ActionTrip: Automating Egocentric Video Production through Gestural Interaction and Landmark Awareness

Ioannis Deliyannis[†]
Department of Audiovisual Arts
inArts Research Laboratory
Ionian University
Corfu Greece
yiannis@ionio.gr

Sifis Symianakis
School of Science and Technology
Hellenic Open University
Patras Greece
sifis.symianakis@gmail.com

Konstantinos Chorianopoulos
Department of Informatics
Ionian University
Corfu Greece
choko@ionio.gr

ABSTRACT

The proliferation of smartphones with video recording and action cameras has empowered users to effortlessly capture everyday moments and activities in video format. This process often results into long video sequences that need to be edited down and summarized in order to be properly archived, shared within social media, or simply projected to family and friends. For those who do not possess video-editing skills, isolating the most appropriate moments is a difficult and time-consuming task. This work presents "ActionTrip", a system that provides gestural user interaction during video-capture, in order to seamlessly tag personal points of interest. In addition, the system captures contextual information about nearby landmarks and information from social networking systems. This combination of personal with social preferences has enabled a single-step final editing process, which presents to the user a small list of points interest (public landmarks and personal preferences), instead of the actual recorded video that might be very long to comprehend. The benefits of this system have been demonstrated through a small-scale case study during the visit and tour in a European city.

CCS CONCEPTS

•H.5.1 Multimedia Information Systems (Video, Evaluation, Navigation, Maps) •H.5.2 User Interfaces (User centered design, Interaction Styles)

KEYWORDS

Interactive Video Capture, Location-based Metadata, Automatic Video Editing, Location Aware Content, Gestural Interaction, Automatic Video Summarization, Action Camera.

ACM Reference format:

Ioannis Deliyannis, Sifis Symianakis and Konstantinos Chorianopoulos. 2018. Actiontrip: Automating Egocentric Video Production through Gestural Interaction and Landmark Awareness. *In Proceedings of 22nd Panhellenic Conference on Informatics (PCI'18), ACM, New York, NY, USA, 6 pages*, https://doi.org/10.1145/3291533.3291585

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

PCI '18, November 29-December 1, 2018, Athens, Greece
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6610-6/18/11...$15.00
https://doi.org/10.1145/3291533.3291585

1 Introduction

The research-based system developed in this work lies within the areas of video summarization and video-based life logging. We approach the process from the user and content perspectives, as they are intrinsically connected, with the main strategy following the main rule: collect now, use later. Therefore for a visitor that records while navigating within a city, the actual movement paths are also recorded as the user films through various locations. Although user interaction is supported, it is not necessary for the completion of the process, mainly for safety reasons. As such, we have developed a user interface enabling users to declare their preference or interest, using the touch screen of their mobile device, through a seamless process that requires little or no attention. Comparing the user trajectory with locative media applications, and user-sourced feedback, the system simplifies the editing process. The final video-editing procedure automatically detects and identifies content of importance by collecting and evaluating location-based interest expressed by other visitors on locative media, such as "*Foursquare*".

2 Video Summarization and Life Logging

In the not so distant past, moments of people's lives were filmed using traditional video cameras and those videos were rarely edited professionally [6]. This has changed today as the availability of video capture devices is such that users continuously carry on and are used to record a large number of everyday moments, even realising traditional tasks spiced up with new technologies, such as the development of video diaries [7]. In terms of new uses, social software often prompts users to capture and create videos that in turn are used to create cross-user interactions within each network [16]. Online or offline video editing tools are employed select the preferred video sections and produce short summary videos [9]. Since this process often seems complicated and requires users to acquire appropriate knowledge and commit sufficient time to learn how to edit and produce quality content, video editing tools have attempted to simplify the editing process. The major editing suites focusing on inexperienced users, including Windows Movie Maker and Apple's iMovie, follow a specific workflow process. Although simplified, these editing suites often suffer from the above deficiencies, where users have to scan through their videos and no contextual information is added in the video to aid in the editing process.

3 Commercial Applications and Research

The automated video summarization research area is a highly active domain involving interdisciplinary approaches. These include the development of context-aware systems [21], user profiling [19], video analysis to identify recognisable objects and keyframes [10], spatial and temporal analyses [4], multi-view content evaluation [3] and user attention [8]. New systems emerge combining human memory and content [20]. Existing systems focus on video processing using algorithms based on both the video content analysis and exploit contextual information [5]. The more data from the context of use collect the more flexible the processing of the final video [1]. Xu et al [18] suggest the use of external video frame information for a summary of football whose duration is usually large, and few places really are the best scenes, or otherwise «highlights» of the match. Connecting each highlight scene with specific keywords (tags), a second time distinguish scenes and include the best ones in the final summary. Efforts such as Peng et al [13] focus on user reactions when watching a video and produce a summary of what the algorithm evaluates as the user's best moments. "Video summagator" [11] provides an alternative approach by providing a volume-based video representation by combining time and space information. In this representation, the user views all the scenes simultaneously using a 3D representation as an overview, enabling selection and scene isolation.

According to Abigail J. Sellen and Steve Whittaker [15], within video-based life logging the key insights are summarized by the following points: "*focusing on "total capture," current approaches to lifelogging have failed to explore what practical purpose such exhaustive personal digital records might actually serve. Evaluating new approaches, psychology has emerged as an underexploited resource in defining the nature of human memory and its key processes and weaknesses. Psychology as design framework could help define the types of memory such systems should support, along with their key interface properties and need to work in synergy with human memory, rather than as its replacement.*"

Considerable research is traced on life-logging systems [17] featuring a plurality of sensors and the collection of context for the identification of individual video scenes. Hori et al [5] implemented a system combining a wide range of sensors including "*brain-wave analyzers, GPS, acceleration sensors, and gyro sensors*" to name a few, enabling users to retrieve specific scenes by asking questions relating to the context.

Our approach targets the problem by placing the context-based perspective on the interaction forefront, while other systems employ a content-based strategy. This is clearly an informed choice made, justified as we will see further by various reasons. As such, reaching a state that requires limited user attention is achieved, while at any point during the trip the user is able to assume full control of the system if that is either desired or demanded.

4 System Design and Development

The "*ActionTrip*" system has been developed to address the problem of video summarization.

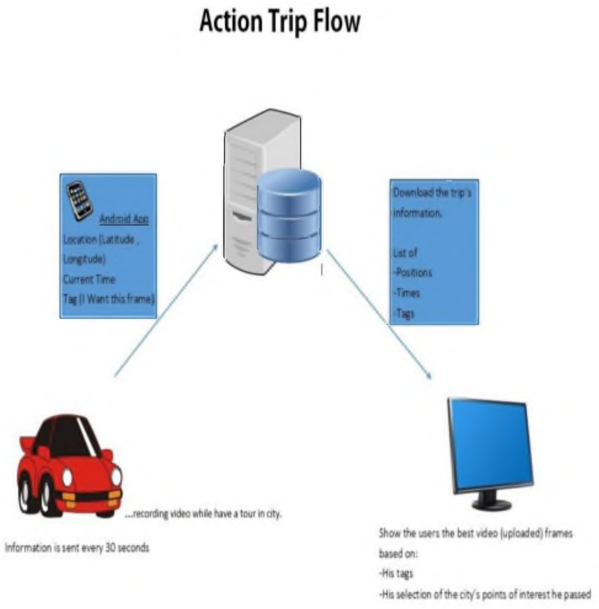

Action Trip Flow
Android App
Location (Latitude , Longitude)
Current Time
Tag (I Want this frame)
Download the trip's information.
List of
-Positions
-Times
-Tags
....recording video while have a tour in city.
Information is sent every 30 seconds
Show the users the best video (uploaded) frames based on:
-His tags
-His selection of the city's points of interest he passed

Figure 1: ActionTrip Data Content & Flow

It consists of three information-exchanging subsystems where a mobile-based interface, a database and a web-based interface are specifically designed to allow users to *capture video and contextual information on the move,* either using vehicles or on foot are combined.

Attributing distinct objectives and roles for each individual subsystem enabled the setup of a clear and functional workflow. Figure 2 displays the system components and linking.

4.1 Server

The main data flow control element is a server responsible for storing information and its on-demand recovery. The data is stored in a remote database based on Mongo, a technology that supports large volumes of multimedia data stored in Quick Time format. The database and the server support encryption and user-access through certification validation, ensuring content safety.

The mobile application subsystem operates during the tour and it stores the user's path including the duration that the user spends at each location. Capturing location-based and user-generated contextual information during filming was enabled via utilization of the mobile a phone's sensors and the touch screen interface.

The data are stored on the server during the trip. The user is able through the mobile interface to select or in the term of the system "*TAG*" a route that is interesting. Since the majority of users are in motion, the interface features a simplistic interaction design mechanism in order to minimize distraction. The monochrome background of the interface changes colour depending on the application's operation and notifies the user at a glance about its status. Interaction design choices and gestures are described within the user interaction section below.

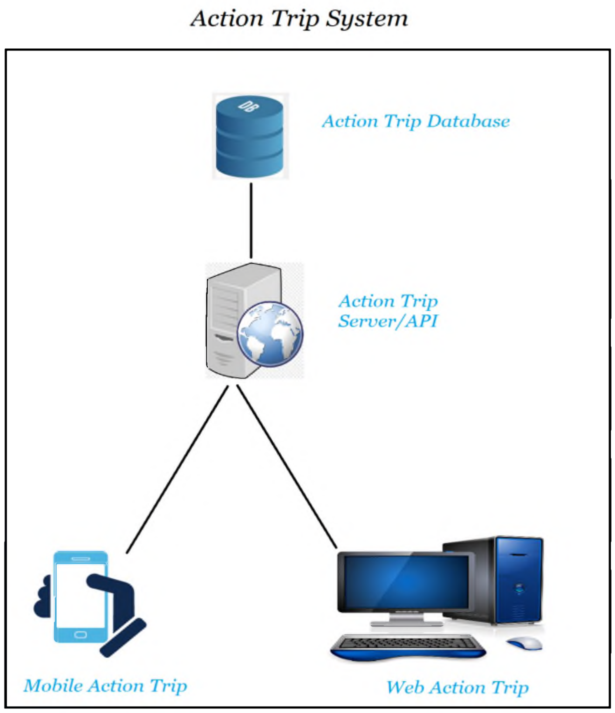
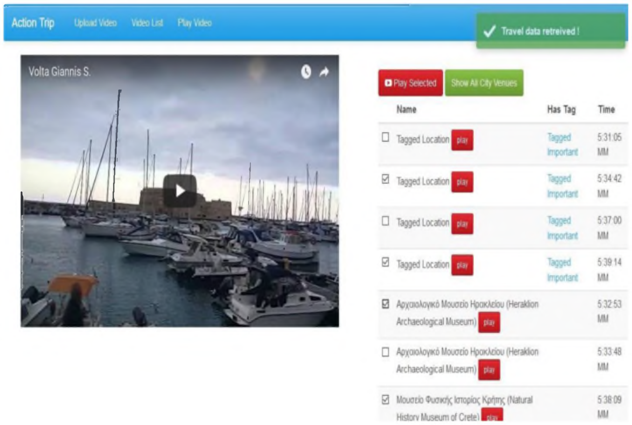


Figure 2: ActionTrip Application Architecture

4.2 Mobile Application

The web application module addresses the video summarisation, editing and final production processes. Following the trends in video editing it is implemented on web-based basis and in order to enable content privacy, it requires user login using a Gmail account in order to allow access to the video content. At this prototype level, the social software environment *Foursquare* is utilised to source location-based information regarding the areas of interest and the attractions, as it compares the navigation path recorded to socially-sourced information captured by other users. Figure 3 displays the web-based user interface where video information classified as "*important*" is placed at the top of the list, followed by captured information of less importance.



	Name	Has Tag	Time
□	Tagged Location play	Tagged Important	5:31:05 MM
☑	Tagged Location play	Tagged Important	5:34:42 MM
□	Tagged Location play	Tagged Important	5:37:00 MM
☑	Tagged Location play	Tagged Important	5:39:14 MM
☑	Αρχαιολογικό Μουσείο Ηρακλείου (Heraklion Archaeological Museum) play		5:32:53 MM
□	Αρχαιολογικό Μουσείο Ηρακλείου (Heraklion Archaeological Museum) play		5:33:48 MM
☑	Μουσείο Φυσικής Ιστορίας Κρήτης (Natural History Museum of Crete) play		5:38:09 MM

Figure 3: Web-based Clip Organisation and Classification of Importance

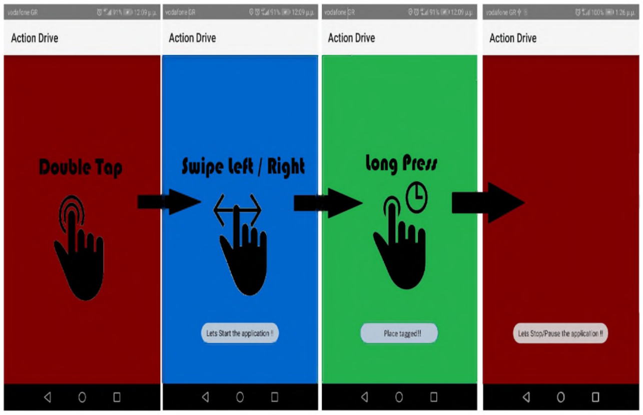
Content/context analysis reveals clips of interest that are identified and assume a key role into the content selection/editing process. The density of information is used to identify particular locations of interest, in conjunction with user comments and ratings. The algorithm employs a mechanism that synchronises the captured content and data with frame-based precision. This process enables the editing routines to isolate video-sections, which contain a number or more of points of interest, as the public marked them. In addition, the algorithm includes the sections where the user declared interest through the user interface.

5 Capturing Video and Location-based Content

The modular design of the end-system enables various device types to be employed for video recording. Integrated mobile-device cameras, traditional digital video and action cameras can be used, providing flexibility and economy to the end-user who can utilise their existing equipment. Further to the captured video stream, the mobile-application dynamically collects information including the navigation path and user interactions, permitting the web-based system to discover areas of interest using location-based social application posts by other users. When the video capture task is completed, the automated web-based video processing process commences. Ultimately, frame data recorded during filming are combined with user and location-driven information in order to create a video summary featuring the most interesting points within this route.

5.1 Seamless User Interaction

Within the interaction aspects of the process, we know that the majority of mobile applications do not take into the account the state of attention of the user during operation, often linked to their kinetic state [2]. Often user attention and public safety are compromised, as they cannot cope with a complex interaction demanded by the application, focus on its use and loses concentration on other more important tasks. Systematic studies may be referenced [12] which highlight the design factors that are not taken into consideration within interaction design. In this work we isolated the factors that introduce what we have termed "*annoying user interfaces*" that demand user attention at states where they are unable to multitask. It is clear that there is a vital need to design appropriate interfaces that respect the state of the user and minimize distraction [2]. The principal application characteristics and the user-interfacing functionality were both decided during the design-phase of the system. From the user perspective the target-application should enable users the freedom to control it whenever possible and with minimum interaction requirements in order to minimise memory load and avoid the loss of attention from complex tasks such as driving. The learning curve of the complete end-to-end process should be short while the worst-case scenario includes the condition that the application should be able to function without user interaction. The user interface instance displayed here uses red to indicate that the system is recording information and does not provide any other information in order not to attract user attention.



Lets Start the application !!

Place tagged!!

Lets Stop/Pause the application !!

Figure 4: User Interface – Application states by gesture input

Hence the end-system design focuses on minimizing distractions during user browsing. First the user environment was examined in order to design a passive-interactive process that furnishes the system with user feedback while it does not require or dynamically attract user attention. The prototype developed utilises a simple user interaction mechanism employing gestures on the touch screen while the interface outputs are multimodal using related sounds, vibration and short messages for potential passengers. These gestures include: dragging the finger towards a certain screen direction, single tap on the screen, double tab on the screen and the use of two fingers moving in the opposite direction. The decision to use gestures was made in order to avoid diverting the user's attention from the road while driving. The negative characteristic of this approach is that the driver will have to memorize the gestures to be made for each function, something that as we have seen from other applications requires a short learning curve.

6 Content Summarization and Editing

Information editing is a simple and straightforward process, involving minimal information exchange. In fact the system is able to produce a solution automatically, without user intervention, in order to speed up the process. Then the user may inspect the end-video and be able to tweak the settings in case they wish to include or exclude particular content.

From the information-domain perspective, the application is programmed to collect as much data as possible, in order to furnish the video timeline with additional information regarding the navigational path, locations of interest, particular user interests and also the public's interest in those locations.

Clearly, all the above interdisciplinary prerequisites demand advanced interaction design techniques to be developed for the user interface, in order to minimise information loss, ensure user safety and enable high system usability.

Within video editing, the system displays the city's points of interest that the user has visited. Those data derive from a social networking system, which provides the points of interest based on the GPS coordinates captured by the mobile device. Four well-known social networking systems that served our purpose were compared in our attempt to derive useful information: Foursquare, Google Places, Flickr and Panoramio. The comparison was made using the following criteria:

• It should cover the working hypothesis, and function works in cities

• Provide crowd city POI information

• The points of interest should be linked to coordinates

• The content should be accessible without restrictions

• The data should be validated

• It should provide the data filtering capabilities

The "*Foursquare*" system met all the above criteria and it was selected for experimentation and in Figure 6 it displays the information gathered and the points of interest during one of the test runs.

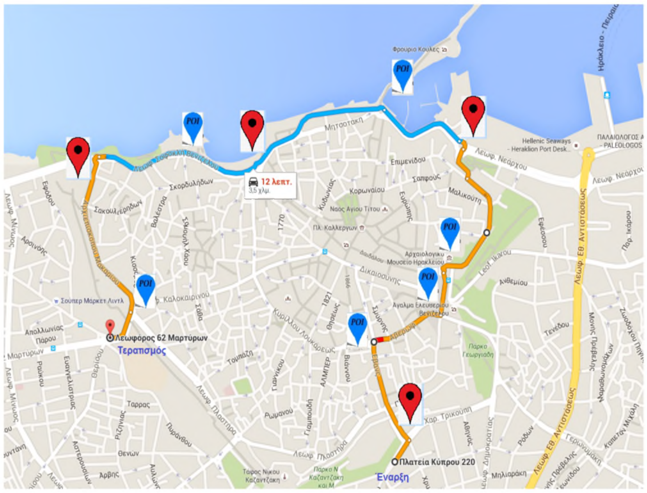


Figure 5: Points of Interest on the Map

It is possible to employ all other social networking systems in order to source additional content and context, a process that is planned for future system versions. Location-based social networking systems including "*Foursquare*" and "*Google Places*", have developed algorithms for capturing POI GPS-based snapshots. The algorithm matches site-based information with user preferences. Some systems [14] actively employ this feature and combine the above information in practice. They think that the greater the information of a number of the more valid is the result. The Zhang et al [22] present a video summary of the best user-selected routes based on location popularity, enabling a camera-based smart device to record the video enrich it with GPS information of the geographic coordinates. The system aims to be targeted by tourists who visit places of interest and proposes scenes based on popularity.

6.1 User Testing and Evaluation

Usability of the functional prototype was evaluated by a group of six users aged between 27 and 42 year old, all familiar with mobile phone technologies, with variable editing experience on video editing and summarization as it is displayed in Table 1. We developed a specific use case that could take place in controlled circumstances and decided for the participants to follow a specific route driving through their own city. Their test case study involved driving and capturing video using a car-mounted camera through a specific driving path across Heraklion city in Crete, Greece. The marked route included six points of interest and four points that they had to mark as liked by putting a "Tag" on the scene, producing a video containing the tour with a duration of about 12 minutes. Each participant received a short introduction of all the functionality features of the application. All the tours were realised via the use of their personal vehicles.

Participant	Techno-logical Level	Video-editing Experience	Time Movie Maker	Time ActionTrip
Giannis Z.	Excellent	Good	6'21"	1'59"
Nikoleta M.	Interm.	Poor	7'12"	2'28"
Nikos Thr.	Good	Good	5'36"	1'38"
Giannis S.	Good	Poor	6'22"	1'29"
Pavlos D.	Good	Average	5'48"	1'38"
Haris M.	Excellent	Average	6'02"	1'48"

Table 1: Test-group Members and their Experience

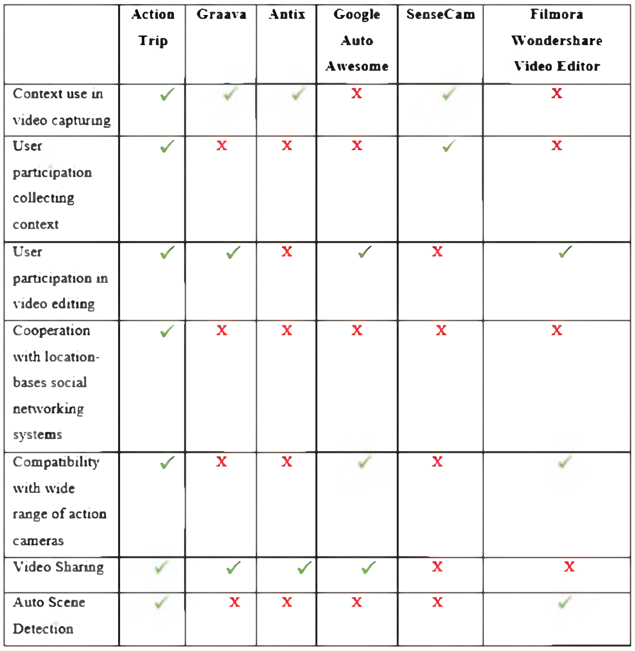
Successful video recording and context collection was followed by the video-editing session, where participants were asked to edit the video both manually and through the application of automated video summarization. Participants completed editing procedure using both a traditional editing software (Movie Maker) and the ActionTrip web application prototype for the editing process. Regardless the editing experience of each participant, they received basic training on the video editing software in use so that a smooth operation and comparison between the two systems could be achieved. Their manual video editing end-result should including three points of interest and three marked points which they marked as "liked" (out of the 10 recorded). Comparison was based on the required time to fulfil the procedure and the usability of each system.

Evaluation of the results showed that users faced no difficulties using ActionTrip. Mobile application didn't distract their attention while driving and recording as they had already memorized the required gestures to use as inputs, a process that did not require them to look at the interface, enabling them to concentrate on driving and the video-capture task in hand, enabling us to claim that ActionTrip may be categorized as a system that features minimum distraction characteristics. The summarization process indicated that editing using ActionTrip required less time compared with the traditional editing process. Users preferred the proposed system and stated that they would use in their everyday life. To their view it is described it as a "*simplified and usable application with clear functions*". Another interesting note is that all the participants are residents in the town that the evaluation took place. So, we can safely assume that for a visitor there can be significant time saving gain as they are not able to identify the areas of interest with ease and they are required to review the complete content again before editing.

7 Discussion and Conclusion

On the commercial level, video-editing software evolves and new software products employ different strategies designed to suggest to the user new approaches in order to perform video editing. The "*Antix*" app is a mobile application that evaluates video and sensor-based data in order to suggest the best video moments while the "*Graava*" application when accompanied by a sensor-enabled action camera produces a fully automated video summary. The "*Sense Cam*", on the other hand, is a life-logging technology with a variety of sensors to retrieve selected images and scenes while "*Filmora*" is a video editor that supports automatic detection and separation of scenes. Google *"Photos Assistant"* produces a movie based on selected pictures and videos located in user's device, but it doesn't support the contextual input during recording or the real-time user interactivity. It also lacks the summarization functionality based on context while the produced video contains the whole selected videos' duration.

The "ActionTrip" system presented in this work provides a complete and integrated approach for the evaluation of video summarization combined with life logging technologies and techniques. Our further work targets the information domain area where additional underlying context databases may be employed to enrich further the precision of the system, enabling each user to identify different domains of interest that match their personal requirements. Finally, user interaction might become more seamless through wearable devices, such as the smart watch.



	Action Trip	Graava	Antix	Google Auto Awesome	SenseCam	Filmora Wondershare Video Editor
Context use in video capturing	✓	✓	✓	✗	✓	✗
User participation collecting context	✓	✗	✗	✗	✓	✗
User participation in video editing	✓	✓	✗	✓	✗	✓
Cooperation with location-bases social networking systems	✓	✗	✗	✗	✗	✗
Compatibility with wide range of action cameras	✓	✗	✗	✓	✗	✓
Video Sharing	✓	✓	✓	✓	✗	✗
Auto Scene Detection	✓	✗	✗	✗	✗	✓

Table 2: Comparison with Commercial Applications

REFERENCES

[1] DE SILVA, G.C., YAMASAKI, T., and AIZAWA, K., 2005. Evaluation of video summarization for a large number of cameras in ubiquitous home. In *Proceedings of the 13th annual ACM international conference on Multimedia* ACM, 820-828.

[2] ENDSLEY, M.R., 2016. *Designing for situation awareness: An approach to user-centered design*. CRC press.

[3] FU, Y., GUO, Y., ZHU, Y., LIU, F., SONG, C., and ZHOU, Z.-H., 2010. Multi-view video summarization. *IEEE Transactions on Multimedia 12*, 7, 717-729.

[4] GUO, Z., GAO, L., ZHEN, X., ZOU, F., SHEN, F., and ZHENG, K., 2016. Spatial and Temporal Scoring for Egocentric Video Summarization. *Neurocomputing*.

[5] HORI, T. and AIZAWA, K., 2003. Context-based video retrieval system for the life-log applications. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval* ACM, 31-38.

[6] KIRK, D., SELLEN, A., HARPER, R., and WOOD, K., 2007. Understanding videowork. In *Proceedings of the SIGCHI conference on Human factors in computing systems* ACM, 61-70.

[7] LEE, H., SMEATON, A.F., O'CONNOR, N.E., JONES, G., BLIGHE, M., BYRNE, D., DOHERTY, A., and GURRIN, C., 2008. Constructing a SenseCam visual diary as a media process. *Multimedia Systems 14*, 6, 341-349.

[8] MA, Y.-F., HUA, X.-S., LU, L., and ZHANG, H.-J., 2005. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia* 7, 5, 907-919.

[9] MARQUES, O., 2016. Image and Video Everywhere! In *Innovative Technologies in Everyday Life* Springer, 45-58.

[10] MENG, J., WANG, H., YUAN, J., and TAN, Y.-P., 2016. From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1039-1048.

[11] NGUYEN, C., NIU, Y., and LIU, F., 2012. Video summagator: an interface for video summarization and navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* ACM, 647-650.

[12] OVIEDO-TRESPALACIOS, O., HAQUE, M.M., KING, M., and WASHINGTON, S., 2016. Understanding the impacts of mobile phone distraction on driving performance: A systematic review. *Transportation Research Part C: Emerging Technologies 72*, 360-380.

[13] PENG, W.-T., CHU, W.-T., CHANG, C.-H., CHOU, C.-N., HUANG, W.-J., CHANG, W.-Y., and HUNG, Y.-P., 2011. Editing by viewing: automatic home video summarization by viewing behavior analysis. *IEEE Transactions on Multimedia 13*, 3, 539-550.

[14] POSTEL, M., 2013. Point-of-interest recommendation in location based social networks with topic and location awareness.

[15] SELLEN, A.J. and WHITTAKER, S., 2010. Beyond total capture: a constructive critique of lifelogging. *Communications of the ACM 53*, 5, 70-77.

[16] WAKEFIELD, R. and WAKEFIELD, K., 2016. Social media network behavior: A study of user passion and affect. *The Journal of Strategic Information Systems 25*, 2, 140-156.

[17] WHITTAKER, S., KALNIKAITĖ, V., PETRELLI, D., SELLEN, A., VILLAR, N., BERGMAN, O., CLOUGH, P., and BROCKMEIER, J., 2012. Socio-technical lifelogging: Deriving design principles for a future proof digital past. *Human–Computer Interaction 27*, 1-2, 37-62.

[18] XU, C., WANG, J., WAN, K., LI, Y., and DUAN, L., 2006. Live sports event detection based on broadcast video and web-casting text. In *Proceedings of the 14th ACM international conference on multimedia* ACM, 221-230.

[19] YIN, Y., THAPLIYA, R., and ZIMMERMANN, R., 2016. Encoded Semantic Tree for Automatic User Profiling Applied to Personalized Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology*.

[20] ZHANG, K., CHAO, W.-L., SHA, F., and GRAUMAN, K., 2016. Video Summarization with Long Short-term Memory. *arXiv preprint arXiv:1605.08110*.

[21] ZHANG, S., ZHU, Y., and ROY-CHOWDHURY, A.K., 2016. Context-Aware Surveillance Video Summarization. *IEEE Transactions on Image Processing 25*, 11, 5469-5478.

[22] ZHANG, Y., MA, H., and ZIMMERMANN, R., 2013. Dynamic multi-video summarization of sensor-rich videos in geo-space. In *International Conference on Multimedia Modeling* Springer, 380-390.