Upper Body Gestures in Lecture Videos: Indexing and Correlating to Pedagogical Significance

John R. Zhang
Department of Computer Science
Columbia University
New York, NY 10027
jrzhang@cs.columbia.edu

ABSTRACT

The growth of digitally recorded educational lectures has led to a problem of information overload. Semantic video browsers present one solution whereby content-based features are used to highlight points of interest. We focus on the domain of single-instructor lecture videos. We hypothesize that arm and upper body gestures made by the instructor can yield significant pedagogic information regarding the content being discussed such as importance and difficulty. Furthermore, these gestures may be classified, automatically detected and correlated to pedagogic significance (e.g., highlighting a subtopic which may be a focal point of a lecture). This information may be used as cues for a semantic video browser.

We propose a fully automatic system which, given a lecture video as input, will segment the video into gestures and then identify each gesture according to a refined taxonomy. These gestures will then be correlated to a vocabulary of significance. We also plan to extract other features of gestures such as speed and size and examine their correlation to pedagogic significance. We propose to develop body part recognition and temporal segmentation techniques to aid natural gesture recognition. Finally, we plan to test and verify the efficacy of this hypothesis and system on a corpus of lecture videos by integrating the points of pedagogic significance as indicated by the gestural information into a semantic video browser and performing user studies. The user studies will measure the accuracy of the correlation as well as the usefulness of the integrated browser.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; D.2.8 [**Content Representation**]

General Terms

Algorithms

MM'12, October 29–November 2, 2012, Nara, Japan.

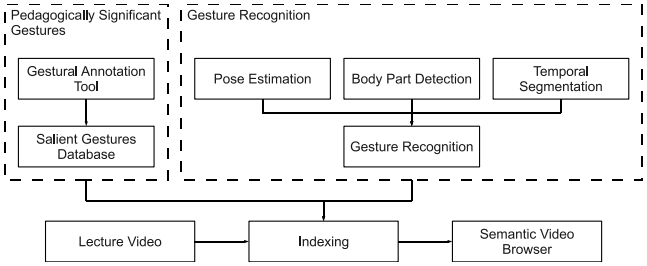

Figure 1: A high level overview of the proposed system. Video lectures are taken as input and gestures (predetermined to be semantically relevant) are identified and indexed. Finally, the results are combined into a semantic video browser for users.

Keywords

gesture, pose, pedagogy, lecture, body model

1. INTRODUCTION

The increasing online availability of digitally recorded video has led to a problem of information overload. Automatically determining the relevance of segments of video through the use of multimedia features present one solution that would reduce the search space for human users seeking informational content. Toward this goal, we study the correlation between semantic relevance and gestures in the domain of single-instructor lecture videos, although results could be directly applied to any single-speaker presentation video such as television journalism, talk shows (both guest and host) and interviews of public figures. We hypothesize that arm and upper body gestures made by the speaker can yield significant pedagogic information regarding the content being discussed such as importance and difficulty. Furthermore, these gestures may be classified, automatically detected and correlated to pedagogic significance (e.g., highlighting a subtopic which may be a focal point of a lecture).

Our proposed system will involve three parts as illustrated in Figure 1: the study and collection of pedagogically significant gestures, a subsystem for the recognition of the aforementioned gestures in the input video and indexing them, and finally displaying the results of the indexing in a useful manner in a video browser such as [5].

This paper is organized as follows. In Section 2, we discuss work related to our area of interest. In Section 3, we describe the progress we have made so far in this research

direction. In Section 4 we propose solutions for approaching unmet challenges. Finally, we conclude in Section 5 with a discussion of other possible applications of our work.

2. RELATED WORK

We briefly review the existing literature exploring the representation of gestures, the semantic relevance of gestures and automatic pose and gesture recognition as our proposed work lies at the intersection of these fields.

Three prominent schemes for the representation of gestures are: 1) FORM [8], which represents gestures as transitions between timestamped nodes on a graph, 2) ANVIL [6], which was designed for maximum extensibility and represents gestures as time-anchored hierarchical attribute-value pairs, and 3) CoGesT [4], which represents gestures according to a number of properties including phase, hand shape and hand symmetry. We assert that due to the difference in domain, the proposed schemes and tools are insufficient for our purposes and justify the need for further development, with inspiration taken from the literature.

We also reviewed literature examining the relationship between instructors' gestures during lectures and semantic significance. Linguistics studies by Bavelas et al. [1] have demonstrated a correlation between gestures and meanings in communication (such as turn-taking or seeking a response). Roth et al. [11] examined gestures in education and demonstrated that instructors' upper body and arm gestures affect the learning of students. Many of these studies build upon the taxonomies proposed by McNeill [9], whereby gestures are divided into five classes: iconics, metaphorics, beats, cohesives and deictics. The proposed taxonomies were designed from a linguistic and educational perspective and are insufficiently robust for automatic recognition.

Finally, we review the state of the art for gesture and pose recognition in computer vision. The iterative parsing method by Ramanan et al. [10] used in conjunction with pictorial structures [3] have been shown to provide robust pose estimation techniques. Robust gesture recognition from video has also been demonstrated by Buehler et al. [2] in the domain of sign language recognition. Many of these methods are robust and will be applied to our work. We will explore alternate approaches in some areas such as the evaluation of temporal segmentation and the indexing of gestures. We will further work in the area of natural gesture recognition, which is far less constrained than gesture recognition for control or sign language.

3. RESEARCH PROGRESS

In an initial study we gathered ground truth annotations of gesture appearance using a 27-bit pose vector. We manually annotated and analyzed the gestures of two instructors, each in a 75-minute computer science lecture yielding 866 gestures and identifying 126 fine equivalence classes which we further divide into 9 semantic classes. We observed these classes encompassing pedagogical gestures of punctuation and encouragement, as well as traditional classes such as deictic and metaphoric. We introduced a tool with an avatar-based annotator (Figure 2) to facilitate the manual annotation of gestures in video, and presented initial results on their frequencies and co-occurrences; in particular, we found that pointing (deictic) and spreading (pedagogical) predominate, and that 5 poses represented 80% of the variation in the collected ground truth. This work was published and presented at CVPR4HB in 2009 [12].

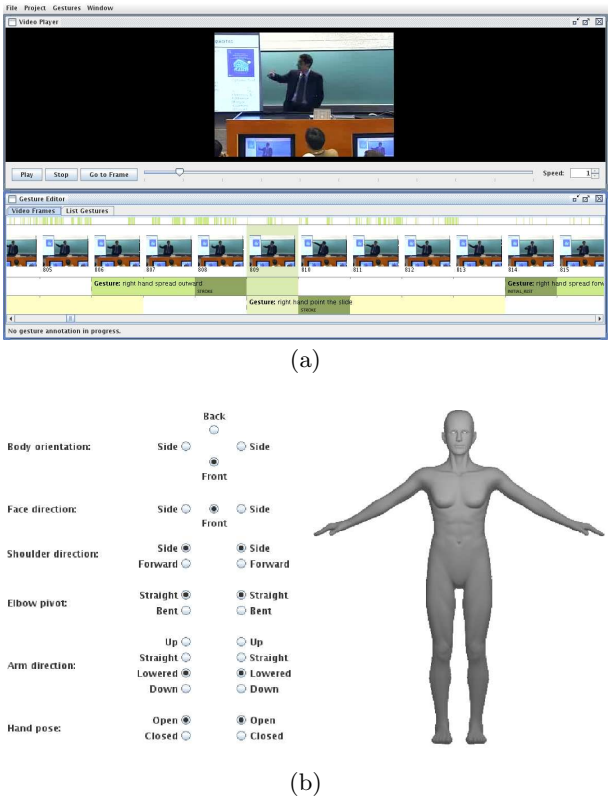

Figure 2: (a) The main user interface of the gesture annotator tool. (b) The avatar poser controls in the default configuration, along with the corresponding avatar preview image.

We also performed preliminary work on automatically classifying poses belonging to the point and spread classes (Figure 3 shows examples as well as automatically estimated poses) as a preliminary step toward their identification in video. We use a joint-angle descriptor derived from an automatic pose estimation framework to train an SVM in order to classify extracted video frames of an instructor giving a lecture. Ground-truth is collected in the form of 2500 manually annotated frames covering approximately 20 minutes of a video lecture. Cross validation on the ground-truth data showed initial classifier F-scores of 0.54 and 0.39 for point and spread poses. This work was published and presented at ICIP in 2010 [13].

We have also attempted to correlate content with the degree of variation of arm gestures. Intuitively, we demonstrated that the extent of "arm waving" correlated to inflection points in presentation videos as indicated by the presence of a class of contrasting conjunctions in corresponding subtitles. We experimented on 4243 video clips, each with time-synced subtitles and averaging 3.25 seconds long. The task was framed as a binary classification problem, with positive labels assigned to those video segments with subtitles intersecting a given class of conjunctions (negative, otherwise). Arm gesture variation were used as features, and were derived from the circular variance of the orientations

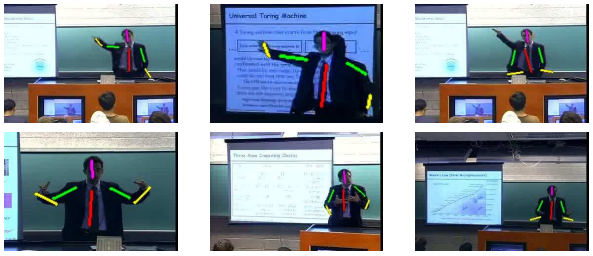

Figure 3: Examples of point poses (top row) and spread poses (bottom row) with automatically estimated poses overlayed.

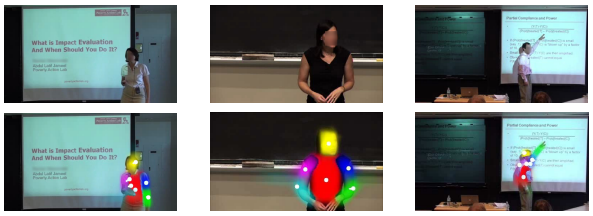

Figure 4: Examples of pose estimations (i.e., estimating the position and orientations of the head, torso, upper and lower left and right arms, which are shaded) and part-based optical flows (visualized by the white arrows from the centroids of each part). The originals are shown in the top row while the parts and flows are visualized in the bottom row. The rightmost column shows an example of a poor pose estimation.

of part-wise image flow across frames in a video segment, as shown in Figure 4. An AdaBoost-based classifier was trained and tested using 4-fold cross validation on multiple sets of conjunctions, including: a set of 45 commonly used conjunctions and a subset of 11 conjunctions which indicated contrasts in discourse. Our experiments showed that the classifier performed with a classification accuracy of 0.508 on the class of all conjunctions (no better than chance) and 0.549 on the class of contrasting conjunctions indicating a slight but existent correlation. This work has been submitted for peer-review.

4. PROPOSED WORK

We propose to gather more gestural data with regards to both appearance as well as correlation to pedagogic significance. The former can be done on a large scale e.g., via Amazon Mechanical Turk. We will need to address the challenge of crowdsourcing the task of annotation of complex videos, as such a long task would reduce the efficacy of human raters, as we demonstrated in a related work [7]. For the latter, we propose to hire trained raters to evaluate segments of video based on the semantic significance of its content.

The data collected will be used for training and testing our machine learning methods, as well as to refine our taxonomy of gestures. We also propose a fully automatic system which, given a presentation video as input, will segment the video into gestures and then identify each gesture according to a refined taxonomy. One possible approach we intend to explore for the problem of gesture indexing and retrieval is to build on our work for near-duplicate video retrieval based on motion histograms which has been accepted for publication at ICME 2012 [14] and for which a patent is currently pending. Indexed gestures will then be correlated to a vocabulary of significance. We also plan to continue our research into other features of gestures such as speed and size and examine their correlation to pedagogic significance. We propose to develop body part recognition and temporal segmentation techniques to aid natural gesture recognition.

Finally, we plan to test and verify the efficacy of this hypothesis and system on a corpus of informational videos by integrating the points of pedagogic significance as indicated by the gestural information into a semantic video browser and performing user studies.

Additional research directions may be pursued such the creation of gestural profiles for individual speakers based on the pedagogical gesture taxonomy. This would allow us to catalog good speaker techniques, and possibly gain other psychological insights into individual speakers (e.g., such as public figures being interviewed or giving speeches).

5. CONCLUSION

A hypothesis relating upper body gestures to pedagogic significance is presented here, and a system demonstrating the hypothesis in the form of a semantic video browser using gesture information as cues is proposed. We propose work on collecting gesture data from lecture videos, analyzing them for relevant classes of gestures, developing methods for their automatic recognition and indexing, and finally user studies to assess their effectiveness. Applications of this work can be directly applied to any single-speaker informational video such as lectures, news reports, interviews and speeches.

6. REFERENCES

[1] J. Bavelas, N. Chovil, L. Coates, and L. Roe. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, 21(4):394–405, 1995.

[2] P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *Proc. Computer Vision and Pattern Recognition*, 2009.

[3] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 66–73, 2000.

[4] U. Gut, K. Looks, A. Thies, T. Trippel, and D. Gibbon. Cogest âĂŞ conversational gesture transcription system. Technical report, University of Bielefeld, 1993.

[5] A. Haubold and J. Kender. Vast mm: multimedia browser for presentation video. In *Proc. ACM International Conference on Image and Video Retrieval*, pages 41–48, New York, NY, USA, 2007. ACM.

[6] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proc. EUROSPEECH-2001*, pages 1367–1370, 2001.

[7] T. Leung, Y. Song, and J. R. Zhang. Handling label noise in video classification via multiple instance

learning. In *Proc. International Conference in Computer Vision*, Nov 2011.

[8] C. Martell. Form: An extensible, kinematically-based gesture annotation scheme. In *Proc. 3rd International Conference on Language Resources and Evaluation*, 2002.

[9] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University Of Chicago Press, 1992.

[10] D. Ramanan. Learning to parse images of articulated bodies. *Advances in Neural Information Processing Systems*, 19:1129, 2007.

[11] W. Roth. Gestures: Their role in teaching and learning. *Review of Educational Research*, 71(3):365–392, 2001.

[12] J. R. Zhang, K. Guo, C. Herwana, and J. R. Kender. Annotation and taxonomy of gestures in lecture videos. In *Proc. CVPR Workshop on Human Communicative Behavior Analysis*, June 2010.

[13] J. R. Zhang and J. R. Kender. Identifying salient poses in lecture videos. In *Proc. IEEE International Conference on Image Processing*, Sept 2011.

[14] J. R. Zhang, J. Ren, F. Chang, T. Wood, and J. R. Kender. Fast near-duplicate video retrieval via motion time series matching. In *International Conference on Multimedia & Expo, to appear*, July 2012.