Wearable Cameras for Real-time Activity Annotation

Jiang Zhou[1], Aaron Duane[1], Rami Albatal[1], Cathal Gurrin[1,2], and Dag Johansen[2]

[1] Insight, Dublin City University, Dublin, Ireland
[2] Computer Science, UiT, The Arctic University of Norway

Abstract. Google Glass has potential to be a real-time data capture and annotation tool. With professional sports as a use-case, we present a platform which helps a football coach capture and annotate interesting events using Google Glass. In our implementation, an interesting event is indicated by a predefined hand gesture or motion, and our platform can automatically detect these gestures in a video without training any classifier. Three event detectors are examined and our experiment shows that the detector with combined edgeness and color moment features gives the best detection performance.

1 Introduction

Video annotation has played a very important role in multimedia information retrieval, medical image processing, sports performance analysis and many other domains. However, the real-time annotation has received little consideration [7]. In this paper, we present a platform built with Google Glass which enables us to capture and annotate videos in real-time. There are two major reasons motivating us applying the real-time annotation. First, post-capture annotation is laborious and error-prone. Second, it would be an intractable task to build special purpose detectors for each interesting event a priori due to a wide variety of potential events [6]. Our platform manages to provide the real-time annotation by using a predefined hand gesture or motion such that the play itself is not disturbed on the field and rewinding a video for manually tagging is not required.

The fundamental principle that enables real time video annotation with Google Glass is the *hindsight recording* [4]. A person observes an entire situation unfolding and determines afterwards whether it was a notable event worth capturing or not. The annotation works as a stop button in a video, indicating the end of a sequence worth capturing. The net effect of this hindsight evaluation process in real-time is that it is less likely important events can be missed, and there is no need for labor-intense manual tagging of entire videos. Only a detector needs to be built to extract footage of interesting events fully automatically by exploiting the annotation as input a definition of interesting events.

Event detection has been intensively studied in recent decades and many methods depend on building classifiers of specific instances to infer the events

[3]. For example, Lai [5] learns an instance-level event detection model based on video-level labels, and Aarflot [1] applies face detection to reduce the need of ever deleting digital objects from a digital library. However, building a prior classifier for each interesting event would not be practical or necessary for real-time annotation. Therefore, our platform identifies signals provided from real-time annotations as an indication of interesting events, without building any classifiers. A 10 second video segment before each annotation point is extracted as the footage of the interesting event. The footage will then be viewed through a web interface to recall activity performance, diagnose problems and give feedback.

The major contribution of our work presented in this paper is combining the Google Glass, a state of the art hardware, with event detection techniques for real-time annotation, which can be used for many real-world applications. Taking the specialty of Google Glass into consideration, a region importance mask is created to reduce noisy information in the event detection. An event detector built on the combination of edgeness and color moment features is proposed and has shown very promising detection performance.

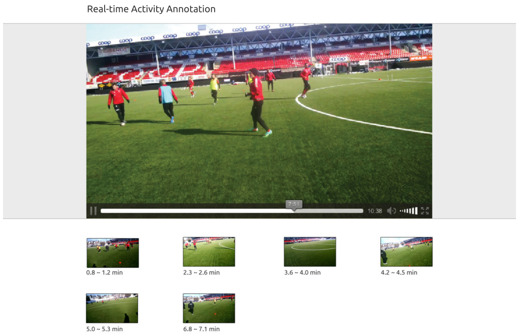

Fig. 1: *The platform web interface*

2 The Platform

Our platform consists of video acquisition using Google Glass, an interesting events detection tool and a web interface. Our platform is evaluated with the coach of a soccer team in Norway. In the training, whenever there was a ”good” or ”bad” example of play, the coach would wave his hand close to the camera of Google Glass. This would insert an annotation point to indicate an event of interest and generate a keyframe. As shown in the figure 1, the extracted event footage can be reviewed by the coach clicking any keyframe.

2.1 Interesting Event Detection

Our interesting event detection is designed to identify hand-wave gestures. A hand-wave gesture would cause sudden changes in color and significant decrease in edges. Therefore, the event detector is built with a combination of grid edgeness and color moment of each frame, which can be computed very efficiently. We noticed that the coach's fixated point is almost around a quarter of the height from the top in the picture. Hence, a region importance mask is applied to each frame, with weight 1 at a quarter of the height from the top and gradually decreases to 0 toward the top and bottom of the image, to relatively enhance the potential important information in each frame. Hence, the edgeness and hue deviation are calculated in a cell-grid. The edgeness and second order hue moment of each frame are then obtained by summing up the values from all cells and the negative edgeness value represents the final edgeness.

As we can see in figure 2, neither the edgeness nor the color moment can give a good indication of interesting events alone. Therefore, we propose to fit both features into a sigmoid function with equal weight 0.5. Both features are first scaled into a range of $(-6, 6)$ such that the function would have an approximate probability output. As shown in figure 2, the combination of edgeness and color moment is discriminative. The peaks in the plot strongly suggest the occurrence of interesting events. However, non-standard hand-wave movement may introduce fluctuation on the top of a peak or even double peaks in a very shot time period due to the hand wave-up and wave-down movement. Thus, we apply a maximum filter on the function output to make sure there is only one rise and fall for an event of interest. The middle frame within a rise and fall window is then regarded as the annotated point.

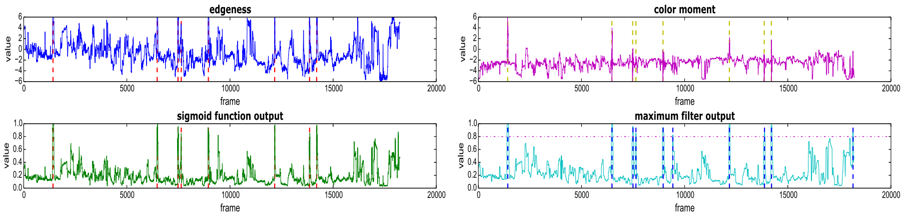

Fig. 2: *Event detection with multiple features; red vertical dotted lines indicate ground truth and blue vertical dot lines indicate annotation points (threshold 0.8).*

Another impression from the conspicuous changes over frames due to the hand-wave movement is that the changes have effects analogous to the fade-in and fade-out shots in movies. Therefore, we also experiment on re-deploying the camera-shot detection algorithms [2] for our interesting event detection. Table 1 shows the results of the three detection methods we proposed. From the results, it can be seen that the combined-feature detector gives the best performance. The combined-feature detector finds all interesting events and only a small amount of false positive events are introduced. A camera-shot detection algorithm can struggle to make a good balance between precision and recall; our experiment also shows that the camera-shot detection algorithms are vulnerable when the Glass camera is in an unstable state or during white balance adjustment.

Table 1: *Experimental results*

event detection method	precision	recall
combined-feature	88.9%	100.0%
fade-in shots	80.0%	69.6%
fade-out shots	61.1%	95.7%

3 Conclusions

In this paper we propose an annotation platform for real-time activity logging using Google Glass. Hand-wave signals are detected in a Glass video stream using three detectors: a combined-feature detector, a fade-in shot detector and a fade-out shot detector. Our experimental results show that the combined-feature detector can give very promising detection performance with 100% recall rate and 88.9% precision rate. Our current platform was a proof-of-concept application. We envisage that this can be employed to many real-time annotation tasks, where the annotation requirement is sufficiently straightforward to be represented by one, or a small number of, hand gestures.

Acknowledgement

This work has been performed in the context of the iAD center for Research-based Innovation project number 174867 funded by the Norwegian Research Council. A special thanks to TromsøIL and coach Morten Pedersen.

References

1. Aarflot, T., Gurrin, C., Johansen, D.: A framework for transient objects in digital libraries. In: Digital Information Management, 2008. ICDIM 2008. Third International Conference on. pp. 138–145 (Nov 2008)
2. Boreczky, J.S., Rowe, L.A.: Comparison of video shot boundary detection techniques. Journal of Electronic Imaging 5(2), 122–128 (1996)
3. Jiang, Y.G., Bhattacharya, S., Chang, S.F., Shah, M.: High-level event recognition in unconstrained videos 2(2), 73–101 (2013)
4. Johansen, D., Stenhaug, M., Hansen, R., Christensen, A., Hogmo, P.M.: Muithu: Smaller footprint, potentially larger imprint. In: Digital Information Management (ICDIM), 2012 Seventh International Conference on. pp. 205–214 (Aug 2012)
5. Lai, K.T., Yu, F.X., Chen, M.S., Chang, S.F.: Video event detection by inferring temporal instance labels. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), oral. Columbus, OH (June 2014)
6. Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Kraaij, W., Smeaton, A.F., Quenot, G.: Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2013. NIST, USA (2013)
7. Stenhaug, M., Yang, Y., Gurrin, C., Johansen, D.: Muithu: A Touch-Based Annotation Interface for Activity Logging in the Norwegian Premier League 8326, 365–368 (2014)