Presented at the ICCV Event 2001 Workshop

Temporal Events in All Dimensions and Scales

Malcolm Slaney
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
malcolm@almaden.ibm.com

Dulce Ponceleon
IBM Almaden Research Center
650 Harry Road
San Jose CA, 95120
dulce@almaden.ibm.com

James Kaufman
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
kaufman@almaden.ibm.com

Abstract

This paper describes a new representation for the audio and visual information in a video signal. We reduce the dimensionality of the signals with singular-value decompositions (SVD) or mel-frequency cepstral coefficients (MFCC). We apply these transforms to word, (word transcript, semantic space or latent semantic indexing), image (color histogram data) and audio (timbre) data. Using scale-space techniques we find large jumps in a video's path, which are evidence for events. We use these techniques to analyze the temporal properties of the audio and image data in a video. This analysis creates a hierarchical segmentation of the video, or a table-of-contents, from both audio and the image data.

1. TABLE OF CONTENTS FOR MEDIA[1]

Browsing videotapes of image and sound (hereafter referred to as "videos") is difficult. Often, there is an hour or more of material, and there is no roadmap to help viewers find their way through the medium. It would be tremendously helpful to have an automated way to create a hierarchical table of contents that listed major topic changes at the highest level, with subsegments down to individual shots. Realization of such an automated analysis requires the development of algorithms which can detect changes in the video and/or semantic content of a video as a function of *time*. We propose a technology that performs this indexing task by representing the all the major sources of data — images, non-speech audio and words — from the video into a common representations.

This paper describes an approach to event detection based on dimensionality reduction and scale-space segmentation. We are looking for events marked by large changes in the audio or image data at many different time scales. We demonstrate this idea based on the semantic information in the video's transcript, color information in the sequence of images, and acoustic information in a musical accompaniment.

Our technique is analogous to one that detects edges in an image. Instead of trying to find similar regions of the multimedia signal, called segments, we think of the audio–visual content as a signal and look for "large" changes in this signal or peaks in its derivative. The location of these changes are events; they represent the nodes in a table of contents.

1.1 Temporal Properties of Multimedia

The techniques we describe in this paper allow us to characterize the temporal properties of both the audio and image data in the video. The color information in the image signal and the non-speech acoustic and semantic information in the audio signal provide different information about the content.

Color provides robust evidence for a shot change in a video signal. A straightforward way to convert the color data into a signal that indicates scene changes is to compute each frame's color histogram and to note the frame-by-frame differences [7][21]. In general, however, we do not expect the colors of an image to tell us anything about the global structure of the video. The color balance in a video does not typically change systematically over the length of the film.[2]

Random words from a transcript, on the other hand, do not reveal much about the low-level features of the video. Given just a few words from the audio signal, it is difficult to define the current topic. But the words indicate a lot about the overall chronological structure of the story. A documentary script may, for instance, progress through topic 1, then topic 2, and finally topic 3.

We describe any time point in the video by its position in an acoustic–color–semantic vector space. We represent the information in the signals as separate vectors as a func-

1. This paper is similar to a paper we have submitted to the ACM Multimedia conference. In this work, we have focused on events in multimedia signals and extended the ideas to apply to musical signals.

2. The 1997 movie Titanic is a notable exception. The first half of the movie is warm and bright; until the ship strikes the iceberg, when the movie turns dark. Although we were not looking for such changes, the techniques we describe are sensitive to such long-range color changes.

tion of time. Using scale-space techniques we can then talk about the changes that the combined acoustic–color–semantic vector undergoes as the video unwinds over time. We label as segment boundaries large jumps in the acoustic–color–semantic vector. "Large jumps" are defined by a scale-space algorithm that we describe in Section 3.

1.2 Literature Review

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

LSI has a long history, starting with Deerwester's paper [8], 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.

Hearst [13] 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 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 [11], 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 instead find the edges of these regions and characterize their strength.

Segmentation is a popular topic in the signal and image processing worlds. Witkin [22] introduced scale-space ideas to the segmentation problem and Lyon [17] extended Witkin's approach to multi-dimensional signals. A more theoretical discussion of the scale-space segmentation ideas was published by Leung [15]. This work extends the signal processing approach by using LSI as a basic feature and changing the distance metric to fit semantic data.

Current video shot detectors [21][7] look at local changes in the color histogram and luminance patterns to detect shot boundaries. We use the same color information but extend their techniques by analyzing the changes over many different time scales.

Using subspace representations of video signals is not new. Kobla and his colleagues advocate a low-dimensional representation of image data [14]. Their algorithm calculates an efficient approximation to the multi-dimensional scaling problem, which is related to the singular-value decomposition we use in this work. They use a minimum bounding box to find the shots that should be clustered. We extend their work by finding a hierarchy of clusters.

Neti and his colleagues [18] extended the subspace ideas and apply them to joint audio–video speech recognition. They use linear-discriminant analysis to reduce the dimensionality of the combined space, in such a way that minimizes the within class scatter, while maximizing the between class differences. This approach makes sense when we have labeled classes and it is not clear how to apply these ideas to unstructured video.

1.3 Overview of paper

This paper proposes a unified representation for the audio–visual information in a video. We use this representation to compare and contrast the temporal properties of the audio and images in a video. We form a hierarchical segmentation with this representation and compare the hierarchical segmentation to other forms of segmentation.

As we have explained, we combine two well-known techniques to find the edges or boundaries in a multimedia. We use a singular-value decomposition (SVD) to reduce the dimensionality of the data and to put them all into the same format. The SVD and its application to color and word data are described in Section 2.

Scale-space techniques give us a way to analyze temporal regions of the video that span a time range from a few seconds to tens of minutes. We describe properties of scale spaces and their application to segmentation in Section 3.

We discuss several simple segmentation results in Section 4. We perform a hierarchical segmentation of a multimedia signal, automatically creating a table of contents for the several types of signals.

We conclude in Section 5 with some observations about this representation.

1.4 Test Material

We studied two different videos and a musical signal to evaluate our algorithm.

The shorter video test was the manual transcript of a 30 minute CNN Headline News television show [16]. This transcript is cleaner than those typically obtained from closed-captioned data or automatic speech recognition.

We also looked at the words and images from a longer documentary video, "21st Century Jet," about the making of the Boeing 777 airplane [19]. We analyzed the color information from the first hour of this video, and the words from all six hours.

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 groups or "sentences." We believe that a sta-

tistical 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.

Finally, we evaluated our approach to music segmentation using a recording of a folk song [12].

2. DIMENSIONALITY REDUCTION

We use two types of transformations to reduce raw audio and video signals into meaningful spaces where we can find edges or events. We use the singular-value decomposition (SVD) for signals where we have little prior information about their structure (Sections 2.1–2.3). We use an auditory representation known as mel-frequency cepstral coefficients (MFCC) to model musical signals (Section 2.4).

The SVD provides a principled way to reduce the dimensionality of a signal in a manner which is optimum, in a least-squared sense. In this section, we describe how we apply the SVD to color and semantic information. The SVD transformation allows us to represent disparate video data. We will connect this wide range of data using the SVD and scale space in Section 3 of this paper.

2.1 SVD Principles

We express both audio and image data as vector-valued functions of time, $\vec{x}(t)$. We collect data from an entire video and put the data into a matrix, $\boldsymbol{X}$, where the columns of $\boldsymbol{X}$ represent the signal at different times. Using an SVD, we rewrite the matrix $\boldsymbol{X}$ in terms of three matrices, $\boldsymbol{U}$, $\boldsymbol{S}$ and $\boldsymbol{V}$, such that

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T. \quad (1)$$

The columns of the $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices are orthonormal; $\boldsymbol{S}$ is a maximally compact, diagonal matrix.

Given the left-singular vectors $\boldsymbol{U}$ and our original data $\boldsymbol{X}$, we project our data into the optimal k-dimensional subspace by multiplying

$$\boldsymbol{X}^k = (\boldsymbol{U}^k)^T \boldsymbol{X} \quad (2)$$

where $\boldsymbol{U}^k$ contains only the first k columns of $\boldsymbol{U}$, and $\boldsymbol{X}^k = \vec{x}^k(t)$ is a k-dimensional function of time. We compute a new SVD and a new $\boldsymbol{U}$ matrix for each video, essentially creating movie-dependent subspaces with all the same advantages of speaker-dependent speech recognition.

We use the SVD to reduce the dimensionality of our audio and image data. The reduced representation is nearly as accurate as the original data, but is more meaningful (the noise dimensions have been dropped) and is easier to work with (the dimensionality is lower).

2.2 Color Space

Color changes provide a useful metric for finding the boundary between shots in a video [7][21]. We can represent the color information by collecting a histogram of the colors within each frame and noting the temporal positions in the video where the histogram indicates large frame-to-frame differences.

We collected color information by using 512 histogram bins. We converted the three red, green, and blue intensities — each of which range in value from 0 to 255 — to a single histogram bin by finding the log, in base 2, of the intensity value, and then packing the three colors into a 9-bit number:

$$\text{Bin} = 64\text{floor}(\log_2(R)) + 32\text{floor}(\log_2(G)) + \text{floor}(\log_2(B)) \quad (3)$$

We chose this logarithmic scaling because it equalizes the counts in the different bins for our test videos.

The process of histogramming the video frames converts the original video images into a 512-dimensional signal that is sampled at 29.97 Hz. The order of the dimensions is arbitrary and meaningless; the SVD will produce the same subspace regardless of how the rows or columns of the $\boldsymbol{X}$ matrix are arranged.

The power of the SVD to analyze the color information is shown in Figure 1. This figure shows the first 4 dimensions of the reduced-dimensionality color-histogram data. In the reduced representation, 2 cuts and a dissolve are readily apparent as the 4-D vector changes over time. This representation is effective for finding shot boundaries, but is computationally more expensive than just computing the difference between adjacent histograms. The advantage of this approach is that it lets us combine the color information with the words.

2.3 Word Space

Latent semantic indexing (LSI), a popular technique for information retrieval [8], uses an SVD in direct analogy to the color analysis described above. As we did with the color data, we start analysis of the audio data by collecting a histogram of the words in a transcript of the video. The video's transcript is a document.

Normally, in information retrieval, each document is one of a large collection of electronically-formatted documents from which we want to retrieve the best match. In our case, as with our previous work [20], we want to study only a single document, so we consider portions of that document — sentences. The sentences of a document define a semantic space; each sentence, in general, represents a specific point in the semantic space.

Ando proposed an enhancement to LSI with properties that are advantageous for data that is not uniformly distributed [2]. We used LSI in this work because it is well known

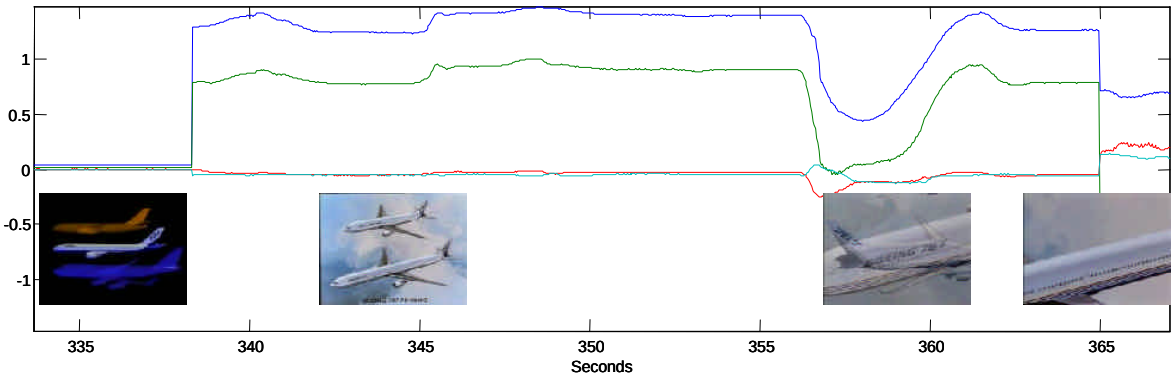

Figure 1: This figure shows the first four dimensions of the reduced-dimensionality color signal. It shows shot boundaries at 338 and 365 seconds; and a dissolve from 357 to 360 seconds. Sample frames from each section are also shown (from "21st Century Jet" [19]).

but other representations can be substituted as long as they allow us to think about the document's structure as a point in a subspace.

Two difficult problems associated with semantic information retrieval are posed by synonyms and polysemy. Often, two or more words have the same meaning. For information retrieval, we want to be able to use any synonym to retrieve the same information. Conversely, many words have multiple meanings: for example, *apple* in a story about a grocery store is likely to have a denotation different from *Apple* in a story about a computer store. LSI handles synonyms and polysemy by finding axes in the term–document space that best explain the histogram data, thus connecting words that are linked in different documents.

We describe the similarity of two points in semantic space by the angle between the two vectors. We compute this value by finding the cosine of the angle between the two vectors,

$$\cos\varphi = (v_1 \cdot v_2) / (|v_2||v_2|). \quad (4)$$

The angle metric is important because documents will often different lengths, but have the same relative proportions of words. Thus their vector space representations will have very different magnitudes (and a high Euclidean distance) but will both have the same angle in semantic space.

2.4 Acoustic Space

MFCC (mel-frequency cepstral coefficient) is a popular technique in the automatic speech recognition community to analyze speech sounds. Based on auditory perception, the MFCC representation captures the overall spectral shape of a sound, while throwing away the pitch information. This allows MFCC to distinguish one vowel from another, or more musically, one instrument from another, but to ignore the melody. We use MFCC to reduce an audio signal from it's original representation (>22kHz sampling rate) to a 13-dimensional vector sampled 89 times a second.

MFCC has been used in earlier music segmentation work [11], but it is not the ideal representation. In this test it allows us to capture the overall timbral qualities of the sound, but ignore the detailed pitch information. The dimensionality reduction is similar to that performed by an SVD, but takes into account specialized knowledge about audio and dimensions that are safe to ignore. The MFCC representation does not give us any high-level information about rhythm or other musical properties of the signal. Eventually we hope that more sophisticated acoustic features will allow us to segment musical accompaniment at phrase or beat boundaries, or even to detect mood changes in movie scores.

3. SCALE SPACE

Witkin [22] introduced the idea of using 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

$$s_\sigma(t) = \int s(\tau) g(\sigma, t-\tau) d\tau \quad (5)$$

where $g(\sigma, t-\tau)$ is a Gaussian kernel with a variance of σ^2. As σ approaches zero, $s_\sigma(t)$ is nearly equal to $s(t)$. For larger values of σ, the resulting signal, $s_\sigma(t)$, is smoother because the kernel is a lowpass filter. We have transformed a one-dimensional signal into a two-dimensional image.

An important feature of scale space is that the resulting analysis is a continuous function of the scale parameter, σ. Because the location of a local maximum in scale space is well behaved [3], we can start with a peak in the signal at the largest scale and trace it back to the exact point at zero

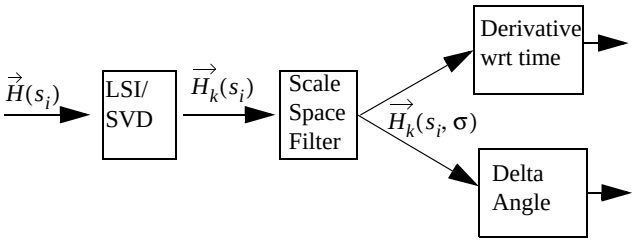

Figure 2: The LSI-SS algorithm. The top path shows the derivative based on euclidean distance. The bottom path shows the proper distance metric for LSI based on angle. See Section 3 for definitions.

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 look for changes in the signal over time. We do so by calculating the derivative of the signal with respect to time and then finding 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 previously 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, we trace with the new signal a smoother path through this 92-dimensional space. To segment the signal, we look for the local peaks in the magnitude of the vector derivative.

Combining LSI analysis with scale-space segmentation is straightforward [20]. This process is illustrated in Figure 2. We describe the scale-space process as applied to semantic content. The analysis of the acoustic and color data is identical.

The semantic data is first grouped into a time sequence of sentences $\{s_i\}$. From these groups, we create a histogram of word frequencies, $\vec{H}(s_i)$ a vector function of sentence number s_i. LSI/SVD analysis of the full histogram $\boldsymbol{X} = \vec{H}(s_i)$ produces a k-dimensional representation, $\overrightarrow{H_k}(s_i) = \boldsymbol{X}^k$ of the document's semantic path (where the dimensionality k is much less than the original histogram.)

We use a lowpass filter on each dimension of the reduced histogram data $\overrightarrow{H_k}(s_i)$, replacing s in Equation (5) 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 replacement gives $\overrightarrow{H_k}(s_i, \sigma)$, a k-dimensional vector function of sentence number and scale.

We calculate a joint audio-video segmentation by concatenating all the data before calculating the SVD. If needed, we resample the word, audio and video data so they have the same sampling rate. It is also important to scale the two sets of data so they have similar power before they are combined (EigenPoints has a similar problem [6]). $\overrightarrow{H_k}(t)$ is now a function of time and includes all the information — image, words and acoustics — we have about the signal.

We are interested in detecting large changes in multimedia scale spaces. The distance metric in Witkin's original scale-space work [22] was based on Euclidean distance. When we use 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 value into a distance measure by subtracting the cosine from 1.

4. RESULTS

We evaluated our approach with several pilot studies. First, to quantify the results of our segmentation algorithm, we performed scale-space hierarchical segmentation on a 1-hour videotape of a documentary and compared the results to several types of segmentations (Section 5.2).

We evaluated our hierarchical representation's ability to segment a musical signal, the 30-minute Headline News television show and the first hour of the Boeing 777 documentary. We describe qualitative results and a quantitative metric, and show how our results compare to those obtained with automatic shot-boundary and manual topical segmentations. We have not found a good test of joint audio–video segmentation.

In this section we illustrate our algorithm by showing some intermediate results (Section 4.1) and qualitative results (Section 4.2) on semantic segmentation. We show similar qualitative results on the music that might accompany a video (Section 4.3). Then we describe a quantitative measure of segmentation accuracy (Section 4.4) and apply it to shot-boundary segmentation (Section 4.5) and semantic segmentation (Section 4.6)

4.1 Intermediate Results

We first illustrate our hierarchical segmentation algorithm by showing intermediate results using just the semantic information from the Headline News video. We compared the results of hierarchical segmentations and of the ground truth. The LDC [13] provided story boundaries

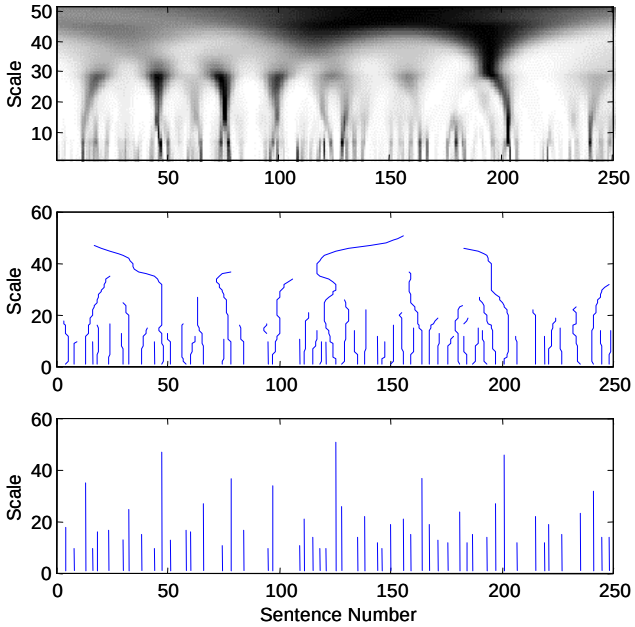

Figure 3: Representations of the semantic information in the Headline News video in scale space. The top image shows the cosine of the angular change of the semantic trajectory with different amounts of lowpass filtering. The middle plot shows the peaks of the scale-space derivative for the tomography chapter. The bottom plot shows the peaks traced back to their original starting point. These peaks represent topic boundaries

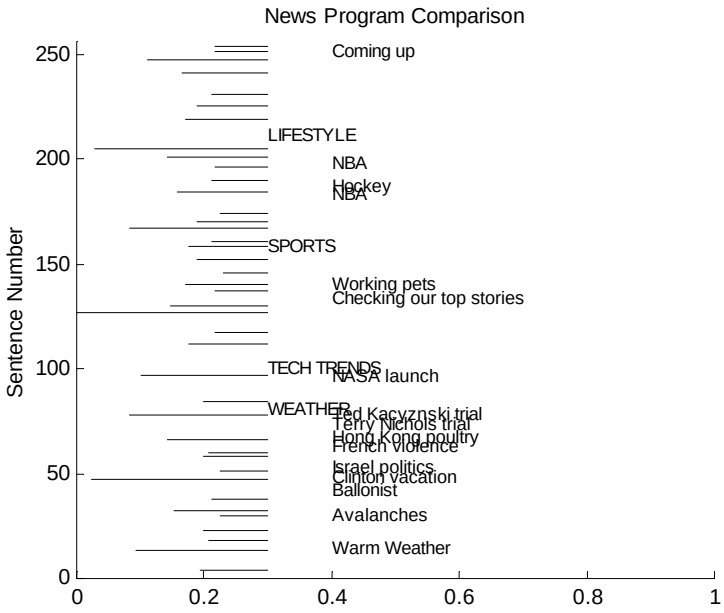

Figure 4: A comparison of ground truth (right) and the size of boundaries for the Headline News video as determined by scale-space segmentation. The major headings are in all capitals, and the sub-headings are in upper and lower case.

for this video, but we estimated the high-level structure based on our familiarity with this news program. We removed the timing and other meta information from the transcript before analysis. We found 257 sentences in this broadcast transcript; after removal of stop words, we counted 1032 distinct words.

Our segmentation algorithm measured the changes in a signal over time as a function of the scale size. A scale-space segmentation algorithm produced a boundary map showing the edges in the signal, as shown in Figure 3. At the smallest scale there were many possible boundaries; at the largest scale, with a long smoothing window, only a small number of edges remained.

The location of the peak is not fixed in time for all scales. A peak (or segmentation boundary) at scale 0 moves to different times as we lowpass filter the underlying data. Thus we traced the boundary back to its true location (at zero scale) and drew the straightened boundary map shown at the bottom of Figure 3. For any one boundary — indicated by its vertical lines — strength is represented by line height, and is a measure of how significant this topic change is to the document.

4.2 Qualitative Measure of Semantic Segmentation

The classic measures for the evaluation of text-retrieval performance [1] do not extend easily to a system that has hierarchical structure. Instead, we evaluated our results by examining a plot that compared headings and the scale-space segmentation strength. The scale-space analysis produced a large number of possible segmentations; for each study, we plotted only twice the number of boundaries indicated by the ground truth.

Our results of calculating the hierarchical segmentations of the Headline News are shown in Figure 4. On the right, the major (leftmost text) and the minor (rightmost text) headings are shown. The left side of the plot shows the strength of the boundary. The “Weather,” “Tech Trends” and “Lifestyles” sections are indicated within a few sentences, yet there are large peaks at other locations in the transcript. Interestingly, there is a large boundary near sentence 46, which neatly divides the softer news stories at the start of this broadcast from the political stories that follow.

4.3 Qualitative Measure of Musical Segmentation

Figure 5 shows the qualitative results of our music segmentation test. We hand labeled the major events in a live performance. We used only the acoustic features in this

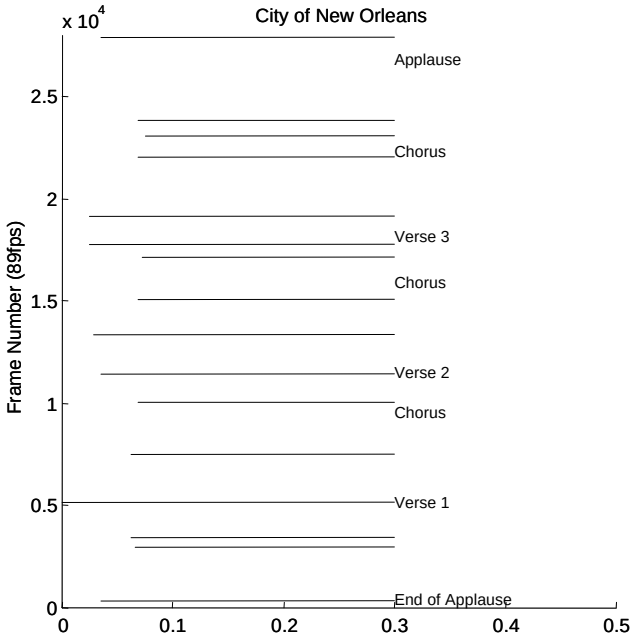

Figure 5: Qualitative segmentation results from the musical test signal. The labels on the right show manual segmentations, while the lines on the left show the size of the discontinuity.

test. We were able to easily find the end of the applause and the start of verse 1 and 2. The other verses and choruses are close to segment boundaries.

4.4 Measuring Segmentation Accuracy

We used Lafferty's segmentation metric (Section 8 of [4]) to summarize our quantitative results. Lafferty's measure is appropriate in that it allows us to characterize our segmentation performance with a single number, and it takes into account close matches without adding undue complexity.

We measured the degree of agreement between two segmentations by looking at both ends of a fixed window passed over two sets of segmentation data. Figure 6 summarizes the process for one set of data labeled with ground truth and for another set labeled "experimental." The segmentation, at this point, is successful if both ends of the window fall within the same segment or if each is in a different segment. The segmentation has made a mistake at some point if the window falls entirely within one segment

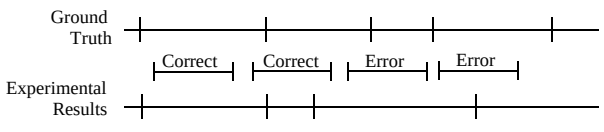

Figure 6: We evaluate accuracy by measuring whether the ends of a fixed-size window fall in the same or different segments.

for one set of data and the window for the second set of data covers two or more segmentation boundaries. We then moved the window over the entire document and calculated the fraction of correct windows. An especially important property of Lafferty's measure for semantic segmentations is that small offsets in the segmentation lower the performance metric, but do not cause complete misses. As suggested by Doddington [4], we used a fixed window size that was 50 percent of the length of the average segment calculated using the ground-truth segmentation.

Lafferty's segmentation metric has several properties that were reflected in our data. Assume that the probability that any particular frame is a boundary is independent and is fixed at $p = 1/(2N)$ where N is the window length or half the average segment length. If we measure the accuracy of an experimental result with just one (large) segment, then Lafferty's measure is equal to (for large N)

$$(1-p)^N \cong 0.606 . \quad (6)$$

Conversely, if we measure the accuracy of a segmentation that puts a boundary at every time step, then Lafferty's measure is equal to

$$1-(1-p)^N \cong 0.394 . \quad (7)$$

Finally, if we compare two random segmentations, each with the same probability of a boundary, Lafferty's measure indicates that the segmentation accuracy is

$$(1-p)^{2N} + [1-(1-P)^N]^2 \cong 0.523 . \quad (8)$$

4.5 Shot Boundary Segmentation

We used the segmentation produced by a state-of-the-art commercial segmenter, designed by YesVideo [7], as our shot-boundary "ground truth." Covell reported that, on a database of professionally produced wedding videos, their segmenter had an overall precision of 93% and a recall of 91%. For their test set, most of the errors, both false positives and false negatives, were due to uncompensated camera motion.

We performed quantitative tests on the first hour of the Boeing 777 video. This video had 102,089 frames. The semantic analysis found 1314 distinct words in 537 sentences. There were shot boundaries on average every 242 frames. The standard deviation of the Gaussian blur used in scale-space filtering is $\sigma = 1.1^{s-1}$, where s is the scale number.

We show the results here using only the color data, only the word data, and the combination. We evaluate Lafferty's measure for the segmentation boundaries predicted at each scale, effectively assuming that a single scale would produce the best segmentation. Assuming all segmentation boundaries are at the same scale is not the best solution; instead, the information from the scale-space segmentation

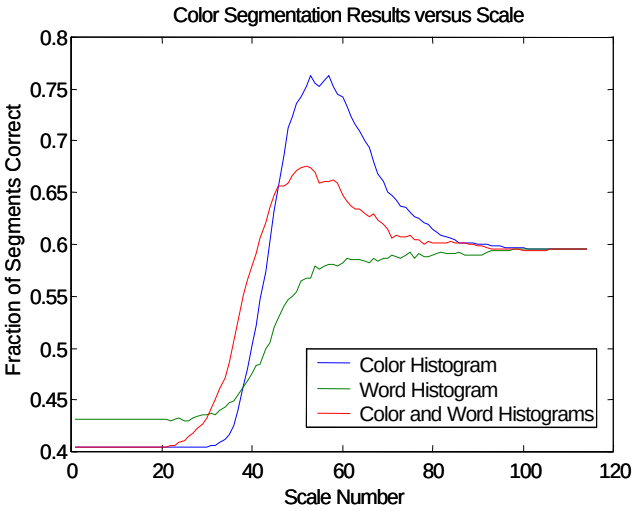

Figure 7: This figure shows the accuracy of the scale-space segmentation algorithm, for each scale, at finding shot boundaries. Video was the first hour of the Boeing 777 video, compared to "ground truth" from YesVideo's segmenter [7]. The upper peak is from color data, the middle peak is from color and word data, and the lower curve is from words alone.

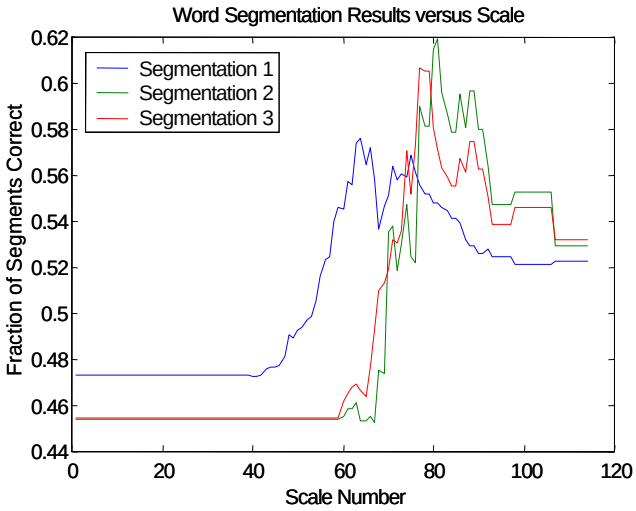

Figure 8: Scale-space segmentation using the word histograms compared to three different manual segmentations, tested with Lafferty's measure. Video from the Boeing 777 video.

metric should be used as input to a higher-level model of video transitions, as suggested by Srinivasan [21].

Figure 7 shows how segmentation accuracy varied with scale, comparing the segmentation at each scale to the YesVideo results. At small scales the probability, as predicted, was 40%. At large scales, only one or two boundaries were found, and, as predicted, the accuracy was 60%. At the middle scale, the segmentation accuracy was 77% — well above that of random segmentations (52%). As expected, the semantic signal does not predict the color boundaries. Adding the semantic information to the color information does reduce the highest accuracy at any one scale to 67%.

4.6 Semantic Segmentation

We also compared our algorithm's semantic segmentations to those of humans. Two of the authors of this paper and a colleague segmented the transcript of the first hour of the Boeing 777 video. There was a wide range in what these three readers described as a segment; they chose to segment the text with 29, 37, and 122 segment boundaries. They found it difficult to produce a hierarchical segmentation of the text. The video was designed to be watched in one sitting; it transitions smoothly from topic to topic, weaving a single story.

Figure 8 shows the how the scale-space segmentation compares, across scale, to the manual segmentations. As expected, the best scale is larger then that shown for the color segmentations in Figure 7. The scale-space segmentation algorithm matches each of the human's segmentations, equally well. Perhaps most surprisingly, the color information is a good predictor of the semantic boundaries. This correlation may indicate that the color signal carries information regarding content changes that is richer than we assumed.

5. CONCLUSIONS

We have demonstrated a new framework for combining into a unified representation and for segmenting information from multiple types of information from a video. We used the SVD to reduce the dimensionality of the combined signal. Then, we applied scale-space segmentation to find edges in the signals that corresponded to large changes. We demonstrated how these ideas apply to words, music and color information from a video.

These techniques are an important piece of a complete system. The system we have described does not have the domain knowledge to know that, for example, when it is considering a videotape of a news broadcast, the phrase "coming up after the break" is a pointer to a future story and is not a new story in its own right. Systems that include domain knowledge about specific types of video content [9] show how this knowledge is incorporated.

We described our hierarchical segmentation results by comparing them to conventional segmentations. Qualitatively, the automatic segmentation has many similarities to a manual segmentation. It is hard to evaluate the quantitative results, but we were surprised by the amount of information that was available in the color information for topical segmentation.

It is difficult to know when to stop when segmenting data, or equivalently, how small a change to detect. Hierarchical segmentation, as we have described here, does not solve that problem, but does make it easier for a human user to manage the decision. Witkin [22] describes a means to automatically to form a single segmentation using segments that are stable across many scales. We have not applied this algorithm to our data.

Many videos are not organized in a perfect hierarchy. In text, the introduction often presents a number of ideas, which are then explored in subsequent sections; a graceful segue is used for transitions between ideas. This lack of hierarchy is much more apparent in a news show, the structure of which may be somewhat hierarchical, but is designed to be watched in a linear fashion. For example, 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. We do not know whether more videos are organized linearly, as a news program, or hierarchically.

7. ACKNOWLEDGMENTS

We appreciate thoughtful discussions we had with Michele Covell, Byron Dom and Arnon Amir. The anonymous reviewers had many good questions and suggestions. Lyn Dupré helped with the editing.

8. 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 Understanding Workshop* (Sponsored by DARPA), Feb. 1998.

[2] Rie Kubota Ando. "Latent semantic space: iterative scaling improves precision of inter-document similarity measurement." *Proceedings of the ACM SIGIR*, Athens, Greece, pp. 216–223, 2000.

[3] 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.

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

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

[6] M. Covell, C. Bregler. "Eigenpoints." *Proc. Int. Conf. Image Processing*, Lausanne, Switzerland, Vol. 3, pp. 471–474, 1996.

[7] Michele Covell, Subutai Ahmad, Jeff Edwards. "Transition detection for home-video browsing." Submitted to *ACM Multimedia 2001*, Ottawa, Ontario, Canada, Sept. 30, 2001.

[8] 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.

[9] S. Dharanipragada, M. Franz, J.S. McCarley, K. Papineni, S.Roukos, T.Ward, W.-J. Zhu, "Statistical models for topic segmentation." *Proceedings of ICSLP-2000*, Beijing, 2000.

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

[11] Jonathan Foote. "Visualizing music and audio using self-similarity." *Proceedings of ACM Multimedia'99*, pp. 77–80, Orlando, Florida, November 1999.

[12] Arlo Gutherie, "City of New Orleans," on *Tribute to Steve Goodman*, Audio CD, Red Pajamas Records, Nashville, TN, 1991.

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

[14] V. Kobla, D. Doermann, C. Faloutsos. "Video Trails: Representing and visualizing structure in video sequences," *Proc. of ACM Multimedia*, pp. 335–346, 1997.

[15] 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.

[16] Linguistic Data Consortium. "1997 english broadcast news speech (Hub-4)." LDC catalog no.: LDC98S71, File ed980104.

[17] Richard F. Lyon. "Speech recognition in scale space," *Proc. of 1984 ICASSP*. San Diego, March, pp. 29.3.1–4, 1984.

[18] C. Neti, G. Potamianos, J. Luettin, I. Matthews, D. Vergyri, J. Sison, A.Mashari, and J. Zhou. "Audio–visual speech recognition," *Final Workshop 2000 Report*, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD (Oct. 12, 2000).

[19] PBS Home Video, "21st century jet: The building of the 777." Channel 4, London, 1995.

[20] 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.

[21] Savitha Srinivasan, Dulce Ponceleon, Arnon Amir, Dragutin Petkovic. "'What *is* in that video anyway?' In search of better browsing." *Proceedings IEEE International Conference on Multimedia Computing and Systems*, pp. 388–393. Florence, Italy, 7-11 June 1999.

[22] 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.