Published in the *Proceedings of the SIAM Text Mining 2001 Workshop,* Chicago, IL, pp. 6–13, April 7, 2001

Hierarchical Segmentation: Finding Changes in a Text Signal

Malcolm Slaney and Dulce Ponceleon
IBM Almaden Research Laboratory
650 Harry Road, San Jose, CA 95120
malcolm@almaden.ibm.com   dulce@almaden.ibm.com

ABSTRACT[1]

This paper describes a signal processing algorithm which discovers the hierarchical organization of a document or media presentation. We use latent semantic indexing to describe the semantic content of the signal, and scale-space segmentation to describe its features at many different scales. We represent the semantic content of the document as a signal that varies through the document. We low-pass filter this signal to compute the document's semantic path at many different time scales and then look for changes. The changes are sorted by their strength to form a hierarchical segmentation. We present results from a text document and a video transcript.

1. THE PROBLEM

As prices decline and storage and computational horsepower increase, we will soon be swamped in multimedia data. Unfortunately, given an audio or a video signal there is little information readily available that can help us find our way around such a time-based signal. Technical papers are structured into major and minor headings, imposing a hierarchical structure. Often professional or high-quality audio–visual (AV) presentations are also structured. However, this information is hidden in the signal. Our goal is to use the semantic information in the AV signal to create a hierarchical table of contents that describes the associated signal. Towards this end we combine two powerful concepts: scale space (SS) filtering and Latent Semantic Indexing (LSI).

We use LSI to provide a continuously valued feature that describes the semantic content of an AV signal. By doing this we reduce the dimensionality of the problem and, more importantly, we address synonymy and polysemy as LSI does. The combined approach remains language independent.

We use scale-space techniques to represent the semantic signal over many different time scales. We are looking for changes in the signal and scale space allows us to talk about features of the document that span from a single sentence to entire chapters. The scale parameter specifies the level of detail for our analysis. Intuitively, at small scales we are looking at the individual trees, and at large scales we are seeing the entire forest. We look at a wide range of scales to determine when the content of the signal has changed. In Section 5 we use Latent Semantic Indexing (LSI) as a means to describe the semantic content of a signal.

This paper is organized as follows. In Section 2 we describe how our approach builds on previous work in this field. In Section 3 we describe our test data so that we can use it as an example in the description of our algorithm. In Section 4 we introduce scale space and describe an algorithm that looks at a signal at many different scales. We describe an algorithm that combines scale-space analysis and LSI in Section 6. Finally, in Section 7 we present the results obtained on two segmentation tests.

2. PREVIOUS SOLUTIONS

Our work extends previous work on text analysis and segmentation in several different ways.

LSI has a long history, starting with Deerwester's paper [6], as a powerful means to summarize the semantic content of a document and measuring the similarity of two documents or a query and a document. We use LSI to capture the synonymy and polysemy, but, more importantly, LSI allows us to describe the position of a portion of the document in a multi-dimensional semantic space.

This paper differs from previous information retrieval work, for example Kurimo's work [11] on clustering, in two ways. First, we are looking for differences *within* a document using LSI. Second, we are using scale-space as a

[1] This paper is an expanded version of a paper to be published at the 2001 International Conference on Acoustics, Speech and Signal Processing [16].

principled way to smooth in time the semantic content of the document.

Hearst [9] proposes to use the dips in a similarity measure of adjacent sentences in a document to identify topic changes. Her method is powerful because the size of the dip is a good indication of the relative amount of change in the document. We extend this idea by using scale-space techniques to allow us to talk about similarity or dissimilarity over larger portions of the document.

Choi [5], for text, and Foote [8], for audio, represent a document in terms of its self-similarity matrix. Their task is then to search for and identify the square regions of this matrix that are self-similar. Using scale-space methods, we automatically find the edges of these regions and characterize their strength.

This work assumes a different model than change point analysis [4]. Change point analysis scans the text and builds a model of what it has seen. It then identifies when this model of the text no longer fits the data and consequently it is necessary to change the model.

Beeferman and his colleagues [3] apply change point analysis to text segmentation, using an exponential model to capture the current semantic state of the document and find features that indicate topic changes. They explicitly focus on finding large topic boundaries and not the subtle changes within a story.

Change point analysis is a forward-looking algorithm. Our work, on the other hand, looks for points in a smoothed version of the data where the difference between neighboring topics reaches a local maximum. Since our filters are symmetric, our decisions are based on the semantic content both before and after the proposed segmentation boundary. We do not know how this difference from change point analysis affects our performance.

Segmentation is a popular topic in the signal and image processing worlds. Witkin [18] introduced scale-space ideas to the segmentation problem and Lyon [14] extended Witkin's approach to multi-dimensional signals. A more theoretical discussion of the scale-space segmentation ideas was published by Leung [12]. This work extends the signal processing approach by using LSI as a basic feature and changing the distance metric to fit semantic data.

Finally, the signal processing analysis proposed in this paper is just one part of a complete system. We use LSI to do the basic semantic analysis, but more sophisticated techniques are also applicable. The key concept in this paper is to think about the document's path through semantic space, and detect the topic jumps at multiple scales. Any method which allows us to summarize the semantic content of the document can be used with the techniques described here.

3. TEST DATA

Typically, LSI indexes a collection of documents. In this work the target is a single document, so we use each sentence as one sub-document in our tests. We used a list of 398 stop words and removed any words that included digits. We did not do any stemming. We expect that stemming will be important in this application since our databases, each a single document, are relatively small.

We used two different texts in our study: a long chapter from a book on tomography and a comparatively shorter transcript from CNN Headline News.

The long test was text decoded via optical character recognition (OCR) of Chapter 4 from a scanned book on tomography [10]. This text has errors due to the OCR. Each page of the book was scanned in raster order, so figure captions and equations are included inline with the text. This makes the segmentation job harder since the text and the corresponding figure captions are sometimes separated by pages. We did not include the reference section in our analysis since it is organized alphabetically and not topically structured. We found 1093 sentences in this chapter and after removing stop words there were 1830 distinct words.

The shorter test was the manual transcript of a 30 minute CNN Headline News television show [13]. We removed the timing and other meta information before analysis. This transcript is cleaner than those typically obtained from closed-captioned data or automatic speech recognition. We found 257 sentences in this broadcast and after removing stop words there were 1032 distinct words.

In these two cases we have relatively clean transcripts and the ends of sentences are marked with periods. We can also use automatic speech recognition to provide a transcript of the audio, but then sentence boundaries are not available. However, we could divide the text arbitrarily into 20 word "sentences." We believe that a statistical technique such as LSI will fail gracefully in the event of word errors. In addition, LSI can take into account multiple word hypothesis as produced by speech recognition engine.)

4. SCALE SPACE SEGMENTATION

Witkin [18] introduced the idea of scale-space segmentation to find the boundaries in a signal. In scale space, we analyze a signal with many different kernels that vary in the size of the temporal neighborhood that is included in the analysis at each point in time. If the original signal is $s(t)$, then the scale-space representation of this signal is given by

$$s_\sigma(t) = \int s(\tau) g(\sigma, t - \tau) d\tau \qquad (1)$$

where $g(\sigma, t-\tau)$ is a Gaussian kernel with a variance of σ^2. With a σ approaching zero, $s_\sigma(t)$ is nearly equal to $s(t)$. For larger values of σ, the resulting signal, s_σ, is smoother because the kernel is a lowpass filter, removing the high-frequency details in the signal. We have transformed a one-dimensional signal into a two-dimensional image, where the analysis scale is a continuous and explicit parameter of the analysis.

An important feature of scale space is that the resulting analysis is a continuous function of the scale parameter. Because a local maximum in scale space is well behaved [2], we can start with a peak in the signal at the very largest scale and trace it back to the exact point at zero scale where it originates. The range of scales over which the peak exists is a measure of how important this peak is to the signal.

In scale-space segmentation we are looking for changes in the signal. We do this by calculating the derivative of the signal with respect to time and look for the local maximum of this derivative. Because the derivative and the scale-space filter are linear we can exchange their order. Thus the properties of the local maximum described above also apply to the signal's derivative.

Lyon [14] extended the idea of scale-space segmentation to multi-dimensional signals and used it to segment a speech signal. The basic idea remains the same: we filter the signal by a Gaussian kernel with a range of scales. By performing the smoothing independently on each dimension, the new signal traces out a smoother path through his 92-dimensional space. To segment the signal, we now look for the local peaks in the magnitude of the vector derivative.

Cepstral analysis transforms each vocal sound into a point in a high-dimensional space. This makes it easy to recognize each sound (good for automatic speech recognition, ASR) and to perform low-level segmentation of the sound (as demonstrated by Lyon). Unfortunately, there is little information in the cepstral coefficients about high-level structures. We improve this situation by considering the semantic content of the signal.

5. LATENT SEMANTIC INDEXING (LSI)

LSI is an important statistical tool for describing the semantic content of a collection of text. As originally defined [6], we collect a histogram of all the words in a document, where $\vec{H}(d)$ describes the *d*-th document and is the *d*-th column of a matrix. We used local (log of term frequency + 1) and global term weighting (entropy of the term frequency) as suggested by Dumais [7]. After collecting the histograms for a number of documents, a singular-value decomposition (SVD) is used to summarize the words in the collection of documents by projecting on to the first k left-singular vectors and scaling by the inverse of the associated k singular values. This gives us $\vec{H}_k(d)$, a k-dimensional representation of the semantic content of the documents. In the experiments described here we arbitrarily reduced the semantic space to 10 dimensions.

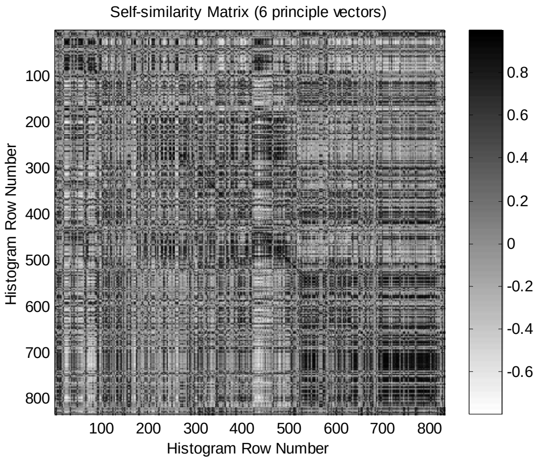

Figure 1: This plot shows the semantic similarity between the 832 sentences in Chapter 4 of a tomography book [10]. The large block in the lower right-hand corner (starting near sentence 500) corresponds to a change in topic from different types of tomography (first 3 subsections of the chapter) to magnetic resonance imaging.

The angle between two documents in the LSI space is a measure of the similarity of the two documents. This angle is measured by computing the dot product of the two (normalized) vector: this gives the cosine of the angle between the two points. This idea is the basis of a simple but effective document retrieval system.

We extend LSI analysis to describe the semantic content within a document. We do this by breaking the document into pieces and thinking about each piece as a separate sub-document. The angle between two sub-documents is the "distance" in semantic space.

Figure 1 shows a self-similarity matrix [8] for the tomography chapter. In our work, the self-similarity matrix *S(i,j)* shows the semantic distance between the *i*'th and *j*'th sentences. For illustrative purposes, the texture differences were most striking for this document using just the first six singular dimensions. As expected along the diagonal, each sentence is identical to itself, but more importantly, the matrix exhibits a block-diagonal structure. These blocks vary in size and indicate a group of sentences that are on the same topic and are related to each other.

Scale space enables us to use this self-similarity measure, group sub-documents and talk about their boundaries at different scales.

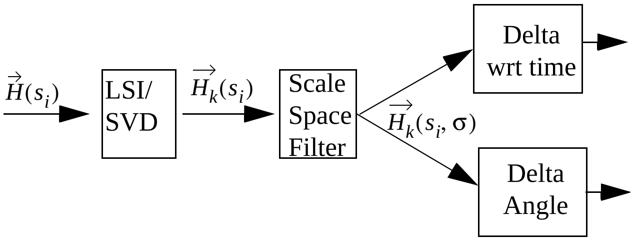

Figure 2: The LSI-SS algorithm. The top path shows the derivative based on euclidean distance. The bottom path shows the proper distance metric for LSI based on angle. See Section 4 for definitions.

6. COMBINING LSI AND SS

Combining LSI analysis with scale-space segmentation is straightforward. This process is illustrated in Figure 2.

We use LSI to convert the histograms of the sub-documents, $\vec{H}(s_i)$ a vector function of sentence number s_i, into a k-dimensional representation of the document's semantic path, $\overrightarrow{H_k}(s_i)$. A lowpass filter is used on each dimension of the reduced histogram data $\overrightarrow{H_k}(s_i)$, replacing s in equation (1) with each component of $\overrightarrow{H_k}(s_i) = [H_1(s_i)H_2(s_i)\ldots H_k(s_i)]^T$ to find a lowpass filtered version of the semantic path. This gives $\overrightarrow{H_k}(s_i, \sigma)$, a k-dimensional vector function of sentence number and scale. We then compute the vector magnitude of the temporal derivative and identify the local peaks to find the segmentation as a function of time and scale.

An important property of the scale-space segmentation is that the length of the boundary in scale-space is a metric for the importance of the boundary. It is useful to think about a point representing the document's local content wandering through the LSI space in a pseudo-random walk. Each sentence is a slightly different point in space and we are looking for large jumps in the topic space. As we (lowpass) filter the LSI representation, the point moves more sluggishly. It eventually moves to a new topic, but small variations in the topic do not move the point very much. Thus the boundaries that are left at the largest scales are the biggest changes within the document.

There are two different phases in this analysis. In the first phase, a model of the current text is built using LSI and its SVD. Then in the second phase the histogram data for the same document is projected into the LSI subspace and scale-space filtering is done on this data. Now we can identify the local peaks in the magnitude of the vector derivative.

The distance metric in the original scale-space work [14] was based on Euclidean distance. When using LSI as input to a scale-space analysis our distance metric is based on angle. The dot product of adjacent (filtered and normalized) semantic points gives us the cosine of the angle between the two points. We convert this into a distance metric by subtracting the cosine from 1.

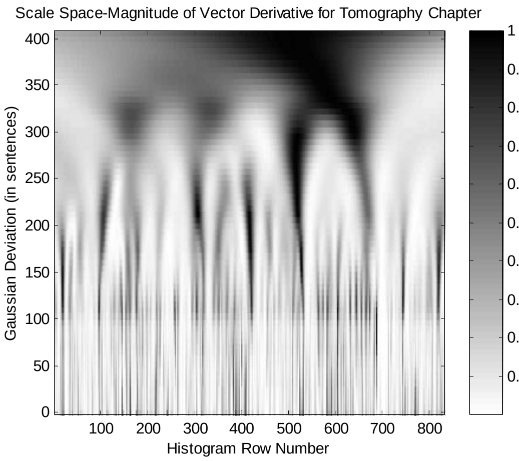

Figure 3: Change in semantic content of the tomography chapter in scale space. This image shows the cosine of the angular change of the semantic trajectory with different amounts of lowpass filtering.

Figure 3 shows the scale-space representation of the LSI data for the tomography chapter. This plot shows the cosine of the angle of the vector derivative as a function of sentence number (horizontal axis) and scale (vertical axis). At the bottom, where the scale is small, there are many small changes in topic. These topic changes are gradually filtered out as we move to the larger scales. The largest peak, which starts around sentence 500 in the coarsest scale, leads us back to the point in the chapter where the text moves from talking about different forms of tomography to how tomography and magnetic resonance imaging (MRI) are related. (The sentence numbers in Figure 3 are not the same as those in Figures 4, 5 and 6 because some sentences have no content words after dropping stop words, and are deleted from the SVD analysis. The sentence counts are adjusted to the true numbers after we find the peaks.) The scale-space filtered semantic path forms the basis of our hierarchical segmentation algorithm.

The big question when using LSI within a document is how to choose the appropriate block size. Placing the entire document into a single histogram gives us little information that we can use to segment the document. On the other hand, splitting the document into one-word chunks is too fine; each sub-document is a single word and we have no way to link one word to another. The power of LSI is available when we use a small chunk of text, where words that occur in close proximity are linked together by the histogram data.

Choosing the proper segment size is easy during the segmentation phase since projecting onto a subspace is a linear operator. Thus even if we start with single-word histograms, the projection of the (weighted) sum of the histograms is the same as the (weighted) sum of the projections of the histograms. The story is not so simple with the SVD calculation. In this work, we chose a single sentence as the basic unit of analysis since a sentence contains one thought. But it is possible that larger sub-documents, or documents keyed by other parameters of a video, such as color information, might be more meaningful.[1]

There are many ways to choose the segmentation to use when analyzing the input text. When OCR'ed text is available we use single sentences (or a small number of sentences) as the input documents. In a video transcript we can use a fixed number of words, look for pauses, or look for scene breaks as determined by the color histogram data.

The computational requirements for this algorithm is reasonable. There are many ways to calculate the SVD in an LSI algorithm, while the scale-space calculation requires O(Nlog$N$) operations. More importantly, the number of sentences, N, in a document is small compared to the number of documents in a large collection used for information retrieval.

7. RESULTS

This section illustrates our algorithm by showing intermediate results and compares the results of hierarchical segmentations and the ground truth (manual segmentation of segments and hierarchy.) The ground truth for the tomography chapter was the locations of the headings and the sub-headings in the printed chapter. The LDC [13] provided story boundaries for the news video but the high-level structure was estimated based on our familiarity with this news program.

Most media are not organized in a perfect hierarchy. In text, the introduction presents a number of ideas, which are then explored in more detail, and then a graceful segue is used to transition between ideas. This is much more apparent in a news show, which has some hierarchy, but is designed to be watched in a linear fashion. Thus the viewer is teased with information about an upcoming weather segment, and the “top of the news” is repeated at various stages through the broadcast.

The peaks in the LSI-SS analysis are tracked back to their origin to determine the original point of change in the document. This result is shown in Figure 4 for the tomography chapter. The length of the line represents the range of scales where this peak exists and is a measure of how significant this topic change is to the document.

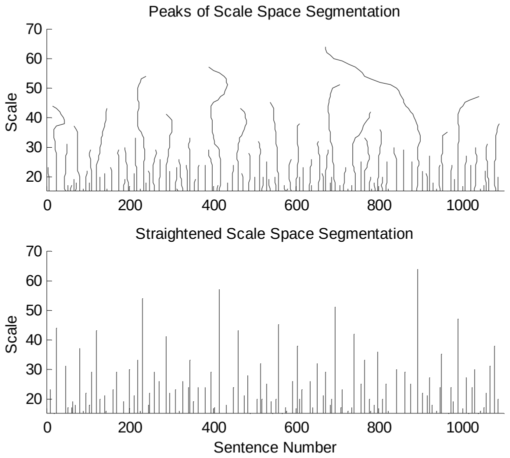

Figure 4: The top plot shows the peaks of the scale-space derivative for the tomography chapter. The bottom plot shows the peaks traced back to their original starting point.

We have used the chapter headings and sub-headings, and their titles, as a form of ground truth. The classic measures for the evaluation of text-retrieval performance [1] do not easily extend to a system with hierarchical structure. Instead we demonstrate our results with a plot that compares heading titles and the scale-space segmentation strength. The scale-space analysis produces a large number of possible segmentations, in this work we are only plotting twice the number of boundaries indicated by the ground truth.

Figure 5 shows a comparison of the ground truth and the scale-space segmentation results for the tomography chapter. On the right, the major (left-most text) and the minor (right-most text) are shown. The left side of the plot shows the strength of the boundary. As expected, the start of the MRI section at sentence 891 is the most important change. The other section headings are marked by segment boundaries with significant strength.

Our results with the CNN Headline news are shown in Figure 6. While the “Weather”, “Tech Trends” and “Lifestyles” sections are indicated within a few sentences, there are large peaks at other locations in the transcript. Interestingly, there is a large boundary around sentence 46, which neatly divides the softer news stories at the start of this broadcast from the political stories that follow.

8. CONCLUSIONS

This paper presents a signal-processing algorithm to hierarchically segment a text or AV signal. By using LSI,

[1] Preliminary results reported elsewhere [17] indicate that combining 4 to 8 sentences increases the correlation between one chunk of text and the next. This result suggests that 4 to 8 sentences, a paragraph, might be the smallest meaningful semantic chunk.

we have a statistical model of the semantics that spans the entire document. As we view the document, the sentences trace a trajectory in the semantic space. We use scale-space filtering to analyze each document's path through space and then look for the points of greatest change, at all different scales, to determine the document's segment boundaries. We demonstrated our algorithm's performance on a text document and the transcript from a television news show.

There are many ways to combine scale-space ideas with different representations. Color histograms are a common metric in video segmentation. A color metric can be combined with scale-space filtering, with or without the SVD dimensionality reduction. Similarly, musical features [8] or emotional measures [15] can be used to segment audio signals. Finally, the most interesting possibilities are a combination of features. Thus eventually we would like to combine color histogram data, giving evidence of the finest segmentation points, and the semantic content to provide the high-level information. We have not integrated such disparate metrics.

Most importantly, we need better ways to quantify our results. The large databases used in segmentation studies are labeled with story boundaries, but we don't know how to quantify the difference between two hierarchical segmentations.

Not all AV presentations have a clear hierarchical organization. In one video we examined from the television documentary "The Making of a 21st Century Jet" the story proceeded from topic to topic, but the segment boundaries were blurred to give the story more continuity. An evaluation metric needs to account for these fuzzy boundaries. We do not know whether a strict or blurred hierarchy is more common in audio-visual documents.

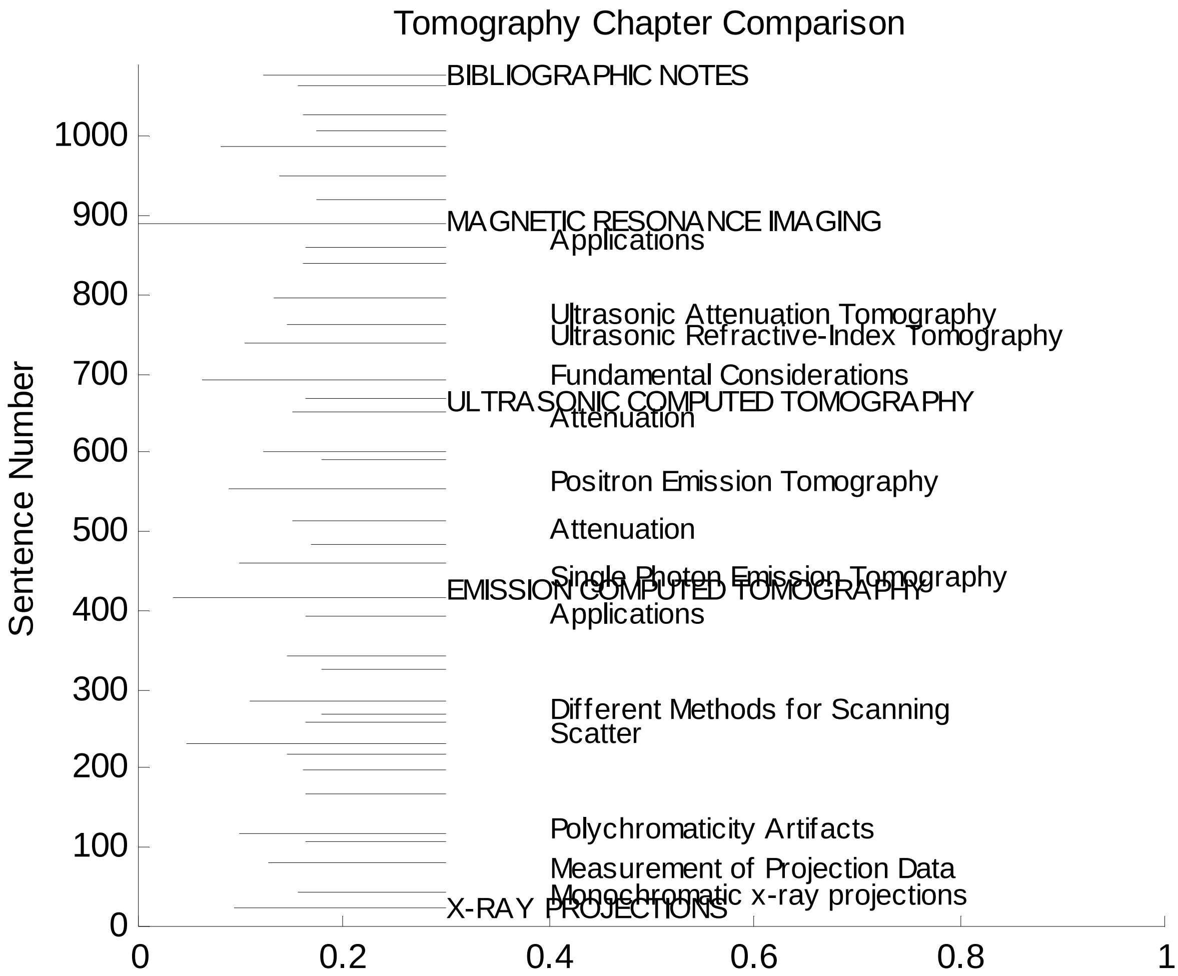

Figure 5: A comparison of ground truth (right) and the size of boundaries for the tomography chapter as determined by scale-space segmentation. The major headings are in all capitals, and the sub-headings are in upper and lower case.

ACKNOWLEDGEMENTS

We appreciate the ideas and feedback we have received from Byron Dom, Shivakumar Vaithyanathan, Savitha Srinivasan, David Beymer, Myron Flickner, Mike Wirth and Michele Covell.

REFERENCES

[1] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. “Topic Detection and Tracking Pilot Study Final Report.” *Proceedings of the Broadcast News Transcription and Understranding Workshop* (Sponsored by DARPA), Feb. 1998.

[2] Jean Babaud, Andrew P. Witkin, Michel Baudin, Richard O. Duda. “Uniqueness of the Gaussian Kernel for Scale-Space Filtering.” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-8, No. 1, pp. 26–33, January 1986.

[3] Doug Beeferman, Adam Berger, and John Lafferty. “Statistical models for text segmentation.” *Machine learning,* Special issue on Natural Language Processing, 34(1-3):177-210. C. Cardie and R. Mooney (editors), 1999.

[4] Jie Chen, Arjun K. Gupta. *Parametric Statistical Change Point Analysis.* Birkhauser, Boston, 2000.

[5] F. Choi. “Advances in domain independent linear text segmentation.” *Proceedings of NAACL’00,* Seattle, USA, April 2000.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman. “Indexing by latent semantic analysis.” *Journal of the American Society for Information Science,* 41, pp. 391–407, 1990.

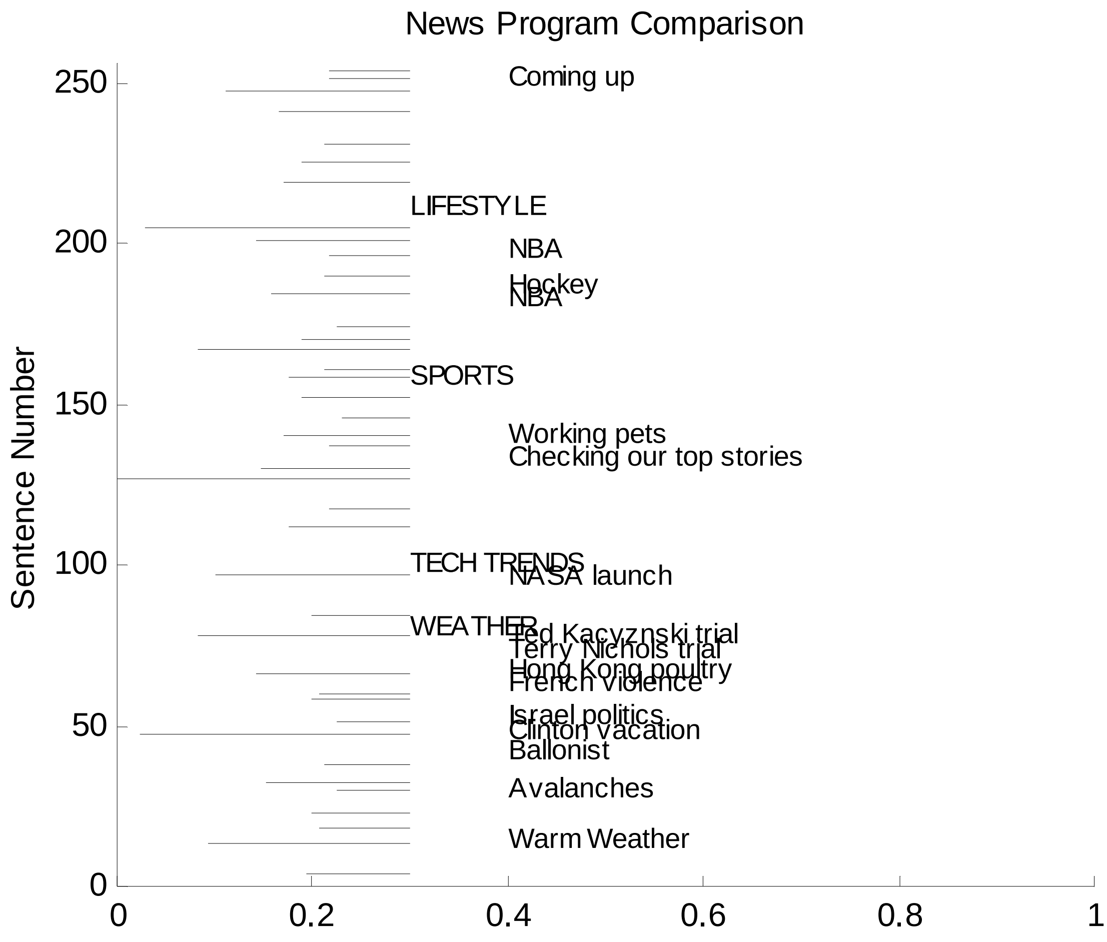

Figure 6: A comparison of ground truth (right) and the size of boundaries for the news show as determined by scale-space segmentation. The major headings are in all capitals, and the sub-headings are in upper and lower case.

[7] S. T. Dumais. “Improving the retrieval of information from external sources” *Behavior Research Methods, Instruments, & Computers,* 23, pp. 229–236, 1991.

[8] Jonathan Foote. “Visualizing Music and Audio using Self-Similarity.” *Proceedings of ACM Multimedia’99,* pp. 77–80, Orlando, Florida, November 1999.

[9] M. A. Hearst. Multi-paragraph segmentation of expository text. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* 1994.

[10] A. C. Kak and Malcolm Slaney. *Principles of Computerized Tomographic Imaging.* IEEE Press, 1988 (also available at http://www.slaney.org/pct).

[11] Mikko Kurimo. “Fast latent semantic indexing of spoken documents by using self-organizing maps.” *Proc. of ICASSP,* Istanbul, Turkey, pp. 3781–3784, June 2000.

[12] Yee Leung, Jiang She Zhang, Zong Ben Xu. “Clustering by Scale-Space Filtering.” *IEEE Transactions on PAMI*, Vol.22(12), pp. 1396–1410, Dec. 2000.

[13] Linguistic Data Consortium. “1997 English Broadcast News Speech (Hub-4).” LDC catalog no.: LDC98S71, File ed980104.

[14] Richard F. Lyon. “Speech Recognition in Scale Space,” *Proc. of 1984 ICASSP.* San Diego, March, pp. 29.3.1–4, 1984.

[15] Malcolm Slaney, Gerald McRoberts. “BabyEars: A recognition system for affective vocalizations.” *Proc. ICASSP,* Seattle, WA, pp. 985–988, May 12–15, 1998.

[16] Malcolm Slaney and Dulce Ponceleon. “Hierarchical Segmentation using Latent Semantic Indexing in Scale Space.” To be published in the *Proceedings of the 2001 ICASSP,* Salt Lake City, Utah, May, 2001.

[17] Malcolm Slaney, Dulce Ponceleon, James Kaufman. “Multimedia Edges: Finding Hierarchy in all Dimensions.” Submitted to *ACM Multimedia 2001.*

[18] Andrew P. Witkin. “Scale-Space Filtering: A New Approach to Multi-Scale Description.” *Proc. of ICASSP,* San Diego, CA March, pp. 39A.1.1–39A.1.4, 1984.