The CAMETRON lecture recording system: High quality video recording and editing with minimal human supervision *

Dries Hulens (✉), Bram Aerts, Punarjay Chakravarty, Ali Diba, Toon Goedem, Tom Roussel, Jeroen Zegers, Tinne Tuytelaars, Luc Van Eycken, Luc Van Gool, Hugo Van Hamme, and Joost Vennekens**

University of Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
firstname.lastname@kuleuven.be

Abstract. In this paper, we demonstrate a system that automates the process of recording video lectures in classrooms. Through special hardware (lecturer and audience facing cameras and microphone arrays), we record multiple points of view of the lecture. Person detection and tracking, along with recognition of different human actions are used to digitally zoom in on the lecturer, and alternate focus between the lecturer and the slides or the blackboard. Audio sound source localization, along with face detection and tracking, is used to detect questions from the audience, to digitally zoom in on the member of the audience asking the question and to improve the quality of the sound recording. Finally, an automatic video editing system is used to naturally switch between the different video streams and to compose a compelling end product. We demonstrate the working system in two classrooms, over two 2-hour lectures, given by two lecturers.

Keywords: Lecture recording, Smart camera systems, Virtual cameraman, Virtual editor, Sound source localization, Active speaker detection

1 Introduction

Wouldn't it be nice if we could make video-lectures of all classes taught at our universities, at minimal cost yet high quality? Well, now we can ! This paper describes a system to record lectures and process multiple video-streams yielding a professional looking montage with minimal human intervention.

The benefits for students of recording lectures and broadcasting them on-line have already been widely discussed and proven in the literature [7, 12, 21] and several (semi-)automatic lecture recording systems have been developed, focusing on capturing regions of interest [11, 13, 23] (e.g. projected slides), automated capture of lecturer notes taken in class [3] or (live) viewing over the internet

* This work is supported by the Cametron Project grant.
** Excluding the corresponding author, authors are listed in alphabetical order

[17, 18, 23]. A common disadvantage of most of those systems, however, is the typically low quality of the viewing experience, which is very different from the one obtained when a human camera crew records the lecture. Indeed, a human cameraman does more than just tracking the speaker positions: he positions the speaker cinematographically correct in the frame, he zooms in on the blackboard when the lecturer is writing, he zooms in on a person in the audience asking a question, etc. This requires a high level of semantic understanding of what is going on in the scene. Lecture recording systems on the market today lack such a semantic layer, resulting in lecture captures that are rather static and not very engaging for learners, which in turn leads to ineffective learning.

The goal of this work is to improve the viewing experience with an automatic lecture recording system, while keeping the cost down. To this end, we built custom hardware that synchronously records multiple video and audio streams of both the lecturer and the audience. We use computer vision and audio processing techniques to detect any relevant information in these data streams, such as action recognition, sound source localization and face detection. This information is fed to a virtual editor that is designed to switch between video-streams depending on any actions that were detected and following standard cinematographic rules. Further, to make the video more visually appealing, a virtual cameraman zooms in on the most relevant image region, mimicking the camera motions a human cameraman would make, by tracking the lecturer while he is moving.

The resulting CAMETRON system is highly flexible thanks to its modular design. It makes minimal assumptions about the class room layout, the number of cameras used, etc. The hardware can be set up in any class room in a couple of minutes, and needs no interaction from the lecturer other than starting and stopping the recordings.

The main contributions of this paper are: i) custom hardware for synchronized video and audio recording, ii) a virtual cameraman making the video more visually attractive, iii) action recognition helping the virtual editor choosing the best shot, iv) sound source localization used to zoom-in into the audience, and v) a virtual editor choosing the best shot and making the final video.

The remainder of this paper is organized as follows. After a discussion of related work, we describe the audiovisual analysis of the recordings of the lecturer (Section 2) and of the audience (Section 3). Next, we discuss the virtual editor (Section 4) and the hardware setup (Section 5). Section 6 discusses the experimental validation of our system and section 7 concludes the paper.

2 Analysis of the lecturer recordings

The actual video-stream chosen by the editor depends on different factors. One of the most important ones is if the lecturer is visible in that stream or not. A shot of the lecturers desk without a lecturer is not visually attractive to watch. Therefore we first run a person detector and tracker. In addition, a new video-stream is created out of the original video-stream. This new stream is a close-up shot of

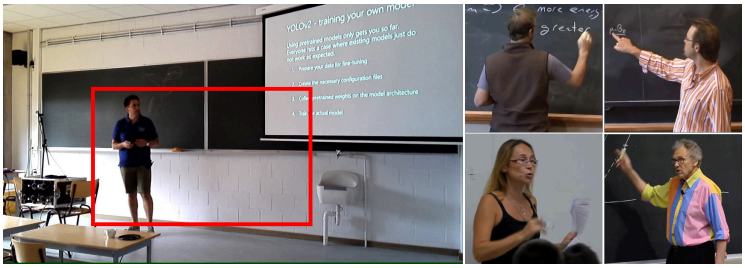

Fig. 1: Left: The original frame (1280×960) and the cut-out (640×480) in red with the lecturer positioned on the left due to the rule of thirds and with a bit of headroom. Right: Actions like writing, pointing and gesturing that trigger a close-up shot.

the lecturer where the lecturer is placed at the correct cinematographic position. Furthermore, several actions (like pointing to the blackboard) of the lecturer are detected to give the virtual editor cues whether to switch to a different video-stream or not.

2.1 Virtual zoom of the lecturer

The lecturer is captured by two or more HD overview cameras. To make the recordings visually more attractive, a virtual zoom is applied generating a close-up shot of the lecturer. This close-up shot is actually a cut-out (640×480 pixels) of the original image (overview shot of 1280×960 pixels). We chose not to use a Pan-Tilt-Zoom camera because there is always a transition between e.g. an overview shot and a close-up shot (zooming), which takes time and is not visually attractive to watch.

To make the close-up shot visually pleasing, some basic cinematographic rules are applied, namely the *rule of thirds* and the *head room*. The *rule of thirds* states that the lecturer should be positioned on the left of the screen (on 1/3rd of the width) when his face is looking to the right and vice versa, to leave some empty space (the other 2/3rd of the width) for the action to take place. The *head room* rule imposes that there should be some empty space between the head and the top of the frame. An example cut-out is shown in Figure 1 in red, together with some actions that trigger the virtual editor to switch to a close-up shot.

To apply these rules, first the position and gaze direction of the lecturer needs to be determined. The position is found by running a DPM (Deformable Part Model) upper-body detector [5] over the entire image. Next, the algorithm searches for the head in the upper part of the detection window. For the gaze direction estimation, we use the method described in [8], which is optimized for speed. It is based on three Viola and Jones face models (frontal, left- and right-looking) and a linear interpolation of the viewing angles of the two training images with the most similar detector scores. Based on the estimated gaze

direction, the desired position of the lecturer in the zoomed-in frame can be determined. Because the face angle is estimated in every frame and the zoomed-in region is adapted to that, the zoomed-in region will follow the lecturer. A PID control loop is used to move the zoomed-in region in a smooth way.

Since a log file is generated with the position and angle of the face for every frame, the virtual editor knows in which frames a face is visible or not and can switch to a different shot when the lecturer is not visible. The algorithm runs at 15fps while the video is captured at 25fps. This does not induce a problem in our setup since processing happens off-board and after the recordings are finished.

2.2 Action Recognition

While at times zooming in on the lecturer may be a good strategy, it is sometimes also better to show a wider view, e.g. including the projected slides when the lecturer is pointing to them, or the blackboard when the lecturer is writing on it. To enable the virtual editor to choose the right setup, we analyze the behavior of the lecturer and recognize a number of different actions.

The number of actions performed by lecturers is limited. We focus on writing on the blackboard, pointing, and talking gestures. Since there is no publicly available dataset including these categories of actions in a class environment, we created our own. It uses free on-line courses and lectures content and contains samples of the mentioned actions by different lecturers. Our method for action classification in these videos is inspired by recent progress in the field of deep convolutional neural networks [10, 6] and especially two stream networks for human action recognition [19, 4]. The two stream approaches utilize two modalities of data: RGB frames and motion Optical-Flow [22] frames. Each of the stream networks are trained separately on their corresponding data to recognize the actions [19].

To achieve the goal in this task accurately, we need to find the person beforehand. For this, we use the Fast-RCNN [6] object detector which is just trained to detect human bodies in video frames. After localizing the person in the video, we feed the cropped image of the person in the video frames to the action recognition neural network. Based on the score of lecturer actions, the CAMETRON system is able to decide for further steps and video editing action points.

3 Analysis of the audience recordings

In most lectures, a couple of questions are asked from members of the audience (questioners) and sometimes even a small discussion occurs. Most traditional lecture recording systems, however, fail to capture both the audio and the video from the questioner, which can be very frustrating for someone watching the recording. To tackle this problem we first localize the sound source, then zoom in on the active speaker. We also use this info to improve the quality of the audio.

3.1 Sound source localization

We propose a far distance beamforming solution together with a single wide-view camera. A microphone array (see section 5) is used to locate the direction of the sound by detecting spatial correlations, using the far-field approximation. The spatial setup of the microphone array (see figure 2 left) allows for an estimate of both the azimuth (θ) and elevation (ϕ) angle. An angular spectrum is created over (θ, ϕ) using a Generalized Cross-Correlation with PHAse Transform (GCC-PHAT) [9], since it was shown to perform best in source localization tasks [2]. A nonlinear transform of the GCC-PHAT is taken, to emphasize larger values in the angular spectrum. Source localization is done by detecting the peak in the angular spectrum. A questioner is assumed stationary while he talks.

For sound source localization, the audio stream is divided into segments of one minute (with one second overlap) for computational reasons. For each frame in each segment a Voice Activity Detector (VAD) is used to detect voice activity from the audience [20]. Consecutive audience active frames are grouped into a fragment, and for each fragment a spatial location is estimated. Fragments shorter than 1 second are regarded as noise or false positives and thrown away.

3.2 Active speaker detection

Whenever the sound source localization detects a speaker in the audience, it is necessary to localize them in the video as well. To find the active speaker, we use a fast and state-of-the-art face detector: the Multi-Task Cascaded Convolutional Neural Network of [24]. This has been trained for both face detection and alignment. The resulting detections are candidates for being the active speaker. We translate the azimuth and elevation angles into a position in the image frame using a third order polynomial: $x_i(\theta) = p_1 \cdot \theta^3 + p_2 \cdot \theta^2 + p_3 \cdot \theta + p_4$, and similarly for $y_i(\phi)$. Whichever candidate is closest to this result is assumed to be the active speaker. From there we can crop the frame around the speaker. The polynomial coefficients are determined through a calibration process that needs to be repeated once in each new environment. It consists of having a single person walk around the environment and clap his hands from several positions. As there is only one face in the frame when doing this, the face detector only has a single correspondence with the clapping sound. This allows us to fit the polynomial to translate the sound source localization angles to image coordinates. Clapping is detected by a simple energy threshold on the signal of one of the microphones from the microphone array.

3.3 Speech enhancement

Aside from positioning, the sound source localization technique can also be used to enhance the speech signal of the questioner. Since the distance of the questioner to the recording device can be up to the size of the class room, reverberation and a noisy sound source can reduce audio intelligibility for the listener. The speech signal is estimated by minimum variance distortionless response (MVDR)

beamforming with diagonal loading [14]. The multichannel covariance matrix of noise is estimated over 2 seconds of non-speech segments in the near context. The steering vector is determined on the Time Difference Of Arrival (TDOA), which was estimated for the sound source localization. Notice that this technique is only used for speech from the audience, since the lecturer uses a wireless close-talk microphone. Moreover, the speech of a questioner close to the microphone is less influenced by reverberation and the need for beamforming is reduced.

4 Virtual Editor

Once the autonomous cameras have captured high quality footage of the event, it's time for the virtual editor system to combine the footage of the different cameras into a single, coherent and qualitative montage. Creating a video that is both interesting and easy-to-follow is not a straightforward task. Human editors typically follow a number of different cinematographic rules to accomplish this task. To develop our virtual editor, we follow a declarative approach, in which we explicitly represent these rules. This approach has the benefit that it offers a great deal of flexibility in deciding which rules should be taken into account and how they should take priority over each other. Rules can be added and removed with relative ease. For example, using this framework we can easily add rules that take new actions into account. To represent the rules, we need a suitable knowledge representation language. A particular challenge in this application is that cinematographic rules are not strict: they are guidelines that are typically followed, but not always. Indeed, the rules may sometimes contradict each other, and even if they do not, a human editor may still choose to ignore a rule, simply because the result "feels" better. A virtual editor should therefore not rigidly follow the rules, but it should sometimes deviate from them in order to give the montage a more interesting and natural flavor, thereby mimicking the creativity of a human editor. For this reason, we have chosen to make use of a Probabilistic Logic Programming (PLP) language, which allows us to represent these rules in a non-deterministic way. This has the additional benefit that – just like a human editor – the system is able to produce different montages from the same input streams.

A detailed overview of the full virtual editor system can be found in [1]. A number of adaptations have been made to the described system in order to make it more applicable in the current lecture recording setting. Below, we briefly discuss these adaptations, which at the same time serves as illustration of how rules are used in our system.

In particular, we focus on a new aspect of the lecture not present in the system of [1]: questions from the audience. When a member of the audience asks a question, the viewer should see that person speaking. For this purpose we introduce an additional shot type, the interaction shot. Switching to this shot is possible from any other shot. A question from a member of the audience triggers the switch to the interaction shot. After an interaction shot, the reaction of the

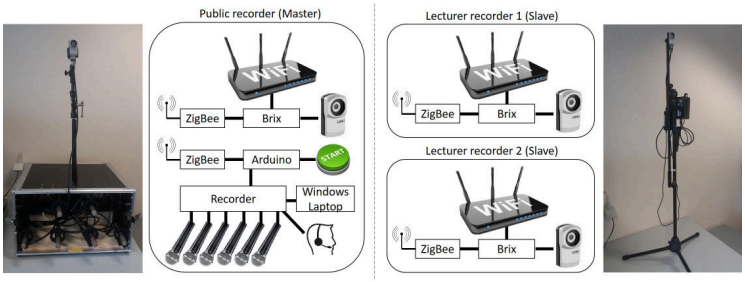

Fig. 2: Left: The main recording box aimed at the audience. Right: The individual recording devices aimed at the speaker.

lecturer is shown in a long shot, medium shot or close up. This approach also allows fluent switching between questioner and lecturer in a conversation.

5 Hardware

To record a lecture three devices have been developed, as shown in Figure 2. The first device is the *Audience Recorder*. It records video and sound of the audience and has everything integrated in a portable flight case. Six T-Bone EM700 microphones are used in the microphone array, while the lecturer is equipped with a Shure MX53 headworn earset connected to a Shure GLXD wireless receiver. The sound of all microphones is captured by an 8 channel Focusrite Scarlett 18i20 USB audio interface and stored on a windows laptop computer. The camera used in the Audience recorder is a Ueye XC 13MP color camera from IDS which provides images at a framerate of 25fps. The video is recorded on a Brix minicomputer which is connected to a WIFI hotspot for easy video transfer. To start and stop the recordings simultaneously on all devices, an Arduino is used which sends a signal to all devices (recorders) via a ZigBee wireless connection. In addition a sound signal is sent to the sound recorder to synchronize video and sound. Furthermore two smaller recorders are developed to record the lecturer. These recorders are a copy of the video recorder in the Audience Recorder box. When a start signal is sent by the Arduino, all recorders start recording simultaneously at a frame-rate of 25fps. When a stop signal is sent, the recorders stop recording and the video of all recorders is automatically transfered to the Windows laptop where they can be processed.

6 Experiments

Before we evaluate our complete system, we zoom in on the performance of some of the components: virtual zoom of the lecturer, action recognition, active speaker detection/sound source localization and speech enhancement.

6.1 Setup

To demonstrate the CAMETRON system, two lectures were recorded, in different class rooms and with different lecturers. The first lecture is called the *seminar recordings* and the second lecture is called *Matthews lecture*. They're respectively 95 and 117 minutes long. For both recordings the *Audience recorder* was placed in front of the classroom facing the audience, a second recorder was placed in the audience facing the lecturer and the third recorder was placed in the back of the room also facing the lecturer in a wider shot.

6.2 Virtual zoom of the lecturer

The virtual zoom of the lecturer was explained in section 2.1. To evaluate the correctness of the virtual zoom procedure we perform two experiments on the *seminar recordings*. In the first experiment we evaluate the detection of the lecturer. We find that the lecturer is not detected in 311 of the 142.900 frames. This yields an excellent detection accuracy of 99.8%. In the second experiment 100 frames are randomly extracted out of the *seminar recordings* and the consistency of the position of the lecturer w.r.t. the frame and the gaze direction is examined. Only in 2 frames the position of the lecturer is chosen wrongly, due to a wrong gaze direction estimation (caused by light of the beamer shining on the lecturers face). In some occasions we find the lecturer positioned in the center of the frame when looking to the left, but this is because the zoomed in frame can not move any further to the left due to the boundaries (size) of the original frame. These frames are also classified as correctly positioned frames. The Virtual zoom algorithm took respectively 158 minutes and 195 minutes to process both videos.

6.3 Action Recognition

For the action recognition, we first run an evaluation on part of the data we collected from the internet. Half of the dataset (400 short clips) is used for training and the other half for evaluation. This setup results in a mean average precision of 87% on the three action classes (88% talking gestures, 91% writing, 82% pointing). Next, we evaluate the module on the recorded lectures. To this end, we randomly select 100 frames (out of both recordings) in which an action is recognized and visually check the correctness. In this case the performance of the module is 84%, i.e. the action class label is correct in 84 out of 100 sample points.

6.4 Active Speaker Detection & Sound Source Localization

Evaluating sound source localization in our use case is more difficult, as we do not have any ground truth values for the azimuth and elevation angles. For this reason we evaluate it jointly with the active speaker detection with a simple experiment. We go through all of the video fragments where an active speaker

is detected in the audience and manually check if the correct speaker is cropped in the frame or not. We do this for both lecture recordings, giving a total of 79 detections. Using this evaluation method we get a true positive rate of 79.75%.

6.5 Speech enhancement

The speech enhancement procedure was explained in section 3.3. Its goal was to enhance the audibility of the question for the listener. Speech enhancement quality will be measured in terms of Word Error Rate (WER) of an Automatic Speech Recognizer (ASR). While optimizing speech enhancement as a preprocessing step to ASR is not identical to optimizing towards improved audibility, it does provide an indication without the need of a close-talk reference microphone for each questioner to measure, for example, signal-to-distortion ratio. For the ASR, a standard Kaldi [16] recipe is used to train an acoustic model on the AURORA-4 [15] database. This database contains utterances of read speech, while the system will be evaluated on spontaneous speech and thus poor results can be expected. Furthermore, no specific guidelines were given to the questioners and sometimes intelligibility is very low. In addition, all test speakers are non-native. The audio signal of a single microphone was used as baseline to compare with the MVDR of the microphone array. The WER for questions of multiple lectures were determined. Experiments were performed on four different lectures. In three lectures average WER was above 90% illustrating the low intelligibility. We found more acceptable results for one lecture (containing five questions). The WER of the baseline was 69.6% and was reduced to 66.1% (a relative reduction of 5.0%) using beamforming, indicating that the speech enhancement does improve audibility.

6.6 Overall system

Some snapshots of the *seminar recording* are shown in Figure 3. The first row shows the raw video input, while the second row shows processed footage. In the first image there is a question from the audience and the system reacts by zooming-in on the person that is asking the question. In the second image nothing is happening and the virtual editor selects the overview camera as output. In the third image the teacher is pointing and a close-up shot of the teacher is shown.

Figure 4 shows a number of shot sequences (one sequence per row). We briefly discuss the underlying reasons for those shot transitions. Sequence 1 and 2 show how the system reacts to a question from the audience. When a question is asked, the system generates a close-up stream of that person. The sudden appearance of this video stream triggers the first switch in sequence 1. The first switch in sequence 2 is triggered directly by a person speaking in the audience. When the person in the audience stops talking, the close-up stream is aborted, which causes a switch back to the lecturer. Sequences 3, 4 and 5 show the behavior when a shot becomes too long. When this is about to happen, a switch to a different shot is triggered. In sequence 3, the selected shot is an overview shot, where the lecturer is still visible. In sequences 4 and 5, the system selects the overview of

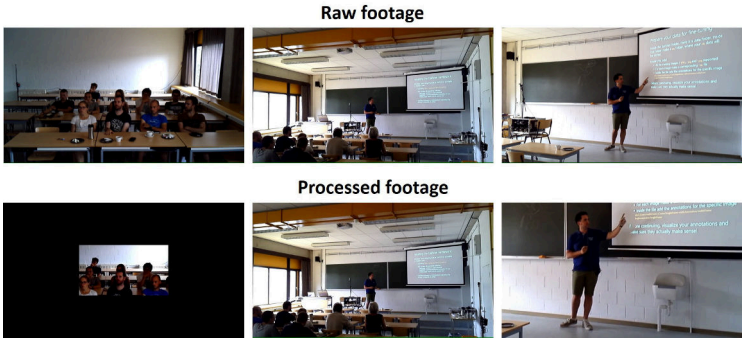

Fig. 3: Result of the overall system. First row: raw input footage of the audience (1), overview (2) and mid-shot (3) camera. Second row: Processed output: Zoom in on audience (1), overview shot(2) and crop of lecturer (3)

the audience. Here the lecturer is no longer visible, but is still talking. This is only suitable for a short period of time, so as soon as the minimal shot length has passed, the system decides to switch back to the lecturer. In sequence 6, an action (writing) is detected and a switch to a close-up shot is triggered.

Because the validation of the overall system is person-dependent we published both processed videos on Youtube[1]. In addition we added a movie of the first 5 minutes of the *seminar recordings* with the reasons of switching shot (determined by the virtual editor) plotted on the video[2].

7 Conclusion

In this paper we described a system to record multiple audio- and video-streams of a lecture and process these into a single video montage with as little human intervention as possible. The best video-stream is chosen by a virtual editor, taking into account the action that is taking place and cinematographic rules. To make the video more attractive, a virtual cameraman is created that zooms in on the lecturer and tracks his movement while considering the *rule of thirds* and *head-room*. When there is a question from the audience, sound source localization is used to zoom in on the person that is speaking. All processes are successfully validated individually as well as in an overall system.

References

1. Aerts, B., Goedemé, T., Vennekens, J.: A probabilistic logic programming approach to automatic video montage. In: ECAI. pp. 234–242 (2016)

[1] Seminar recordings: https://youtu.be/DalAafs38TU Matthew recordings: https://youtu.be/p3ZeFfj238g

[2] https://youtu.be/4Ruzv9jAZ6E

Fig. 4: Result of the overall system. Each row represents a sequence of shots, generated by the system. Sequences 1 to 4 are taken from the *seminar recording*, sequences 5 and 6 are from *Matthews lecture*.

2. Blandin, C., Ozerov, A., Vincent, E.: Multi-source tdoa estimation in reverberant audio using angular spectra and clustering. Signal Processing 92(8), 1950–1960 (2012)
3. Brotherton, J.A., Abowd, G.D.: Lessons learned from eclass: Assessing automated capture and access in the classroom. ACM Transactions on Computer-Human Interaction (TOCHI) 11(2), 121–155 (2004)
4. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (2016)
5. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
6. Girshick, R.: Fast r-cnn. In: IEEE International Conference on Computer Vision (ICCV) (2015)
7. Hahn, E.: Video lectures help enhance online information literacy course. Reference Services Review 40(1), 49–60 (2012)

8. Hulens, D., Van Beeck, K., Goedemé, T.: Fast and accurate face orientation measurement in low-resolution images on embedded hardware. In: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016). vol. 4, pp. 538–544. Scitepress (2016)
9. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing 24(4), 320–327 (1976)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
11. Lampi, F., Kopf, S., Benz, M., Effelsberg, W.: An automatic cameraman in a lecture recording system. In: Proceedings of the international workshop on Educational multimedia and multimedia education. pp. 11–18. ACM (2007)
12. Marchand, J.P., Pearson, M.L., Albon, S.P.: Student and faculty member perspectives on lecture capture in pharmacy education. American journal of pharmaceutical education 78(4), 74 (2014)
13. Mavlankar, A., Agrawal, P., Pang, D., Halawa, S., Cheung, N.M., Girod, B.: An interactive region-of-interest video streaming system for online lecture viewing. In: 18th International Packet Video Workshop (PV), 2010. pp. 64–71. IEEE (2010)
14. Mestre, X., Lagunas, M.A.: On diagonal loading for minimum variance beamformers. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). pp. 459–462. IEEE (2003)
15. Pearce, D.: Aurora working group: DSR front end LVCSR evaluation AU/384/02. Ph.D. thesis, Mississippi State University (2002)
16. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: Workshop on automatic speech recognition and understanding (ASRU). No. EPFL-CONF-192584, IEEE (2011)
17. Rui, Y., Gupta, A., Grudin, J., He, L.: Automating lecture capture and broadcast: technology and videography. Multimedia Systems 10(1), 3–15 (2004)
18. Schulte, O.A., Wunden, T., Brunner, A.: Replay: an integrated and open solution to produce, handle, and distributeaudio-visual (lecture) recordings. In: Proceedings of the 36th annual ACM SIGUCCS fall conference: moving mountains, blazing trails. pp. 195–198. ACM (2008)
19. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
20. Tan, Z.H., Lindberg, B.: Low-complexity variable frame rate analysis for speech recognition and voice activity detection. IEEE Journal of Selected Topics in Signal Processing 4(5), 798–807 (2010)
21. Tugrul, T.O.: Student perceptions of an educational technology tool: Video recordings of project presentations. Procedia-Social and Behavioral Sciences 64, 133–140 (2012)
22. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Joint Pattern Recognition Symposium (2007)
23. Zhang, C., Rui, Y., Crawford, J., He, L.W.: An automated end-to-end lecture capture and broadcasting system. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 4(1), 6 (2008)
24. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23(10), 1499–1503 (Oct 2016)