Int. J. Human-Computer Studies 77 (2015) 10–22

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Evaluating visual query methods for articulated motion video search ☆

Cecilia Mauceri [a], Evan A. Suma [b], Samantha Finkelstein [c], Richard Souvenir [d,*]

[a] Department of Computer Science, University of Illinois at Urbana-Champaign, United States
[b] Institute for Creative Technologies, University of Southern California, United States
[c] Human-Computer Interaction Institute, Carnegie Mellon University, United States
[d] Department of Computer Science, University of North Carolina at Charlotte, United States

ARTICLE INFO

Article history:
Received 27 October 2014
Received in revised form
3 December 2014
Accepted 30 December 2014
Communicated by Andrew Howes
Available online 12 January 2015

Keywords:
Video search
Action recognition
Sketch-based interface

ABSTRACT

We develop and evaluate three interfaces for video search of articulated objects, specifically humans performing common actions. The three interfaces, (1) a freehand interface with motion cues (e.g., arrows), (2) an articulated human stick figure with motion cues, and (3) a keyframe interface, were designed to allow users to quickly generate motion-based queries. We performed both quantitative and qualitative analyses of the interfaces through a formal user study by measuring accuracy and speed of user input and asking the users to complete a free-response questionnaire. Our results indicate that the constrained interfaces outperform the freehand sketch-based interface, in terms of both search accuracy and query completion time. Additionally, the users described strong preferences for the search interfaces containing pre-defined models, and the generated queries were rated higher, in terms of semantic matches to the query concept.

1. Introduction

With the proliferation of video capture devices and inexpensive, large-scale storage, video data is increasingly being aggregated for both entertainment and analytic (e.g., athletics, surveillance, medical) purposes. Developing efficient and robust methods for searching large video repositories is an ongoing challenge. Commercially available solutions (e.g., Google Video) generally match text queries to video metadata (e.g., keywords, title). Searching for data in these repositories typically requires a large investment of manual effort in either annotation or real-time observation, and the possibility of incomplete or incorrect metadata is a well-known limitation (Carson and Ogle, 1996). Even with extensive, accurate annotation, it is still difficult to capture all of the semantic information contained in even short video clips. A number of approaches (e.g., Suma et al., 2008; Chang et al., 1997) focus on non-textual input, or visual queries, for searching video. These approaches not only hold the promise of avoiding the database annotation step required for text-based matching, but also introduce new challenges that cut across multiple areas of computing, including video processing, data representation, and interface design.

Many of the domains in which repositories of data are stored, such as athletics or surveillance, contain video of human activity and would benefit from new methods for video search that accelerate the process of locating relevant videos, potentially aiding in physical therapy training or identifying specific security footage of interest. Therefore, in this paper, we focus on the problem of searching for video clips of humans performing common actions. We design and evaluate three different interfaces for generating visual queries. The first interface follows the sketch-based input paradigm, where the user can draw a stick figure with action arrows to indicate motion. The second interface extends the first by providing a pre-defined template of an articulated human figure (stick figure) for the user to pose and also using action arrows as motion cues. The third interface re-uses the pre-defined template, but avoids the use of motion cues; instead it defines a sequence of poses to represent the visual video query. Fig. 1 shows examples of each interface. These three interfaces span a range of approaches that are applicable to the typical keyboard–video–mouse interface, and also can be applied to touch interfaces found on smartphones and tablets. In order to evaluate the effectiveness of the different visual query interfaces, we conducted a formal user study where we measured the query generation time and accuracy of the resulting search in terms of the number of highly ranked results matching the search concept. Additionally, the users provided feedback on the positives and negatives of each interface through a post-experiment survey.

2. Related work

The literature on automated methods for content-based visual information retrieval (CBVIR) is extensive; see Lew et al. (2006) and

☆ This paper has been recommended for acceptance by Andrew Howes.
* Corresponding author.

E-mail addresses: mauceri2@illinois.edu (C. Mauceri), suma@ict.usc.edu (E.A. Suma), slfink@cs.cmu.edu (S. Finkelstein), souvenir@uncc.edu (R. Souvenir).

http://dx.doi.org/10.1016/j.ijhcs.2014.12.009
1071-5819/© 2015 Elsevier Ltd. All rights reserved.

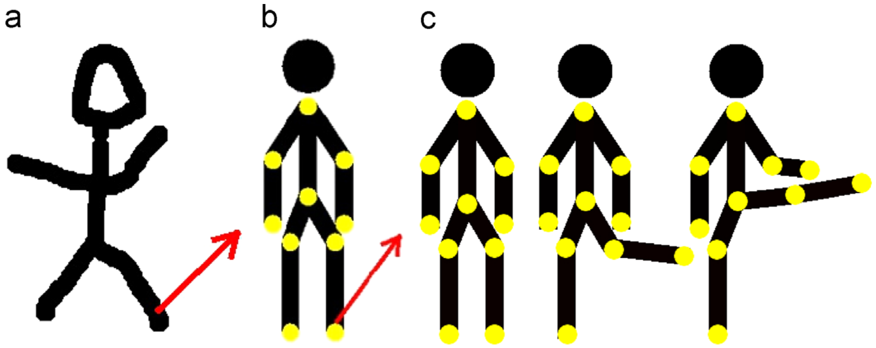

Fig. 1. The three interfaces for generating visual queries provide different methods of representing human motion. In this case, the figures show a user depiction of a kick from each interface.

Marchand-Maillet (2000) for surveys. This body of work includes both methods for image and video search, and various paradigms, such as text-based searching (e.g., Naphade and Huang, 2000; Zha et al., 2009) or search by example (e.g., Taskiran et al., 2004). With text-based approaches, which match the query to metadata associated with images or video, the quality of the retrieval results is a function of the quantity and quality of the annotations. The relative success of image search methods (e.g., Google Images) has not been reached for video search due to the complexity needed to describe even the simplest of videos. Example-based approaches can overcome the limitations of ambiguous searches because a query video is generally more informative than a text label. However, finding representative videos to use for querying other videos can be difficult. Beyond text or example videos are approaches that extend beyond the keyboard and utilize other common user interfaces, such as the mouse or pen. The dominant paradigm of this class is sketch-based approaches, which have been widely adapted. In this section, we focus specifically on methods that use visual queries for image and video search.

Sketches have been used as the underlying model for a number of applications including image matching (Tang et al., 2003; Cao et al., 2011; Wang et al., 2010), queries for GIS data (Egenhofer, 1997), face recognition (Uhl Jr. and da Vitoria Lobo, 1996), and object relationships (Zitnick and Parikh, 2013). Searching video databases using sketches has also previously been explored. Collomosse et al. (2008) have explored the types of sketches a user might produce for a video retrieval system. Other methods (Jacobs et al., 1995; Lew, 2000) use a sketch-based system to search a large static image database. One interface (Fonseca et al., 2012) allows a user to provide a skeleton without motion cues as input and searches video databases for keyposes, rather than video clips. Other related systems (Chang et al., 1997; Collomosse et al., 2009; Hu et al., 2012) query video database using sketches with motion cues, but mainly provide for queries focusing on single object translation. Unlike the interfaces used for our evaluation, these related approaches are not designed to search for the finer-grained articulated motions of humans.

An area related to visual query construction is sketch interpretation (Paulson and Hammond, 2008a). This includes sketch-based modeling (Olsen et al., 2009; Bernhardt et al., 2008; Igarashi and Hughes, 2003; Schmidt et al., 2005), using a 2D sketch to inform the pose of a 3D model (Vaidya et al., 2006), and sketch beautification (Paulson and Hammond, 2008b). However, unlike these methods, using sketching for video search does not require explicit understanding of the sketch input.

The number of examples of existing methods for sketching for video retrieval demonstrates the interest this approach. Compared to text-based or example-based approaches, sketching requires neither database annotation nor pre-collected examples. A downside, however, is that the quality of a query is directly related to the skill of the user in sketching an input that corresponds to the search concept and the extent to which the interface helps the user perform their sketch by providing them with appropriate feedback to improve performance. In this paper, we evaluate three different visual query interfaces, with varying amounts of "sketching" required for the problem of searching videos for common human actions. The first interface, which is an updated version of an existing approach (Suma et al., 2008), follows the commonly described freehand sketch based paradigm, while the other two interfaces provide more guidance (and, thus, less freedom) to the user.

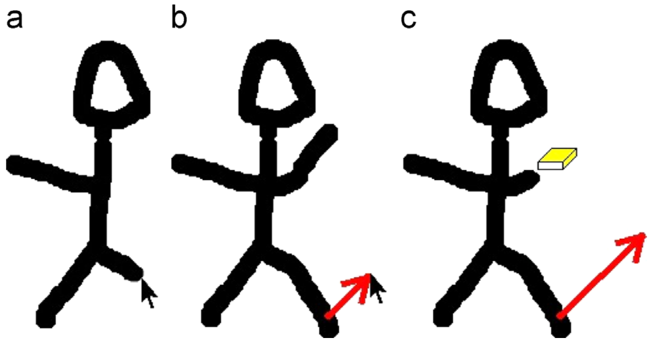

Fig. 2. The Freehand interface provides three tools (pen, arrow, eraser) for generating visual queries.

3. Input interfaces and motion inference

To ground the evaluation, we developed three interfaces for generating visual queries for human actions in video. While the specific interface components (e.g., feature transform and matching algorithm described in Section 4.2) are not the focus of this work and could be replaced with other methods, they serve as means to allow a comparison of the three interface styles for searching for articulated motions. Fig. 1 shows an example of each interface: (1) Freehand (Fig. 1(a)) allows the user to freely sketch an object and add arrows to indicate motion, (2) Template (Fig. 1(b)) provides an articulated human figure that the user can pose using drag gestures on the joints and, similar to Freehand, add arrows for motion, and (3) Keypose (Fig. 1(c)) allows the user to define a series of poses using the same articulated human figure as the Template interface. For each interface, our system interprets the visual query, animates it, and compares the generated video to a database of existing videos. In this section, we describe each interface in detail and explain how motion is inferred.

3.1. Interface #1: freehand sketching

The Freehand interface allows the user to sketch an articulated object using drag gestures. Three tools are provided: pen, arrow, and eraser. The pen tool (Fig. 2(a)) is used to define the figure.