Skimming Video Action Using Annotated 3D Surfaces

Ben Falchuk[1], Chung-Ying Wu[2], Tarek El-Gaaly[3], Akshay Vashist[1]

[1]Telcordia Technologies, Inc., Piscataway, USA
[2]Columbia University, New York, USA
[3]Rutgers University, Piscataway, USA

Abstract

It has become all too clear that despite the ever-growing reams of available media, we face diminishing strategic returns from it unless we craft better tools that not only let us playback media but also get us quickly to media segments of most interest. Witness the everyday frustration of false positives when watching user-generated video content that, with appropriate insight, the user might have otherwise chosen not to watch. In this paper we describe a new and dramatically different new way to both summarize and interact with multimedia information in rapid, 3D, user-in-the-loop, skimming sessions. Our new interaction technique, which can accommodate object recognition algorithms, is a fusion of media summarization and 3D scene generation techniques and runs on mobile, tablet and desktops.

Categories: I.3.6 [Methodology and Techniques]: Interaction techniques, H.5.2 [User Interfaces]: Graphical user interfaces

1. Introduction

Multimedia information consumption and creation – particularly video – has seen an incredible uptake in the last several years. YouTube is, by most metrics, the world's second largest search engine (second only to Google Web search) with more search hits than both Yahoo! and Bing in 2009. In both mass-market and professional contexts, searching and watching networked video is here to stay. It is a major part of Apple's successful mobile strategy and is of increasing importance in an ever-more monitored world (e.g., closed-circuit security cameras, unmanned aerial vehicles). Enormous media upload rates onto popular sites such as YouTube, Hulu, and Vimeo, are a testament to how seemingly infinite and diverse is the supply of user-generated content. Stepping back though, we notice that the sophistication of tools to help us find and browse videos remain, in many ways, unsatisfactory. Consider the tasks: T1: *Quickly determining the essence of a previously unseen video (i.e., will it be interesting?)*; and T2: *Quickly determining if a particular event or scene occurs within a given video (i.e., is it the one you remember?)*. These tasks continue to be difficult to complete, in many situations.

Put another way, if one was asked to find a particular episode of the NBC show *The Office* entitled *Training Day*, it would not be hard to track down that resource by title and play it. If, on the other hand, we asked, "Is the episode *Training Day* the one in which the microwave oven catches on fire?", then we'd likely be faced with the time-consuming task of watching many of the episode's scenes in order to determine if the oven burns. Today's purely algorithmic techniques can be effective if properly trained but video content understanding is AI-hard. No current technology can tell us whether the microwave oven catches fire in a given video and not be confused, for example, by flaming *food* being taken out of the oven. Human poses are similarly difficult to automatically detect (e.g., "Is the one where Astaire dances cheek to cheek with his partner and then leaps into the air?").

To this end, we have designed an interaction technique named Donatello that supports tasks such as T1 and T2. Donatello makes efficient and effective use of graphics while drawing attention to scenes of interest. Donatello is intended to be a new way to skim media before streaming it, and to *complement* (*not* replace) existing algorithmic search and streaming tools. We also describe a methodology for annotating keyframes called pose extraction in which human poses can be automatically highlighted, a practice that, according to users, will be helpful for surveillance analysis and choreography, to name a few. This paper focuses on the design of our new visual technique but also offers the results of a very preliminary user study.

2. Related Work

Video skimming is the visual presentation of sub-segments in order to allow content understanding, and is usually interactive. On the other hand, most sites employ user tags and descriptions as principle indices. At a coarse level, tagging has great practical value [Bal08][YLL10]. At a fine