Discourse Topic and Gestural Form

Discourse Topic and Gestural Form

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

Abstract

Coverbal gesture provides a channel for the visual expression of ideas. While some gestural emblems have culturally predefined forms (e.g., "thumbs up"), the relationship between gesture and meaning is, in general, not conventionalized. It is natural to ask whether such gestures can be interpreted in a speaker-independent way, or whether gestural form is determined by the speaker's idiosyncratic view of the discourse topic. We address this question using an audiovisual dataset across multiple speakers and topics. Our analysis employs a hierarchical Bayesian author-topic model, in which gestural patterns are stochastically generated by a mixture of speaker-specific and topic-specific priors. These gestural patterns are characterized using automatically-extracted visual features, based on spatio-temporal interest points. This framework detects significant cross-speaker patterns in gesture that are governed by the discourse topic, suggesting that even unstructured gesticulation can be interpreted across speakers. In addition, the success of this approach shows that the semantic characteristics of gesture can be detected via a low-level, interest point representation.

Introduction

Coverbal gesture supplements speech with a visual representation of the underlying discourse semantics. While some gestural emblems have culturally predefined forms (e.g., "thumbs up"), the relationship between gesture and meaning is, in general, unstructured. In contrast to language, gesture is not organized grammatically and the majority of gesture patterns do not come from a predefined lexicon (McNeill 1992). Thus, it is natural to ask whether and how gestures reflect the topic of discourse.

While gesture conveys discourse information, its form may be idiosyncratic to the speaker. If multiple speakers employ their own distinct gestural patterns for a given topic, then it would be difficult to leverage gesture in a speaker-general way. Indeed, previous work on multimodal discourse analysis has applied gesture in a speaker-specific fashion (e.g., Eisenstein and Davis 2007). The speaker-specificity of gesture has theoretical implications as well, shedding light on the question of how human viewers extract content from co-speech gesture. Empirical research finds that viewers are sensitive to the relationship between gesture and semantics (Kelly *et al.* 1999), suggesting either that gestural form is not entirely idiosyncratic, or that viewers dynamically build a speaker-specific model of gesture over the course of a conversation. Our research attempts to quantify the topic- and speaker-dependence of gesture from automatically extracted visual features.

The relationship between gestural form and meaning is an open question in psychology and cognitive science. Indeed, the extent to which co-speech gesture affects listeners' comprehension is a subject of debate (Kelly *et al.* 1999; Krauss 2001; McNeill 1992). Many researchers have focused on a micro-scale study of individual gestures and their relationship to discourse structure and semantics (Quek *et al.* 2002). An important complementary approach would be to investigate this phenomenon across a broad range of speakers and discourse topics, which is possible only with automated methods that can easily be applied to multiple speakers and topics.

In this paper, we present an unsupervised technique for automatically quantifying the extent to which the gestural forms in a dataset are shaped by speaker-specific and topic-specific factors. The approach is driven by a new characterization of gesture in terms of spatiotemporal interest points (Laptev 2005): a robust, low-level representation that can be extracted from multiple videos without manual intervention. These interest points are then clustered into a lexicon of gestural "codewords," and the distribution of codewords across speakers and topics forms the backbone of our analysis. We employ a hierarchical Bayesian author-topic model, which learns a lexicon of gestural forms, while jointly learning to associate lexical items with specific speakers and topics.

We find that discourse topic exerts a consistent influence on gestural form, even across speakers. This finding is based on both the Bayesian model and traditional frequentist analysis, using a corpus of 33 short videos in which multiple speakers describe each of five topics. Both tests indicate that a significant proportion of gestural features are produced in a topic-specific, speaker-general fashion.

Related Work

A large body of prior research addresses the relationship between gesture and meaning. For instance, McNeill (1992) and Kendon (2004) explore the semantic relevance of individual gestural forms, as described by human annotators. Automatically extracted visual features can also be used to aid such an analysis (Quek *et al.* 2002). However, we are aware of little prior work quantitatively comparing the distribution of gesture forms across speakers and topics, though the literature contains a few relevant qualitative studies. McNeill (1992) explores variability in gesture location and hand shape across six speakers, demonstrating significant inter-speaker variation. He also presents examples showing that the use of individual handshapes is motivated by the semantics of the utterance, suggesting that a given topic could be predisposed to specific gestural forms. Analyzing the relationship between gestures and visual metaphors among four speakers, Webb (1996) finds commonalities in the gestural expression of certain ideas. We extend this work with a quantitative analysis of the relationship between gestural form and topic.

Author-topic models

The Bayesian model that we employ is inspired by the author-topic model of Rosen-Zvi *et al.* (2004). In their model, each author induces a distribution of topics, which in turn induces a distribution over words. Rather than modeling the distribution of words, we model the distribution of gestural forms, which are organized into a lexicon in a joint clustering step. Another important difference is that for Rosen-Zvi *et al.*, topics were hidden, while in our case they are known. Thus, in our model, both topics and authors induce distributions over codewords, and a hidden auxiliary variable determines the distribution from which each codeword is drawn, given the author and topic.

Spatiotemporal Interest Points for Gesture

We employ visual features that describe motion at a sparse set of *spatiotemporal interest points* (Laptev 2005). Interest points are defined as high-contrast image regions – especially corners and edges – that undergo complex motion. Two examples of interest points detected in our dataset are shown in Figure 1. Both show the hand moving up out of rest position against a black background; these two interest points are clustered together by our model, as described in the next section.

At each detected interest point, the visual, spatial, and kinematic characteristics are concatenated into high-dimension feature vectors. Principal component analysis (PCA) is applied to reduce the dimensionality to a manageable size (Bishop 2006); the resulting transformed feature vectors comprise the final representation of gestural features.

This approach is motivated by techniques from the computer vision task of *activity recognition* (Efros *et al.* 2003). The goal of activity recognition is to classify video sequences into semantic categories: e.g., walking, running, jumping. As a simple example, a classifier may learn that a key difference between videos of walking and jumping is that walking is characterized by horizontal motion and jumping is characterized by vertical motion. Spurious vertical motion in a walking video is unlikely to confuse the classifier as long as the large majority of interest points move horizontally. Our hypothesis 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.

We apply the Activity Recognition Toolbox (Dollár *et al.* 2005)[1] to detect spatiotemporal interest points in our dataset. At each interest point, we consider a small space-time volume of nearby pixels and take the brightness gradient in the temporal and spatial dimensions. This representation captures local spatial and temporal characteristics of the motion in the video. The spatial position of the interest point is added to the feature vector, which is then reduced using PCA.

This visual feature representation is at a lower level of abstraction than the usual 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). A similar set of attributes is used in the formal semantics of gesture of Lascarides and Stone (2006). Working with automatic hand tracking, Quek *et al.* automatically compute perceptually-salient gesture features, such as holds (Bryll, Quek, & Esposito 2001) and oscillatory repetitions (Xiong & Quek 2006).

In contrast, our visual feature representation is 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 still camera and static background. Second, it is robust: while image noise may cause a few spurious interest points, the majority will guide the system towards 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.

Gestural Lexicon and Author-Topic Model

To assess the relative contributions of speaker and topic to the form of the gesture, we employ a hierarchical Bayesian model (Gelman *et al.* 2004). The model induces a clustering over gesture features and determines whether each feature is generated by the speaker or the topic. This determination is governed by a hidden parameter, whose value we infer in an unsupervised fashion.

The model assumes a sparse representation, in which gesture is characterized by a set of spatiotemporal interest

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

Figure 1: The two rows show examples of interest points that were clustered together by our model; both show the hand moving up out of rest position against a dark background. The center column shows the frame in which the interest point is detected; the left and right columns are 5 frames (166 milliseconds) before and after, respectively.

points, each defined by a real-valued tuple. Each interest point is assumed to be generated from a mixture model. Interest points that are generated by the same mixture component should be visually similar. These components serve as cluster centers and are called gestural "codewords" – Figure 1 shows two examples of a single codeword, indicating upward motion against a dark background. The sufficient statistics of the mixture components are shared across all dialogues; however, the component weights vary by both topic and speaker. In this way, the model learns a global lexicon of visual forms, while jointly learning a distribution of visual forms with respect to speaker and topic.

The distribution over components for each speaker and topic is represented by a multinomial distribution; a hidden auxiliary variable decides whether each codeword is drawn from the speaker-specific or topic-specific distributions. The parameter governing this hidden variable indicates the model's assessment of the relative importance of speaker and topic for gestural form.

The plate diagram for the model is shown in Figure 2. Each of the D dialogues is characterized by N_d visual features, which are written $\mathbf{x}_{d,i}$. Each visual feature vector $\mathbf{x}_{d,i}$ is generated from a multivariate Gaussian, $\mathbf{x}_{d,i} \sim \mathcal{N}(\mu_{z_{d,i}}, \sigma_{z_{d,i}})$, where $z_{d,i}$ indicates the codeword and σ is a diagonal covariance matrix. This induces a standard Bayesian mixture model over gesture features (Bishop 2006). Each $z_{d,i}$ is drawn from either a speaker- or topic-specific multinomial, depending on the auxiliary variable $c_{d,i}$. If $c_{d,i} = 0$, then $z_{d,i} \sim \phi_{s_d}$, where s_d is the identity of the speaker for document d. If $c_{d,i} = 1$, then $z_{d,i} \sim \theta_{t_d}$, where t_d is the topic of document d. The distribution of c is governed by a binomial distribution with parameter λ.

Weakly informative conjugate priors are employed for all model parameters (see Gelman *et al.*, 2004). Specifically, the parameters μ and σ are drawn from a Normal-Inverse-

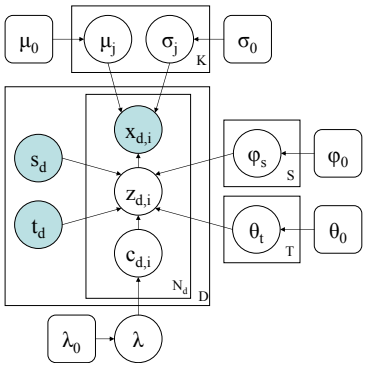

Figure 2: A plate diagram showing the dependencies in our model. Filled circles indicate observed variables, empty circles indicate hidden variables, and rounded rectangles indicate priors.

Wishart distribution centered at the mean and variance of the observed data (Bishop 2006). The multinomials ϕ and θ are drawn from symmetric Dirichlet priors, with parameter $\phi_0 = \theta_0 = .1$. The binomial parameter λ is drawn from a weakly informative beta prior with parameters $(.1, .1)$. As shown below, the use of conjugate priors ensures that standard closed-form posteriors can be easily found.

Our goal is to learn the relative importance of speaker versus topic, captured in the posterior distribution of the parameter λ, given observed data $\mathbf{x}, \mathbf{s}, \mathbf{d}$. We employ Gibbs sampling, a widely-used and easily-implemented technique for inference in hierarchical Bayesian models (Gelman *et al.* 2004), which successively samples over the posterior for each hidden variable with respect to the rest of the model configuration. After initializing the parameters randomly, Gibbs sampling is guaranteed in the limit to con-

verge to the true distribution over the hidden variables, $p(\mathbf{z}, \mathbf{c}, \mu, \sigma, \lambda, \phi, \theta | \mathbf{x}, \mathbf{s}, \mathbf{t})$. We can then use the set of samples to construct Bayesian confidence intervals for λ.

Sampling Distributions

Gibbs sampling requires posterior sampling distributions for all of the hidden variables. Rao-Blackwellization(Bishop 2006) is used to reduce sampling variance by integrating out the parameters $\theta, \phi, \mu, \sigma$ and λ. This is possible through the use of conjugate priors. Thus we need sample only the hidden variables z and c.

We write $p(z_{d,i}| \ldots)$ to indicate the probability distribution of $z_{d,i}$ given all the variables in the model; $z_{-(d,i)}$ denotes all z except $z_{d,i}$. Finally, $N(z_{d,i}, t_d, c_{d,i})$ denotes the count of times the codeword $z_{d,i}$ was drawn from the topic-specific distribution for topic t_d. This is computed as $\sum_{d' \leq D} \delta(t'_d, t_d) \sum_{i' \leq N_{d'}} \delta(z_{d',i'}, z_{d,i}) \delta(c_{d',i'}, c_{d,i})$, where the delta function takes the value one if the arguments are equal, and zero otherwise.

$$p(z_{d,i} = j| \ldots) \propto p(\mathbf{x}_{d,i} | \mu^{(j)}, \sigma^{(j)}) p(z_{d,i} = j | c_{d,i}, \phi^{(s_d)}, \theta^{(t_d)}) \tag{1}$$

$$p(z_{d,i} = j | c_{d,i}, \phi^{(s_d)}, \theta^{(t_d)}) = \begin{cases} \phi_j^{(s_d)} & \text{if } c_{d,i} = 0 \\ \theta_j^{(t_d)} & \text{if } c_{d,i} = 1, \end{cases}$$

where $\phi^{(s_d)}$ is the multinomial distribution indexed by the speaker s_d, and $\phi_j^{(s_d)}$ is the entry for $z_{d,i} = j$ in that distribution. Integrating out the parameters μ and σ from the first part of equation 1, we obtain a student-T distribution, which may be approximated by a moment-matched Gaussian (Gelman *et al.* 2004). Integrating out the parameters ϕ and θ, we obtain,

$$\begin{aligned}
&p(z_{d,i} = j | c_{d,i}, z_{-(d,i)}, s_d, t_d, \phi_0, \theta_0) \propto \\
&\quad \int d\phi d\theta p(z_{d,i} = j | c_{d,i}, \phi^{(s_d)}, \theta^{(t_d)}) \times \\
&\quad p(\phi^{(s_d)} | z_{-(d,i)}, \phi_0) p(\theta^{(t_d)} | z_{-(d,i)}, \theta_0) \\
&= \int d\phi d\theta (\phi_j^{(s_d)} \delta(c_{d,i}, 0) + \theta_j^{(t_d)} \delta(c_{d,i}, 1)) \times \\
&\quad p(\phi^{(s_d)} | z_{-(d,i)}, \phi_0) p(\theta^{(t_d)} | z_{-(d,i)}, \theta_0) \\
&= \delta(c_{d,i}, 0) \int \phi_j^{(s_d)} p(\phi^{(s_d)} | z_{-(d,i)}, \phi_0) d\phi + \\
&\quad \delta(c_{d,i}, 1) \int \theta_j^{(t_d)} p(\theta^{(t_d)} | z_{-(d,i)}, \theta_0) d\theta \qquad (2) \\
&= \delta(c_{d,i}, 0) \frac{N(j, s_d, c_{d,i} = 0) + \phi_0}{N(., s_d, c_{d,i} = 0) + K\phi_0} + \\
&\quad \delta(c_{d,i}, 1) \frac{N(j, t_d, c_{d,i} = 1) + \theta_0}{N(., t_d, c_{d,i} = 1) + K\theta_0} \qquad (3)
\end{aligned}$$

The derivation of line 3 from line 2 follows from standard Dirichlet-Multinomial conjugacy (Gelman *et al.* 2004), enabling us to compute the posterior probability of $z_{d,i}$ in a ratio of counts. Sampling c is more straightforward:

$$\begin{aligned}
&p(c_{d,i} | c_{-(d,i)}, z_{d,i}, \lambda_0) \propto \\
&\quad p(z_{d,i} | c_{d,i}, z_{-(d,i)}, t_d, s_d, \phi_0, \theta_0) \times \\
&\quad \int p(c_{d,i} | \lambda) p(\lambda | c_{-(d,i)} \lambda_0) d\lambda
\end{aligned}$$

The first part of the product is defined above. The integral can be handled analogously, as Beta-Binomial conjugacy is a special case of Dirichlet-Multinomial conjugacy,

$$\int p(c_{d,i} | \lambda) p(\lambda | c_{-(d,i)} \lambda_0) d\lambda = \frac{N(c_{d,i}) + \lambda_0}{N + 2\lambda_0}. \tag{4}$$

As both $z_{d,i}$ and $c_{d,i}$ are categorical variables, we can consider all possible pairs of values, thus sampling from z and c jointly. These parameters are tightly coupled, and sampling them together substantially speeds convergence. The joint sampling distribution is given by,

$$\begin{aligned}
&p(z_{d,i}, c_{d,i} | \ldots) = \\
&\quad p(z_{d,i} | c_{d,i}, z_{-(d,i)}, s_d, t_d, \phi_0, \theta_0) p(c_{d,i} | c_{-(d,i)}, \lambda_0),
\end{aligned}$$

where the first part of the product is defined in equation 3 and the second part is defined in equation 4.

Implementation Details

Our model includes four tunable parameters: the number of iterations of Gibbs sampling to run, the number of interest points to extract, the number of mixture components K, and the dimensionality of the gesture features after PCA.

Gibbs sampling is performed along five parallel runs for 15000 iterations each. The first 5000 iterations are considered a "burn-in" period, and confidence intervals are estimated from the remaining 10000. The number of interest points extracted is set to 1/10 the number of frames in each video; on average, 390 interest points were extracted per video. The number of components was set to 100, and the dimensionality of the gesture features after PCA was set to 5. These parameter settings were determined before the experiments were run, and were chosen based on speed and memory availability. In general, these settings impact the gesture clustering and do not directly affect the assignment of codewords to the speaker or topic; however, alternative settings may be considered in future work.

Dataset

The dataset for our experiments includes 33 short videos of dialogues, in which fifteen speakers describe one of a total of five different topics. There were between one and three videos of each speaker. Speakers were recruited on a university campus, and ranged in age from 18 to 32. All were native or fluent speakers of English.

Dialogues were limited to three minutes in duration and consist of a conversation between friends; the experimenters were not present in the room. One participant, whom we will call *A*, was informed of the discourse topic in advance

and was required to explain it to participant *B*, who was later tested for comprehension. *A* was instructed to stand, while *B* sat. Both participants were permitted to speak, but not to draw examples; other visual references were not provided. Participants were not instructed to gesture, though all did so.

The topics for discussion consisted of a short "Tom and Jerry" cartoon and simulations of four mechanical devices: a piston, a candy dispenser, a pinball machine, and a toy. Only participant *A* knew any details about the topic prior to the conversation. We chose to emphasize mechanical devices because it seemed likely that direct gestural metaphors emphasizing the structure and motion of the device would be shared across speakers. We believe that such concrete discourse topics represent a best case scenario for finding speaker-general gestural forms. Whether our results extend to topics in which the gestural metaphors may be more idiosyncratic is an important question for future work.

Videos were recorded using camcorders; participants wore colored gloves, which were used in another study involving hand tracking, but not in this work. No video post-processing was performed. Only the video of participant *A* was used in our experiments here. The dataset is described in more detail by Eisenstein (2008).

Experimental Setup

The first experiment involves estimating the number of gestural features that are generated in a topic-specific manner, using the model described in the previous sections of the paper. We note that even if there were no topic-specific, speaker-general patterns in gesture in our corpus, the topic-specific model θ might be used to overfit the data. To account for this, we ran five baseline conditions in which the topic indicators for the documents were randomly shuffled, with the goal of isolating the extent to which the topic-specific model can be used to overfit. We then assess whether the topic-specific model of gestural forms is used significantly more frequently in conjunction with the true document topics than with the random topic assignments.

Just as in text, lexical distributions are indicative of discourse topic (Hearst 1994); thus, it may be profitable to examine the distribution of gestural "codewords" across topics. Our model builds a lexicon of gestures by clustering gesture features; this is done jointly with the assignment of gesture features to speakers and topics. Such a joint model is advantageous because it is possible to integrate over uncertainty in the clustering, rather than propagating the effects of a bad clustering decision to the other stages. However, it is illustrative to consider the maximum *a posteriori* (MAP) clustering induced by the model (Bishop 2006) and investigate how the distribution of cluster tokens varies by topic. To this end, our second experiment performs chi-squared analysis of the distribution of cluster membership, with respect to both topic and speaker.

Results

The results of the first experiment are shown in Table 1 and Figure 3. With the correct topic labels, 12% of gestures are classified as topic-specific. When the topic labels are corrupted, this average drops to less than 3%. Thus, the model uses the topic-specific codeword distributions mainly when the topic labels are actually informative, supporting the hypothesis of a connection between discourse topic and gestural form that transcends individual speakers.

condition	mean	upper	lower
true topic labels	.120	.136	.103
random topic labels	.0279	.0957	0

Table 1: Proportion of gestures assigned to the topic-specific model, with 95% confidence intervals

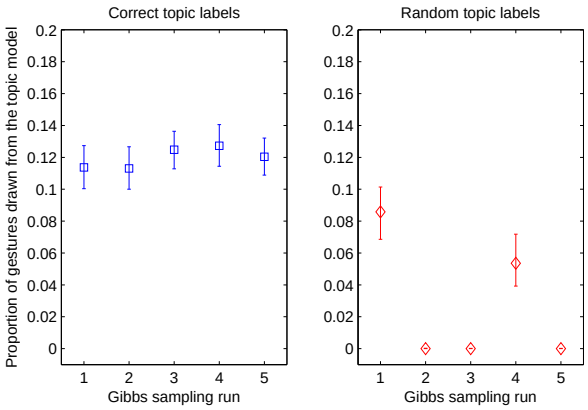

Figure 3: Proportion of gestures assigned to the topic model, per run.

Bayesian confidence intervals constructed from the samples show that these differences are robust. As indicated in Table 1, the confidence intervals for the randomized conditions is much larger. This is expected, as each randomization of topic labels varies from the true labels to a different extent. Figure 3 illustrates this situation, showing the confidence intervals from each randomized run. In the topic-randomized condition, there is substantial between-run variance; in three of the runs, the topic exerts no influence whatsoever. In contrast, in the condition with correct topic labels, the influence of the topic-specific model is consistently in the range of 12%.

Next, we analyze the influence of topic and speaker on gestural form using the classical chi-squared test. We find the maximum *a posteriori* (MAP) gesture feature clustering by selecting the iteration of Gibbs sampling with the highest likelihood.[2] The chi-squared test is used to determine whether the distribution of clusters differs significantly according to topic and speaker (De Groot & Schervish 2001).

[2] In sampling-based inference, MAP estimates are often found by taking a mean or mode over multiple samples. However, in the case of estimating a clustering, this technique suffers from non-identifiability. For example, two data points may appear in many different clusters, though usually together. However, their modal cluster memberships may differ, causing them to be separated in the MAP estimate.

Strong effects were found for both topic and speaker. For topics, $p < .01, \chi^2 = 1.12 * 10^4, \text{dof} = 439$.[3] For speakers, $p < .01, \chi^2 = 5.94 * 10^4, \text{dof} = 1319$. While the chi-squared values are not directly comparable – due to the different degrees of freedom – this experiment indicates an important effect from both the topic and speaker.

Discussion

The results support the existence of topic-specific gestural forms that are shared across speakers. The frequency of such shared gestural forms is likely influenced by both the population of speakers and the topics of discussion. The speakers in our dataset are all American residents and fluent speakers of English. The extent to which gestural forms are shared across cultures is a key topic for future research. We are also interested to study whether gestural forms are shared when the discourse topics are less concrete: do multiple speakers use similar gestures when talking about, say, their circle of friends, or their ideas on politics?

But while our dataset was designed to encourage speaker-general gestures, we believe that any automatic vision-based technique for gesture analysis is likely to *overstate* speaker-specific factors. This is because it is difficult, if not impossible, to abstract away all features of the speaker's appearance. The principal visual features that we leverage are brightness gradients and the location of movement. Brightness gradients are influenced by the speaker's skin tone and clothing; location of movement is influenced by anatomical factors such as the speaker's height. Thus, the likelihood of such visual features being clustered in a speaker-dependent manner is artificially inflated. We believe that with the development of robust vision techniques that describe gesture's visual form on a more abstract level, future work may show that topic exerts a greater influence than reported here.

Other extensions to the interest-point based characterization of gesture are possible. Our representation of gesture focuses on individual, local visual features. This is sufficient to describe a range of gestural forms, such as handshapes and paths of motion. However, it does not account for higher-level phenomena, such as when both hands move in a synchronized or anti-synchronized fashion. Rather than assigning codewords to individual, local visual features, it may be advantageous to consider sets of local features that frequently co-occur. Such an approach may result in a characterization of gesture that better coheres with human perception.

Finally, we consider this research to be an early example of a new methodology for studying the communicative role of gesture. Previous techniques relying on hand annotation have focused on close analysis of a small number of particularly illustrative gestures. We believe that automated techniques, as described in this paper, offer a complementary approach, potentially enabling the study of gesture across large populations under a variety of conditions. In the future, we hope to consider more fine-grained and structured representations of dialogue semantics, which may shed further light on the interaction between gesture, speaker, and meaning.

[3] Clusters with fewer than five examples were excluded, as the chi-squared text is not accurate for small bin values. Thus, the number of degrees of freedom is less than the expected $KT - 1$.

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Bryll, R.; Quek, F.; and Esposito, A. 2001. Automatic hand hold detection in natural conversation. In *IEEE Workshop on Cues in Communication*.

De Groot, M. H., and Schervish, M. J. 2001. *Probability and Statistics*. Addison Wesley.

Dollár, P.; Rabaud, V.; Cottrell, G.; and Belongie, S. 2005. Behavior recognition via sparse spatio-temporal features. In *ICCV VS-PETS*.

Efros, A. A.; Berg, A. C.; Mori, G.; and Malik, J. 2003. Recognizing action at a distance. In *Proceedings of ICCV*, 726–733.

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

Eisenstein, J. 2008. *Gesture in Automatic Discourse Processing*. Ph.D. Dissertation, Massachusetts Institute of Technology.

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

Gelman, A.; Carlin, J. B.; Stern, H. S.; and Rubin, D. B. 2004. *Bayesian data analysis*. Chapman and Hall/CRC.

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

Kelly, S.; Barr, D.; Church, R.; and Lynch, K. 1999. Offering a Hand to Pragmatic Understanding: The Role of Speech and Gesture in Comprehension and Memory. *Journal of Memory and Language* 40(4):577–592.

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

Krauss, R. 2001. Why do we gesture when we speak? *Current Directions in Pschological Science* 7(54-59).

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

Lascarides, A., and Stone, M. 2006. Formal Semantics for Iconic Gesture. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL)*, 64–71.

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

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

Quek, F.; McNeill, D.; Bryll, R.; Duncan, S.; Ma, X.; Kirbas, C.; McCullough, K. E.; and Ansari, R. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction* 9:3:171–193.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. In *Proceedings of Uncertainty in artificial intelligence (UAI)*, 487–494.

Webb, R. 1996. *Linguistic features of metaphoric gestures*. Ph.D. Dissertation, University of Rochester.

Xiong, Y., and Quek, F. 2006. Hand Motion Gesture Frequency Properties and Multimodal Discourse Analysis. *International Journal of Computer Vision* 69(3):353–371.