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

level though, tags fail to index the nuanced essence of video action – for example, conditions leading to a character's fall, the gestures or gait of a character, or any temporal relationships such as one event happening after another. A variety of past work in info-viz (e.g., Perspective Wall [MRC91]) and video summarization [YEO97][RWW10] has been both instructive and inspirational to our work.

In the large, visual interaction techniques to support comprehensive video skimming are sorely lacking. Few Web sites – even the largest and most popular - offer skimming techniques other than the familiar pre-selected shortlist of scenes and the accompanying tip, *Choose a scene to begin playing from there*. This can be effective (if they have pre-selected scenes that you have interest in) but largely it is not effective as a skimming technique. Some sites display video keyframes that begin playback in-place when the mouse hovers over them. The playback, however, is almost always very coarse, non interactive, and limited to a small preset range of keyframes. These limitations also plague mobile video browsing applications (this is *despite* the many reasons that skimming is even more important on mobiles!). Few to none of the many mobile offerings from media and telecom companies enable rich, effective, within-video skimming, instead relying keyword search, simple "TV-guide" like interfaces, and extremely limited "choose a scene"-technique. Recently, an iPad app by ABC News uses a sphere surface to present news headlines, but not for video skimming or playback.

Pose extraction has been an active research topic over the last decade [RBM05][FMZ08]. Many initial approaches were limited to extracting certain characteristic poses from user-defined regions of interest. Recent progress in machine learning has made it possible to overcome these limitations and existing research [RBM05] can now be leveraged to develop automatic techniques for satisfactory pose extraction in uncontrolled video.

3. Donatello

Our project - named Donatello - is a response to the realization that users' ability to quickly understand the essence of stored video remains limited across both mobile and desktop screens, and across casual and professional user-bases. Donatello aims to a) involve the user directly and interactively in skims, b) interwork with a networked media server, and c) effectively and efficiently represent and annotate video. To achieve the above goals, the Donatello mobile application communicates with the Media and Processing servers. The former provides an API to the raw video segments while the latter constructs textures from - and performs pose extraction on - the video data. Textures returned to the client are rendered onto one or more interactive objects in the 3D scene. The scene can be customized and adapted based on both the video content and user preference. Figure 1 depicts this.

The rationale for using an adaptive set of 3D shapes on the client side was to make the experience of skimming familiar, attractive and effective. We have initially experimented with a simple 3D scene comprised only of a sphere. The shape of the sphere, when viewed as shown in the figure on the screen, gives us a natural central region (along "equator") that we use as the focal area (frames of most interest to the user) and the regions of the sphere arcing away from view give a natural place for contextual frames (e.g., scenes before and after the focus).

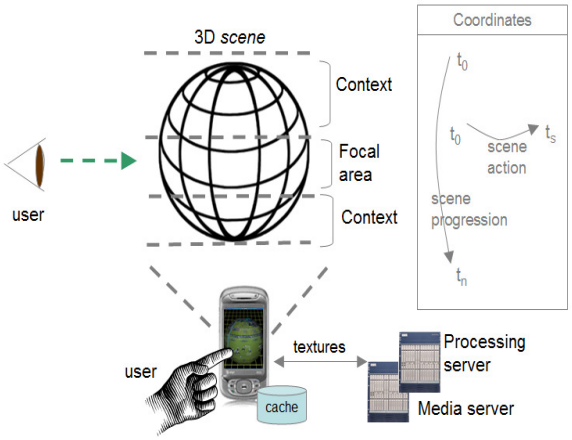

Figure 1: *Donatello schematic; a 3D shape surface is textured and "read" by the user according to two coordinate systems*

We also used enlarged facets on the central part compared to those near poles. Regardless of the video, scene action runs around the sphere's latitude lines while scene progression is ordered down longitude lines form the "north". Donatello employs a navigation scrollbar at the bottom of the GUI (see Figure 2); the width and position of scroll-thumb indicate the range and focus of the current view proportional to the entire length of the video. While a sphere is not necessarily the *optimal* way to use pixels on the screen it is, under some reasonable constraints, comparable in pixel-efficacy to a 2D matrix of frames (e.g., consider that the sphere holds and hides information on the reverse side). We found that a cylinder may be the next best 3D surface, while cube-like shapes and others were intriguing but their angular facets created complications (user studies are in progress). The sphere was, in our opinion, the best to initially prototype and allows for interactions that are, "as intuitive as rolling a ball".

4. Results

Figure 2 illustrates a Donatello skimming session, which is initiated after the user chooses a video to skim. The thin blue strip indicates the temporal reset position on the shape. Figure 2 also shows the kind of menu options available during a skimming session; these include the ability to: step back to the previous view/position on the sphere, show media metadata, reset the view, or start an HD playback of the video. Recursive non-linear skimming is achieved by either spinning the sphere, dragging the scroll thumb or by double-tapping a particular scene, after which the sphere is re-textured by the server with keyframes sampled *solely from the selected subscene*. The yellow scroll thumb resizes and repositions such that it coveys the current range of

view (in the Figure only a small part of the cartoon is being skimmed, at about the 1/3 mark of the video). At any view, the user can drag a finger across the screen to spin it around the z-axis or drag up and down to bring the poles more into view (as in Figure 3).

Figure 3 illustrates a football game skim. When applied to sporting videos, users quickly get the essence of the progression of plays in a match (by dragging and rotating the sphere to see its top and bottom) as well as the essence of the action within plays (by dragging and rotating the sphere along the Z-axis). Furthermore, scoring and other significant plays can be highlighted by the server (see next section). Users double-tap on the action of interest and the 3D scene was iteratively repainted with a subset of frames represented in the previous view. Users see rapid playback of brief segments (e.g., touchdowns) or could launch a high-definition playback.

Figure 2: *Skimming a cartoon textured onto an object in a 3D scene (right); available menu options (left).*

Pose Isolation as an Optional Summarizing Feature

For most human activities, including sports, dancing, and dramatic acting, pose often constitutes a succinct description of activities and therefore a convenient mechanism by which to browse and search action (pose-based vocabularies do exist). Pose isolation can also optionally and beneficially remove potentially distracting backgrounds, streamlining content analysis for users.

Automatic 2D human pose estimation in an uncontrolled video such as the ones typically available from YouTube, is a difficult problem. This is due to large variability and uncertainty about human position, scale, occlusion, clothing, shadows, and so on. However, recent advances in computer vision and machine learning research have made it possible to estimate poses in a wide variety of videos [RBM05]. After automatically subtracting the background and detecting humans in an image, pose detection is treated as the problem of a search for configurations of body parts. While many existing methods rely on interactions and feedback from humans to estimate the pose, we leveraged a recently proposed fully automatic method [FMZ08] for upper body pose detection.

Over the many possible representation of poses including geometric figures (such as rectangles, ellipses, etc.), Donatello currently represents pose using "stickmen" (see Figure 4) which are appealing for any screen size and tend to be visually clear compared to more elaborate figures. To improve the efficacy of the existing pose estimation research, we augmented the upper body pose (head, torso, and upper/lower right/left arms) estimation algorithm in [FMZ08] by a lower body pose estimator, which takes the upper body pose as the seed for searching the configuration of the lower body. Based on the torso position detected by [FMZ08] our algorithm estimates the position of knees and ankles from the body contours with an eye on recovering the full body pose. Since the degrees of freedom in the lower body motion are relatively low compared to the upper body, the algorithm successfully leverages the existing algorithm to produce satisfactory results (Figure 4). In its current version, our extraction is limited to front and back views – a limitation it inherits from the existing algorithm. In future work a retraining of the pose estimator using a richer collection of frontal and profile body poses should overcome this limitation even with partial occlusions.

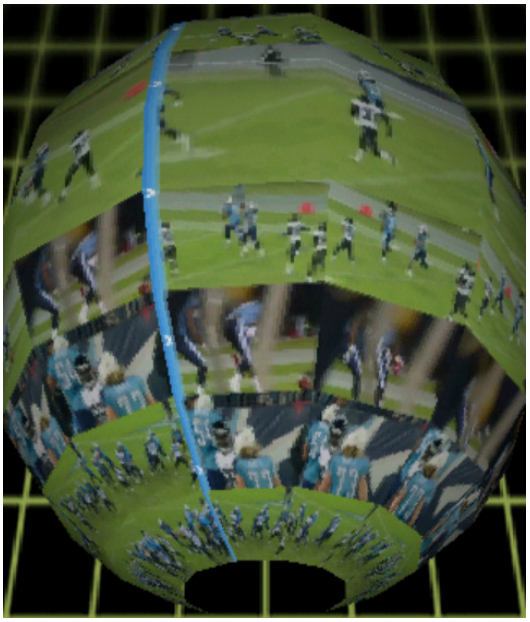

Figure 3: *Skimming a football game. Plays/scenes are ordered sequentially and from the top along "latitudes", while play action proceeds around "longitudes"*

Human pose extraction is an effective and succinct representation of interaction between people which can be extended to other animate (such as pets) and inanimate objects to summarize an even wider range of activities.

Along with automatic object detection made possible by excellent research in computer vision, the proposed extension can be leveraged to tag video with activities and to enable text based browsing and searching of visual content. On the whole, Donatello's visual metaphor was particularly effective for showing pose-isolated keyframes as summarization aids and allowed us to interleave and overlay poses that convey the essence of the action. Pose isolation occurred on the Processing Server (Figure 1) where poses were interleaved with the returned texture.

Technology and User Feedback

Donatello was implemented on an Android *Nexus One* mobile phone (version 2.2 with support for OpenGL ES 2.0). We used Eclipse to develop software in both Java and C. Our Media and Processing servers were staged on a Windows server. To support the 'Play (in HD)' option, Donatello launches the Android YouTube app.

"I would like to use this technique and tool on video"	75%
"Doesn't seem useful – I'd rather use other tools"	20%
"I don't understand the technique"	5%

Table 1. *User feedback (further studies under way)*

Our small-scale user studies to date indicate that a majority of people have found the technology to be potentially useful. At this point, however, we have only engaged users with Donatello (user base in Table 1 is approximately 30 people, mostly technical and highly Internet-literate).

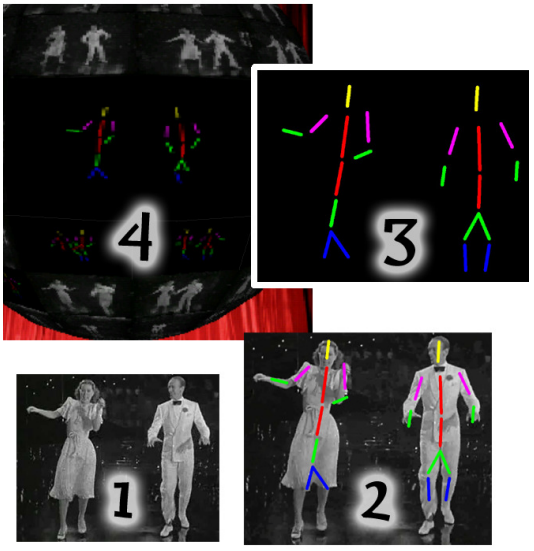

Figure 4: *Pose extractions as overlays. Steps 1-3: server extracts poses and stores them. Step 4: server selectively replaces keyframes with poses to aid the skim session*

In our prototype, textures (.png files) are generated on-demand on the server-side to texture the scene object(s) in question (e.g., sphere). We found that at roughly 512KB the resulting sphere had good visual quality (maxing out at 2MB). Textures are transferred upon each unique view of the media (i.e., at double-taps, not rotations). In practice we have found that about 4 to 9 distinct views often suffice to convey the essence of the media. On the flipside, YouTube bitrate tends to be in the range [75 Kbps, 2.77 Mbps] (see http://bit.ly/dD69B9). Conservatively taking 768 Kbps as an estimate, then, we have:

- Num. unique views in typical Donatello skimming session: 9
- Total data in a single Donatello skim session: 4.5MB
- Total bits transferred for YouTube movie of length:
2 min = 11 MB.. and for length 30 min, total bits = 165 MB

Therefore, from a purely network traffic point of view, we see that – with some assumptions - for (YouTube) movies longer than one minute or so there is great bandwidth-savings potential of first using Donatello to rule out "false positives" as opposed to watching them in full.

5. Conclusions

The main contribution of our work is the design and prototype of a completely new video interaction technique with improved practical ability to "get to the essence" of a given video. There are few such techniques in the large at the moment, on any platform. The advantages are:

- Works with any API-accessible Internet video
- Familiar and intuitive interface (can you roll a ball?)
- Saves the user valuable time (reduces false positives)
- High information-density (versus other techniques)
- Is lightweight and reduces bandwidth usage
- Clearly displays the essence of human poses

Our small-scale studies to date indicate that users are intrigued by this technique. Donatello is less useful when the media clips are extremely short (i.e. those that can be quickly played out in full at low expense without skimming tools). Audio is not yet accounted for. Human pose extraction is an optional - but potentially effective - aspect but is only relevant on videos with human actors in it. In our view, Donatello has great promise in several application domains realms. Our future work includes: improving pose extraction, applying new server-side algorithms, and managing different 3D scenes.

References

[Bal08] BALUJA, S. ET AL, "Video Suggestion and Discovery for YouTube", *Proc. Int'l. WWW Conference*, Beijing, 2008

[YLL10] YANG, J., LUO, D., LIU, Y., "Newdle: Interactive Visual Exploration of Large Online News Collections", *IEEE Comp. Graphics & Applications*, 30(5), Sep. 2010

[MRC91] MACKINLAY, J., ROBERTSON, G., CARD, S., "The perspective wall", *Proc. ACM SIGCHI Conf. on Human factors in Comp. systems*, pp.173-176, New Orleans, 1991

[YEO97] YEO,B., YEUNG, M., "Retrieving and visualizing video", CACM, 40(12), pp.43-52, 1997

[RWW10] ROO, O.D., WORRING, M., WILJK, J., "MediaTable: Interactive Categorization of Multimedia Collections", *IEEE Comp. Graphics & Applications*, 30(5), Sep. 2010

[RBM05] REN, X., BERG, A., MALIK, J., "Recovering Human Body Configurations using Pairwise Constraints Between Parts", *Proc. Intl. Conf. Computer Vision*, pp. 824-831, 2005

[FMZ08] FERRARI, V., MARIN-JIMENEZ, M., ZISSERMAN, A., "Progressive Search Space Reduction for Human Pose Estimation", *Proc. IEEE* CVPR, 2008