Gestural Cohesion for Topic Segmentation

Jacob Eisenstein, Regina Barzilay and **Randall Davis**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge MA 02139
{jacobe, regina, davis}@csail.mit.edu

Abstract

This paper explores the relationship between discourse segmentation and coverbal gesture. Introducing the idea of *gestural cohesion*, we show that coherent topic segments are characterized by homogeneous gestural forms and that changes in the distribution of gestural features predict segment boundaries. Gestural features are extracted automatically from video, and are combined with lexical features in a Bayesian generative model. The resulting multimodal system outperforms text-only segmentation on both manual and automatically-recognized speech transcripts.

1 Introduction

When people communicate face-to-face, discourse cues are expressed simultaneously through multiple channels. Previous research has extensively studied how discourse cues correlate with lexico-syntactic and prosodic features (Hearst, 1994; Hirschberg and Nakatani, 1998; Passonneau and Litman, 1997); this work informs various text and speech processing applications, such as automatic summarization and segmentation. Gesture is another communicative modality that frequently accompanies speech, yet it has not been exploited for computational discourse analysis.

This paper empirically demonstrates that gesture correlates with discourse structure. In particular, we show that automatically-extracted visual features can be combined with lexical cues in a statistical model to predict topic segmentation, a frequently studied form of discourse structure. Our method builds on the idea that coherent discourse segments are characterized by *gestural cohesion*; in other words, that such segments exhibit homogeneous gestural patterns. Lexical cohesion (Halliday and Hasan, 1976) forms the backbone of many verbal segmentation algorithms, on the theory that segmentation boundaries should be placed where the distribution of words changes (Hearst, 1994). With gestural cohesion, we explore whether the same idea holds for gesture features.

The motivation for this approach comes from a series of psycholinguistic studies suggesting that gesture supplements speech with meaningful and unique semantic content (McNeill, 1992; Kendon, 2004). We assume that repeated patterns in gesture are indicative of the semantic coherence that characterizes well-defined discourse segments. An advantage of this view is that gestures can be brought to bear on discourse analysis without undertaking the daunting task of recognizing and interpreting individual gestures. This is crucial because coverbal gesture – unlike formal sign language – rarely follows any predefined form or grammar, and may vary dramatically by speaker.

A key implementational challenge is automatically extracting gestural information from raw video and representing it in a way that can applied to discourse analysis. We employ a representation of *visual codewords*, which capture clusters of low-level motion patterns. For example, one codeword may correspond to strong left-right motion in the upper part of the frame. These codewords are then treated similarly to lexical items; our model identifies changes in their distribution, and predicts topic

Proceedings of ACL-08: HLT, pages 852–860,
Columbus, Ohio, USA, June 2008.

boundaries appropriately. The overall framework is implemented as a hierarchical Bayesian model, supporting flexible integration of multiple knowledge sources.

Experimental results support the hypothesis that gestural cohesion is indicative of discourse structure. Applying our algorithm to a dataset of face-to-face dialogues, we find that gesture communicates unique information, improving segmentation performance over lexical features alone. The positive impact of gesture is most pronounced when automatically-recognized speech transcripts are used, but gestures improve performance by a significant margin even in combination with manual transcripts.

2 Related Work

Gesture and discourse Much of the work on gesture in natural language processing has focused on multimodal dialogue systems in which the gestures and speech may be constrained, e.g. (Johnston, 1998). In contrast, we focus on improving discourse processing on unconstrained natural language between humans. This effort follows basic psychological and linguistic research on the communicative role of gesture (McNeill, 1992; Kendon, 2004), including some efforts that made use of automatically acquired visual features (Quek, 2003). We extend these empirical studies with a statistical model of the relationship between gesture and discourse segmentation.

Hand-coded descriptions of body posture shifts and eye gaze behavior have been shown to correlate with topic and turn boundaries in task-oriented dialogue (Cassell et al., 2001). These findings are exploited to generate realistic conversational "grounding" behavior in an animated agent. The semantic content of gesture was leveraged – again, for gesture generation – in (Kopp et al., 2007), which presents an animated agent that is capable of augmenting navigation directions with gestures that describe the physical properties of landmarks along the route. Both systems generate plausible and human-like gestural behavior; we address the converse problem of *interpreting* such gestures.

In this vein, hand-coded gesture features have been used to improve sentence segmentation, showing that sentence boundaries are unlikely to overlap gestures that are in progress (Chen et al., 2006). Features that capture the start and end of gestures are shown to improve sentence segmentation beyond lexical and prosodic features alone. This idea of gestural features as a sort of visual punctuation has parallels in the literature on prosody, which we discuss in the next subsection.

Finally, ambiguous noun phrases can be resolved by examining the similarity of co-articulated gestures (Eisenstein and Davis, 2007). While noun phrase coreference can be viewed as a discourse processing task, we address the higher-level discourse phenomenon of topic segmentation. In addition, this prior work focused primarily on pointing gestures directed at pre-printed visual aids. The current paper presents a new domain, in which speakers do not have access to visual aids. Thus pointing gestures are less frequent than "iconic" gestures, in which the form of motion is the principle communicative feature (McNeill, 1992).

Non-textual features for topic segmentation Research on non-textual features for topic segmentation has primarily focused on prosody, under the assumption that a key prosodic function is to mark structure at the discourse level (Steedman, 1990; Grosz and Hirshberg, 1992; Swerts, 1997). The ultimate goal of this research is to find correlates of hierarchical discourse structure in phonetic features.

Today, research on prosody has converged on prosodic cues which correlate with discourse structure. Such markers include pause duration, fundamental frequency, and pitch range manipulations (Grosz and Hirshberg, 1992; Hirschberg and Nakatani, 1998). These studies informed the development of applications such as segmentation tools for meeting analysis, e.g. (Tur et al., 2001; Galley et al., 2003).

In comparison, the connection between gesture and discourse structure is a relatively unexplored area, at least with respect to computational approaches. One conclusion that emerges from our analysis is that gesture may signal discourse structure in a different way than prosody does: while specific prosodic markers characterize segment boundaries, gesture predicts segmentation through intra-segmental cohesion. The combination of these two

modalities is an exciting direction for future research.

3 Visual Features for Discourse Analysis

This section describes the process of building a representation that permits the assessment of gestural cohesion. The core signal-level features are based on *spatiotemporal interest points*, which provide a sparse representation of the motion in the video. At each interest point, visual, spatial, and kinematic characteristics are extracted and then concatenated into vectors. Principal component analysis (PCA) reduces the dimensionality to a feature vector of manageable size (Bishop, 2006). These feature vectors are then clustered, yielding a codebook of visual forms. This video processing pipeline is shown in Figure 1; the remainder of the section describes the individual steps in greater detail.

3.1 Spatiotemporal Interest Points

Spatiotemporal interest points (Laptev, 2005) provide a sparse representation of motion in video. The idea is to select a few local regions that contain high information content in both the spatial and temporal dimensions. The image features at these regions should be relatively robust to lighting and perspective changes, and they should capture the relevant movement in the video. The set of spatiotemporal interest points thereby provides a highly compressed representation of the key visual features. Purely spatial interest points have been successful in a variety of image processing tasks (Lowe, 1999), and spatiotemporal interest points are beginning to show similar advantages for video processing (Laptev, 2005).

The use of spatiotemporal interest points is specifically motivated by techniques from the computer vision domain of *activity recognition* (Efros et al., 2003; Niebles et al., 2006). The goal of activity recognition is to classify video sequences into semantic categories: e.g., walking, running, jumping. As a simple example, consider the task of distinguishing videos of walking from videos of jumping. In the walking videos, the motion at most of the interest points will be horizontal, while in the jumping videos it will be vertical. Spurious vertical motion in a walking video is unlikely to confuse the classifier, as long as the majority of interest points move horizontally. The hypothesis of this paper is that just as such low-level movement features can be applied in a supervised fashion to distinguish activities, they can be applied in an unsupervised fashion to group co-speech gestures into perceptually meaningful clusters.

The Activity Recognition Toolbox (Dollár et al., 2005)[1] is used to detect spatiotemporal interest points for our dataset. This toolbox ranks interest points using a difference-of-Gaussians filter in the spatial dimension, and a set of Gabor filters in the temporal dimension. The total number of interest points extracted per video is set to equal the number of frames in the video. This bounds the complexity of the representation to be linear in the length of the video; however, the system may extract many interest points in some frames and none in other frames.

Figure 2 shows the interest points extracted from a representative video frame from our corpus. Note that the system has identified high contrast regions of the gesturing hand. From manual inspection, the large majority of interest points extracted in our dataset capture motion created by hand gestures. Thus, for this dataset it is reasonable to assume that an interest point-based representation expresses the visual properties of the speakers' hand gestures. In videos containing other sources of motion, preprocessing may be required to filter out interest points that are extraneous to gestural communication.

3.2 Visual Descriptors

At each interest point, the temporal and spatial brightness gradients are constructed across a small space-time volume of nearby pixels. Brightness gradients have been used for a variety of problems in computer vision (Forsyth and Ponce, 2003), and provide a fairly general way to describe the visual appearance of small image patches. However, even for a small space-time volume, the resulting dimensionality is still quite large: a 10-by-10 pixel box across 5 video frames yields a 500-dimensional feature vector for each of the three gradients. For this reason, principal component analysis (Bishop, 2006) is used to reduce the dimensionality. The spatial location of the interest point is added to the final feature vector.

[1] http://vision.ucsd.edu/∼pdollar/research/cuboids_doc/index.html

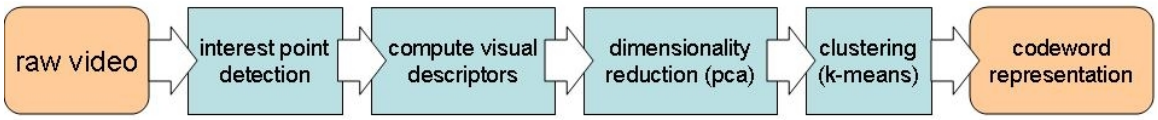

Figure 1: The visual processing pipeline for the extraction of gestural codewords from video.

Figure 2: Circles indicate the interest points extracted from this frame of the corpus.

This visual feature representation is substantially lower-level than the descriptions of gesture form found in both the psychology and computer science literatures. For example, when manually annotating gesture, it is common to employ a taxonomy of hand shapes and trajectories, and to describe the location with respect to the body and head (McNeill, 1992; Martell, 2005). Working with automatic hand tracking, Quek (2003) automatically computes perceptually-salient gesture features, such as symmetric motion and oscillatory repetitions.

In contrast, our feature representation takes the form of a vector of continuous values and is not easily interpretable in terms of how the gesture actually appears. However, this low-level approach offers several important advantages. Most critically, it requires no initialization and comparatively little tuning: it can be applied directly to any video with a fixed camera position and static background. Second, it is robust: while image noise may cause a few spurious interest points, the majority of interest points should still guide the system to an appropriate characterization of the gesture. In contrast, hand tracking can become irrevocably lost, requiring manual resets (Gavrila, 1999). Finally, the success of similar low-level interest point representations at the activity-recognition task provides reason for optimism that they may also be applicable to unsupervised gesture analysis.

3.3 A Lexicon of Visual Forms

After extracting a set of low-dimensional feature vectors to characterize the visual appearance at each spatiotemporal interest point, it remains only to convert this into a representation amenable to a cohesion-based analysis. Using k-means clustering (Bishop, 2006), the feature vectors are grouped into *codewords*: a compact, lexicon-like representation of salient visual features in video. The number of clusters is a tunable parameter, though a systematic investigation of the role of this parameter is left for future work.

Codewords capture frequently-occurring patterns of motion and appearance at a local scale – interest points that are clustered together have a similar visual appearance. Because most of the motion in our videos is gestural, the codewords that appear during a given sentence provide a succinct representation of the ongoing gestural activity. Distributions of codewords over time can be analyzed in similar terms to the distribution of lexical features. A change in the distribution of codewords indicates new visual kinematic elements entering the discourse. Thus, the codeword representation allows gestural cohesion to be assessed in much the same way as lexical cohesion.

4 Bayesian Topic Segmentation

Topic segmentation is performed in a Bayesian framework, with each sentence's segment index encoded in a hidden variable, written z_t. The hidden variables are assumed to be generated by a linear segmentation, such that $z_t \in \{z_{t-1}, z_{t-1} + 1\}$. Observations – the words and gesture codewords – are

generated by multinomial language models that are indexed according to the segment. In this framework, a high-likelihood segmentation will include language models that are tightly focused on a compact vocabulary. Such a segmentation maximizes the lexical cohesion of each segment. This model thus provides a principled, probabilistic framework for cohesion-based segmentation, and we will see that the Bayesian approach is particularly well-suited to the combination of multiple modalities.

Formally, our goal is to identify the best possible segmentation S, where S is a tuple: $S = \langle \mathbf{z}, \theta, \phi \rangle$. The segment indices for each sentence are written z_t; for segment i, θ_i and ϕ_i are multinomial language models over words and gesture codewords respectively. For each sentence, $\mathbf{x}_t$ and $\mathbf{y}_t$ indicate the words and gestures that appear. We will seek to identify the segmentation $\hat{S} = \text{argmax}_S p(S, \mathbf{x}, \mathbf{y})$, conditioned on priors that will be defined below.

$$p(S, \mathbf{x}, \mathbf{y}) = p(\mathbf{x}, \mathbf{y}|S)p(S)$$
$$p(\mathbf{x}, \mathbf{y}|S) = \prod_i p(\{x_t : z_t = i\}|\theta_i)p(\{y_t : z_t = i\}|\phi_i) \quad (1)$$

$$p(S) = p(\mathbf{z}) \prod_i p(\theta_i)p(\phi_i) \quad (2)$$

The language models θ_i and ϕ_i are multinomial distributions, so the log-likelihood of the observations $\mathbf{x}_t$ is $\log p(\mathbf{x}_t|\theta_i) = \sum_j^W n(t,j) \log \theta_{i,j}$, where $n(t,j)$ is the count of word j in sentence t, and W is the size of the vocabulary. An analogous equation is used for the gesture codewords. Each language model is given a symmetric Dirichlet prior α. As we will see shortly, the use of different priors for the verbal and gestural language models allows us to weight these modalities in a Bayesian framework. Finally, we model the probability of the segmentation $\mathbf{z}$ by considering the durations of each segment: $p(\mathbf{z}) = \prod_i p(\text{dur}(i)|\psi)$. A negative-binomial distribution with parameter ψ is applied to discourage extremely short or long segments.

Inference Crucially, both the likelihood (equation 1) and the prior (equation 2) factor into a product across the segments. This factorization enables the optimal segmentation to be found using a dynamic program, similar to those demonstrated by Utiyama and Isahara (2001) and Malioutov and Barzilay (2006). For each set of segmentation points $\mathbf{z}$, the associated language models are set to their posterior expectations, e.g., $\theta_i = E[\theta|\{x_t : z_t = i\}, \alpha]$.

The Dirichlet prior is conjugate to the multinomial, so this expectation can be computed in closed form:

$$\theta_{i,j} = \frac{n(i,j) + \alpha}{N(i) + W\alpha}, \quad (3)$$

where $n(i,j)$ is the count of word j in segment i and $N(i)$ is the total number of words in segment i (Bernardo and Smith, 2000). The symmetric Dirichlet prior α acts as a smoothing pseudo-count. In the multimodal context, the priors act to control the weight of each modality. If the prior for the verbal language model θ is high relative to the prior for the gestural language model ϕ then the verbal multinomial will be smoother, and will have a weaker impact on the final segmentation. The impact of the priors on the weights of each modality is explored in Section 6.

Estimation of priors The distribution over segment durations is negative-binomial, with parameters ψ. In general, the maximum likelihood estimate of the parameters of a negative-binomial distribution cannot be found in closed form (Balakrishnan and Nevzorov, 2003). For any given segmentation, the maximum-likelihood setting for ψ is found via a gradient-based search. This setting is then used to generate another segmentation, and the process is iterated until convergence, as in hard expectation-maximization. The Dirichlet priors on the language models are symmetric, and are chosen via cross-validation. Sampling or gradient-based techniques may be used to estimate these parameters, but this is left for future work.

Relation to other segmentation models Other cohesion-based techniques have typically focused on hand-crafted similarity metrics between sentences, such as cosine similarity (Galley et al., 2003; Malioutov and Barzilay, 2006). In contrast, the model described here is probabilistically motivated, maximizing the joint probability of the segmentation with the observed words and gestures. Our objective criterion is similar in form to that of Utiyama and Isahara (2001); however, in contrast to this prior

work, our criterion is justified by a Bayesian approach. Also, while the smoothing in our approach arises naturally from the symmetric Dirichlet prior, Utiyama and Isahara apply Laplace's rule and add pseudo-counts of one in all cases. Such an approach would be incapable of flexibly balancing the contributions of each modality.

5 Evaluation Setup

Dataset Our dataset is composed of fifteen audio-video recordings of dialogues limited to three minutes in duration. The dataset includes nine different pairs of participants. In each video one of five subjects is discussed. The potential subjects include a "Tom and Jerry" cartoon, a "Star Wars" toy, and three mechanical devices: a latchbox, a piston, and a candy dispenser. One participant – "participant A" – was familiarized with the topic, and is tasked with explaining it to participant B, who is permitted to ask questions. Audio from both participants is used, but only video of participant A is used; we do not examine whether B's gestures are relevant to discourse segmentation.

Video was recorded using standard camcorders, with a resolution of 720 by 480 at 30 frames per second. The video was reduced to 360 by 240 gray-scale images before visual analysis is applied. Audio was recorded using headset microphones. No manual postprocessing is applied to the video.

Annotations and data processing All speech was transcribed by hand, and time stamps were obtained using the SPHINX-II speech recognition system for forced alignment (Huang et al., 1993). Sentence boundaries are annotated according to (NIST, 2003), and additional sentence boundaries are automatically inserted at all turn boundaries. Commonly-occurring terms unlikely to impact segmentation are automatically removed by using a stoplist.

For automatic speech recognition, the default Microsoft speech recognizer was applied to each sentence, and the top-ranked recognition result was reported. As is sometimes the case in real-world applications, no speaker-specific training data is available. The resulting recognition quality is very poor, yielding a word error rate of 77%.

Annotators were instructed to select segment boundaries that divide the dialogue into coherent topics. Segmentation points are required to coincide with sentence or turn boundaries. A second annotator – who is not an author on any paper connected with this research – provided an additional set of segment annotations on six documents. On this subset of documents, the P_k between annotators was .306, and the WindowDiff was .325 (these metrics are explained in the next subsection). This is similar to the interrater agreement reported by Malioutov and Barzilay (2006).

Over the fifteen dialogues, a total of 7458 words were transcribed (497 per dialogue), spread over 1440 sentences or interrupted turns (96 per dialogue). There were a total of 102 segments (6.8 per dialogue), from a minimum of four to a maximum of ten. This rate of fourteen sentences or interrupted turns per segment indicates relatively fine-grained segmentation. In the physics lecture corpus used by Malioutov and Barzilay (2006), there are roughly 100 sentences per segment. On the ICSI corpus of meeting transcripts, Galley *et al.* (2003) report 7.5 segments per meeting, with 770 "potential boundaries," suggesting a similar rate of roughly 100 sentences or interrupted turns per segment.

The size of this multimodal dataset is orders of magnitude smaller than many other segmentation corpora. For example, the Broadcast News corpus used by Beeferman *et al.* (1999) and others contains two million words. The entire ICSI meeting corpus contains roughly 600,000 words, although only one third of this dataset was annotated for segmentation (Galley et al., 2003). The physics lecture corpus that was mentioned above contains 232,000 words (Malioutov and Barzilay, 2006). The task considered in this section is thus more difficult than much of the previous discourse segmentation work on two dimensions: there is less training data, and a finer-grained segmentation is required.

Metrics All experiments are evaluated in terms of the commonly-used P_k (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner and Hearst, 2002) scores. These metrics are penalties, so lower values indicate better segmentations. The P_k metric expresses the probability that any randomly chosen pair of sentences is incorrectly segmented, if they are k sentences apart (Beeferman et al., 1999). Following tradition, k is set to half of the mean seg-

Method	P_k	WD
1. gesture only	.486	.502
2. ASR only	.462	.476
3. ASR + gesture	**.388**	**.401**
4. transcript only	.382	.397
5. transcript + gesture	**.332**	**.349**
6. random	.473	.526
7. equal-width	.508	.515

Table 1: For each method, the score of the best performing configuration is shown. P_k and WD are penalties, so lower values indicate better performance.

ment length. The WindowDiff metric is a variation of P_k (Pevzner and Hearst, 2002), applying a penalty whenever the number of segments within the k-sentence window differs for the reference and hypothesized segmentations.

Baselines Two naïve baselines are evaluated. Given that the annotator has divided the dialogue into K segments, the random baseline arbitrary chooses K random segmentation points. The results of this baseline are averaged over 1000 iterations. The equal-width baseline places boundaries such that all segments contain an equal number of sentences. Both the experimental systems and these naïve baselines were given the correct number of segments, and also were provided with manually annotated sentence boundaries – their task is to select the k sentence boundaries that most accurately segment the text.

6 Results

Table 1 shows the segmentation performance for a range of feature sets, as well as the two baselines. Given only gesture features the segmentation results are poor (line 1), barely outperforming the baselines (lines 6 and 7). However, gesture proves highly effective as a supplementary modality. The combination of gesture with ASR transcripts (line 3) yields an absolute 7.4% improvement over ASR transcripts alone (line 4). Paired t-tests show that this result is statistically significant ($t(14) = 2.71, p < .01$ for both P_k and WindowDiff). Even when manual speech transcripts are available, gesture features yield a substantial improvement, reducing P_k and WD by roughly 5%. This result is statistically significant for both P_k ($t(14) = 2.00, p < .05$) and WD ($t(14) = 1.94, p < .05$).

Interactions of verbal and gesture features We now consider the relative contribution of the verbal and gesture features. In a discriminative setting, the contribution of each modality would be explicitly weighted. In a Bayesian generative model, the same effect is achieved through the Dirichlet priors, which act to smooth the verbal and gestural multinomials – see equation 3. For example, when the gesture prior is high and verbal prior is low, the gesture counts are smoothed, and the verbal counts play a greater role in segmentation. When both priors are very high, the model will simply try to find equally-sized segments, satisfying the distribution over durations.

The effects of these parameters can be seen in Figure 3. The gesture model prior is held constant at its ideal value, and the segmentation performance is plotted against the logarithm of the verbal prior. Low values of the verbal prior cause it to dominate the segmentation; this can be seen at the left of both graphs, where the performance of the multimodal and verbal-only systems are nearly identical. High values of the verbal prior cause it to be oversmoothed, and performance thus approaches that of the gesture-only segmenter.

Comparison to other models While much of the research on topic segmentation focuses on written text, there are some comparable systems that also aim at unsupervised segmentation of spontaneous spoken language. For example, Malioutov and Barzilay (2006) segment a corpus of classroom lectures, using similar lexical cohesion-based features. With manual transcriptions, they report a .383 P_k and .417 WD on artificial intelligence (AI) lectures, and .298 P_k and .311 WD on physics lectures. Our results are in the range bracketed by these two extremes; the wide range of results suggests that segmentation scores are difficult to compare across domains. The segmentation of physics lectures was at a very course level of granularity, while the segmentation of AI lectures was more similar to our annotations.

We applied the publicly-available executable for this algorithm to our data, but performance was poor, yielding a .417 P_k and .465 WD even when both verbal and gestural features were available.

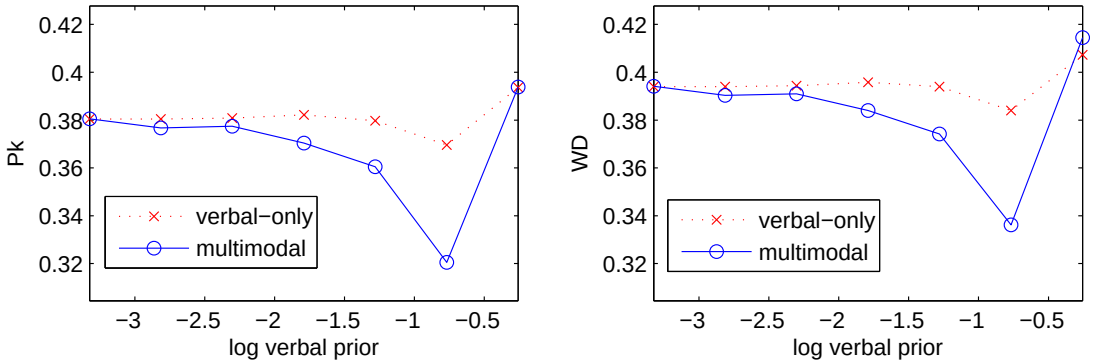

Figure 3: The multimodal and verbal-only performance using the reference transcript. The x-axis shows the logarithm of the verbal prior; the gestural prior is held fixed at the optimal value.

This may be because the technique is not designed for the relatively fine-grained segmentation demanded by our dataset (Malioutov, 2006).

7 Conclusions

This research shows a novel relationship between gestural cohesion and discourse structure. Automatically extracted gesture features are predictive of discourse segmentation when used in isolation; when lexical information is present, segmentation performance is further improved. This suggests that gestures provide unique information not present in the lexical features alone, even when perfect transcripts are available.

There are at least two possibilities for how gesture might impact topic segmentation: "visual punctuation," and cohesion. The visual punctuation view would attempt to identify specific gestural patterns that are characteristic of segment boundaries. This is analogous to research that identifies prosodic signatures of topic boundaries, such as (Hirschberg and Nakatani, 1998). By design, our model is incapable of exploiting such phenomena, as our goal is to investigate the notion of gestural cohesion. Thus, the performance gains demonstrated in this paper cannot be explained by such punctuation-like phenomena; we believe that they are due to the consistent gestural themes that characterize coherent topics. However, we are interested in pursuing the idea of visual punctuation in the future, so as to compare the power of visual punctuation and gestural cohesion to predict segment boundaries. In addition, the interaction of gesture and prosody suggests additional possibilities for future research.

The videos in the dataset for this paper are focused on the description of physical devices and events, leading to a fairly concrete set of gestures. In other registers of conversation, gestural form may be driven more by spatial metaphors, or may consist mainly of temporal "beats." In such cases, the importance of gestural cohesion for discourse segmentation may depend on the visual expressivity of the speaker. We plan to examine the extensibility of gesture cohesion to more naturalistic settings, such as classroom lectures.

Finally, topic segmentation provides only an outline of the discourse structure. Richer models of discourse include hierarchical structure (Grosz and Sidner, 1986) and Rhetorical Structure Theory (Mann and Thompson, 1988). The application of gestural analysis to such models may lead to fruitful areas of future research.

Acknowledgments

We thank Aaron Adler, C. Mario Christoudias, Michael Collins, Lisa Guttentag, Igor Malioutov, Brian Milch, Matthew Rasmussen, Candace Sidner, Luke Zettlemoyer, and the anonymous reviewers. This research was supported by Quanta Computer, the National Science Foundation (CAREER grant IIS-0448168 and grant IIS-0415865) and the Microsoft Research Faculty Fellowship.

References

Narayanaswamy Balakrishnan and Valery B. Nevzorov. 2003. *A primer on statistical distributions*. John Wiley & Sons.

Doug Beeferman, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

José M. Bernardo and Adrian F. M. Smith. 2000. *Bayesian Theory*. Wiley.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Justine Cassell, Yukiko I. Nakano, Timothy W. Bickmore, Candace L. Sidner, and Charles Rich. 2001. Non-verbal cues for discourse structure. In *Proceedings of ACL*, pages 106–115.

Lei Chen, Mary Harper, and Zhongqiang Huang. 2006. Using maximum entropy (ME) model to incorporate gesture cues for sentence segmentation. In *Proceedings of ICMI*, pages 185–192.

Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. 2005. Behavior recognition via sparse spatio-temporal features. In *ICCV VS-PETS*.

Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. 2003. Recognizing action at a distance. In *Proceedings of ICCV*, pages 726–733.

Jacob Eisenstein and Randall Davis. 2007. Conditional modality fusion for coreference resolution. In *Proceedings of ACL*, pages 352–359.

David A. Forsyth and Jean Ponce. 2003. *Computer Vision: A Modern Approach*. Prentice Hall.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. *Proceedings of ACL*, pages 562–569.

Dariu M. Gavrila. 1999. Visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98.

Barbara Grosz and Julia Hirshberg. 1992. Some intonational characteristics of discourse structure. In *Proceedings of ICSLP*, pages 429–432.

Barbara Grosz and Candace Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of ACL*.

Julia Hirschberg and Christine Nakatani. 1998. Acoustic indicators of topic segmentation. In *Proceedings of ICSLP*.

Xuedong Huang, Fileno Alleva, Mei-Yuh Hwang, and Ronald Rosenfeld. 1993. An overview of the Sphinx-II speech recognition system. In *Proceedings of ARPA Human Language Technology Workshop*, pages 81–86.

Michael Johnston. 1998. Unification-based multimodal parsing. In *Proceedings of COLING*, pages 624–630.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Stefan Kopp, Paul Tepper, Kim Ferriman, and Justine Cassell. 2007. Trading spaces: How humans and humanoids use speech and gesture to give directions. In Toyoaki Nishida, editor, *Conversational Informatics: An Engineering Approach*. Wiley.

Ivan Laptev. 2005. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123.

David G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of ICCV*, volume 2, pages 1150–1157.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of ACL*, pages 25–32.

Igor Malioutov. 2006. Minimum cut model for spoken lecture segmentation. Master's thesis, Massachusetts Institute of Technology.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Craig Martell. 2005. *FORM: An experiment in the annotation of the kinematics of gesture*. Ph.D. thesis, University of Pennsylvania.

David McNeill. 1992. *Hand and Mind*. The University of Chicago Press.

Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. 2006. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. In *Proceedings of the British Machine Vision Conference*.

NIST. 2003. The Rich Transcription Fall 2003 (RT-03F) Evaluation plan.

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Francis Quek. 2003. The catchment feature model for multimodal language analysis. In *Proceedings of ICCV*.

Mark Steedman. 1990. Structure and intonation in spoken language understanding. In *Proceedings of ACL*, pages 9–16.

Marc Swerts. 1997. Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America*, 101:514.

Gokhan Tur, Dilek Hakkani-Tur, Andreas Stolcke, and Elizabeth Shriberg. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of ACL*, pages 491–498.