VideoHandles: Replicating Gestures to Search through Action-Camera Video

Jarrod Knibbe, Sue Ann Seah, Mike Fraser
Department of Computer Science, University of Bristol, UK
{Jarrod.Knibbe, s.a.seah, Mike.Fraser}@bristol.ac.uk

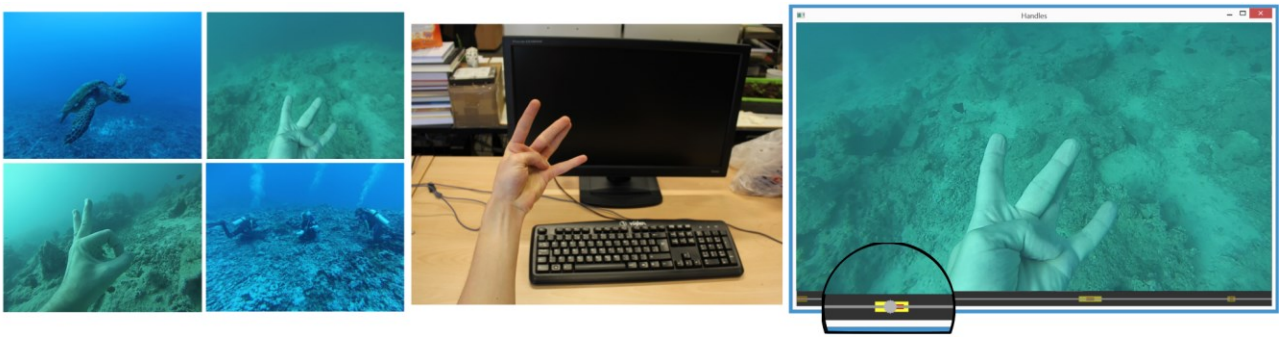

Figure 1. VideoHandles interaction technique: a) Max records his diving activities, b) upon returning home, he repeats a gesture he remembers performing as a search query and c) the system compares the query with the original footage, returning a range of possible results (shown by highlighted yellow spans in the time bar).

ABSTRACT

We present *VideoHandles*, a novel interaction technique to support rapid review of wearable video camera data by re-performing gestures as a search query. The availability of wearable video capture devices has led to a significant increase in activity logging across a range of domains. However, searching through and reviewing footage for data curation can be a laborious and painstaking process. In this paper we showcase the use of gestures as search queries to support review and navigation of video data. By exploring example self-captured footage across a range of activities, we propose two video data navigation styles using gestures: *prospective gesture tagging* and *retrospective gesture searching*. We describe *VideoHandles'* interaction design, motivation and results of a pilot study.

INTRODUCTION

Max, a marine zoologist, is performing a scuba dive to record some underwater footage using an action camera. During the dive Max performs various hand gestures for his buddy, indicating aquatic life of interest. On one occasion, he sees a trigger fish and performs a fish swimming gesture followed by a trigger mime. On another occasion, he sees a puffer fish and performs the gesture (fish swimming gesture followed by a two-handed mimicked inflation) to his dive buddy so that she can identify it too. Upon returning home, Max uploads the footage to his computer in order to analyse some of the key moments. He performs a puffer fish gesture as a search query and VideoHandles produces the puffer fish footage as a top ranking result among other results that include the fish swimming gesture. In the results, Max notices the trigger fish and decides to review that footage as well.

A wide variety of users, from amateurs to professionals, have adopted action cameras across a diverse range of activities, from mountain biking and scuba diving through to professional fieldwork. These cameras, such as the GoPro [6], are frequently mounted on head-gear or fixed to the chest and record throughout an activity, often for 1 – 2 hours, with little or no additional interaction. From these positions, and given a wide field of view (circa 170 degrees), the cameras are able to catch the majority of the wearer's view including any interactions or gestures they may be performing with their hands.

Where professionals may capture footage in order to maintain a clear record of their actions, others (e.g. sports enthusiasts) are more likely capture footage for key exciting moments. Although these are different motivations, all scenarios necessitate review in order to locate desired moments. Current approaches for video review are limited, with the widely adopted traditional method of video scrubbing (i.e. clicking through a timeline) being an inefficient process. As lifelogging becomes more pervasive, this process is clearly not scalable.

SUI '14, October 04 - 05 2014, Hololulu, HI, USA

978-1-4503-2820-3/14/10□$15.00.
http://dx.doi.org/10.1145/2659766.2659784

In this paper we present *VideoHandles*, a novel video search technique to expedite the review of specific moments in wearable video camera data. By exploiting the camera's view of the wearer's interactions and gestures, our system allows users to query their footage by repeating interactions and gestures performed during capture. The user re-attaches the action camera to the original position, and re-performs their target gesture. These reproduced gestures are matched to instances in the original footage.

From observing footage across a range of activities, we propose two video data navigation styles using gestures: *prospective gesture tagging*, where gestures are specifically performed to 'tag' moments in the footage, and *retrospective gesture searching*, where gestures are simply a part of the activity, recalled through muscle memory. We describe *VideoHandles'* interaction design, motivation and results of a pilot study.

RELATED WORK

The most widely adopted method for video data navigation and review is video scrubbing. As a technique, it has been shown to be very fast in low latency systems where the video output updates perfectly in time with a moved slider [11]. Further, user knowledge of the footage (temporal, contextual and spatial) helps to increase the speed of navigation [11]. However, scrubbing has its limitations and these become increasingly apparent as video duration increases. As one example, the mapping between the scrubbing slider and the corresponding video timeline is rarely one-to-one [9]. The slider is limited in size by the window width, itself defined by the video's resolution. As the video length increases, each pixel of slider movement corresponds to a longer time step within the video. For example, given a 2.5 hour 1080p recording, each pixel of slider movement corresponds to 4.69 seconds. Assuming the video is viewed at full resolution on a 24-inch 1080p monitor, a pixel measures 0.28 mm. Moving 1 cm along the slider thus represents a time step of 2 minutes 47 seconds. These time steps make it easy to miss interesting moments of footage even with small slider movements.

VideoHandles aims to assist video navigation by providing location markers in a video based on user's input gesture. In this way, *VideoHandles* not only provides navigation cues, but also functions as a video search interface.

Traditionally, in an attempt to align with other online search forms, video search has been based on a query-by-text approach [13]. This technique requires pre-defined textual annotations that necessitate unacceptable initial time expenditure [12]. Furthermore, the annotations encode user perception, making their sharing difficult between multiple users difficult [7]. Recently, video search has moved to focus on a combination of different approaches, such as visual and audio cues [15], concept search [17], and image search [5]. While video search results are improving, the best results are often achieved using a combination of image processing and human interaction [7].

As gestural interfaces have grown in popularity, so too has research on image processing, specifically gesture segmentation and matching. The approaches typically vary based on camera type [1, 8] or key algorithmic features [e.g. 4, 10]. As yet there is no one-size-fits all solution and the design decisions of any proposed algorithm need to be closely tied to that of the setting.

AN EXPLORATION OF ACTION CAM FOOTAGE

In order to motivate the design of our technique and to identify different styles of candidate gestures, we observed more than 50 hours of footage captured from a range of activities, including snowboarding, cycling, scuba diving and archaeological excavation. The footage was collated from 5 existing users of this type of camera.

Observation 1: Activity-based Gestures

Our first observation is that the style of gestures naturally performed varies significantly across usage scenarios. For example at one end of the spectrum, scuba diving includes frequent sign-language gestures, where meaning is directly encoded in the hand-shape and motion. In the middle of the spectrum are activities such as archaeological excavation, tennis and windsurfing. Whilst these activities do not have a clear gestural vocabulary (like scuba diving), the majority of the skill is performed manually (i.e. embodied action performed specifically with the hands). For example, this could be careful trowel movements in archaeology, the various shots in tennis (forehand vs. backhand) or the different hand-holds for mast and boom positioning in windsurfing. As the motion of the hands plays a key role in these activities, these gestures are also good candidates for *VideoHandles*.

Observation 2: Non-activity-based Gestures

At the other end of the spectrum, where manual variation is limited, are activities such as mountain biking and running. In these activities, the primary execution of the skill is non-hand based and thus the participant's hands perform a limited variety of movements. For this reason, the 'normal' opportunities for gesture or action repetition are more limited without the performance of additional deliberate gestures for use as prospective gesture tags for later searching.

Even during those activities whose skill is less manually-performed, one key similarity observed between all the activities is the frequent and continued use of social or 'pantomime' gestures. These settings showcase frequent language-like gestures, such as congratulatory 'high-fives' or 'fist-bumps.' These same activities also utilized language-tied and deictic gestures, such as *"that time you went left and I went the other way."* Gestures of this kind also provide good candidates for our technique.

VIDEO HANDLES

The *VideoHandles* video search technique enables users to remember, or specifically plan, gestures produced during recording and to reproduce these gestures as search criteria to relocate specific moments in footage. Our technique

reduces the requirement for human time and effort in reviewing vast reams of video data.

Users can search their footage by repeating any hand-actions / gestures from a similar viewpoint to which they were originally captured. By not making any assumptions about the style of gestures performed or their meaning, our technique can support a wide variety of gestures, including sign-language like scuba-diving gestures and manual skill based actions, such as trowelling in archaeology.

Based on our observations in the previous section, we propose two video data navigation styles using gestures: *prospective gesture tagging* and *retrospective gesture searching*.

Prospective Gesture Tagging

As users become accustomed to *VideoHandles*, gaining an understanding of the gestures that are matched most successfully by the system, we foresee increased performance of gestures during recording specifically designed for later retrieval. We term these gestures *prospective tagging.* For example, if a moment of immediate interest occurs during mountain biking, the rider could pre-emptively perform a gesture to 'tag' the moment and increase the accuracy of later retrieval (e.g. a gesture that would not normally occur during the activity). Gesturing has been shown to assist our ability to remember [3], and thus prospective tagging of this kind further aids users' memory of the gesture for search, ensuring more accurate search results.

Retrospective Gesture Searching

VideoHandles is also able to support occasions where gestures are simply a part of the activity. In some instances, prospective marking will not be possible, perhaps because the event only acquires importance and meaning for the user in retrospect rather than at the time, or because the user is not interested only in a single event, but wishes to review and compare all examples of a particular activity (such as trowelling in archaeology). In these cases, the user can use VideoHandles to perform a *retrospective search*. One of the benefits of this mode of searching is that the user may be able to rely on visual and muscle memory to perform the query. If the search is for one or more instances of a well-rehearsed manual skill, then the previous practice will also enhance the consistency of the search query.

Multiple Results

Just as a web search provides multiple results, *VideoHandles* does not intend to provide only 'exact' matches; rather it supports reflection and comparison between hits by returning a ranked range of results. In this way, *VideoHandles* also serves to better support 'chance finds' when clicking through footage and specifically supports 'middle-spectrum' activity-based gestures.

PROTOTYPE SYSTEM

We developed a prototype system to explore the feasibility of our concept and its value from an HCI perspective. We used a combination of existing computer vision algorithms to track, segment, and shape- and motion-match gestures in different videos. The technical approach we adopt is just one of many possible approaches and any appropriate computer vision algorithm could be used.

We segment gestural information from the scene based on motion (using Farneback optical flow [5]) and color (a combination of RGB and HSV skin color detection). Identified skin regions are tracked over a series of frames (typically 10 frames or 0.33 seconds) and subsequently saved as a 'gesture chunk.' When gesture chunks have been located in both the raw and query footage, these chunks are compared (in both shape and motion) to determine suitable matches.

The shape match is calculated using Fourier Descriptors of the contours [16] and chamfer matching [2]. Both of our matching techniques are rotation and scale invariant and by varying the examined 'gesture chunk' window size our techniques are also time invariant. Motion matching is conducted using the $ gesture recognizer [14]. Similarly to our shape matching, this is invariant to time, scale and location. Combined scores (from both shape and motion matching) below a given threshold are considered a match.

Once the footage has been processed, temporally and spatially co-located matched chunks are grouped together and time-period scores are returned based on frequency of matches.

PILOT STUDY OF *VIDEOHANDLES* IN THE WILD

To further explore our interaction technique and to begin to evaluate our prototype, we conducted an initial study of *VideoHandles* in realistic use. A participant wore a GoPro action camera on a chest mount whilst cycling, recording 42 minutes of footage. The participant was asked to perform a gesture of their choosing, indicating: every time they saw a red car, when they were feeling energetic and when they felt tired. After the activity was completed, the participant reviewed their footage using video scrubbing, recording the time and meaning of every gesture they saw. While not all gestures were revisited, a subset were noted to provide correspondences for our *VideoHandles* software. In total, 28 gestures were identified. After review, an example of each type of gesture was recorded as a search query for our system.

Results

Our participant indicated 28 'two-finger gun' style gestures corresponding to red cars. Our prototype algorithm returned 89% of these gestures and 1 false-positive gesture. Our participant performed 3 'OK' gestures and our prototype algorithm returned 2 correct matches and 16 false-positive gestures, including both car gestures and 'tired' gestures.

DISCUSSION

Our exploration and evaluation has highlighted a number of interesting features for further consideration in the design of our approach.

Firstly, our study highlighted the importance of shot framing and camera mounting. Even with the camera's wide field of view, our study participant (an inexperienced action camera user) repeatedly performed their 'tagging' gestures at the very top of the frame, resulting in part of the shape detail being lost and making any matching difficult. Given increasing use of these cameras and our technique, we foresee users becoming more accustomed to the camera's field of view, thus purposefully positioning their gestures more accurately.

Secondly, our study highlighted differences between human and computer interpretation of gestures. For example, where humans understand a clear distinction between 'thumbs up' and 'thumbs down' gestures, these are interpreted the same in our system (due to rotation invariance). The same is true for gestures of similar outer shape, such as 'thumbs up' and 'OK.' While improvements could be made in our prototype, users would benefit from selecting gestures that are shape and rotation unique, rather than human meaning unique.

Thirdly, our sample footage and pilot study demonstrated the variations in lighting and color palette of captured videos. Our technique could be extended such that users could specify a target 'search' color in an example frame. By selecting an example frame from the raw footage, the user could specify their own skin color (thus increasing accuracy given the lighting and color conditions). Further, non-skin color sections could be selected for tracking and comparison, extending our approach to be applicable to gloves or specific equipment details. As an example, an archaeologist could choose whether to search for trowelling by providing a 'troweling-like' action or by selecting the handle color of the trowel as a search criteria.

Due to the continuous capture style of these cameras and the often repetitive nature of the activities recorded, our users benefit from a system supporting both searching and filtering. Our system was not intended to only return correct matches (searching), rather to also highlight similar moments (filtering), thus supporting wider exploration of the footage. This further supports our archaeologist, above, allowing them to not only locate exact gestures but also segment all moments of trowelling.

Finally, our sample footage highlighted the frequently social nature of the activities recorded. There is an opportunity for our work to be extended to support the repetition of any captured gesture, not only that of the wearer. In turn, this would allow for a more complex range of queries from a wider audience and thus support wider results and reflection.

CONCLUSION

VideoHandles is a novel search interaction technique for action camera footage which allows users to search through footage by repeating actions performed during the original recording. *VideoHandles* allows real-time tagging and categorizing of data, thus reducing time spent on post-processing, whilst facilitating wider exploration of recorded footage by supporting comparison between search matches. Our technique also supports a range of usage methods, allowing for both retrospective searching through memory of actions or prospective marking of footage with specific gestures during the initial capture.

We have highlighted and explored the style of footage captured by the action-camera community, described the implementation of our prototype system and explored the feasibility of its use in one study in the wild where our prototype has demonstrated promising initial results.

REFERENCES

1. Baran, J., Gauch, J., Motion Tracking in Video Sequences Using Watershed Regions and SURF Features, in *Proc. Of SE*, 2012.
2. Barrow, HG., et al, Parametric correspondence and chamfer matching: Two new techniques for image matching, in *Proc. 5th Int. Joint Conf. AI*, 1977.
3. Cook, SW, Yip, TK, Goldin-Meadow, S., Gesturing Makes Memories that Last, *Journal of Memory and Language,* 1996.
4. Endres, D., et al., Emulating human observers with Bayesian binning: Segmentation of Action Streams, in *Trans. A.P.,* 2011.
5. Flickner, M. et al, Query by Image and Video Content: the QBIC System, *Computer 28.9*, 1995.
6. GoPro Action Cameras , http://gopro.com/
7. Halvey, M., Joemon, MJ., The role of expertise in aiding video search, in *Proc. CIVR*, 2009.
8. Hanlu, L., et al, Contour Cue Based Particle Filter for Monocular Human Motion Tracking, in *VRCAI*, 2010.
9. Hürst, W., Interactive Audio-Visual Video Browsing, in *Proc. Of MM*, 2006.
10. Li, C., Kitani, K., Pixel-level Hand Detection in Ego-Centric Videos, in *Proc. Of CVPR,* 2013.
11. Matejka, J., Grossman, T., Fitzmaurice, G., Swift: Reducing the Effects of Latency in Online Video Scrubbing, in *CHI*, 2012.
12. Shim, JC., Dorai, C., Bolle, R. Automatic Text Extraction from Video for Content-based Annotation and Retrieval, in *PR,* 1998.
13. Tian, X. et al, Bayesian Video Search Reranking, in *MM,* 2008
14. Wobbrock, JO., Wilson, AD., Li, Y., Gestures without Libraries, Toolkits or Training: a $1 Recognizer for User Interface Prototypes, *Proc. UIST*, 2007.
15. Yuan, J., Tian, Q., Ranganath, S., Fast and Robust Search Methods for Short Video Clips from Large Video Collection, in *Proc. Of PR,* 2004.
16. Zhang, D., Lu, G., A comparative study on shape retrieval using Fourier descriptors with different shape signatures, in *Proc. Of ICIMADE*, 2001.
17. Zhong, D, Chang, SF., Spatio-temporal Video Search Using the Object Based Video Representation, in *Proc of IP,* 1997.