# Characterizing User Behavior for Speech and Sketch-based Video Retrieval Interfaces

Ozan Can Altıok
Koç University
Graduate School of Sciences and Engineering
Rumelifeneri Yolu, Sarıyer
Istanbul, Turkey 34450
oaltiok15@ku.edu.tr

Tevfik Metin Sezgin
Koç University
Graduate School of Sciences and Engineering
Rumelifeneri Yolu, Sarıyer
Istanbul, Turkey 34450
mtsezgin@ku.edu.tr

## ABSTRACT

From a user interaction perspective, speech and sketching make a good couple for describing motion. Speech allows easy specification of content, events and relationships, while sketching brings in spatial expressiveness. Yet, we have insufficient knowledge of how sketching and speech can be used for motion-based video retrieval, because there are no existing retrieval systems that support such interaction. In this paper, we describe a Wizard-of-Oz protocol and a set of tools that we have developed to engage users in a sketch- and speech-based video retrieval task. We report how the tools and the protocol fit together using "retrieval of soccer videos" as a use case scenario. Our software is highly customizable, and our protocol is easy to follow. We believe that together they will serve as a convenient and powerful duo for studying a wide range of multi-modal use cases.

## CCS CONCEPTS

**•Human-centered computing → Systems and tools for interaction design;** *Empirical studies in HCI;*

## KEYWORDS

sketch-based interfaces, human-centered design, motion, multimedia retrieval

**ACM Reference format:**
Ozan Can Altıok and Tevfik Metin Sezgin. 2017. Characterizing User Behavior for Speech and Sketch-based Video Retrieval Interfaces. In *Proceedings of Expressive '17, Los Angeles, CA, USA, July 29–30, 2017,* 2 pages.
DOI: 10.1145/3122791.3122801

## 1 INTRODUCTION

Video retrieval has been evolving from text-based retrieval into content-based retrieval. In the human computer interaction front, this evolution has triggered use of natural interaction modes to articulate search queries. Sketching and speech are two modalities that complement each other to describe a video. Sketching indicates placement of objects together with their motion path. Through speech, users can indicate events and their relationships. However, we do not have adequate information about how speech and sketching can be used for video retrieval since there are no existing video retrieval systems based on these modalities. To gather information regarding the joint use of these modalities, a protocol must be designed to capture participants to talk and draw simultaneously. Also needed are some applications to collect speech and sketching data and to process these data for later use.

This paper presents a protocol together with a set of software to stimulate participants to draw and talk in a video retrieval task. We tested the protocol and the software on soccer match retrieval use case scenario. Our software can be modified for use in different use cases of video search.


*Expressive '17, Los Angeles, CA, USA*
 978-1-4503-5174-4/17/07…$15.00
DOI: 10.1145/3122791.3122801

## 2 RELATED WORK

There are a few studies that have investigated the use of speech and sketching. Adler et. al. have designed a study [Adler and Davis 2007a] to observe how people use speech and sketching simultaneously and to finally implement a multi-modal white-board application based on these observations. They asked students from circuit design class to talk about and draw a floor plan for a room, an AC/DC transformer, a full adder, and the final project they had developed for their circuit design class. Later, Adler et. al. have designed another study [Adler and Davis 2007b] for collecting speech and sketched symbols to develop a multi-modal mechanical system design interface. He asked six participants to talk about and draw a pendulum system on a white board.

Although these studies investigate the joint use of speech and sketching, they did not have multimedia retrieval in focus. Our protocol and a set of tools are focusing on the investigation of collaboratively using speech and sketching for multimedia retrieval.

## 3 WIZARD-OF-OZ PROTOCOL

In each session, the participant and the experimenter sit across a table (see Figure 1). Initially, the experimenter provides a set of sample sketched symbols describing the visual vocabulary for constructing drawings. To practice use of the visual vocabulary, the experimenter opens a video clip, and requests the participant to describe the movement of objects in the clip. Next, they move on to the search tasks. In the search tasks, there are different motion events to describe through speech and sketching. For each motion event, the experimenter picks a video clip of the event randomly, and plays the clip on the participant's screen 3 times. Then, the participant draws the motion of objects in the clip on her tablet while talking at the same time. The experimenter then looks at videos of the event on his screen, finds a video that is similar to the scenario drawn and uttered by the participant, and plays the

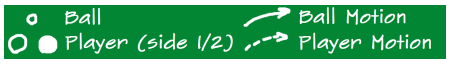


**Figure 2: Visual vocabulary for drawings. The vocabulary was prepared by referring to [Bangsbo and Peitersen 2000].**

| | | | |
|---|---|---|---|
| **Ball** | 88 | **Ball Motion** | 313 |
| **Player Motion** | 250 | **Player (side 1/2)** | 480/396 |

**Table 1: Distribution of sketched symbols among classes**

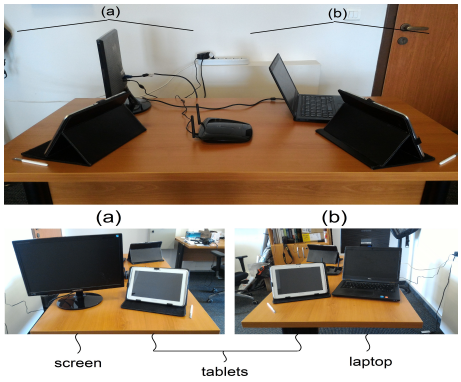


**Figure 1: Physical set up; (a) for the participant, (b) for the experimenter**

video on the participant's screen 3 times. Finally, the experimenter asks whether the video is the one presented at the beginning or not. If the participant says yes, they proceed to the next motion event. Otherwise, the participant tries to produce more accurate description of the scene to be retrieved by improving the drawing and the speech, and the retrieval process repeats.

## 4 SOFTWARE TOOLS

The applications we developed for video retrieval tasks are presented in the subsections below.

### 4.1 Data Collection

We developed two applications running on Android tablets, one for the experimenter [Altıok 2017b] and one for the participant [Altıok 2017c]. These applications use a shared canvas, that is, whatever gets drawn in one tablet is replicated in another tablet and is also written to a text file on both sides for later use. These applications communicate with each other over a Wi-Fi network to replicate the drawings. In addition, these applications record their users' speech using built-in microphones of the tablets.

### 4.2 Video Search Front-End

Another application serves as the video search front-end for the participant [Altıok 2017a]. When the participant finishes describing the motion in the video clip presented, the experimenter can navigate through all video clips using this application to find the clip with the motions described. Additionally, for each motion event, the application allows the experimenter to randomly play a video clip on the participant's screen 3 times.

### 4.3 Sketched Symbol Annotator

We implemented an application to select and annotate individual sketched symbols drawn in sessions [Altıok 2017d]. It allows users to select and annotate symbols in drawings, and next to save each symbol as a list of points and an image.

## 5 OUTCOME OF THE USE CASE DEMONSTRATION

We conducted sessions with 7 soccer fans, 6 soccer players, 3 soccer coaches and 2 soccer referees. Altogether, the sessions took 661 minutes and 24 seconds. Average duration of a session was 36 minutes and 44 seconds. More information about the data collected is presented in the next subsections.

### 5.1 Sketched Symbols

We collected 1527 symbols in the sessions. The visual vocabulary used in the demonstration is presented in Figure 2. Distribution of the symbols among classes is given in Table 1.

### 5.2 Utterances

We also collected 9296 words in the sessions. We collected minimum of 233 words and maximum of 925 words in a session. On the average, we obtained 516.4 words in a session.

## 6 CONCLUSION

We presented a protocol and a set of software for investigating joint use of speech and sketching for motion-based video retrieval. The protocol and the software were also demonstrated on soccer match retrieval use case. The amount of data collected in the demonstration implies that the protocol design and the applications make a good couple for user engagement in motion-based video retrieval tasks. In the long run, we believe that our tools and protocol will open way for further investigation of the hybrid use of speech and sketching in a wide spectrum of multimedia retrieval use cases.

## ACKNOWLEDGMENTS

This work was supported by the CHIST-ERA project iMotion with contributions from the Scientific and Technological Research Council of Turkey (TÜBITAK, grant no. 113E325).

## REFERENCES

Aaron Adler and Randall Davis. 2007a. Speech and sketching: An empirical study of multimodal interaction. In *Proceedings of the 4th Eurographics workshop on Sketch-based interfaces and modeling*. ACM, 83–90.

Aaron Adler and Randall Davis. 2007b. Speech and sketching for multimodal design. In *ACM SIGGRAPH 2007 courses*. ACM, 14.

Ozan Can Altıok. 2017a. Ad-Hoc Video Search Application for Multi-Modal Data Collection. (2017). https://github.com/ozymaxx/multimodal_datacollection_adhocsearch

Ozan Can Altıok. 2017b. Multi-Modal Collector (Experimenter's Side). (2017). https://github.com/ozymaxx/multimodal_collector_experimenter

Ozan Can Altıok. 2017c. Multi-Modal Collector (Participant's Side). (2017). https://github.com/ozymaxx/multimodal_collector

Ozan Can Altıok. 2017d. Sketch Annotator for Multi-Modal Collector. (2017). https://github.com/ozymaxx/multdatacol_sketch_annotator

Jens Bangsbo and Birger Peitersen. 2000. *Soccer systems and strategies*. Human Kinetics.