# Summarizing Video using Non-Negative Similarity Matrix Factorization

Matthew Cooper and Jonathan Foote
FX Palo Alto Laboratory
3400 Hillview Avenue Bldg. 4
Palo Alto, CA 94304 USA
Email: [cooper, foote]@fxpal.com

*Abstract*— **We present a novel approach to automatically extracting summary excerpts from audio and video. Our approach is to maximize the average similarity between the excerpt and the source. We first calculate a similarity matrix by comparing each pair of time samples using a quantitative similarity measure. To determine the segment with highest average similarity, we maximize the summation of the self-similarity matrix over the support of the segment. To select multiple excerpts while avoiding redundancy, we compute the non-negative matrix factorization (NMF) of the similarity matrix into its essential structural components. We then build a summary comprised of excerpts from the main components, selecting the excerpts for maximum average similarity within each component. Variations integrating segmentation and other information are also discussed, and experimental results are presented.**

## I. Introduction

As digital media collections grow in size and number, summarization has become an increasingly important research area. Media summarization technologies have numerous applications in e-commerce and in information retrieval. Many such applications use summaries and/or proxies of longer works, because of the large file sizes and high bandwidth requirements of multimedia data. Thus it is desirable to have a summary of the media work that is reduced in some manner, typically by excerpting a segment that is a good representation of the longer work. Many existing segmentation algorithms can't guarantee that the segment is at all representative of the larger work. For example, some approaches use the first 30 seconds of an audio track to represent the whole track. This can be highly unsatisfactory if the bulk of a particular track bears little resemblance to its idiosyncratic introduction.

Most summarization approaches start by analyzing the structure or semantics of the source material. Statistical text summarization typically uses term frequency/inverse document frequency (tf/ idf) to select paragraphs [1], sentences [2], or key phrases that are both representative of the document and differentiate it from other documents. Audio summarization approaches typically segment the audio, then select a representative portion of each segment. These are concatenated to serve as a summary [3]. Video has been summarized by "scene transition graphs" [4], among other methods. After clustering, the keyframes closest to each cluster centroid are chosen to represent that cluster. Other approaches summarize video using various heuristics, typically derived from an analysis of accompanying closed captions [5]. Gong and Liu have used the singular value decomposition (SVD) of a feature-frame matrix to construct video summaries [6]. The SVD is used to reduce feature space dimension for clustering, and to assess the novelty of individual frames.

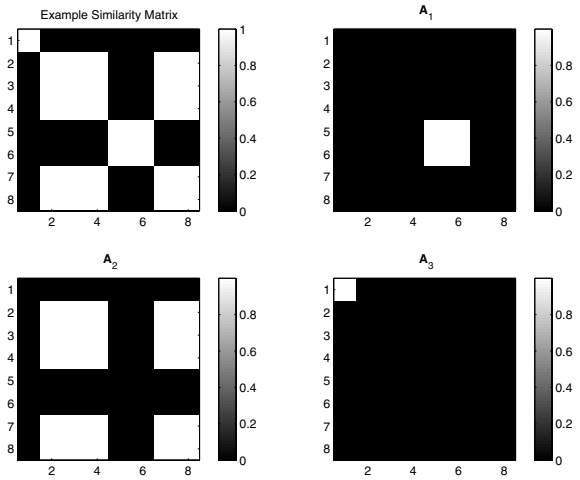


Fig. 1. The upper left panel shows $\mathbf{S}$ from (3). The remaining panels show $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ computed via nmf and (8).

We present a method for automatically producing summaries of linear media, where linear means a function of a one-dimensional variable. Examples of linear media are audio and video, which are functions of time, and ASCII text, which is a discrete function of file position. We construct our summaries by analysis of the self-similarity across all time instants embedded in a similarity or affinity matrix (e.g. [7], [8]). We then identify the contiguous segment with maximum average similarity to the piece as a whole. This can be found by maximizing the sum of the similarity matrix over the support of the segment.

We further extend this approach to construct video summaries. Video, especially home video, typically contains heterogeneous segments from diverse locations. Attempting to summarize this with a single contiguous segment will be less satisfactory than a summary comprised of multiple excerpts from the different segments. To select these excerpts, we calculate the non-negative matrix factorization (NMF) of the similar-

ity matrix. The NMF is an unsupervised technique for building a "parts-based" representation of a data set [9]. We employ the NMF to determine the *essential structural components* of the source stream, as represented in the similarity matrix.

## II. SELF-SIMILARITY ANALYSIS

### A. Parameterization

The first step is to parameterize the video. We compute feature vectors based on low-order discrete cosine transform (DCT) coefficients. We sample frames at 1 Hz and transform the individual RGB frames into the Ohta color space in which the three channels are approximately decorrelated [10]. The DCT of each transformed channel is computed and a feature vector is formed by concatenating the resulting 25-49 low frequency coefficients of the three channels. The sole requirement is to quantify similarity; similar frames must have similar feature vectors. The transform method is optimized for analysis (and, if desired, computational complexity) rather than dimension reduction or fidelity.

### B. Distance Matrix Embedding

Once the signal has been parameterized, it is embedded in a two-dimensional representation. The key is a measure $d$ of the (dis)similarity between a pair of feature vectors $v_i$ and $v_j$ (calculated from frames $i$ and $j$ in the previous step). A useful similarity measure is the cosine of the angle between the parameter vectors:

$$d_c(v_i, v_j) = \frac{< v_i, v_j >}{\|v_i\|\|v_j\|} \ . \tag{1}$$

This measure has the property that it yields a large similarity score even if the vectors are small in magnitude. Herein we use an exponential variant of this measure to limit its range to $(0, 1]$:

$$d_e(v_i, v_j) = \exp(d_c(v_i, v_j) - 1) \ . \tag{2}$$

The distance measure is a function of two frames, hence instants in the source signal. To consider the similarity between all possible instants in a signal, we embed the distance measure in the similarity matrix $\mathbf{S}$ such that $\mathbf{S}(i, j) = d_e(v_i, v_j)$. In general, $\mathbf{S}$ will have maximum values on the diagonal (because every frame will be maximally similar to itself); furthermore if $d$ is symmetric then $\mathbf{S}$ will be symmetric as well. Example similarity matrices are shown in the upper left panels of Figs. 1 and 3.

## III. AUTOMATIC SUMMARIZATION

For summarization, we aim to determine the excerpt of a desired length with maximum similarity to the work as a whole. In a video stream with repeated similar segments, we would expect that group of similar segments with the maximum total duration would be represented in the summary. Representative elements from predominant clusters of similar segments can be found from the similarity matrix.

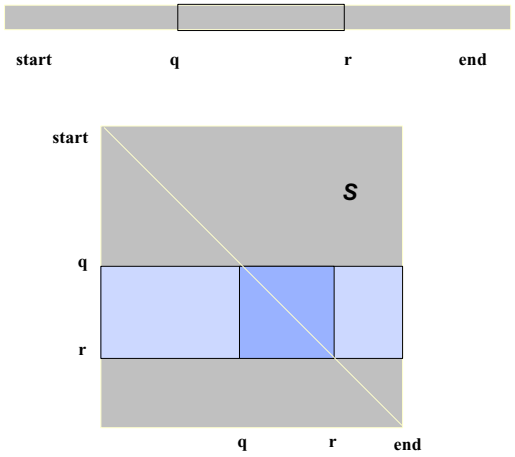


Fig. 2. Evaluating summary score $\bar{\mathbf{S}}(q, r)$ by summing similarity matrix over the interval $(q, \cdots, r)$

### A. Selecting a Single Excerpt

A small example will motivate the following discussion. Given the sequence `ABBBCCBB`, it is desired to find the subsequence of length three with maximal average similarity. We compute the similarity to be one if the sequence members match and zero otherwise. The similarity matrix, shown in the upper left panel of Fig. 1, is:

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \ . \tag{3}$$

Calculating the average similarity for any subsequence is a simple matter of summing the rows (or columns for symmetric $\mathbf{S}$) corresponding to that subsequence, and normalizing by the total sequence length. Thus, the second element in the sequence, `B`, has an average similarity of $5/8 = 0.625$. The possible length three subsequences are: `ABB`, `BBB`, `BBC` and `BCC`, `CCB`, and `CBB`. By adding the corresponding columns we determine that `BBB`, with average subsequence similarity $1.875/3 = 0.625$, is the optimal three-element contiguous summary of the sequence. Using maximum average similarity as our summary criterion also allows us to compare subsequences of different lengths.

The previous example can be generalized to arbitrary sequence lengths as in Fig. 2. Given a segment that starts at time $q$ and ends at time $r$, the average similarity of the segment can be calculated as the total average similarity between the segment and the entire work, normalized by the segment length:

$$\bar{\mathbf{S}}(q, r) = \frac{1}{N(r-q)} \sum_{m=q}^{r} \sum_{n=1}^{N} \mathbf{S}(m, n) \ , \tag{4}$$

where $N$ is the length of the entire work (width and height of $\mathbf{S}$).

In Section IV we compute summaries of a desired length $L$. We determine the summaries by optimizing a score based on (4):

$$Q_L(i) = \bar{\mathbf{S}}(i, i+L) = \frac{1}{NL} \sum_{m=i}^{i+L} \sum_{n=1}^{N} \mathbf{S}(m,n) \qquad (5)$$

for $i = 1, \cdots, N-L$. We select the start point for the summary, $q_L^*$, as

$$q_L^* = \underset{1 \le i \le N-L}{\text{ArgMax}}\, Q_L(i) \ . \qquad (6)$$

The resulting summary is then the excerpt of the source stream from start time $q_L^*$ to end time $q_L^* + L$.

### *B. Selecting Multiple Excerpts*

While a single piece of music often exhibits some global coherence, video is commonly comprised of multiple shots of unrelated scenes. For this reason, a single excerpt may fail to provide an adequate representation of an entire video. One solution is to construct a summary or "skim" from multiple excerpts. These excerpts must then be selected to represent the video's contents while avoiding redundancy.

To satisfy these criteria, we factor the similarity matrix via NMF. The NMF of an $N \times N$ matrix $\mathbf{S}$ is a linear approximation to $\mathbf{S}$ formed by the product of an $N \times K$ matrix $\mathbf{W}$ and a $K \times N$ matrix $\mathbf{H}$:

$$\mathbf{S} \simeq \mathbf{WH} \ = \ \sum_{k=1}^{K} \mathbf{A}_k \ , \qquad (7)$$

$$\text{where } \mathbf{A}_k(i,j) \ = \ \mathbf{W}(i,k)\mathbf{H}(k,j) \ . \qquad (8)$$

The columns of $\mathbf{W}$ are basis vectors for the columns of $\mathbf{S}$. The columns of $\mathbf{H}$ are the projections of the columns of $\mathbf{S}$ on to the basis of $\mathbf{W}$. NMF is distinguished from more common linear approximations such as the SVD by the fact that $\mathbf{W}$ and $\mathbf{H}$ are non-negative[1] These non-negativity constraints cause $\mathbf{W}$ and $\mathbf{H}$ to form an additive, or parts-based, representation in which the basis vectors of $\mathbf{W}$ combine to approximate the columns of $\mathbf{S}$.

NMF has been successfully used to build low-dimensional, additive representations for facial imagery and text document collections [9]. NMF minimizes a generalized divergence between $\mathbf{S}$ and $\mathbf{WH}$, and can be implemented as a simple and efficient iterative procedure [11]. In our context, the basis vectors of $\mathbf{W}$ represent the significant "parts" of $\mathbf{S}$: the significant blocks of high similarity. We use this factorization to generate the terms, $\mathbf{A}_k$ in (8), that represent a structural decomposition of $\mathbf{S}$ and hence, of the source stream itself.

[1] In contrast, the SVD constrains the columns of $\mathbf{W}$ to be orthonormal and the rows of $\mathbf{H}$ to be orthogonal. As a result, they combine to both add and cancel, whereas the combinations of NMF basis vectors and coefficients are strictly additive.

To select multiple excerpts, we process each of the terms in the sum of (7) as in the single excerpt case described in Section III A. We estimate the effective rank, $K$ of $\mathbf{S}$, and compute the optimal length $L$ summaries for each $\mathbf{A}_1, \cdots, \mathbf{A}_K$ using

$$Q_L^{(k)}(i) = \frac{1}{NL} \sum_{m=i}^{i+L} \sum_{n=1}^{N} \mathbf{A}_k(m,n) \qquad (9)$$

as in (5). We select the start point for each excerpt by substituting these scores into (6). Returning to the previous example, the upper right panel of Fig. 1 shows the similarity matrix of (3). The remaining panels show the three terms $\mathbf{A}_1, \mathbf{A}_2$, and $\mathbf{A}_3$, computed via (8), clearly elucidating the structure of the original sequence.

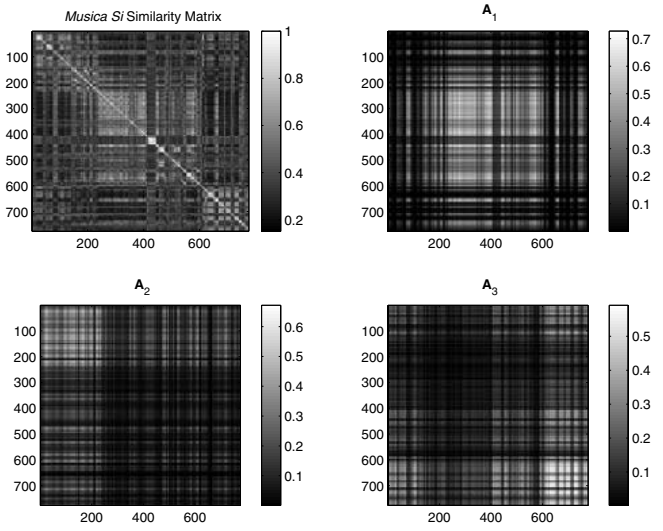


Fig. 3. The upper left panel shows the similarity matrix for "*Musica Si*". The upper right, lower left, and lower right panels show $\mathbf{A}_1, \mathbf{A}_2$,and $\mathbf{A}_3$ of (8), respectively.

## IV. EXPERIMENTS

In [13], we applied the single excerpt technique to summarization of digital music. In limited experiments, we successfully summarized popular, classical, and jazz pieces. For brevity, we focus here on experiments with video in which we produce summaries comprised of multiple excerpts. We use the similarity matrix to determine the optimal summaries of a desired length, $L$, in three steps. First, we estimate the rank of $\mathbf{S}$ by discarding singular vectors with singular values less than one tenth of the maximum singular value (no singular vectors need be computed). Denote the estimated rank $K$. Next, we compute the $K$-term NMF of $\mathbf{S}$ to determine $\mathbf{A}_1, \cdots, \mathbf{A}_K$ of (7). We then compute the score of (9) and the excerpt start point, $q_L^*$ for each $k = 1, \cdots, K$. The upper left panel of Fig. 3 shows the similarity matrix computed from "*Musica Si*", a Spanish show featuring musical performances from [12].

We have analyzed several videos from [12]. While it is difficult to objectively characterize the results, we have informally found that the excerpts do emphasize major segments and segment clusters, when present. "*Musica Si*" contained an introduction and two longer musical performances separated by interludes with the show's hosts. The first performance starts near 280 seconds after a 70 second onstage interview with the musicians. The second performance starts at 530 seconds. For this video, the automatically selected excerpts included both songs and the introduction. We shows the similarity matrix and the results of NMF in Fig. 3. The three major segments of the video are clearly represented by $\mathbf{A}_1, \mathbf{A}_2$ and $\mathbf{A}_3$. The "Home Video of Lisa" exhibited less structure, and the excerpts include different scenes of Lisa in a gymnastics class, and multiple segments of a hot air balloon race. Although the balloon race appeared in two of the excerpts, one excerpt showed the balloons being filling on the ground, while the other showed them in flight. The results appear in Table I. In addition, selected results can be viewed on the web [2].

TABLE I

SUMMARIZATION RESULTS FOR TEST VIDEOS. TIMES ARE IN SECONDS.

| Results: Home Video of Lisa | | | | | | |
|---|---|---|---|---|---|---|
| Length | Excerpt 1 | | Excerpt 2 | | Excerpt 3 | |
| 10 | 180 | 190 | 449 | 459 | 695 | 705 |
| 30 | 182 | 212 | 492 | 522 | 549 | 579 |
| Results: *Musica Si* | | | | | | |
| Length | Excerpt 1 | | Excerpt 2 | | Excerpt 3 | |
| 10 | 160 | 170 | 386 | 396 | 630 | 640 |
| 30 | 154 | 184 | 374 | 404 | 616 | 646 |

## V. CONCLUSION

We have presented a quantitative approach to automatic media summarization which makes minimal assumptions regarding the characteristics of the source video. The approach is founded on similarity analysis, in which inter-frame similarity data is embedded in a matrix which reveals the major segment-level structure of the original video. By summing the columns of the similarity matrix the most representative contiguous portions of the video are determined, and used for summaries of arbitrary length. Using the NMF of the similarity matrix, we decompose the video into major structural components. We have presented the technique and results on test videos. In each case, the resulting summaries provided satisfying summaries of the original videos and their major segment clusters.

The analysis can be customized to build summaries of varying length, or to weight specific features by selection of appropriate feature vectors. Similarly, by use of a weighting function, $w$, specific portions of a video can be emphasized for summary construction, by modifying (4):

$$\bar{\mathbf{S}}_w(q,r) = \frac{1}{N(r-q)} \sum_{m=q}^{r} \sum_{n=1}^{N} w(n)\mathbf{S}(m,n) \ . \qquad (10)$$

This leads to several extensions we hope to investigate in future work. Weighting functions can be designed to emphasize high quality video segments, for instance, segments with little estimated camera motion. In contexts where summaries of varying length are desired, or the summary length is unknown in advance, $\mathbf{S}$ can be discarded after calculating the inner sum of (5) for each $i = 1, \cdots, N-L$. This inner sum is simply the sum of the columns of $\mathbf{S}$. In other extensions, we plan to integrate the summarization with similarity-based video segmentation techniques [14] to constrain the summaries to begin and end at shot boundaries. We plan to evaluate our current approach and these extensions more formally with more extensive testing and user evaluation.

[2] `http://www.fxpal.com/media/videosummaries.html`

## REFERENCES


[1] Abracos, J. and Lopes, G.P., Statistical methods for retrieving most significant paragraphs in newspaper articles, In *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, pages 51-57, 1997.

[2] Zechner, K. Fast generation of abstracts from general domain text corpora by extracting relevant sentences, In *Proc. of the Intl. Conf. on Computational Linguistics*, 1996.

[3] Logan, B. and Chu, S., Music Summarization Using Key Phrases, in *Proc. IEEE ICASSP*, 2000.

[4] Yeo, B.-L., Yeung, M.M., Classification, Simplification, and Dynamic Visualization of Scene Transition Graphs for Video Browsing, in *Storage and Retrieval for Image and Video Databases (SPIE)* pp. 60-70, 1998.

[5] Christel, M., Stevens, S., Kanade, T., Mauldin, M., Reddy, R., and Wactlar, H. Techniques for the Creation and Exploration of Digital Video Libraries. In *Multimedia Tools and Applications*, B. Furht, ed. Kluwer Academic Publishers, 1996.

[6] Gong, Y.H, Liu, X. Video Summarization Using Singular Value Decomposition, in *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recoginition*, June, 2000.

[7] Eckman, J.P., et al., Recurrence Plots of Dynamical Systems, in *Europhys. Lett.* **4**(973), 1987.

[8] Foote, J. Visualizing Music and Audio using Self-Similarity. In *Proc. ACM Multimedia*, pp. 77-80, 1999.

[9] Lee, D. and Seung, H. S., Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature* **401**:788-791, 1999.

[10] Y-I Ohta, T. Kanade, and T. Sakai. Color Information for Region Segmentation. *Comp. Graphics & Image Processing*, **13**:222-241, 1980.

[11] Lee, D. and Seung, H. S., Algorithms for Non-negative Matrix Factorization. *Proc. Conf. on Neural Information Processing Systems*, 2000.

[12] MPEG Requirements Group. Description of MPEG-7 Content Set, Doc. ISO/MPEG N2467, 1998.

[13] Cooper, M. and Foote, J., Automatic Music Summarization via Similarity Analysis, to appear, *Proc. Intl. Symposium on Music Information Retrieval*, 2002.

[14] Cooper, M. and Foote, J., Scene Boundary Detection Via Video Self-Similarity Analysis. *Proc. IEEE Intl. Conf. on Image Processing*, pp. 378-81, 2001.