistance systems per se, but may have the potential to be used as such. We have identified NudgeCam [3] to be the historically first system for affective user assistance in mobile video creation. It is presented as a mobile application executed on a user’s Android phone. NudgeCam categorizes its assistance capabilities into three categories: show (present a demonstration), tell (relay instructions), and make (provide feedback). NudgeCam follows a template approach, where users first look at an example video, storyboard or other media. The teaching media can also contain hints and goals for a user to clear, for example, a list of shots to take for a certain video. Furthermore, template creators may also pick from a number of rules a user has to fulfill. Possible rules are defined regarding the shot length, faces (e.g., position of the most salient face; too centered; too small), audio volume, camera tilt, detection of erratic motion and brightness. In different fashion, The Director’s Lens [13] provides directive assistance in a purely virtual environment. Given the moving virtual scene, the system is able to give suggestions for matching camera work which appropriately captures all actions. Possible suggestions for camera movement are further ranked by fulfilling a number of cinematic continuity rules. Mitarai and Yoshitaka [21, 22] have created a system that assists users in creating videos that communicate preselected emotions. While capturing a shot, the system simultaneously analyzes the result, and gives feedback to the user. Graphical and textual guidance is displayed for the user to correct his/her actions. In particular the system features the following atmospheres: emotion, strength, weakness, tension/excitement, closeness/intimacy, loneliness and liberation.

The affective display uses icons to show the user which of the atmospheres are currently represented in the taken video. Furthermore, a so-called affective navigation system can guide the user to reach a preferred atmosphere.

2.2 Video Stabilization
Stabilizing shaky camera movement is one of the key factors for high quality mobile video. The majority of video stabilization algorithms follow a common scheme described by Morimoto and Chellappa [23] in 1997, which involves three steps: motion estimation, motion compensation, and image composition. Video stabilization methods can be roughly divided into 2D and 3D techniques. 2D video stabilization methods such as addressed in [6, 15–18] use 2D transformations to describe the movement between two video frames. 3D video stabilization methods as described in [4, 5, 14] estimate 3D transformations between consecutive frames usually by employing a structure-from-motion algorithm to reconstruct the camera motion. Common drawbacks of the more complex 3D approaches are higher computing times and oftentimes less robust results compared to 2D methods. We are therefore concentrating on 2D approaches to be able to apply the algorithms also on mobile platforms.

2.3 Commercial systems
Most of the smartphones currently available come with built-in video applications. This is true for the Google Android as well as the Apple iOS eco systems. Although these applications enable users to record videos in various formats, there is little assistance regarding shot-size, scene composition and story boards. For these needs few special applications exist, such as Apple iMovie 1, YouTube Director for Business 2 and Cyberlink PowerDirector 3. These applications provide users with predefined story boards as well as support regarding scene composition and shot length. Features and functionality vary heavily between applications, we therefore conducted an in-depth analysis of existing applications (See Chapter 3.2 Human-Centered Design for details).

3 RESEARCH METHODOLOGY
3.1 Overview of Concept
The overall structure of the proposed intelligent film assistant system is shown in Figure 1. The goal is to support the user in the acquisition of video material that meets the standards established by experts in the field. Rules on scene composition established by film directors may, for example, be used to derive recommendations on the suitable positioning of a story’s main character or product to be highlighted inside the scene. Such stage directions are provided to the user in the form of story boards that are tailored to particular use cases (e.g., video of a wedding ceremony, or marketing video) and guide the user through the video recording process. At the same time, low-level video analysis algorithms are used to extract relevant information on camera motion and scene dynamics or

1Apple iMovie https://www.apple.com/imovie/
2YouTube Director for Business https://www.youtube.com/yt/advertise/de/make-video-ads.html
3Cyberlink PowerDirector https://www.cyberlink.com/stat/product/CyberLink_app/PowerDirector-mobile/enu/PowerDirector-mobile.jsp
One of the main challenges of all the applications was the small smartphone display and therefore missing overview. Since even a small film project might consist of more than 10 or even 20 scenes, keeping the user informed about the overall progress and status has to be a main goal. A second important decision has to be made regarding the flexibility of storyboards. Streamlined story boards with little possibilities for variation are easy to begin with but tend to become unattractive on the long run due to repetition and narrow usage scenarios. Story boards which allow great variation of scene composition, shot length and layout can be used in a much wider range of scenarios but are hard to begin with and also complicated to use on small smartphone displays.

We combined these and all the other findings from our evaluations as well as our expert reviews into an internal interaction design guideline for our application. One of the most important design decisions we took was that we want to support both, novice as well as frequent users. Our application therefore needs to come with a range of simple story boards for beginners but also allows customization of existing story boards in a second step for advanced users. An additional “free mode” allows the creation of personal storyboards and storyboard templates. Our internal guideline in combination with the official user experience guidelines of Apple iOS [8] and Google Android [9] help us design our application prototypes as close to our users’ needs as possible.

3.3 Evaluation and Test Data
For evaluation purposes we have built an extensive video dataset. The dataset is a combination of multiple sources. 1) Videos from other publications in the field of video stabilization are used to evaluate our algorithms, and compare with others. 2) To test the algorithms on realistic examples for our application, we are recording videos with consumer smartphones. 3) A virtual platform is used to stage situations and obtain ground truth data for camera movement. In particular we use a similar method as Richter et al. [26]. We use the video game Grand Theft Auto V (GTA5) to collect video and meta data [8]. A video recording pipeline enables us to grab the frame buffer during execution, and further process the render target contents with post-processing shaders. Ultimately we transfer the post-processed images to host memory and store them into a video file. Compared to [26] our approach does not require access to single render calls, and can therefore run without a slowing down graphics debugger, which was noted to take about 30 seconds per frame of processing time. Our approach is able to record colour and depth images, as illustrated in Figure 2. Finally, resulting movement of the rendering camera can be stored along with the video to obtain ground truth for camera trajectories.

4 CURRENT STAGE OF THE PROJECT
4.1 Design Aspects
Based on our set of user experience guidelines we are currently developing several mid- to high fidelity user interface (UI) prototypes in fast iterations [11, 19, 20]. They are being evaluated regularly by a small number of users (three to five users) [24]. We are testing
4.2 Development of Algorithms

Regarding the algorithmic aspects of the project, we are focusing on video stabilization and shot-size detection. Currently all computations for video stabilization and shot-size detection are performed offline. For the final application tradeoffs must be found to run them in appropriate runtimes on the mobile platform itself. To maintain high framerates and reduce delay on computation results, our focus is on algorithms with a low computational complexity. In the domain of video stabilization we therefore prefer 2D video stabilization over more time-consuming methods based on structure-from-motion that work in 3D space. Although 3D methods would be beneficial for cinematographic camera path planning, their runtime is out of scope for current mobile devices. For initial experiments we have chosen to implement a motion estimator based on Harris feature detector and homography estimation. The resulting camera trajectories are then low-pass filtered, and video frames are warped according to the smoothed trajectory. In addition we have implemented a straightforward approach based on Phase Correlation for successive image pairs. Although the latter algorithm is only able to produce a translational image registration, we found it to be a good enough approximation when only little camera motion is present. An advantage of Phase Correlation is its efficient implementation on parallel computing platforms like graphics processing units. A further related approach we have taken for image registration is Fourier-Mellin transformation, which extends the Phase Correlation algorithm to also compute a scaling and rotational offset between image pairs. Currently we are working on the completion of Subspace Video Stabilization [15] and Content Preserving Warps [14] with a view to future scaling to mobile devices.

Our mobile application is heavily based on storyboards. For each scene in the storyboard the user is supposed to take a video shot. To introduce diversity into the shots, even if the scene is quite simple, changes in camera position and distance should happen frequently. Each scene in the storyboard typically has a shot-size (e.g. close-up) attached to it. While taking a shot, the app communicates to the user how the scene is supposed to look like, and what shot-size should be chosen. Since that is a non-trivial task for a novice user, an automatic detection of the shot-size should assist the user to position the camera according to cinematographic recommendations. To determine the shot-size, we are employing state of the art (multi-)person keypoint detection algorithms [2, 25]. For our current experiments we have chosen OpenPose [2] for its performance, accuracy and ease of use [12]. From an RGB input video the OpenPose algorithm infers a descriptor with 18 keypoints per person. Each descriptor is then classified based on the visibility of joints, to determine the shot-size of the scene. We differentiate between four shot-sizes as displayed in Figure 4. We are currently computing the shot-size on a per person level in each frame, but ultimately the shot-size should be determined on a per frame level instead, by identifying the most dominantly displayed person per frame. Figure 5 depicts the output of OpenPose next to our derived shot-size classification.

5 CONTRIBUTIONS AND OUTLOOK

The described project comprises contributions to Mobile Interaction and Mobile Multimedia, which are two of the main topics supported by MoMM2017. In the context of Mobile Interaction, we deal with the subtopic Mobile user interfaces and interaction techniques. Suitable user interfaces have been developed in order to present the developed guidelines for scene composition and camera movement on the limited space of a mobile display in a compact and intuitive way, without distracting the user from his/her main task (i.e., the video recording). The project’s focus on human-centered design
entails the collection of user requirements and evaluations by different user groups, including amateur users and film professionals, at various stages of the project, it thus also contributes to the subtopic Evaluation and usability of mobile devices and services. Under the second main topic Mobile Multimedia, our development of video analysis algorithms for, e.g., motion estimation, video stabilization and shot-size detection, falls into the field of Audio and video analysis, modeling, processing and transformation. Besides that, the envisioned support system for the creation of high-quality video content on mobile devices also includes aspects of Interfaces for multimedia creation.

In a next step, we plan to refine the low-level algorithms as well as the high-level assistant tools to further improve the quality and capabilities of our mobile film assistant. Especially on the capturing side, where user assistance typically requires real-time and low-latency feedback and errors during acquisition can hardly be corrected, additional or refined techniques to guide the user shall be developed and evaluated. We will continue to perform frequent iterative evaluations of the user behavior and quantitative analysis of the user experience to guarantee real-world applicability and tight connection between low-level algorithmic results and high-level goals. In this context, we also plan to narrow down the use cases in the mobile application and the corresponding storyboards to specific target user groups (e.g. business vs. consumer, short vs. mid/long story narrative) to allow a more detailed analysis of use case and user behavior as well as the identification of specific user requirements.

ACKNOWLEDGMENTS

The work is performed under the project "Intelligentes Assistenzsystem zur Videoerstellung auf Mobilgeräten (PersonalFilmAssistant)", which is supported by Vienna Business Agency under the Program "Users in Focus" (ID 1571390).

REFERENCES


