

Creating and Evaluating a Video Vocabulary for Communicating Verbs for Different Age Groups

Xiaojuan Ma
Dept. of Computer Science
Princeton University, NJ USA
xm@cs.princeton.edu

Perry R. Cook
Dept. of Computer Science (also Music)
Princeton University, NJ USA
prc@cs.princeton.edu

ABSTRACT

Icons and digital images used in augmentative and alternative communication (AAC) are not as effective in illustrating verbs, especially for people with cognitive degeneration or impairment. Realistic videos have possible advantages for conveying verbs, as verified in our studies with young and old adults comparing single image, multiple images, animations, and video clips. Videos are especially more effective for verbs that show concrete movements or actions. Based on our studies, we propose rules for filming video verb representations, exploring possible visual cues and other factors that may affect people's perception and interpretation.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – *Videos, Animations, Artificial, augmented, and virtual realities.*

General Terms

Human Factors, Languages.

Keywords

Visual communication, verbs, aging effects, video, AAC

1. INTRODUCTION

Augmentative and alternative communication (AAC) introduces non-verbal facilities to augment/replace normal written or spoken language to enhance communication for particular populations, such as people who suffer from language impairment, with low literacy, and who speak different languages. For example, Lingraphica [8] is a commercial laptop device that assists people with aphasia (a language disability caused by a stroke or brain injury [10]), where users compose sentences from a given set of words represented in a triplet of text, icon, and sound.

Nouns are names for concrete or abstract concepts, while verbs convey ideas associated with actions, status, and process. Though there are fewer verbs than nouns in the general vocabulary [4][9], verbs contain richer messages, since the ratio of sense to unique word form is 1.19 for verbs compared to 0.70 for nouns. Also, according to statistics drawn from the British National Corpus (BNC) [3], there are more verbs (21) than nouns (5) in the top frequently used word list (over 1000 occurrences) in both written and spoken English. Unlike static nouns, most verbs have a sense of time, (sequence of movements, changes of status, progress in

process), making verbs hard to capture in a single frame. Realizing the importance of verbs as well as the difficulty in visually illustrating them, Lingraphica created animated icons for verbs in particular. As the need to expand a vocabulary increases, the strategy of artist-designed icons for words is challenged because of time and high cost. Thus, Lingraphica allows users to upload photos and create thumbnails as new visual entries. This method serves well for nouns, but not as well for verbs. We propose videos, a flow of frames of static images, for more effective visual representations of verbs.

In order to test the use of videos in AAC, we evaluated their efficacy in illustrating verbs compared to other visual representations. We used one static image, a panel of four static images, and an animated icon as competitors for videos (Figure 1), since images can be generated by importing photos from a camera, and animated icons are what some other systems use. Comparisons across visual modes in previous research were usually between icons and images or icons and animations [2], and the focus was on nouns or action verbs only [5].

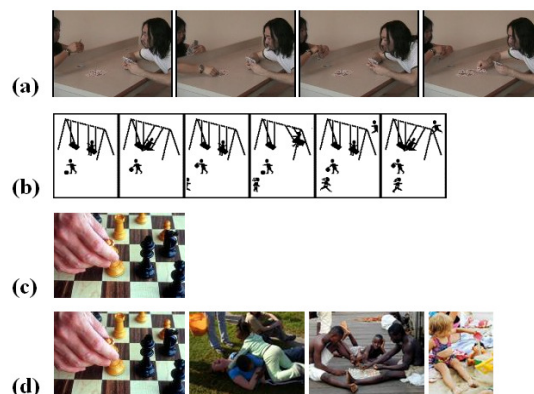


Figure 1: (a) Video frames (b) Animation frames (c) One static image (d) and Four static images for the verb “play.”

2. VERB LIST, IMAGES, AND VIDEOS

Focusing on AAC for daily life, we selected 48 verbs from the BNC according to frequency in spoken language such as conversations [7]. The initial 60 verbs came from sorting all verbs in different forms in frequency descending order. Then, with a linguist and a speech-language pathologist, we combined verbs with similar senses (“tell” and “say”), removed uncommon spoken American English verbs (“reckon”) and the two verbs “know” and “like” (often used as in “like,” “you know”), and came up with our final list of 48 verbs¹.

¹ See <http://www.cs.princeton.edu/aphasia/48verbs.txt>

Single-images and pictures for four-image-panels came from online public domain resources. Image search engines like Google Image Search [6] are not ideal for this purpose since they retrieve images based on surrounding text instead of image content. The top four images were selected based on the ranking of seven raters. Most animations came from Lingraphica, with a few new ones created using the same style guide. Existing computer vision video databases only depict actions such as waving and running, and are too limited to satisfy communication needs. Online video resources [11] cover a wider range of everyday scenes, but most are of low quality. Thus we decided it was necessary to shoot and edit our own verb videos². Five people each wrote a short script for each intended verb sense, describing how it could be acted out. Four other people voted for the best scripts. To minimize perceptual workload for participants, we adopted the following rules: single leading actor; extra hands/feet or supporting actress when interaction required; minimal use of props; blank white background; no sound or text, no lip movement so subjects only interpret the scene. Each video clip was edited to the length of 3.2 ± 0.05 seconds. Effects like fast forwarding were applied to some verbs (“work,” “make”) to address the length restriction.

3. EXPERIMENTS

The first study compared the effectiveness of video versus the other three modes with 16 younger participants between 20 and 39 and 16 older participants over 50. Subjects were asked to name single verbs based on the visual mode given, and to justify their responses with a simple explanation. The 48 verbs were divided evenly into four blocks so as to align to the four visual modes. The second study with 25 people of mixed ages explored how videos convey verbs in sentence contexts. The 65 sentences tested in the study were crawled from senior citizen blogs (Ageless Project [1]). All 48 verbs appeared twice in different sentences. Sentences were divided evenly into five blocks: the four visual modes as well as an additional baseline mode, which removes the target verbs from the sentences. Subjects were asked to interpret the whole sentence word by word. At the end of each study, participants were asked a set of questions related to their preferences and ease of use.

4. RESULTS AND CONCLUSIONS

We computed % correct, response diversity, and WordNet [4] score from the raw data. The first study showed that videos performed significantly better than all other modes for older subjects, while for younger subjects the three multi-frame modes were significantly better than the single image mode (Figure 2). However, such differences were decreased with the help of context (Figure 3). Video mode outperformed other representations in general, and showed clear advantages in conveying contact and motion verbs.

It is important, however, to address the weakness of visual representations to improve the representation of verbs. Based on subjects’ feedback, we looked into possible influences introduced by three visual cues: **gestures**, **facial expressions**, and **symbols** in animations (related symbols like ? with “ask” and ♥ with “love” or indirect symbols like ? with “wonder” and ♥ with “want”). **Gestures** (the classic thinker pose, waving for “leave”) had a positive though not strong influence. Video mode suffered a

slight (not significant) drop in accuracy when the actor displayed observable **facial expression**. Verbs with **indirect symbols** were interpreted poorly in all modes.

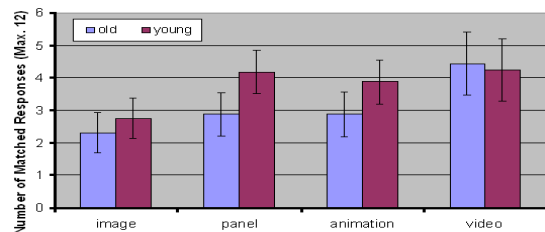


Figure 2: 95%CI of age-related effect.

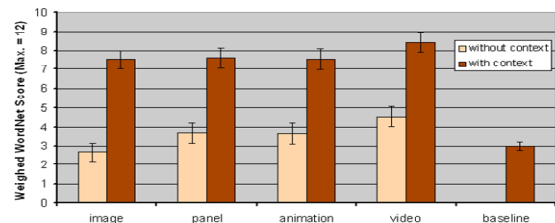


Figure 3: 95%CI with and without context.

From our observations as well as participants’ comments afterwards, we list other causes for failures in perceiving intended verbs. These include **Distraction** (from text in images, background objects, and other sources), **Misapprehension** (consistent wrong answers across a subset of subjects), lack of **Familiarity** (not recognizing an object, situation, or symbol), and **Imagination** (users concocting a story that leads them off track).

We feel our analyses can guide designers in the creation or selection of visual representations for verbs, aiming for low cost and high effectiveness in conveying concepts.

7. ACKNOWLEDGMENTS

We thank Ge Wang, Sonya Nikolova, Jordan Boyd-Graber, Peter Graf, Joanna McGrenere, Rock Leung, and PU SoundLab for help in developing videos and studies. Also the Princeton Senior Resource Center for subjects.

8. REFERENCES

- [1] Ageless Project, 2001. <http://jenett.org/ageless>.
- [2] Baecker, R. Small, I. and Mander, R. Bringing icons to life. *Human Factors in Computing Systems*, ACM, 1991.
- [3] British National Corpus <http://www.natcorp.ox.ac.uk>.
- [4] Fellbaum, C. *WordNet: Electronic Lexical Database, A semantic network of English verbs*. MIT Press, 1998.
- [5] Fiez, J. and Tranel, D. Standardized stimuli procedures for investigating the retrieval of lexical and conceptual knowledge for actions. *Memory & Cognition*, 25, 1997.
- [6] Google Image Search. <http://images.google.com/>.
- [7] Kilgariff, A. BNC database and word frequency lists. <http://www.kilgariff.co.uk/bnc-readme.html>.
- [8] Lingraphica. <http://www.lingraphicare.com/>.
- [9] Miller, G. Nouns in WordNet: A lexical inheritance system. *Intl. Journal of Lexicography*, 3(4), 1990.
- [10] National Aphasia Association. <http://www.aphasia.org>.
- [11] YouTube. <http://youtube.com>.

² See <http://www.cs.princeton.edu/aphasia/wmv>