Turning Lectures into Comic Books Using Linguistically Salient Gestures

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

Creating video recordings of events such as lectures or meetings is increasingly inexpensive and easy. However, reviewing the content of such video may be time-consuming and difficult. Our goal is to produce a "comic book" summary, in which a transcript is augmented with keyframes that disambiguate and clarify accompanying text. Unlike most previous keyframe extraction systems which rely primarily on visual cues, we present a linguistically-motivated approach that selects keyframes that contain salient gestures. Rather than learning gesture salience directly, it is estimated by measuring the contribution of gesture to understanding other discourse phenomena. More specifically, we bootstrap from multimodal coreference resolution to identify gestures that improve performance. We then select keyframes that capture these gestures. Our model predicts gesture salience as a hidden variable in a conditional framework, with observable features from both the visual and textual modalities. This approach significantly outperforms competitive baselines that do not use gesture information.

Introduction

Producing video recordings of lectures and meetings is increasingly easy, but reviewing video material is still a time consuming task. A summary presenting key points of the video could enhance the user's ability to quickly review this material.

We propose a method for summarizing a specific type of video, in which a speaker makes a presentation using a diagram or chart. Examples include academic lectures, business presentations, weather reports, and instructional videos for operating appliances. Such videos are often shot without cuts, camera pans and zooms, and other edits, so the primary source of interesting visual information is the gesture and body language of the speaker.

Summarization of this material is appealing from a modeling perspective: neither the images, nor the transcript in isolation is sufficient for understanding the content of the recording. To see why a textual transcript may be insufficient, consider the following sample, from a presentation about a mechanical device:

> This thing is continually going around, and this thing is continually going around. So these things must be like powered separately.

Without some form of contextual information, such as visual cues, this text is of little use. Our goal is to produce a "comic book" summary, in which a transcript is augmented with salient *keyframes* – still images that clarify the accompanying text. For example, in Figure 1, the references in the text are disambiguated by the pointing gestures shown in the keyframes. Ideally, we would select keyframes that avoid redundancy between the visual and verbal modalities, while conveying all relevant information. Many linguists and psychologists posit that gesture supplements speech with unique semantic information (McNeill 1992). If so, keyframes with salient gestures would be a valuable addition to the transcript text. Therefore, we seek to identify such gestures.

One possible approach is to formulate gesture extraction as a standard supervised learning task, using a corpus in which salient gestures are annotated. However, such annotation is expensive, and we prefer to avoid it. Instead, we learn gesture salience by bootstrapping from another task in language understanding. The task must have the property that salient gestures improve performance, such as in coreference resolution. Gesture salience is treated as a hidden variable that allows the gesture features to be included only when they are likely to improve performance; otherwise only textual features are used. This hidden variable is learned jointly with the task labels in a conditional model.

We evaluate our keyframe selection method on a collection of recordings in which students explain diagrams of mechanical devices. The set of automatically extracted frames is compared against a manually annotated ground truth. Our method yields statistically significant improvement over competitive baseline systems that use visual or textual features, but do not model gesture directly. These results confirm our hypothesis about the contribution of gesture cues in identification of informative non-redundant keyframes.

In the next section, we describe gesture salience in greater detail, and give a model to estimate it. We then summarize our feature set and keyframe selection algorithm. Next, we describe our experimental setup, and provide results and error analysis. Finally, we review related work and summarize our contributions.

Appeared in *Proceedings of AAAI 2007*, pp.877 - 882.

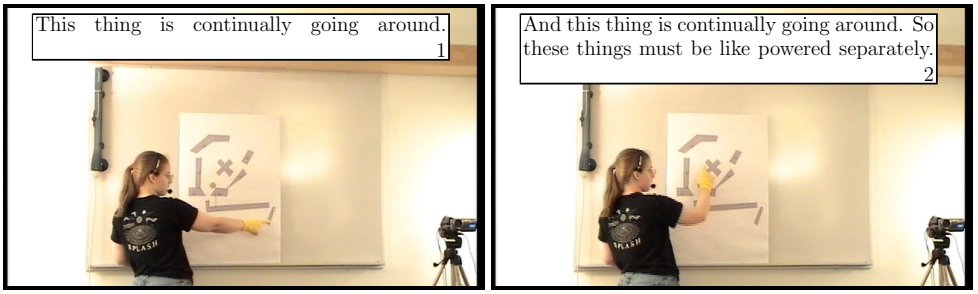

Figure 1: An excerpt of output generated by our keyframe selection system.

Identifying Salient Gestures

When humans communicate face-to-face, they convey information in both verbal and visual modalities. Coverbal hand gestures form one important class of visual cues. Psychologists believe that such gestures can provide relevant information that is not redundant with the speech (McNeill 1992). Gestures are commonly used to convey visual information that is difficult to describe verbally, such as shapes and trajectories. Through the tight semantic connection between gesture and speech, gesture can also be used referentially to bring other visual entities into discourse. For example, in Figure 1, the speaker resolves the ambiguous noun phrase "this thing" by pointing to a relevant part of the diagram.

Thus, keyframes that capture meaningful gesture can play an important role in resolving ambiguities in speech. But not all gestures are essential for understanding. For example, some hand movements are irrelevant, such as adjusting one's glasses; others are redundant, merely repeating information that is fully specified in the speech. To use gestures as a basis for keyframe extraction, we must identify the ones that are salient.

One way to measure the relevance of gesture is by its contribution to understanding the semantics of the discourse. In practical terms, we select a well-defined language understanding task in which gestures could potentially improve performance. An example of such a task is multimodal coreference resolution. By learning to predict the specific instances in which gesture helps, we can obtain a model of gesture salience. For example, we expect that a pointing gesture in the presence of an anaphoric expression would be found to be highly salient (as in Figure 1); a more ambiguous hand pose in the presence of a fully-specified noun phrase would not be salient. This approach will not identify *all* salient gestures, but will identify those gestures that occur in the context of the selected language understanding task. For example, in coreference resolution, only gestures that co-occur with noun phrases can be selected. Since noun phrases are ubiquitous in language, this should still cover a usefully broad collection of gestures.

We build a model for coreference resolution that incorporates both verbal and gestural features. We then augment this model with a hidden variable that controls whether gesture features are taken into account for each specific instance. In this model, the hidden variable is learned jointly with coreference, so that the resulting model not only predicts coreference resolution, but also outputs the probability distribution for the hidden variable. Instances in which gesture features are included with high likelihood are thought to correspond to salient gestures. The gestures rated most salient by this method are included as keyframes in the summary.

A Hidden Variable Model for Gesture Salience

Consider a linear model, with observable features in both the gestural ($\mathbf{x}_g$) and verbal ($\mathbf{x}_v$) modalities. We have labels $y \in \{-1, 1\}$, which describe some phenomenon that we would like to learn to predict, such as noun phrase coreference. When the speaker is gesturing meaningfully, we can improve prediction of y by including the gesture features $\mathbf{x}_g$; when the speaker is not gesturing meaningfully, we would be better off not including them. This decision can be built into the model by the incorporation of a hidden variable $h \in \{-1, 1\}$, which controls whether the gesture features are included. Our hypothesis is that h will also serve as an indicator of gesture salience, so that keyframes should be selected when $Pr(h = 1)$ is high. Our approach is to learn to predict h jointly with y, and then later leverage our model $p(h|\mathbf{x})$ to select keyframes that include meaningful gestures.

In a conditional model parametrized by weights $\mathbf{w}$, we have:

$$p(y|\mathbf{x};\mathbf{w}) = \sum_h p(y,h|\mathbf{x};\mathbf{w}) \quad (1)$$

$$= \frac{\sum_h exp(\psi(y,h,\mathbf{x};\mathbf{w}))}{\sum_{y',h} exp(\psi(y',h,\mathbf{x};\mathbf{w}))} \quad (2)$$

Here, ψ is a potential function representing the compatibility between the label y, the hidden variable h, and the observations $\mathbf{x}$; this potential is parametrized by a vector of weights, $\mathbf{w}$. The numerator expresses the compatibility of the label y and observations $\mathbf{x}$, summed over all possible values of the hidden variable h. The denominator sums over both h and all possible labels y', yielding the conditional probability $p(y|\mathbf{x};\mathbf{w})$. For more discussion of hidden variables in conditionally-trained models, see (Quattoni, Collins, & Darrell 2004).

Coreference Resolution

The specific labeling problem that we choose is coreference resolution: the binary classification problem of determin-

ing whether each pair of noun phrases in a document refers to the same semantic entity. It has previously been shown that the gestures accompanying coreferent noun phrases are more likely to be similar than the gestures accompanying noun phrases that are not coreferent (Eisenstein & Davis 2006). Thus, features that quantify gesture similarity may improve performance on coreference resolution, when the gestures during both noun phrases are meaningful.

Coreference is a property of *pairs* of noun phrases – for gesture similarity to provide meaningful features for coreference resolution, the gesture must be relevant at both noun phrases. We define two hidden variables, h_1 and h_2, representing the salience of gesture at the first (antecedent) and second (anaphor) noun phrases, respectively. The gesture features are included only if $h_1 = h_2 = 1$. This yields the following definition of the potential function:

$$\begin{aligned}\psi(y, h_1, h_2, \mathbf{x}; \mathbf{w}) \quad \equiv \quad & y(\mathbf{w}_v^T \mathbf{x}_v + \delta_1(h_1)\delta_1(h_2)\mathbf{w}_g^T \mathbf{x}_g) \\ & + h_1 \mathbf{w}_h^T \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^T \mathbf{x}_{h_2}\end{aligned} \tag{3}$$

Here, $\delta_1(h)$ is an indicator function, which is 1 when $h = 1$, and zero otherwise. $\mathbf{x}_v$ and $\mathbf{x}_g$ are the verbal and gestural features, respectively. $\mathbf{x}_{h_i}$ is a subset of verbal and gestural features that are used to predict the value of the hidden variable h_i, with $i \in \{1, 2\}$; these features do not measure the similarity between pairs of noun phrases or gestures, but rather, are properties of individual noun phrases or gestures.

An example of a feature in x_g is the Euclidean distance between the average hand positions during the two gestures, with a smaller distance predicting coreference. An example of a feature in x_h is the distance of the hand from the speaker's lap, where meaningful gestures are unlikely to occur. The features used in our implementation are discussed in greater detail below.

Training Procedure

To learn the weight vector $\mathbf{w}$, we employ a gradient-based search to optimize the conditional log-likelihood of all labels in our dataset, given the observations:

$$\begin{aligned} l(\mathbf{w}) \quad &= \quad \sum_i ln(p(y_i|\mathbf{x}_i; \mathbf{w})) \\ &= \quad \sum_i ln \frac{\sum_h exp(\psi(y_i, h, \mathbf{x}_i; \mathbf{w}))}{\sum_{y',h} exp(\psi(y', h, \mathbf{x}_i; \mathbf{w}))} \end{aligned}$$

Taking partial derivatives with respect to the weights, we obtain the following gradient function of the likelihood, given a training example $\langle \mathbf{x}_i, y_i \rangle$.

$$\begin{aligned} \frac{\partial l_i}{\partial w_j} \quad = \quad & \sum_h p(h|y_i, \mathbf{x}_i; \mathbf{w}) \frac{\partial}{\partial w_j} \psi(y_i, h, \mathbf{x}_i; \mathbf{w}) \\ & - \sum_{y',h} p(h, y'|\mathbf{x}_i; \mathbf{w}) \frac{\partial}{\partial w_j} \psi(y', h, \mathbf{x}_i; \mathbf{w}) \end{aligned}$$

Finally, we derive the gradients that are specific to our potential function (Equation 3). We take the partial derivatives of ψ with respect to the weights for each type of feature:

$$\begin{aligned} \frac{\partial \psi}{\partial (w_j \in \mathbf{w}_v)} \quad &= \quad y x_j \\ \frac{\partial \psi}{\partial (w_j \in \mathbf{w}_g)} \quad &= \quad \delta_1(h_1)\delta_2(h_2) y x_j \\ \frac{\partial \psi}{\partial (w_j \in \mathbf{w}_h)} \quad &= \quad h_1 x_{j_1} + h_2 x_{j_2} \end{aligned}$$

This objective function and set of gradients are optimized using L-BFGS, a quasi-Newton numerical optimization technique (Liu & Nocedal 1989). Standard L2-norm regularization is employed to prevent overfitting, using cross-validation to select the regularization constant. Although the objective function for linear conditional models is convex, the inclusion of the hidden variable renders our objective non-convex. Thus, convergence to a global minimum is not guaranteed.

Features

The performance of our implementation depends on selecting features that effectively predict coreference and gesture salience. We describe the verbal and gesture features used for coreference ($\mathbf{x}_v$ and $\mathbf{x}_g$ in Equation 3), and the features used for predicting gesture salience ($\mathbf{x}_h$).

Features for Coreference

Verbal Features for Coreference The set of verbal features is drawn from state-of-the-art coreference resolution systems that operate on text.[1] Pairwise verbal features that predict the compatibility of two noun phrases (NPs) include: several string-match variants; distance features, measured in terms of the number of intervening noun phrases and sentences between the candidate NPs; and some syntactic features that can be computed from part of speech tags. Single-phrase verbal features predict whether a noun phrase is likely to participate in coreference relations as an antecedent or anaphoric noun phrase. For example, pronouns are likely to participate as anaphoric NPs in coreference relations, and are unlikely to be antecedents; indefinite noun phrases (e.g., "a ball") are not likely to participate in coreference relations at all.[2]

Gesture features for Coreference Similar gestures are thought to suggest semantic similarity (McNeill 1992). For example, two noun phrases are more likely to corefer if they are accompanied by identically-located pointing gestures. In cases in which these features successfully predict coreference, gesture is likely to be salient; this suggests a promising location for keyframe extraction.

Three types of gesture similarity are included in our feature set. The most straightforward is Euclidean distance, which captures cases in which the speaker is performing a

[1] See (Daumé III & Marcu 2005) for a detailed analysis of verbal features for coreference.

[2] All verbal features are computed over manual transcriptions of the speech, although in principle the output of a speech recognition system could be used. We apply forced alignment to obtain accurate time-stamps for each transcribed word.

gestural "hold" in roughly the same location. Euclidean distance may not correlate directly with semantic similarity – when gesturing at a detailed part of a diagram, very small changes in hand position may be semantically meaningful, while in other regions, positional similarity may be defined more loosely. For this reason, we use a hidden Markov model (HMM) to perform a spatio-temporal clustering on hand positions, and report whether gestures are clustered together. These features capture the similarity between static gestures, but similarity in gesture trajectories may also indicate semantic similarity. Dynamic time warping is used to quantify the similarity of gesture trajectories (Darrell & Pentland 1993).

All features are computed from hand and body pixel coordinates, which are obtained via computer vision, using a system modeled after (Deutscher, Blake, & Reid 2000). From informal observation, we estimate the system tracks the hands correctly roughly 90% of the time. Temporally, gesture features are computed over the duration of the associated noun phrase. Only the hand that is farthest from the body center is considered in computing the gesture features.

Features for Gesture Salience

The meta features $\mathbf{x}_h$ in Equation 3 are a subset of the verbal and gestural features. They predict *gesture salience* – whether or not the gesture is necessary to determine coreference. Our underlying hypothesis is that salient gestures make useful keyframes. While coreference is a property of *pairs* of noun phrases, gesture salience is a property to be evaluated at individual points in time; consequently, only single-gesture and single-phrase features are permitted to be meta-features.

Verbal Features for Gesture Salience Meaningful gesture has been shown to be more frequent when the associated speech is ambiguous (Melinger & Levelt 2004). Kehler (2000) finds that fully-specified noun phrases are less likely to receive multimodal support. These findings suggest that pronouns should be likely to co-occur with meaningful gestures, while definite NPs and noun phrases that include adjectival modifiers should be less likely to do so. To capture these intuitions, all single-phrase verbal features are included as meta-features.

Non-verbal Features for Gesture Salience Research on gesture has shown that semantically meaningful hand motions usually take place away from "rest position," which is located at the speaker's lap or sides (McNeill 1992). Effortful movements away from these default positions can thus be expected to predict that gesture is being used to communicate. We identify rest position as the center of the body on the x-axis, and at a fixed, predefined location on the y-axis. Our feature set includes the average Euclidean distance of the hands from this rest position. In addition, we use an HMM to perform a spatio-temporal clustering on hand positions and velocities; using parameter tying, we identify a specific cluster that corresponds to rest position, and another cluster for gestures that are merely transitional. Gestures in these clusters are less likely to be salient for noun phrase coreference.

Keyframe Selection

By jointly learning a model of coreference resolution and gesture salience, we obtain a set of weights $\mathbf{w}_h$ that can be used to estimate gesture salience at each noun phrase. To quantify the gesture salience for the antecedent noun phrase, we sum Equation 3 over all possible values for y and h_2, obtaining $\sum_{y,h_2} \psi(y, h_1, h_2, \mathbf{x}; \mathbf{w}) = h_1 \mathbf{w}_h^T \mathbf{x}_{h_1}$. We find the potential for the case when the gesture is salient by setting $h_1 = 1$, yielding $\mathbf{w}_h^T \mathbf{x}_{h_1}$.[3] Our working assumption is that this potential is a reasonable proxy for the informativeness of a keyframe that displays the noun phrase's accompanying gesture.

The potential provides an ordering on all noun phrases in the document. We select keyframes from the midpoints of the top n noun phrases, where n is specified in advance (the number of keyframes returned by our system is assumed to be governed by the user's preference for brevity or completeness). Each keyframe is given a caption that includes the relevant noun phrase and accompanying text, up to the noun phrase in the next keyframe. A portion of the output of the system is shown in Figure 1.

Evaluation Setup

Our intrinsic evaluation setup is similar to the methodology developed for the Document Understanding Conference.[4] We assess the quality of the automatically extracted keyframes by comparing them to human-annotated ground truth.

Dataset We use a dataset in which participants explained the behavior of mechanical devices to a friend, with the aid of a pre-printed diagram. The dataset includes sixteen videos. The videos were limited to three minutes in length; most participants used all the allotted time. The average presentation was 437 words long.

Training Coreference Resolution As described above, our approach to keyframe extraction is based on a model for gesture salience that is learned from labeled data on coreference resolution. The training phase is performed as leave-one-out cross-validation: a separate set of weights is learned for each presentation, using the other fifteen presentations as a training set. The learned weights are used to obtain the values of the hidden variable indicating gesture salience, as described in the previous section.

This training procedure was effective for coreference resolution. Evaluation of coreference resolution is typically quantified in terms of Constrained Entity-Alignment F-measure (CEAF), a global quality metric measuring the degree of overlap between the ground truth and the coreference clusters returned by the system (Luo 2005). Using average-link clustering to partition noun phrases into global clusters, our system achieves a score of 56.4. On a unimodal textual corpus of news broadcasts and articles, Ng reports a CEAF of 62.3. Since the corpus differs, the results are not

[3] Note that if we consider the same noun phrase as the anaphor ($\mathbf{x}_{h_2}$) and sum over all possible values of h_1, the resulting potential is identical.

[4] http://duc.nist.gov

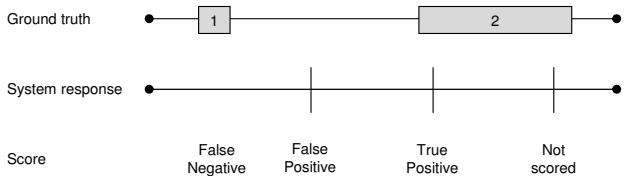

Figure 2: An example of the scoring setup.

directly comparable, but they suggest that our system's performance on global metrics like CEAF is not vastly different from state-of-the-art alternatives.

Ground Truth Annotation For evaluation, a set of ground truth annotations was created. Of the sixteen videos in the dataset, nine were annotated for keyframes. Of these, three were used in developing our system and the baselines, while the remaining six were used for final evaluation.

The goal of the ground truth annotation was to select keyframes that capture all visual information deemed crucial to understanding the content of the video. There were no specific instructions about selecting salient gestures. The number of selected frames was left to the discretion of the annotator; on average, 17.8 keyframes were selected per document, out of an average total of 4296.

One important difference between our dataset and standard sentence extraction datasets is that many frames may be nearly identical, due to the high granularity of video. For this reason, rather than annotating individual frames, the annotator marked *regions* with identical visual information. These regions define equivalence classes, such that any keyframe from a given region would be equally acceptable. If a single keyframe were selected from every ground truth region, the result would be the minimal set of keyframes necessary for a reader to fully understand the discourse. On average, the 17.8 regions selected per document spanned 568 frames. An example of the scoring setup is shown in Figure 2. The top row in the figure represents the ground truth; the middle row represents the system response, with vertical lines indicating selected keyframes; the bottom row shows how the response is scored.

Evaluation Metric The system returns a set of individual keyframes. For all systems the number of keyframes is fixed to be equal to the number of regions in the ground truth annotation. If the system response includes a keyframe that is not within any ground truth region, a false positive is recorded. If the system response fails to include a keyframe from a region in the ground truth, a false negative is recorded; a true positive is recorded for the first frame that is selected from a given ground truth region, but additional frames from the same region are not scored. The system is still penalized for each redundant keyframe, since it has "wasted" one of a finite number of keyframes that it is allowed to select. At the same time, such an error seems less grave than a true substitution error, in which a keyframe not containing relevant visual information is selected.

Model	F-Measure	Recall	Precision
GESTURE-SALIENCE	**.404**	**.383**	**.427**
POSE-CLUSTERING	.290	.290	.290
NP-SALIENCE	.239	.234	.245
RANDOM-KEYFRAME	.120	.119	.121

Table 1: Experimental results

Baselines We compare the performance of our system against three baselines, which we present in order of increasing competitiveness.

The RANDOM-KEYFRAME baseline selects n keyframes at random from throughout the document. The number of keyframes selected is equal to the number of regions in the ground truth. This baseline expresses a lower bound on the performance that any reasonable system should achieve on this task. Our results report the average of 500 independent runs.

The NP-SALIENCE system is based on frequency-based approaches to identifying salient NPs for the purpose of text summarization (Mani & Maybury 1999). The salience heuristic prefers the most common representative tokens of the longest and most homogeneous coreference clusters.[5] This provides a total ordering on NPs in the document; we select keyframes at the midpoint of the top n noun phrases, where is n is the number of keyframe regions in the ground truth.

Our final baseline, POSE-CLUSTERING, is based purely on visual features. It employs clustering to find a representative subset of frames with minimum mutual redundancy. Traditionally (e.g., (Uchihashi *et al.* 1999)), a clustering is performed on all frames in the video, using the similarity of color histograms as a distance metric. In our dataset, there is a single fixed camera and no change in the video except for the movements of the speaker; thus, the color histograms are nearly constant throughout. Instead, we use the tracked coordinates of the speaker's hands and upper-body, normalize all values, and use the Euclidean distance metric. In this setting, clusters correspond to typical body poses, and segments correspond to holds in these poses. As in (Uchihashi *et al.* 1999), the video is divided into segments in which cluster membership is constant, and keyframes are taken at the midpoints of segments. We use Uchihashi *et al.*'s importance metric for ranking segments, and choose the top n, where n is the number of keyframes in the ground truth.

Results

Using paired t-tests, we find that our approach – GESTURE-SALIENCE in Table 1 – significantly outperforms all alternatives ($p < .05$ in all cases). The POSE-CLUSTERING and NP-SALIENCE systems are statistically equivalent; both are significantly better than the RANDOM-KEYFRAME baseline ($p < .05$).

Error analysis A manual inspection of the system output revealed that in many cases, our system selected a noun phrase that was accompanied by a relevant gesture, but the

[5] Coreference clusters are based on manual annotations.

specific keyframe was slightly off. Our method always chooses the keyframe at the midpoint of the accompanying noun phrase; often, the relevant gesture is brief, and does not overlap with the middle of the noun phrase. Thus, one promising approach to improving results would be to "look inside" each noun phrase, using local gesture features to attempt to identify the specific frame in which the gesture is most salient.

Some crucial gestures are not related to noun phrases. For example, suppose speaker says "it shoots the ball up," and accompanies only the word "up" with a gesture indicating the ball's trajectory. This gesture might be important to understanding the speaker's meaning, but since it does not overlap with a noun phrase, the gesture will not be identified by our system. We believe that our results show that focusing on noun phrases is a good start for linguistically-motivated keyframe extraction, but in the future, our system could be supplemented by identifying gestures that accompany other types of phrases.

Related Work

One typical video summarization approach is to cluster frames by an image similarity metric, and then return representative frames from each cluster. As noted, Uchihashi *et al.* (1999) quantify image similarity in terms of color histograms. Other systems attempt to quantify the keyframe salience using more domain-specific features such as motion, face detection, and audio cues (Ma *et al.* 2002). Closed-caption transcript information may also be used to select frames coinciding with salient keywords (Smith & Kanade 2001).

Such techniques are usually applied to videos that contain multiple shots and camera movements. In contrast, we are interested in unedited video taken from a single, fixed camera – for example, a video recording of a class lecture. In such videos, the changes between frames may be very hard to detect using only image features. The relevant visual information consists mainly of the speaker's body language; to select keyframes, the system must identify periods in which the speaker is gesturing meaningfully. We argue that visual features alone are not sufficient in this setting, and that a multimodal linguistic analysis is required.

Some existing video summarization systems try to identify meaningful gestures directly. Ju *et al.* (1997) describe work in a specialized domain, in which the videos include only static, printed slides, and occasional gestures of the presenter's hand. The system selects keyframes that include pointing hand shapes. Wilson *et al.* (1997) work in a more general domain of face-to-face conversations, and attempt to identify gestures that have a "tri-phasic" temporal structure. These approaches differ from ours in at least two important ways. First, they specify *a priori* the type of gestures to be recognized; for Ju *et al.* it is pointing handshapes, while for Wilson *et al.* it is gestures with a tri-phasic temporal structure. In contrast, our system attempts to discover the properties of salient gestures automatically. Second, our system uses both visual and linguistic features to assess the semantic importance of gesture, while these prior efforts incorporate only visual features.

Conclusion

We described a novel approach to video summarization, and present a system that generates keyframe summaries augmented with captions containing the accompanying speech. Our system learns to identify salient gestures by leveraging coreference resolution and treating gesture salience as a hidden variable. Experimental evaluation shows that this is an effective approach to selecting relevant keyframes.

Keyframes are appropriate for capturing the meaning of static gestures; we plan to extend this work to other presentation techniques that are better suited to capture dynamic gestures. This could be done by annotating keyframes with graphics indicating the trajectory of salient dynamic gestures, or with a more interactive system, in which the user clicks a keyframe to see a short "skim" of the dynamic gesture.

References

Darrell, T., and Pentland, A. 1993. Space-time gestures. In *CVPR*, 335–340.

Daumé III, H., and Marcu, D. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *HLT/EMNLP*, 97–104.

Deutscher, J.; Blake, A.; and Reid, I. 2000. Articulated body motion capture by annealed particle filtering. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 126–133.

Eisenstein, J., and Davis, R. 2006. Gesture improves coreference resolution. In *HLT/NAACL: Companion Volume*, 37–40.

Ju, S. X.; Black, M. J.; Minneman, S.; and Kimber, D. 1997. Analysis of gesture and action in technical talks for video indexing. In *CVPR*, 595–601.

Kehler, A. 2000. Cognitive status and form of reference in multimodal human-computer interaction. In *AAAI*, 685–690.

Liu, D. C., and Nocedal, J. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45:503–528.

Luo, X. 2005. On coreference resolution performance metrics. In *HLT/EMNLP*, 25–32.

Ma, Y.-F.; Lu, L.; Zhang, H.-J.; and Li, M. 2002. A user attention model for video summarization. In *ACM MULTIMEDIA*, 533–542.

Mani, I., and Maybury, M. T., eds. 1999. *Advances in Automatic Text Summarization*. Cambridge, MA, USA: MIT Press.

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

Melinger, A., and Levelt, W. J. M. 2004. Gesture and communicative intention of the speaker. *Gesture* 4(2):119–141.

Quattoni, A.; Collins, M.; and Darrell, T. 2004. Conditional random fields for object recognition. In *NIPS*, 1097–1104.

Smith, M. A., and Kanade, T. 2001. Video skimming and characterization through the combination of image and language understanding techniques. In *Readings in multimedia computing and networking*. Morgan Kaufmann Publishers Inc. 370–382.

Uchihashi, S.; Foote, J.; Girgensohn, A.; and Boreczky, J. 1999. Video manga: generating semantically meaningful video summaries. In *ACM MULTIMEDIA*, 383–392.

Wilson, A.; Cassell, J.; and Bobick, A. 1997. Temporal classification of natural gesture and application to video coding. In *CVPR*, 948–954.