Virtual Cinematography and Tracking Heuristics for Presenter Tracking: A Literature Review

Mohamed Tanweer Khatieb
Department of Computer Science
University of Cape Town
Rondebosch, 7701, South Africa
khtmoh003@myuct.ac.za

ABSTRACT

The rise in technology has brought with it the increase in camera quality. The latest camera resolution has become known as 4K and is twice as big and detailed as a 1920x1080 High Definition resolution. The University of Cape Town is currently considering using a wide-angled 4K camera to capture lectures where everything of the presentation space is always in view. This literature review details what post-production operations need to be applied to the recording to make a smaller sized video of a lesser resolution using professional cameraman tactics to produce a smooth video which is pleasant to view without jumping around too fast or losing track of the presenter.

CCS Concepts

• **Computing methodologies →Computer vision problems**

• **Computing methodologies →Video segmentation**

• **Computing methodologies → Tracking**

Keywords

Virtual Cinematographer; 4K video; Panning; Scope; Tracking

1. INTRODUCTION

The production of high quality cameras and recording equipment becomes cheaper and easier as technology progresses [3]. Since it is cheap and easy to acquire high definition cameras and recording equipment it is useful to implement a system in universities which records the lectures in each venue so students are able to consult the information as originally presented at a later date[1]. It is also useful for those who were unable to attend the lecture as a method of making up for lost time [5].

The recordings alone are insufficient for users, however, since the camera is either static or does not behave in a smooth and eye-pleasing manner [3, 5]. It is also expensive and impractical to consider a dedicated filming crew for venues[15].

The concept of Virtual Cinematography (VC), also known as Virtual Videography and Computational Cinematography [3, 5], is the process of taking the video stream and making a new stream from the existing frames of a smaller size and lesser quality or by making the camera operate in such a way (through physical movements) that the viewer believes a person is manning it [2-6, 12] such as a Pan Tilt Zoom camera (PTZ camera).

In order to achieve camera tracking of a similar quality to a professional camera crew, there needs to be a set of rules or heuristics in place to guide the camera when changes occur or the camera encounters a certain problem (e.g. Lecturer position lost or inability to discriminate between multiple lecturers). These heuristics allow the camera to pan across the scene (i.e. lecture venue), zoom in towards the presenter or zoom out in order to gain a better idea of what the lecturer is referring to [5].

The Centre for Innovation in Learning and Teaching (CILT) is working in conjunction with the University of Cape Town (UCT) to create (and improve on) a lecture tracking and recording system which is easily accessible for users and which makes viewing these recording as user friendly as possible.

CILT has recently decided to make use of ultra-high definition 3840 x 2160 (4K) resolution cameras with a wide-angled view of the lecture venue. The goal is to use the large angle of the camera to get all the boards and the lecturer in the frame at all times. The problem with this approach is that the resultant video is far too large to stream and download. In order to alleviate this issue CILT proposes a method where a smaller sized frame is sampled from the large frame with a 720p resolution.

This literature review serves as a tool for comparing the existing implementations of VC in lecture tracking and to find the best approach for solving the problem proposed by CILT.

2. VIRTUAL CINEMATOGRAPHER

The production of recorded videos that track the presenter (lecturer) requires three main steps: input processing to make sense of the environment and to note any changes to that environment, identifying and keeping track of the lecturer and differentiating the lecturer from any other individuals who enter the field of view and the VC for making the “camera” (i.e. the smaller sampled frame) move around on the bigger frame in a way that resembles the movements of a professional camera crew.

In order to achieve this level of professionalism many research papers have developed guidelines on how the camera should behave in various contexts by making use of heuristics [2-6, 15].

Odhabi et al [13] suggest making use of a special set of eyeglasses that records the students during a lecture to monitor their participation and the level of interest shown in the lecture.

Nagai et al [12] suggest using a high-definition camera to record the entire lecture venue without any VC to sample to frames into a smaller resolution video.

Burelli [2] discusses the need for VC in games in order to improve the overall visual quality of the game for the users. The paper discusses camera configurations and heuristics involved in creating cinematic displays and the player experience from playing the cinematic style of the game.

Jones [6] mentions using VC in movies with a lot of Computer Generated Imagery (CGI) to create more realistic and pleasing camera shots for improving the overall video quality.

Lai et al [7] discuss an interactive storytelling system allowing interactive story designers to make use of VC to amplify the communication of the story.

Gleicher et al [4] suggest a system that can automatically create good quality video presentations out of an already recorded video. They mention the need for understanding what is happening in the video before any VC actions can be made. The VC needs to know where the presenter is in the environment and whether the current setup is good enough to display the presenter correctly. If it is not positioned correctly it needs to know what to do based on what it understands from the processing of the raw image data. This paper also makes accommodation for the use of special effects in the final video. It is intended as an additional option for shot creation and allows more possibilities. This paper has found that even with the simplest implementations for this problem there could potentially be useful outcomes but there is still a need for further research in many aspects of the problem.

Rui et al [15] explain their system as an automatic lecture tracking and recording system designed to relieve the necessity for a personal recording crew per lecture venue. The system described in this paper is capable of tracking the lecturer, the members in the audience, the slides used during a presentation, a zoomed out view of the whole scene (lecturer and presentation environment) as well as a Virtual Director (VD) with the ability to select a particular video form a collection of VC video cameras. The paper describes the four components that every VC requires: every VC system must make use of sensors to pick up the data from the world; cameras are required to capture the data for future use; there should be rules in place detailing how the camera should move inrelation to the context; their needs to be a communication between the VC and the VD (see Lampi et al [8] for more on VDs). This paper concludes by stating that their system works almost as well as a manned camera and that the future holds a lot of potential for making a system like the one proposed something to be installed in every lecture venue.

Gleicher et al [3] discuss an improvement on their earlier work , namely Gleicher et al [4], which allows lectures to be recorded inexpensively and without being intrusive on the lecture by necessitating a whole recording crew to be present at the time of the recording or a dedicated team to perform post-production on the captured lecture through manual manipulation. The paper mentions important heuristics required for VC: the system should attract the users' attention to the appropriate focal point (i.e. the lecturer), it must use space and time effectively by pacing the lecture in an artful way, the system needs to maintain a visual interest since the display size is limited and the viewer's attention span is not necessarily long and the system should communicate the message to the viewer in an easily understood and intuitive way. This paper finds that a survey is required in order to test the quality and value of the output produced by the proposed methods.

Heck et al. [5] build onto their earlier work in 2002 [3] creating a VC that processes video's in a post-production stage and adheres to the rules of cinematography mentioned in their earlier papers. The system uses stationary cameras to capture the videos, which is then processed by using computer vision and signal processing methods that gather information about the scene. A planning algorithm is responsible for choosing the best shot to use given the information from the scene and finally an image synthesis method creates new images as output from the original video. They identify four key points for creating realistic VC results that users appreciate. Syntactic cues (such as content written on the board or lecturer gestures and position) are used to assist in deciding how to virtually pan the camera. Offline (post-production) processing allows the ability to look ahead and plan the best shots. Constrained optimization over space and time (or objective function) is a technique that aims to optimize shots in a pleasing manner over the entire video instead of using a per frame approach. Finally, instead of moving a physical camera, an image synthesis algorithm is used to segment a view for output. They note that evaluating the technique is highly subjective, however, to overcome this challenge, they used undergraduate film students and an inexperienced video editor (because experienced cinematographers often break the rules) to manually edit a source video to meet a chosen style. The output of the VC is then compared on how well it can mimic the chosen editing style. The results showed that the VC produced very similar results to the film students and to further verify the results, 150 participants from various disciplines were surveyed and although the study was not qualitative, the results showed an overwhelmingly positive feedback. Processing a ten-minute video using the VC system took approximately one hour and fifteen minutes.

3. DISCUSSION

The project proposed by CILT requires the input to be processed in such a way that the lecturer is identified and any lighting issues are normalised; all the lecturer positions and orientations need to be stored; each frame needs to be reduced by sampling a smaller frame from it and the VC needs to decide which part of each frame to sample.

A lot of the aforementioned papers do not provide an implementation which is relevant to requirements declared by CILT [2, 6, 7, 9-14]. While they mention VC heuristics, they do not explain how to implement them.

The three papers written by Gleicher et al [3-5] come closest to the requirements of the CILT project and explain the VC heuristics well and the latest paper [5] explains them in such detail that it is possible to implement them despite the unavailability of the solution described in the paper.

The solution described by Rui et al [15] also offers at least some of the functionality required by CILT and makes heavy use of VC and the heuristics associated with it.

While both systems above are capable of tracking the lecturer and recording the lecture using VC they require manual guidance on where to find the board and they provide a way to separate the board and slide displays into separate videos feeds. CILT has an infrastructure in place that displays the projector and it is possible to remove the board into a separate feed

4. CONCLUSIONS

Since there was no uniform way to test all of these papers a proper comparison could not be drawn between the papers mentioned in this review. The papers also do not go into much detail about how to implement the solutions described in the papers and any implementation done for CILT is going to have to be made from scratch using the guidelines described in these papers.

The best solution so far is the solution proposed by Heck et al [5] since it covers all the details comprehensively and explains the VC heuristics in such detail that it would be easy to implement the heuristics from scratch.

In conclusion there is an opportunity for further research for a solution that meets the requirements proposed by CILT where some of the work can be borrowed from Heck et al [5] and the rest can build on top of the implementation described in that paper.

5. REFERENCES

[1] Al-Nashash, H. and Gunn, C. Lecture Capture in Engineering Classes: Bridging Gaps and Enhancing Learning. Educational Technology & Society, 16, 1 (January, 2013), 69-78.

[2] Burelli, P. Virtual cinematography in games: investigating the impact on player experience. Foundations of Digital Games, (2013).

[3] Gleicher, M. L., Heck, R. M. and Wallick, M. N. A framework for virtual videography. In Smartgraph '02 Proceedings of the 2nd international symposium on Smart graphics. (June 11, 2002). ACM, New York, 2002, 9-16.

[4] Gleicher, M. and Masanz, J. Towards virtual videography (poster session). In Multimedia Proceedings of the eighth ACM international conference on Multimedia. (October 30, 2000). ACM, New York, 2000, 375-378.

[5] Heck, R., Wallick, M. and Gleicher, M. Virtual videography. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 3, 1 (February 1, 2007), 4.

[6] Jones, N. Quantification and Substitution: The Abstract Space of Virtual Cinematography. Animation, 8, 3 (November, 2013), 253-266.

[7] Lai, P., Wu, H., Sanokho, C., Christie, M. and Li, T. A Pattern-based Tool for Creating Virtual Cinematography in Interactive Storytelling. In Anonymous Smart Graphics. (August 27-29, 2014). Springer, Taiwan, 2014, 121-132.

[8] Lampi, F., Kopf, S., Benz, M. and Effelsberg, W. An automatic cameraman in a lecture recording system. In Anonymous Proceedings of the international workshop on Educational multimedia and multimedia education. (September 28, 2007). ACM, Augsburg Germany, 2007, 11-18.

[9] Lino, C. and Christie, M. Efficient composition for virtual camera control. In Anonymous Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation. (July, 2012). Eurographics Association, Aire-la-Ville Switzerland, 2012, 65-70.

[10] Lino, C., Christie, M., Lamarche, F., Schofield, G. and Olivier, P. A real-time cinematography system for interactive 3d environments. In SCA '10 Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. (July, 2010). Eurographics Association, Aire-la-Ville Switzerland, 2010, 139-148.

[11] Liu, T. and Kender, J. R. Lecture videos for e-learning: Current research and challenges. In Multimedia Software Engineering, 2004. Proceedings. IEEE Sixth International Symposium on. (December 13-15, 2004). IEEE, 2004, 574-578.

[12] Nagai, T., Toyota, T., Nagoya, T., Nishizawa, K. and Imai, M. Implementation of high-definition lecture recording system for daily use. In Anonymous Global Engineering Education Conference (EDUCON), 2013 IEEE. (March 13-15, 2013). IEEE, Berlin, 2013, 520-525.

[13] Odhabi, H. and Nicks-McCaleb, L. Video recording lectures: Student and professor perspectives. British Journal of Educational Technology, 42, 2 (November, 2011), 327-336.

[14] Ranon, R., Christie, M. and Urli, T. Accurately measuring the satisfaction of visual properties in virtual camera control. In Anonymous Smart Graphics. (June 24-26, 2010). Springer, Banff Canada, 2010, 91-102.

[15] Rui, Y., He, L., Gupta, A. and Liu, Q. Building an intelligent camera management system. In Multimedia '01 Proceedings of the ninth ACM international conference on Multimedia. (October, 2001). ACM, New York, 2001, 2-11.