Annotation and Taxonomy of Gestures in Lecture Videos

John R. Zhang Kuangye Guo Cipta Herwana
John R. Kender
Columbia University
New York, NY 10027, USA
{jrzhang@cs., kg2372@, cjh2148@, jrk@cs.}columbia.edu

Abstract

Human arm and body gestures have long been known to hold significance in communication, especially with respect to teaching. We gather ground truth annotations of gesture appearance using a 27-bit pose vector. We manually annotate and analyze the gestures of two instructors, each in a 75-minute computer science lecture recorded to digital video, finding 866 gestures and identifying 126 fine equivalence classes which could be further clustered into 9 semantic classes. We observe these classes encompassing "pedagogical" gestures of punctuation and encouragement, as well as traditional classes such as deictic and metaphoric. We note that gestures appear to be both highly idiosyncratic and highly repetitive. We introduce a tool to facilitate the manual annotation of gestures in video, and present initial results on their frequencies and co-occurrences; in particular, we find that pointing (deictic) and "spreading" (pedagogical) predominate, and that 5 poses represent 80% of the variation in the annotated ground truth.

1. Introduction

Increasingly, post-secondary institutions have been making their recorded lectures for select courses available online. This increase in the availability of recorded lectures has many positive implications but also leads to additional challenges, including the need to efficiently browse through it.

To this end, much work has been done on developing video browsers which allow users to browse video in a non-linear fashion [3]. Identifying semantically significant cues in video is a multimedia problem which can make use of verbal, audible and visual signals [9] [10].

In this paper, we begin the work to explore the feasibility of using the arm, head and upper body gestures of the instructors in video lectures as semantic clues. That is, we attempt to collect statistics and identify patterns in the gestures of the instructors to see how they relate to the material they are teaching and the structure of the lecture itself. Significant correlations could lead to the incorporation of the data into existing non-linear video browsers such as Vast MM [3]. For example, gestures of encouragement or emphasis can be sought to locate difficult concepts, or gestures of pointing can indicate the subparts of a concept.

One of the distinctions of our annotations, compared to existing work, is a consideration for future computer vision work. Hence, poses are collected in ways that we believe have a high likelihood of successful detection should we attempt to extract them in an automatic way using pose or parts recognition techniques such as [1], [11]. In contrast, much of the existing analysis has been done within the fields of psychology or education, and gestures were identified from a more intuitive human perspective.

The paper is divided into the following sections. In Section 2 we review the existing research on the relevance of human gesture in the context of teaching, and on the tools and methods for collecting data. In section 3, we provide a brief overview of an annotation tool we have developed, as well as our justifications for designing a new tool, as opposed to using one of the many existing tools already available. Section 4 reviews our annotation methodology. In section 5 we present a statistical analysis of the ground truth we collected through manual annotation of two 75-minute computer science lectures each featuring a single lecturer. We also discuss the methodology for our analysis as well as our own observations relating to patterns and meanings identified. Finally, in Sections6 and 7 we conclude by discussing future work and the highlights of our contributions.

2. Related Work

We review existing literature relating gestures to meaning with respect to teaching, and the representation, annotation and taxonomy of gestures as our work lies at the intersection of these fields.

2.1. Gestures in Teaching

A number of existing works in the fields of education and psychology have identified the importance of gestures in human communication, especially in the context of teaching.

Seminal work on the relationship between gestures and language done by McNeill identifies five classes: iconics, metaphorics, beats, cohesives and deictics [8]. Iconic gestures attempt to illustrate the semantic content of speech, *e.g*. holding a fist in front and slightly turning it when talking about a steering wheel. Metaphorics are similar to iconics, but whereas iconics describe concrete objects or events, metaphorics are used to depict abstract ideas. Beat gestures are typically simple gestures of emphasis, *e.g*. a light "beat" of a hand in the air. McNeill describes cohesives as composite gestures (*i.e*. they consist of the other types of gestures) which signal continuities in thematically related but temporally separated discourse; *e.g*. a speaker makes a certain gesture when describing an event, makes a different gesture when making a side note, and then returns to the original gesture to signal that they have returned to the original topic. The last class, deictic gestures, are pointing gestures.

Roth *et al*. apply the gestural models of McNeill in their studies on the role of gestures in teaching. Roth particularly discusses the importance of hand and arm gestures relative to body position and motion in [12]. Roth cites the work of Kendon identifying three phases for gestures: a position of rest (*preparation*), a peak structure (*stroke*) and a return to a position of rest (*retraction*) [4]. Roth then continues to argue the importance of gesture in teaching, with his work finding that gestures can sometimes convey information that is not conveyed in speech alone, as well as the finding that some children could express understanding of taught material through gesture even if they could not describe it using words. In [13], Roth *et al*. studied the relationship between talk and gesture of an instructor in an ecology lecture. Of the five gestural classes identified by McNeill, Roth *et al*. note only the three that were apparent in their analysis: deictic, iconic and metaphoric.

The results of McNeill and Roth *et al*. hint at the feasibility of the usage of gestures as semantic cues. In our work, we also apply the models discussed here (*e.g*. the broad gestural classes, the multi-phase gestural model) in our representations of gestures.

2.2. Annotation

A number of efforts have been made to annotate and analyze gestures from recorded video for various purposes.

Kipp *et al*. introduced a gesture annotation scheme and tool specifically aimed at providing gestural data for animated characters [6]. They also resolve the problem of choosing the appropriate level of granularity (*i.e*. how much detail to capture) by choosing the middle ground between purely descriptive data that resembles motion capture techniques, and free-form written descriptions.

They start by isolating hand and arm gestures, which they contend captures sufficient gestural information from conversations. Hand and arm gestures from eighteen minutes of conversational video were annotated manually. Their proposed scheme focuses on positional and temporal data and does not record qualitative observations. Their gesture annotation tool builds upon the generic annotation tool introduced by Kipp [5] and uses predefined text labels, but is augmented to allow the user to graphically illustrate the positions of hands and shoulders. While the tool is able to capture significant hand and arm gestural detail in conversational videos, it cannot account for body orientation (*i.e*. if the speaker were facing sideways, the annotator would not be able to record the spatial information of the arms). Also, the authors identify that the tool is currently incapable of capturing hand shape, nor is it able to capture different gestures for each hand. We address both in our work.

Another key challenge encountered during annotation is gestural segmentation, *i.e*. when a gesture begins and ends, or the identification of the specific phases within a gesture. Previous work involving the analysis of manually annotated gestures including [2], [7] showed the low rate of agreement between manual annotators, although Martell was able to increase that rate by training the annotators [7]. This challenge is also recognized by Wilson *et al*. during evaluation of their technique for the automatic segmentation of gestures [14]. The problem is exacerbated by a lack of agreement within the gesture research community as to what constitutes a gesture. We do not seek to resolve this problem for now, but we do address it by providing data from two independent novice annotators and discuss the results in Section 5.

2.3. Taxonomies

Martell introduces FORM, a gesture annotation scheme in [7]. FORM is designed to encapsulate both kinematic information about gestures as well as conversational information. In the scheme, gestures are represented using *annotation graphs*, which consist of arcs and nodes sharing the same timeline. Nodes represent timestamps, and arcs represents events spanning the time between two nodes. Furthermore, each arc consists of a series of tracks, with two tracks per movable body part: a track describing the location, scale and orientation of a part when static, and a track describing the movement of a part. Objects placed in tracks also include temporal data (*i.e*. start and end times) as well as attributes describing the physical properties. The attributes are assigned according to a given taxonomy. For example, the "upper arm lift" can be assigned one of nine values, roughly dividing the angles between 0 and 180 degrees. The problem of granularity is clearly encountered

but not discussed. FORM is designed to be extensible, so attributes and tracks may be added for conversational information. Martell provides a sample annotation using the ANVIL tool [5], as well as an evaluation of inter-annotator and intra-annotator agreement.

Martell's work provides an interesting structure and insights for the design of a gestural annotation scheme but few specifics (since it is meant to be extensible). Also, by separating body parts, it becomes more difficult to associate more complex gestures to meanings.

Gut *et al*. present another scheme called CoGesT [2] for the annotation of conversational gestures. In terms of granularity, it is quite well defined and provides a system for classifying hand poses. The CoGesT scheme allows annotators to assign quantitative values to spatiotemporal gestural properties such as time and location, and to describe the motion between keyframes of a gesture. CoGesT also clearly defines a separation of the form and the function of gestures.

The authors also perform a preliminary evaluation of their scheme by having three independent users annotate a 15-minute video of a single speaker telling a story. They find that the users agree strongly in terms of gestural segmentation (as much as 86%) but poorly with respect to the specific annotations (as low as 23%).

Like CoGesT, we also separate the form and function of gestures. However, we find that CoGesT provides greater granularity than is necessary in a teaching environment. Furthermore, CoGesT appears to focus on hand gestures, whereas our preliminary findings and existing literature suggest that teaching also involves head and arm gestures.

3. Gesture Annotation Tool

We introduce a novel tool designed for annotation of gestures in video. In this section, we focus on a discussion of the tool's usage and user interface design.

3.1. Overview

The tool takes as input a sequence of still images, an optional audio file, as well as an index file stored in a directory. The audio and still images are usually extracted from a video. This was done mainly to increase the ease of integration between the annotation tool and many implementations of computer vision algorithms, which often process still images or sequences of still images rather than video files directly. This has the added benefit that the tool becomes less concerned with video formats. Producing the requisite files from a video is simplified through the use of a script (available as part of the tool). Video frames are usually stored at a rate of 30 frames per second but we find they may be extracted at a rate as low as 2 per second without loss of significant gestural information, for memory efficiency.

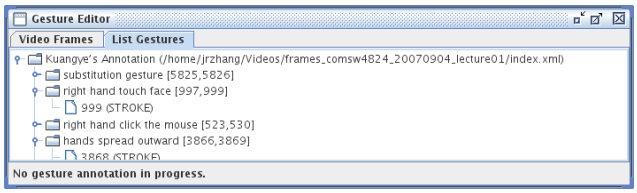

Figure 2. The tree-view tab of the gesture editor internal window, which lists the existing annotations in a project in a hierarchical format.

Once the appropriate files are available, the user can create a new project in the annotator tool, specify generic metadata (*e.g*. project author, comments) as well as the index to the video, and begin the process of annotation. The annotations and associated metadata can be exported to XML.

Gestures in the tool are represented as a collection of keyframes within a subsequence of the images where the poses are specified in detail. As we generally follow the three-phase (or multi-phase) model of gestures as described in [4], [14], the use of keyframes allows us to roughly identify the phases in addition to the distinguishing poses of the gesture and their temporal relationships. The representation was inspired by existing work, but modified to acknowledge their restriction on upper body gestures, and to gestures that preferentially occur in one-sided communications (teacher monologues).

3.2. User Interface

The main user interface (Figure 1) is divided into two sections: the video player, and the gesture editor. The video player gives users the ability to watch the sequence of images in rapid succession as a video, and optionally provides audio if an audio stream is available and the operating system is capable of supporting the codec. The user is capable of jumping to specific frames, speed up and slow down playback, and other common features.

The gesture editor itself is divided into two tabs: video frames and a list of gestures. The video frames tab is visible in Figure 1 and shows a sequence of the video frames in a timeline format. This feature was developed after we observed that it facilitated the identification of the various phases of a gesture as well as the exact frames those phases occur as the user can see "across" time. We also observed that at least two gestures may sometimes overlap. Specifically, out of 372 annotated gestures in our collected data, 26 of them were overlapping with another gesture. In one case, the lecturer simultaneously shrugged while making hand/arm gestures. Therefore, the user is capable of specifying sequences of frames for different gestures, which are shown as different gestural tracks. The list of gestures tab is shown in Figure 2 and contains a tree UI structure which displays hierarchical data and provides the user with a tex-

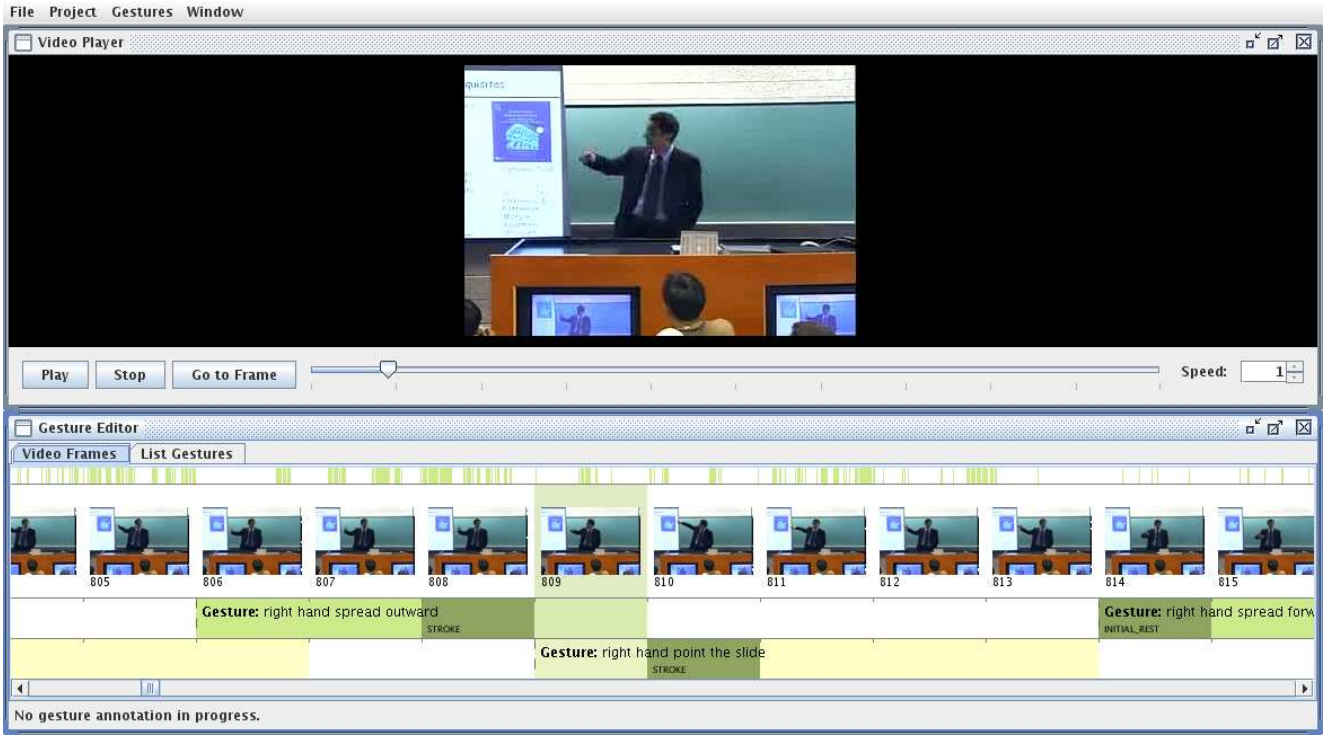

Figure 1. The main user interface of the gesture annotator tool.

tual overview of the current annotations in the project.

To mark a sequence of frames as belonging to a gesture, the user can select the sequence and use the popup-menu that appears. The user is then asked to provide a description of the gesture. This highlights the sequence and makes other options available, particularly the ability to mark individual frames (within the newly marked sequence) as a keyframe, which are highlighted as a darker color in the gesture sequence (see the bottom of Figure 1).

An alternate way to mark the start and end timestamps of a gesture is to play the video and mark the endpoints with hotkeys.

A third interface is shown when the user identifies a keyframe and wishes to specify the pose of the instructor. This interface allows the user to choose the best way to describe the pose, according to their judgment. The user may choose to use the avatar poser (as seen in Figure 3, provide a textual description, or specify that there is no human visible in the frame. The justification for these options, as well as a discussion of the avatar poser in detail, is provided in Section 3.3. The user may also specify the phase of the keyframe (*i.e.* in deference to the three-phase gestural model) as well as provide an optional comment.

3.3. Annotating Poses By Avatar

Once a user has identified a keyframe and wishes to further illustrate the pose of the lecturer, the graphical poser can be used.

In our preliminary findings, we observed that most significant gestures in teaching can be represented using simple upper body, arm and head movements. We chose this as a starting point which is reflected in the granularity of our poser. The state of the poser can be represented in 27 bits, with all possible selections shown in Figure 3. Some examples of gesture and their approximate avatar representations are shown in Figure 4. A discussion on the appropriate level of granularity is given in Section 5.

The user interface is defined to balance the user's ability to describe the pose accurately and quickly. The radio buttons in the graphical UI are positioned in a way as to correspond to the parts of the body and also to minimize the distance between one another, so users may select them faster.

The avatar control radio buttons are placed beside a preview window, which changes to reflect the latest pose selected by the user. The avatar in the preview window will always face forward regardless of body orientation, as we noticed it was easy for annotators to mirror the lecturer's pose, even when they are turned around. We also considered other avatar representations, including the possibility

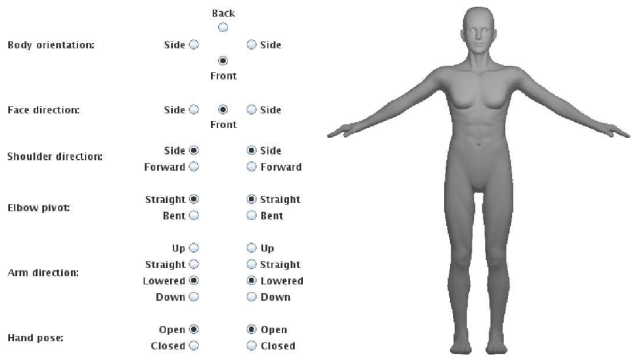

Figure 3. The avatar poser controls in the default configuration, along with the corresponding avatar preview image.

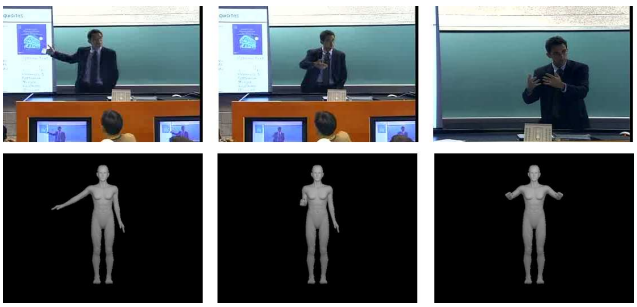

Figure 4. Examples of gestures and their avatar representations below.

of using two separated avatars to represent the lecturer from different perspectives; our present version seems sufficient.

4. Annotation and Analysis

Two 75-minute computer science video lectures have been manually annotated for gestures. In following with Martell's observation of strong intra-annotator but weak inter-annotator consistency [7], both videos were annotated by the same person (one of the authors). Each video captures a different instructor from different cultural backgrounds, presenting topics from different areas of computer science (one lecture is on machine learning, the other is on computer architecture). During preprocessing, the video frames were extracted and collected as a sequence of still images at a rate of 2 frames per second. The videos were provided by the Columbia Video Network: cameras were human operated, there was no post processing, and the video and audio quality are poor. The videos both have a resolution of 352×240. The lighting conditions were varied, as were the clothes and overall appearance of the instructors. The videos do not focus solely on the instructor but sometimes switch to a view of the slides presented for a period of time (for the computer architecture video and the machine learning video, 24% and 41% of the frames extracted were marked as belonging to a gesture, respectively).

Part of one of the videos was also annotated by a second person (another one of the authors) to explore inter-annotator consistency; see Section 5.5.

Finally, observations were collected from both annotators regarding the level of granularity for the avatar poser, the frame rates of the extracted video, and high-level patterns noticed in the gestures.

5. Results

We analyzed the annotated data, and present our results along with qualitative and quantitative observations here.

5.1. Annotation and Taxonomy

The first lecture video (video A by instructor A) presents an introduction to computer architecture, an outline of the course, and an overview of the material without elaborating on the theory.

The second video (video B by instructor B) provides an introduction to machine learning but goes directly into a detailed explanation of linear regression, presenting a lot of mathematics.

During annotation, gestures were assigned a textual label according to the template "*body part, semantic class, orientation*." For example, a gesture where the instructor points with his right hand would be labeled as *right hand point right*, where *right hand* is the body part, *point* is the semantic class and *right* is the orientation (*i.e.* the direction in which he is pointing). We identified 126 unique labels falling into nine semantic classes. We defined a new semantic class whenever we noticed that the gesture was frequently repeated or that the gesture was semantically relevant to the lecture content.

Nine semantic classes were identified as follows. We note that some of them do not cleanly fall into the four or five classes commonly assumed in the literature. We introduce the class of "pedagogic" gestures to label those gestures whose purpose seems to be to structure the lecture or to encourage or remind the students. This category has not been documented in the prior literature, but is apparent in this context, since much teaching depends on developing and maintaining a supportive but asymmetric relationship with the students.

- *Put*. These can be iconic or metaphoric gestures, where the instructor "puts" abstract concepts or objects somewhere into the visible space to help describe their relationships to one another.

- *Spread*. These are gestures where both hands and arms are extended in front of the body and spread outward

in a circular fashion. Spread gestures may be iconic or metaphoric, and often correspond to an important point in the discourse. However, they often serve as pedagogical commentary, independent of lecture content, indicating the difficulty of the content.

- *Swipe*. These occur when one or both arms are moved simultaneously in one direction. These tend to be metaphoric gestures, *e.g*. an instructor makes a swipe gesture to indicate that an abstract object has moved.
- *Close & Open*. These encompass a set of gestures that are visually similar to spread gestures, *i.e*. hands and arms are spread outward or inward in a circular motion, however, arms are generally not extended and therefore they form a much smaller spread. They are considered a separate class since they are less semantically relevant than spreads and are best considered as beats.
- *Flip & Swing*. These are gestures where one or both hands are flipped in a small circle. These pedagogical gestures indicate the continuation of a theme in the discourse. These gestures can also be considered as a beat (two phase) form of a cohesive gesture, a kind of pedagogic punctuation or backward reference.
- *Touch*. These are simple beat gestures where the instructor touches an object (usually the table, glasses, *etc*.) as a beat or as a pedagogic "timeout".
- *Pointing*. These are clearly deictic gestures and accounted for the majority of gestures in both videos (see Table 1). When an instructor points, it generally means that they wish the students to pay attention to a specific region of the slide or blackboard.
- *Hold*. In between gestures, instructors are sometimes noticed to stay relatively motionless. Some of the existing literature may consider this non-gesture to be a phase separating the preparation, stroke and retraction phases. Holds usually indicate that the discourse is focused on a specific point, and it can often be a deliberate pedagogical gesture.
- *Others*. A number of gestures were observed but held no noticeable semantic significance or did not occur frequently enough to merit their own class. These gestures were assigned the "others" class.

5.2. Observations

We observed 372 and 494 gestures from videos A and B respectively. These gestures were broken down into the nine classes as summarized in Table 1. We noticed in these lecture videos three observations about which the literature is basically silent.

Semantic Class		A	A (%)	B	B (%)
Put	I, M	16	4.30	10	2.02
Spread	I, C, P	81	21.77	24	4.86
Swipe	M	8	2.15	5	1.01
Close & Open	B	42	11.29	49	9.91
Flip & Swing	B, C, P	21	5.65	4	0.81
Touch	B, P	5	1.34	7	1.41
Point	D	123	33.06	292	59.11
Hold	P	33	8.87	71	14.37
Others		43	11.56	32	6.48
Total		372		494	

Table 1. Counts and distributions of gestures according to the nine semantic classes for videos A and B. The abbreviations I, M, B, C, D, P stand for *iconic*, *metaphoric*, *beat*, *cohesive*, *deictic* and *pedagogic* respectively. Four of the gesture classes (hold, spread, flip & swing, touch) appear to be pedagogic.

First, we noticed that gestures are highly idiosyncratic. For instance, instructor B seldom does the spread gesture and tends to do more point and hold gestures than the instructor A. The lecture content clearly impacts the gesture distribution. For example, instructor B uses two hands to point at slides to explain details of matrices, while instructor A points with just one hand since discourse was mostly about theoretical topics. Nevertheless, habits of each instructor clearly exist. In video B, the instructor relies on slides more, so deictic gestures occur more frequently. In video A, the instructor refers to the slides less, and so relies on iconic or metaphoric gestures more.

Second, we observed that the gestures are often pedagogic and are correlated to the difficulty and pacing of the lecture material. Explanatory gestures, such as swinging, spreading, suggested that key points were being told. More intense gestures indicated that the material was more difficult or an important concept, while slower gestures seemed to indicate content that was less important.

Third, we noticed that successive gestures tend to overlap on their ends, and do not completely follow the three-phase model of gestures. This has made it difficult to tag adjacent gestures, because there is no hard boundary between when one gesture ends and the next gesture begins. Our tool was modified to allow overlapping gestures, shown as separate layers.

5.3. Avatar Poser Granularity

One of the lecture videos was used to examine and improve the completeness of the gesture grammar. If a pose could not be expressed by the current grammar, the annotator verbally described possible additions to the grammar that would enable it to express that pose. From the video, 183 poses were encoded using the current tool, whereas 91 poses could not be expressed by the grammar. From ana-

lyzing the necessary additions for these 91 poses, we explored five additions that significantly increased the expressiveness of the grammar. Extra precision on shoulder direction and elbow angle helped encode 51 of the poses; 22 poses needed shoulder joint rotation; and 44 needed forearm pronation/supination. Otherwise, the grammar appeared well-matched to what was observed. Future iterations of the taxonomy and pose representation will be modified according to the observations made here.

We also found several ambiguities when proposing additions to increase the expressiveness of the gesture grammar, since different joint configurations can lead to almost the same overall pose. The main source of ambiguities occur when two rotation axes coincide, such as the forearm and shoulder when the arm is straight.

5.4. Dimensionality Reduction

We applied Principal Component Analysis (PCA) to the pose data to gain additional data which can help us refine the tool and pose representation, as well as provide insights regarding the pattern of poses in gestures.

Examining the entire corpus of poses for one instructor (instructor A), we compressed each pose using the annotation tool, into a ten-dimensional vector whose components encoded the quantized positions of: "*body, face, left hand, right hand, left arm, left shoulder, left elbow, right arm, right shoulder, right elbow*." We map each component of the pose to a value either between -1 to 1 or 0 to 1, evenly divided. We used PCA for dimensionality reduction, and found that the first two principal components account for more than half of the variance of poses (51%), and the first five account for nearly all (81%). These eigengestures can be roughly interpreted as:

- Right arm raised with elbow straightened versus right arm lowered with elbow bent, which is basically a point versus a rest gesture (33%, see Figure 5).
- Both arms used symmetrically from the shoulder, either both to the side or both forward, which is basically a spread versus a rest gesture (18%).
- Right elbow used anti-symmetrically from the left elbow in a "Mr. Roboto dance"-like chop (12%).
- Both hands opened or closed symmetrically (9%).
- Right arm raised, but with bent elbow (9%).

We note that the position of body and face did not contribute much to the gesture variance, which is expected, since the body of the lecturer is usually turned towards the class. Also, due to low granularity in the hand annotation, independent hand information also does not significantly contribute to the variance.

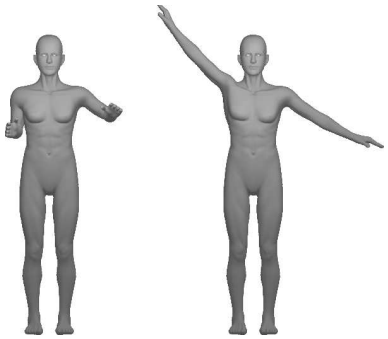

Figure 5. Example of an eigengesture. The left and right poses correspond to the maximum and minimum values and basically represent a point versus a rest.

5.5. Inter-Annotator Analysis

Approximately 60% of video A was annotated by two independent, novice annotators. We attempted to compare these results. As previously stated, there is no standardized method for comparing gesture annotations, so we approached this intuitively.

As a rough metric, we compared the work of the two annotations in terms of segmentation. A visualization of the comparison is shown in Figure 6. Colored regions represent frames that are marked as belonging to a gesture. It can be seen from the figure that, using this metric, inter-annotator agreement is strong: roughly 74% agreement, not too far from reports in the existing literature.

More precise segmentation however is notably more difficult. In Figure 6, green tick marks indicate the start of gestures, and red ticks mark the end of gestures. From this perspective, inter-annotator agreement is very low and is difficult to compare. As previously mentioned, what one annotator may mark as one long gesture, another may break into several smaller gestures.

6. Future Work

Our results suggest the possibility that gestures may be valuable indicators of both the segmentation and relationship of lecture content and the difficulty of the underlying concepts. Future work will explore the integration of such gestural data into non-linear, semantic video browsers such as Vast MM [3].

Changes to the user interface are contemplated. For instance, our current version of the gesture annotator tool uses a single avatar view, but multiple avatars may be implemented in future versions to allow users to specify poses from different perspectives.

The data we collected will be used as ground truth gesture recognition. We attempted to build the taxonomy with consideration to existing computer vision algorithms, *e.g.*

Figure 6. Inter-annotator comparison. The colored regions indicate parts of (roughly half) of video A that have been marked as a frame belonging to a gesture. The line in the middle separates the work of the two independent annotators: one on top, one below. Red and green ticks mark the boundaries of gestures: green ticks indicate the beginning of a gesture, and red ticks indicate the end.

the separation of parts may be applicable to existing pose recognition techniques such as [1].

We will also explore the possibility of using "gestural signatures" to identify lecturers, based on our observations that lecturers appear to have fixed gestural styles.

7. Conclusion

We have introduced a novel gesture annotation tool for digital videos. We have also gathered a significant amount of ground truth data from lecture videos and performed a preliminary analysis. Novel observations relating gestures to content and pedagogy are a first step to exploring the feasibility of using gestures as semantic cues for non-linear video browsers, as well as for other possible applications.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1014–1021, June 2009. 1, 8

[2] U. Gut, K. Looks, A. Thies, T. Trippel, and D. Gibbon. Cogest conversational gesture transcription system. Technical report, University of Bielefeld, 1993. 2, 3

[3] A. Haubold and J. Kender. Vast mm: multimedia browser for presentation video. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 41–48, New York, NY, USA, 2007. ACM. 1, 7

[4] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. In *The relationship of verbal and nonverbal communication*, pages 207–227. Mouton Publishers, 1980. 2, 3

[5] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, 2001. 2, 3

[6] M. Kipp, M. Neff, and I. Albrecht. An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation*, 41(3-4):325–339, 2007. 2

[7] C. Martell. Form: An extensible, kinematically-based gesture annotation scheme. In *Proc. 3rd International Conference on Language Resources and Evaluation*, 2002. 2, 5

[8] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University Of Chicago Press, 1992. 2

[9] M. Merler and J. Kender. Semantic keyword extraction via adaptive text binarization of unstructured unsourced video. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 261–264, November 2009. 1

[10] M. Morris and J. Kender. Sort-merge feature selection and fusion methods for classification of unstructured video. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 578–581, July 2009. 1

[11] D. Ramanan. Learning to parse images of articulated bodies. In *In NIPS 2007*. NIPS, 2006. 1

[12] W. Roth. Gestures: Their role in teaching and learning. *Review of Educational Research*, 71(3):365–392, 2001. 2

[13] W. Roth and G. Bowen. Decalages in talk and gesture: Visual and verbal semiotics of ecology lectures. *Linguistics and Education*, 10(3):335–358, 1998. 2

[14] A. Wilson, A. Bobick, and J. Cassell. Temporal classification of natural gesture and application to video coding. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, page 948, 1997. 2, 3