Hierarchical Segmentation using Latent Semantic Indexing in Scale Space

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

This paper describes a new 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 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 AV presentations are also structured. However, this information is hidden in the signal. Our goal is to use the intrinsic 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.

This paper differs from previous information retrieval work, for example Kurimo's work [1], in two ways. First, we are looking for differences within a document using LSI. Second, we are using scale-space as a principled way to smooth across time the semantic content of the document.

This work assumes a slightly different model than change point analysis [2]. Change point analysis aims to identify when does one model of the data no longer fits the data and consequently it is necessary to change the model. This work, on the other hand, looks for points in a smoothed version of the data where the difference between neighboring topics reaches a maximum. We do not know how this difference affects our task.

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. In each case we have a relatively clean transcript and the ends of sentences are marked with periods. 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. While we did remove words found on a list of 398 stop words and any words that included digits, we did not do any stemming

The long test was OCR'ed text of Chapter 4 from a scanned book on tomography [3]. 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 captions are sometimes separated by pages. We did not include the reference section in our analysis since it is organized alphabetically and not by topic. 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 [4]. We removed the timing and other meta information before analysis. This transcript is cleaner than those typically obtained from closed-captioned data or a speech recognition engine. We believe that a statistical technique such as LSI will fail gracefully in the event of word errors. (LSI also can easily take into account multiple word hypothesis as produced by speech recognition engine.) We found 257 sentences in this broadcast and after removing stop words there were 1032 distinct words.

This paper is organized as follows. In Section 2 we introduce scale space and describe an algorithm that looks at a signal at many different scales. 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 3 we use Latent Semantic Indexing (LSI) as a means to describe the semantic content of a signal. We describe the algorithm that combines scale-space analysis and LSI in Section 4. Finally, in Section 5 we present the results obtained on two segmentation tests.

2. SCALE SPACE SEGMENTATION

Witkin [5] introduced the idea of scale-space segmentation to find the boundaries in a signal. We analyze a signal in scale space with many different kernels, varying 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 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 low-pass filter. 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 [6], 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, described above, of the local maximum also apply to the signal's derivative.

Lyon [7] 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 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 address this problem by considering the semantic content of the signal.

3. LATENT SEMANTIC INDEXING (LSI)

LSI is an important statistical tool for describing the semantic content of a collection of text. As originally defined [8], 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 [9]. 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 reduced the semantic space to 10 dimensions.

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) vectors: this gives the cosine of the angle between the two points. This idea is the basis of a simple but effective document retrieval system.

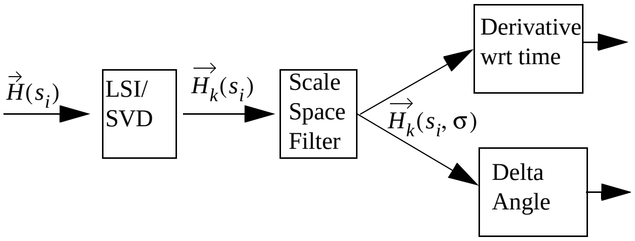


Figure 1: The LSI-SS algorithm. The top path shows the derivative based on euclidean distance. The bottom path shows the proper distance metric for LSI based on angle. See Section 4 for definitions.

This paper extends 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. Scale space enables us to group sub-documents and talk about their boundaries at different scales.

4. COMBINING LSI AND SS

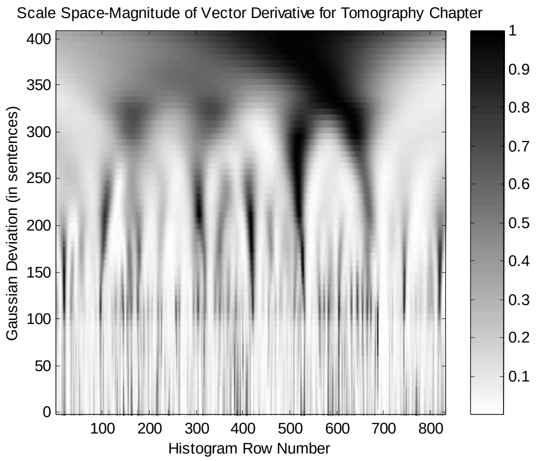
Combining LSI analysis with scale-space segmentation is straightforward. This process is illustrated in Figure 1.

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, $\vec{H}_k(s_i)$. A lowpass filter is used on each dimension of the reduced histogram data $\vec{H}_k(s_i)$, replacing s in equation (1) with each component of $\vec{H}_k(s_i) = [H_1(s_i)H_2(s_i)\ldots H_k(s_i)]^T$ to find a lowpass-filtered version of the semantic path. This gives $\vec{H}_k(s_i, \sigma)$, a k-dimensional vector function of sentence number and scale.

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 [7] 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 2 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 2 are not equivalent to



igure 2: Change in semantic content of the tomography chapter n scale space. This image shows the cosine of the angular hange of the semantic trajectory with different amounts of low-ass filtering.

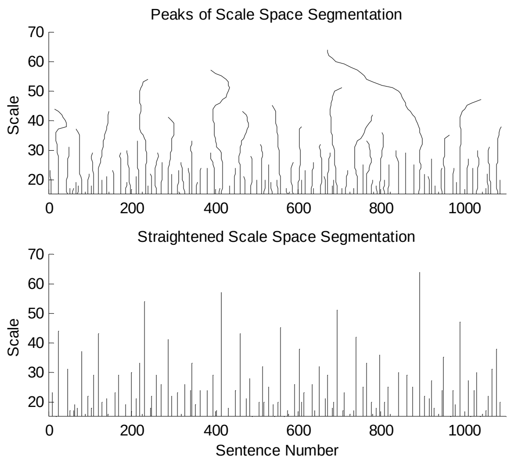


Figure 3: The top plot shows the peaks of the scale-space derivative for the tomography chapter. The bottom plot shows the peaks traced back to their original starting point. These peaks represent topic boundaries

those in Figures 3, 4 and 5 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 might give better results.

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.

5. 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 published chapter. The LDC [4] provided story boundaries for the news video but the high-level structure was estimated based on our familiarity with this news program.

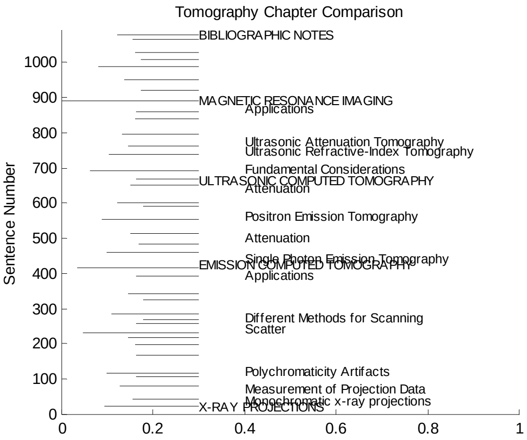
Most media are not organized in a perfect hierarchy. In text, the introduction presents a number of ideas, which are then explored in more detail, and then a graceful segue is used to transition between ideas. This is much more apparent in a news show, which has some hierarchy, but is designed to be 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 3 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.

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 [11] 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 4 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 MRI section change at sentence 891 is the most important change. The other section headings are found by segment boundaries with significant strength.

Our results with the CNN Headline news are shown in Figure 5. While the "Weather", "Tech Trends" and "Lifestyles" sections



igure 4: A comparison of ground truth (right) and the size of oundaries for the tomography chapter as determined by scale-pace segmentation. The major headings are in all capitals, and he sub-headings are in upper and lower case.

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, with the political stories that follow.

6. CONCLUSIONS

This paper presents a signal-processing algorithm that hierarchically segments a text or AV signal. We use LSI to form a statistical model of the entire document's semantic content. As we scan the document, the sentences trace out a curve in semantic space. We use scale-space filtering to analyze this document's path through semantic 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.

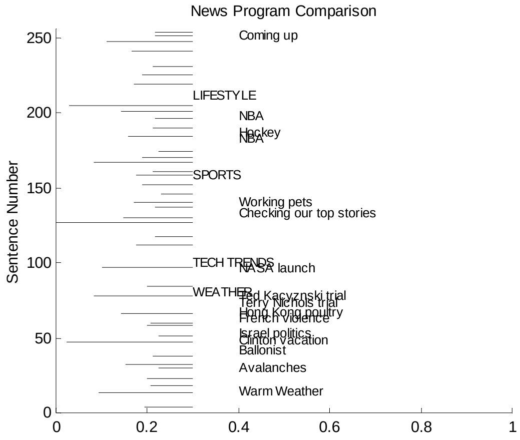


Figure 5: A comparison of ground truth (right) and the size of boundaries for the news show as determined by scale-space segmentation. The major headings are in all capitals, and the sub-headings are in upper and lower case.

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 [10] or emotional measures [12] can be used to segment musical pieces. Finally, the most interesting possibilities are a combination of features. Thus eventually we would like to imagine combine color histogram data, giving evidence of the finest segmentation steps, and with the semantic content to provide the high-level information. We have not integrated such disparate metrics.

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

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

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

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

[5] 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.

[6] 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.

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

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

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

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

[11] 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.

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