Solving the Divided Attention Problem in Lecture Recordings

Gerald Friedland
Institut für Informatik, Freie Universität Berlin
fland@inf.fu-berlin.de

Abstract. The work presented here is part of a dissertation project regarding audio/video recording and transmission of chalkboard based university lectures. The following article presents an approach to detect and extract the biggest moving foreground object in a video. The biggest foreground object is assumed to be the instructor acting in front of a chalkboard. During lecture replay the extracted instructor video is laid over the evolving board content.

1 Introduction

Two approaches are possible for recording and transmission of chalkboard lectures. Either one takes a video of the board content or one uses special input devices, such as electronic whiteboards or tablets to gather stroke based input. In order to record or transmit the information, it is common to either use standard Internet video broadcasting systems \[1,2,3\] or to create software that records and/or transmits the stroke based information \[4,5,6\]. The advantage of using off-the-shelf video broadcasting software is their easy availability and that the impressions of the classroom are recorded. The disadvantages are the high bandwidth and storage demands\(^1\) and the inadequate compression techniques used by the software\(^2\). Using pen tracking devices offers the possibility of a further online processing of the strokes, such as using a handwriting recognition or other semantic operations \[8\]. However, only handwriting is transmitted while mimics and gestures of the instructor are disregarded. For this reason, many lecture recording systems do not only transmit the slides or the board content but also an additional video of the instructor \[9,10\]. The problem that results is known as the divided attention problem: Although every human being only has a single locus of attention\(^3\) \[13,14\], the attention of the remote student is demanded by

---

\(^1\) See for example \[7\]. A 90 minutes talk in MPEG-4 format needs 657 MB.

\(^2\) DCT or Wavelet based codecs assume higher frequency parts of images as irrelevant which results in either an unreadable blurring of the board handwriting or a bad compression ratio.

\(^3\) Some results from research on spatial attention suggest that it is actually possible to attend to two spatial locations simultaneously under certain circumstances \[11,12\]. The model of assuming only a single locus may be simplistic, but concerning its application in user interface design, it is preferred for practical purposes.
two areas of the screen: the video window showing the instructor and the board or slides area. Hence it was tried to separate the video image of the lecturer from the background. The image of the instructor can then be laid over the board, creating the impression that she or he is directly working on the screen of the remote student. Mimics and gestures of the instructor would appear in direct relation to the board content. Ideally, the image of the lecturer can be made semi-transparent or even turned off. After a short section on related work, the remainder of this article briefly describes and discusses an algorithm used to extract the instructor from a PAL or NTSC video that was recorded in addition to the digitized pen input.

2 Related Work

Tracking of objects and segmentation of videos is currently a hot research area. Work seems to concentrate on either tracking of objects for computer vision (like robotic soccer [15], surveillance tasks [16], or traffic applications [17]) or interactive segmentation for image processing applications [18]. In computer vision real time performance is more important than segmentation accuracy as long as the important features can be extracted from each video frame. Therefore several approaches also use specialized hardware like omnivision or stereo cameras. For photo editing applications, accuracy is more important than realtime performance and known algorithms rely on information gathered by user interaction, see discussion in [19]. For the task presented here the segmentation should be as accurate as possible and must not require user interaction. A realtime solution is desirable for live transmission of lectures, however, not required for the post-processing of recorded lectures. It seems obvious to search for a solution in trying to create a combination of the techniques known from computer vision and interactive image processing.

3 Extraction of the Instructor

State of the art foreground extraction techniques for image editing applications get a coarse approximation of the foreground object provided by the user, sometimes only a rectangular area [19]. The segmentation is then refined by using color and intensity information and exploiting color distribution differences between foreground and background object. The idea of the algorithm presented here is to create a coarse grain cut of the foreground objects by exploiting the temporal differences between several frames and then exploiting color and color distribution information to improve the segmentation.

3.1 Temporal Foreground and Background Classification

The input for the algorithm is a sequence of digitized RGB video frames. Each frame is subdivided into 8x8 blocks. The algorithm uses two central datastructures: A foreground block buffer that is filled with any blocks that have a high
probability to be foreground and a background buffer that contains those blocks classified as sure background. A block is put into the foreground buffer if the block has changed more than twice during a sequence of \( n \) frames (where \( n \) equals half the framerate). A block is considered to have changed when it differs significantly from the block at the same position in the previous frame (according to the euclidean distance). The background buffer contains all blocks that have never changed during the considered sequence and where never classified as foreground during later operations. Both foreground buffer and background buffer are organized as a FIFO queue and are thus ageing.

3.2 Color Distribution Classification

For any video frame a color distribution classification is also applied. The frame is color quantized and again divided into 8x8 blocks. For each block in the frame, the histogram is calculated. The block histograms are now classified into foreground and background by comparing each histogram with block histograms of the foreground and background buffer. This is done by calculating a histogram for each block in the foreground and background buffer and using the Earth Movers Distance [20] as a metric.

3.3 Combining the Classifications

Given a frame and the results of the two classifications, all background objects are found as follows. Any block in the foreground buffer is considered foreground. In addition, any block that was classified as foreground according to the color distribution is also considered foreground. A union-find is applied to identify the connected components. The biggest blob is considered to be the instructor and all other blocks are put into the background buffer. An edge detection using the Sobel Operator [21] helps to smooth the edges of the blob, as they appear stepped because of the resolution reduction to 8x8 blocks. Smaller holes are filled up and the corresponding blocks taken out of the background list.

The resulting segmented video is scaled to fit the board resolution and is laid over the board content upon replay. Figure 1 shows the result.

4 Conclusion and Further Work

The presented approach improves in accuracy over approaches that concentrate more on exploiting temporal differences [22,23] where a lack of movement makes the foreground object disappear. The performance of the segmentation is still to be evaluated. In our test video set, about 20% of the frames contained foreground regions that were not part of the instructor. The case that the instructor leaves the screen is not handled yet. Still another problem is that if the instructor points at a rapidly changing object, the two corresponding blobs are merged by union find. Due to the ageing of the buffers, the algorithm is quite resistable against illumination changes. Interlacing effects, reflections, and shadows lead
to unclean cuts. Especially the separation of skin color is quite difficult [24]. In the future, time performance has to be improved by a factor of about 20 to be able to apply the algorithm for live transmissions. This seems quite possible when using the CPU’s built in multimedia instructions sets. Domain specific assumptions, like the instructor always touches the ground, can improve the precision of the segmentation. Introducing special hardware, like a stereo or 3D camera [25] might improve time performance as well as accuracy.

References

Report B-02-11, Fachbereich Mathematik und Informatik, Freie Universität Berlin
(2002)
2289–2296
Feature-based selection of non-contiguous locations. Perception & Psychophysics
61 (1999) 403–423
Cambridge, UK (1988)
14. Raskin, J.: The Humane Interface: New Directions for Designing Interactive Sys-
tems. Addison-Wesley, Boston (MA), USA (2000)
15. Simon, M., Behnke, S., Rojas, R.: Robust real time color tracking. In: RoboCup
Their Activities. IEEE Transactions on Pattern Analysis and Machine Intelligence
22 (2000) 809–831
17. Beymer, D., McLauchlan, P., Coifman, B., Malik, J.: A real-time computer vision
system for measuring traffic parameters. In: Proceedings of the IEEE International
18. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region
segmentation of objects in n-d images. In: Proceedings of the International Con-
ference on Computer Vision, Vancouver, Canada (2001) 105–112
19. Rother, C., Kolmogorov, V., Blake, A.: Grabcut - interactive foreground extraction
University, Department of Computer Science (1999)
USA (1992)
22. Jantz, K., Friedland, G., Rojas, R.: Trennung von Dozenten und Tafel in einem E-
Kreide Video. Technical Report B-04-07, Fachbereich Mathematik und Informatik,
23. Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Foreground object detection from videos
containing complex background. In: Proceedings of ACM Multimedia 2003, Berke-
ley, California, USA (2003)
application to objectionable image filtering. In: Proceedings of ACM Multimedia
(2004)