Audio Similarity Matrices Enhancement in an Image Processing Framework

Florian Kaiser, Marina Georgia Arvanitidou and Thomas Sikora
Communication Systems Group
Technische Universität Berlin
{name}@nue.tu-berlin.de

Abstract

Audio similarity matrices have become a popular tool in the MIR community for their ability to reveal segments of high acoustical self-similarity and repetitive patterns. This is particularly useful for the task of music structure segmentation. The performance of such systems however relies on the nature of the studied music pieces and it is often assumed that harmonic and timbre variations remain low within musical sections. While this condition is rarely fulfilled, similarity matrices are often too complex and structural information can hardly be extracted. In this paper we propose an image-oriented pre-processing of similarity matrices to highlight the conveyed musical information and reduce their complexity. The image segmentation processing step handles the image characteristics in order to provide us meaningful spatial segments and enhance thus the music segmentation. Evaluation of a reference structure segmentation algorithm using the enhanced matrices is provided, and we show that our method strongly improves the segmentation performances.

1 Introduction

Music structure segmentation aims at drawing the temporal map of a music piece and is the core of many Music Information Retrieval (MIR) applications such as music synchronization, music summarization, music transcription, cover detection, score following, etc. A popular approach for this task consists in visualizing the structure of audio signals by means of an audio similarity matrix [2]. Such a visualization indicates segments of high acoustical self-similarity in the audio signal, and boundaries between musical sections within the music piece can then be retrieved. Furthermore, segments of homogeneous acoustical information or repetitive patterns can be detected within the music piece and its structure can be explained.

Similarity matrices have thus become a popular tool in the MIR community, and its introduction considerably improved the music segmentation performances. However, the assumptions on which it relies for the modeling of structural parts are rather restrictive. For a good visualization of musical structures, the intrinsic acoustical properties of musical sections must have very low- variance, whereas the acoustic properties of two different segments should allow for discrimination. Considering large-scale music collections, this condition is rarely fulfilled and the complexity of similarity matrices for some music pieces remains too high to allow for their segmentation.

Some work have been proposed for matrix enhancement, and strengthening of structural information in similarity matrices. In [7], the authors propose to reduce the complexity of the visualization by defining a contextual similarity measure. Considering a larger observation horizon for the measure of similarity, one favors the visualization of repetitive motives. However, while repetitions are highlighted, the temporal resolution of sections boundaries is reduced. A similar approach is used in [10] where dynamic features are extracted by modeling the temporal evolution of the spectral shape over a short time window to capture local timbre properties. Varying the window size, authors derive similarity matrices that either relate to short-term or long-term structures. Finally, matrix transformations designed for the particular properties of repeated elements in the similarity matrices are proposed in [12]. The transformation aims at reinforcing off-diagonals, and therefore facilitate repetitions detection.

Most of the approaches thus introduce a local modeling of audio features to enhance the measure of similarity. While sections might be better characterized, boundaries between sections are however blurred, which is prejudicial to the segmentation performance.

In this paper, we propose an alternative approach for the similarity matrix enhancement problem by considering it in an image processing framework. The idea is to favor regions of high self-similarity with image filtering techniques that preserve the sharpness of boundaries. Musical objects in the music piece, as audio segments parameterized and embedded in a similarity matrix, indeed relate to visual ob-

jects that can be segmented in the image formed by the similarity matrix. Image processing thus seems to be a rather natural approach for the study of similarity matrices, and techniques for audio segmentation with mean of similarity matrices such as in [3] were already inspired from the image segmentation research. Authors defined the pattern of an ideal boundary in the similarity matrix and used visual pattern matching techniques to segment the audio. We propose here to apply further visual object segmentation techniques to enhance the structure visualization and thus improve the performance of the segmentation.

The approach is as follow. We consider similarity matrices as intensity images. These image representations of musical information are processed exploiting their low pass frequencies as well as their spatial geometric structures and are segmented. We show that these segmentation masks can highly discriminate musical structural parts. Thus the original similarity matrix is enhanced using this information, yielding a representation of the original audio data where structural parts are strengthened.

In the following section the similarity matrices are described, the image processing framework and the matrix enhancement are presented. In section 3 we show the benefits of our approach for the task of structure detection in music pieces and in section 4 we conclude this paper and discuss about further work for image-oriented processing of audio similarity matrices.

2 Enhancing Similarity Matrices

In this section we briefly introduce the kind of similarity matrices we are working with. We will then present the image processing tools we use to generate a segmentation mask that reinforces the structural information.

2.1 Similarity Matrices

2.1.1 Audio Features Extraction

For a good visualization of musical structures, it is needed to extract acoustic properties that may distinguish musical sections within a music piece. Studying music perception, Bruderer showed in [1] that boundaries between structural parts of a song are mainly determined by a combination of changes in timbre, tonality and rhythm. Popular audio features for computing similarity matrices are thus the Mel Frequency Cepstral Coefficients (MFCC's) for the description of timbre, and the Chroma features to embed harmonic-related properties of sounds. In this paper, similarity matrices shown in the examples were computed on the MFCC features. In order to contain the size of matrices in a reasonable range, features are sampled at 4Hz. Features are also normalized to zero mean and unit variance.

2.1.2 Similarity Matrix Computation

After parametrization of the audio, the similarity between each signal frame is measured and embedded in the self similarity matrix S. Each element $S(i, j)$ is defined as the distance between the feature vectors v_i and v_j, extracted over frames i and j. The cosine angle is used as a similarity measure :

$$d(v_i, v_j) = \frac{< v_i, v_j >}{||v_i||||v_j||} \quad (1)$$

An exponential variant of this distance is used to limit its range to [0, 1] :

$$de(v_i, v_j) = exp(d(v_i, v_j) - 1) \quad (2)$$

2.2 Image Processing

Image segmentation is the problem of partitioning an image into regions. Depending on the application, the segmentation algorithm can be based on various image features such as color, edges, texture and shape. In this case, the properties of the similarity matrices images are low frequency rectangular shape formations distorted by high frequencies noise, existence of off-diagonals and repetitions of various patterns. Based on these properties, we are going to exploit the existence of low/high frequencies, maintaining the edges, and try to detect homogenous regions of the similarity matrix image that relate to musical objects.

We consider the similarity matrix as an intensity image. The image segmentation algorithm applied on the similarity matrix image is based on [5] and is described as follow: Smoothing of the diagonal, anisotropic filtering, thresholding and morphological post-processing.

2.2.1 Pre-processing of the images

The elements on diagonal of a similarity matrix are $S(i, i) = 1$, providing no useful information for the scope of image segmentation. Thus, the diagonal is smoothed to avoid connecting the objects on it and regarding them as one unified object after the low pass filtering. This is realized by replacing the values on the diagonal of the similarity matrix with the average value of the six neighboring positions.

For maintaining the image's major features, such as edges and corners, Perona and Malik made a significant contribution in the area of noise filtering by proposing a nonlinear diffusion algorithm [13] . Anisotropic diffusion filtering provides smoothing of intra-region areas preferentially over inter-region areas, thereby providing a good prospective for removing unwanted noise and preserving edges. The basic idea behind anisotropic diffusion is to evolve a family of smoothed images $S_t = S(i, j, t)$ from

an initial noisy image $S_0 = S(i, j)$ converging to a solution of the partial differential equation

$$S_t = div(g(|\nabla S|)\nabla S) = g(|\nabla S|)\Delta S + \nabla g \cdot \nabla S \quad (3)$$

The diffusion coefficient function $g(|\nabla S|)$ is selected in this particular application to favor wide regions over smaller ones [13], thus it is selected as:

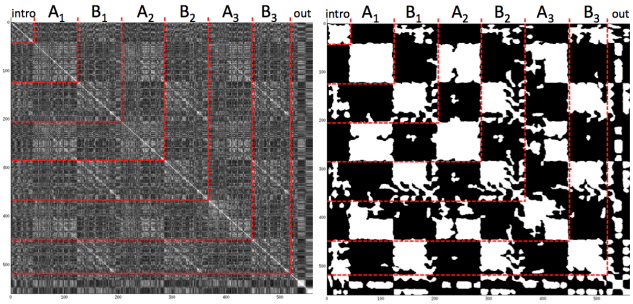
$$g(|\nabla S|) = \frac{1}{1 + (|\nabla S| / K)^2} \quad (4)$$

where *K* is a constant that controls the sensitivity of the algorithm to objects' edges.

2.2.2 Segmentation mask generation

After filtering, the similarity matrix image is binarized. The *Otsu* thresholding [8] is employed effectively, since the images have unimodal histograms. Following, morphological operations [14] are performed, for refining the binary masks. Morphologic processing considers close sets and exploits the geometric structure of them to remove small outlier regions (open operation) and closing small holes inside the segments (close operation). Fig. 1 shows an example of such a segmentation mask computed for the song *Help!* by *The Beatles*. Musical parts of the song (intro, A_1, B_1, A_2, B_2, A_3, B_3, outro) are also annotated on the matrices.

Elements of high similarity in the original matrix form compact regions in the mask. On the other hand, morphological operators are able to remove regions that mainly consisted in dissimilarity. Musical parts form therefore coherent and distinguishable objects in the segmentation mask. We were thus able to provide a much sparser image representation of the audio in which the musical structure is clearly enhanced.



(a) Original Matrix (b) Segmentation Mask

Figure 1: Original Similarity Matrix S (a) and the obtained segmentation mask M (b)

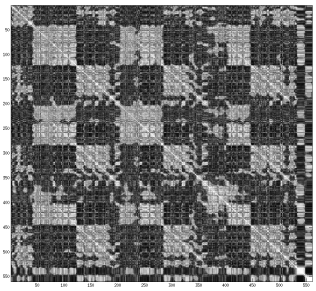


Figure 2: Enhanced similarity matrix S_e with mean of the segmentation mask

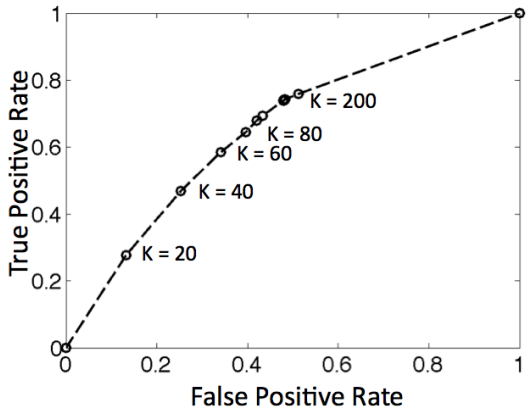
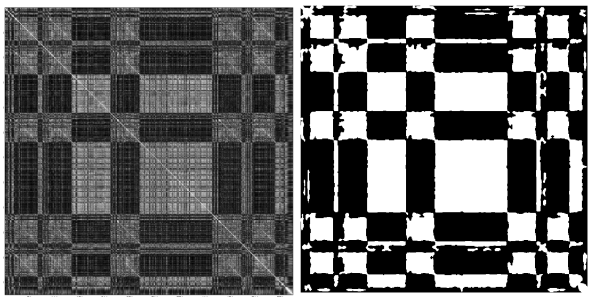


Figure 3: True positive rate vs false positive rate for different values of K

2.2.3 The Parameter K

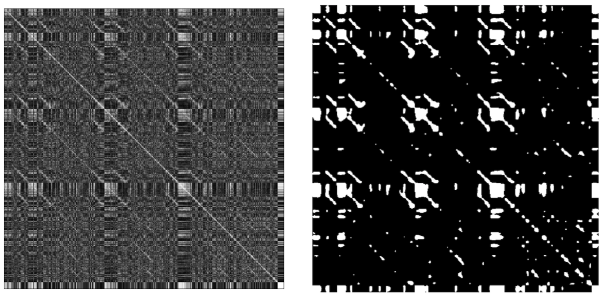
The parameter K controls the sensitivity of the algorithm to objects' edges and should thus be carefully chosen. We ran a small experiment on our dataset (see section 3.3) for evaluating the quality of the segmentation masks for different values of K. The groundtruth mask is generated with mean of the annotated audio segmentation. Each structural part is thus displayed as a block in the mask. The estimated and groundtruth masks are then compared using the true positive and false positive rates (see Figure 3).

In order to have a good compromise between the coverage of the structural parts (true positive rate) and oversegmentation (false positive rate) we set the value of K to 40. Indeed, groundtruth masks are generated from a rough annotation of the audio data and do not account for dissimilarity regions within structural parts. They thus shouldn't be taken as the ideal masks. However, our masks should contain as less energy as possible in non-structural regions. Setting K to 40 we are still able to cover 50% of the groundtruth masks while keeping the false positive rate under 30%.



(a) Original Matrix (b) Segmentation Mask

Figure 4: Similarity matrix for the song *Everybody is trying to be my baby* performed by *The Beatles* (a) and the corresponding segmentation mask (b)



(a) Original Matrix (b) Segmentation Mask

Figure 5: Similarity matrix for the song *Think for your self* by *The Beatles* (a) and the corresponding segmentation mask (b)

2.2.4 Matrix Enhancement

As shown above, the segmentation mask contains very relevant information regarding musical objects in the original image. It is therefore very tempting to directly segment the objects defined by the mask for further audio processing. However, the segmentation mask is binary and even though indicating meaningful regions, it loses information about the similarity of the structural part with itself. In other terms, temporal evolution of similarity within the parts is lost. The binarization can be prejudicial to the description of music pieces and to the final application. In order to maintain this temporal information, we consider the segmentation mask as a weighting matrix for the enhancement process. Elements in the original matrix $\mathbf{S}$ that were retained in the mask $\mathbf{M}$ are multiplied by a certain weight w, whereas unretained elements remain unchanged. In the resulting matrix $\mathbf{S}_e$, regions of musical interest are strengthened and the variation in similarity levels is kept.

$$\mathbf{S}_e = \mathbf{S} \cdot (w-1) \cdot (\mathbf{M} + \mathbf{O}_1) \tag{5}$$

where $\mathbf{S}_e$ and $\mathbf{M}$ are the enhanced similarity matrix and the mask respectively. $\mathbf{O}_1$ is a matrix of the same size as $\mathbf{S}$ whose elements all equal to one. It ensures that the original matrix information $\mathbf{S}$ will be retained. w is a scalar that weights the elements of the segmentation masks in the original similarity matrix.

Figure 2 shows the final enhanced matrix for our example using a weight of 3.

2.2.5 Interpretation for the task of structure segmentation

Among the proposed similarity matrix based structure analysis methods, two definitions of structure are distinguished in [11]: the state representation and the sequence representation. For the state representation, musical sections are assumed to be rather self-similar and therefore form blocks in the similarity matrix. It is the case for the music piece example shown in figures 1 and 2. And the matrix enhancement clearly strengthened the state representation. We show an other example of a well defined state representation in Figure 4. The song example is *Everybody is trying to be my baby* performed by *The Beatles*. Musical sections are very homogenous in timbre in the piece, thus yielding well-defined structural blocks in the similarity matrix. In that case the segmentation mask perfectly retains musical sections and discards regions of dissimilarity.

The sequence representation on the other hand considers series of frames that are repeated over the music piece, frames within a section not necessarily being similar. Structure is then displayed in the similarity matrix by dominant repetitive motives on the off-diagonals. For audio material that fits the sequence representation and as shown in Figure 5, our enhancement processing is not adequate yet. Indeed, while segments represented as states are retained in the segmentation mask, most of the sequence motives on the off-diagonals are discarded. Image segmentation techniques that fit the sequence representation should be considered in further developments.

3 Structural Segmentation

Structural segmentation of music pieces aims at extracting basic structural parts such as verse and chorus. Many of the proposed approaches for this task are based on a segmentation of similarity matrices [9]. In order to evaluate the pertinence of our matrix enhancement method, we will compare in this section the performance of the structure detection algorithm described in [4], using standard similarity matrices and enhanced matrices.

3.1 Approach

The structure segmentation proposed in [4] detects musical sections in the similarity matrix by its factorization with the Non-negative Matrix Factorization (NMF, [6]) technique. Authors show that if the musical structure is displayed as a state representation in the similarity matrix, musical sections can easily be modeled and classified over the dimension of such a factorization. We therefore hope that our enhancement approach, by strengthening the state representation, will also improve the performance of the structure segmentation.

To illustrate the NMF-based structure segmentation approach, we consider the similarity matrix $\mathbf{S}$ $(n \times n)$ for the song *Help* by *The Beatles*. After its NMF decomposition of rank r, $\mathbf{S}$ can be written as:

$$\mathbf{S} \approx \mathbf{WH} \tag{6}$$

with $\mathbf{W}$ $(n \times r)$ and $\mathbf{H}$ $(r \times n)$ the two non-negative matrix factors that best estimate $\mathbf{S}$.

Each element s_{ij} of $\mathbf{S}$ can be written as:

$$s_{ij} \approx \sum_{k=1}^{r} \mathbf{W}(i,k)\mathbf{H}(k,j) \tag{7}$$

To show how structural parts can be discriminated with such a decomposition, we perform the factorization of $\mathbf{S}$, and set the rank of decomposition to 2, the music piece having two main structural parts. We can display the contribution of each dimension of the factorization considering the matrices $\mathbf{D}_k$:

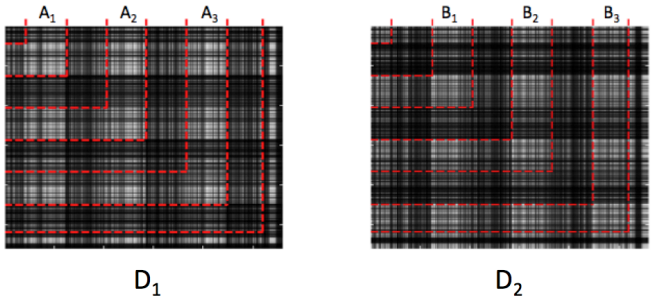
$$\mathbf{D}_k = \mathbf{W}(:,k)\mathbf{H}(k,:) \tag{8}$$

$\mathbf{D}_1$ and $\mathbf{D}_2$ for our song example are shown in Figure 6, using the original similarity matrix in (a), and the enhanced matrix in (b). In order to show the correlation between the NMF dimensions and the structural parts, we also annotated the segments A and B that are repeated over the song.

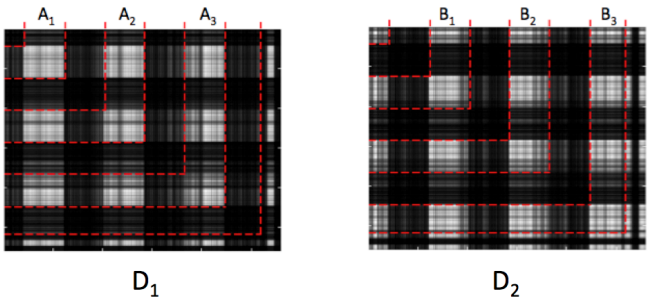
Decomposition of the original matrix yields a reasonable separation of the two parts. Nevertheless, the separation is not complete, and both dimensions of the NMF contain energy in regions corresponding to the two parts. Decomposing the enhanced matrix yields a much sparser and preciser separation of the structural parts. This is of great help for the structure detection step. Indeed, the sparser the decomposition is, the preciser is the structure clustering.

3.2 Structure explanation

Boundaries between musical sections are first retrieved in the similarity matrix with mean of the audio novelty score [3]. Similarity between potential musical sections is then measured in the NMF decomposed matrices and a hierarchical clustering is applied to merge segments belonging to the same section together.



(a) 2^{nd} order NMF Decompositon of the original Matrix



(b) 2^{nd} order NMF Decompositon of the Enhanced Matrix

Figure 6: Separation of structural parts of the song *Help* by *The Beatles* with mean of the NMF decomposition of the original matrix (a), and of the enhanced matrix (b)

3.3 Evaluation

For performance evaluation of the proposed algorithm, we considered the album *Help!* by *The Beatles* (14 songs) and its annotation in the *TUT Beatles*[1] dataset. For each song, the corresponding similarity matrix $\mathbf{S}$, segmentation mask $\mathbf{M}$ and enhanced similarity matrix $\mathbf{S}_e$ are computed.

We compare the performance of the structure analysis using the reference algorithm [4] on the original similarity matrices (*Reference*), the structure segmentation incorporating binary segmentation mask only (*Proposed 1*) and the structure segmentation algorithm using the enhanced matrices obtained with different weights (*Proposed 2*). The pairwise F-measure (F), Precision (P) and Recall (R) are employed for the performance evaluation. The mean F, P and R for all 14 songs for the various scenarios are reported in table 1.

Introducing enhanced similarity matrices clearly improved the over-all performance of the structure detection,

[1] http://www.cs.tut.fi/sgn/arg/paulus/structure.html

Algorithm	F	P	R
Reference [4]	63.5%	**64.7%**	65.7%
Proposed 1	64.8%	62.5%	69.5%
Proposed 2 ($w = 2$)	65.0%	61.7%	72.3%
Proposed 2 ($w = 3$)	**66.9%**	63.4%	**74.0%**
Proposed 2 ($w = 4$)	64.3%	62.1%	71.1%
Proposed 2 ($w = 5$)	63.8%	62.0%	69.9%

Table 1: Performance of the proposed structure detection algorithms against the state-of-the-art algorithm described in [2].

gaining up to 3.4% in terms of F-measure. While the precision is not increased but remain in the same range as the reference, we achieve much better recall rates (up to 8.3 %) in any of the proposed cases. This results in increasing the F-measure, which is the harmonic median of precision and recall.

The strong increase in the recall rates means that our proposed algorithm deals better with over-segmentation issues. Indeed, change of instrumentation within structural parts often leads to a division of parts in a set of sub-segments. As structure can be explained at several hierarchical levels, this does not affect the general quality of the estimated structure as long as sub-segments are affected relevant labels. The results also confirm that it is not worth using the binary segmentation mask alone for the analysis.

4 Conclusion and Perspectives

In this paper we have presented an image-oriented preprocessing algorithm for audio similarity matrices enhancement. By strengthening structural information in the original matrices, we reduce the complexity of the structure visualization, and the discrimination of musical sections is improved. Evaluation shows that this approach consistently improves the performances of the music structure segmentation. Especially, the enhanced matrices seem to cope with over-segmentation issues. The enhancement procedure is however more appropriate for music pieces that fit the state representation of structure. In further work, segmentation of sequence representations of structure should be included in the framework. Evaluation over larger and more content diverse databases will also be conducted. We believe that audio visualization and its applications can strongly benefit from an image-oriented analysis. Therefore, further image analysis based on texture or shape features should be applied to similarity matrices in order to build robust representations of audio signals. We will also consider to have an hybrid approach and enhance either the state or sequence representations with further modeling of audio features before the computation of the similarity matrix, and complete the enhancement in our image processing framework.

5 Acknowledgment

This work was supported by the European Commission under contract FP7-21644 PetaMedia.

References

[1] M. J. Bruderer, M. F. McKinney, and A. Kohlrausch. Structural boundary perception in popular music. In *ISMIR*, pages 198–201, 2006.

[2] J. Foote. Visualizing music and audio using self-similarity. In *ACM Multimedia (1)*, pages 77–80, 1999.

[3] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo (I)*, page 452, 2000.

[4] F. Kaiser and T. Sikora. Music structure discovery in popular music using non-negative matrix factorization. In *ISMIR*, 2010.

[5] A. Krutz, M. Kunter, M. Mandal, M. Frater, and T. Sikora. Motion-based object segmentation using sprites and anisotropic diffusion. In *WIAMIS*, 2007.

[6] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, oct 1999.

[7] M. Mueller and F. Kurth. Enhancing similarity matrices for music audio analysis. In *in Proc. IEEE ICASSP*, 2006.

[8] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62 –66, 1979.

[9] J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. In *ISMIR*, 2010.

[10] G. Peeters. Toward automatic music audio summary generation from signal analysis. In *Proc. ISMIR*, pages 94–100, 2002.

[11] G. Peeters. Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach. In U. K. Wiil, editor, *CMMR*, volume 2771 of *Lecture Notes in Computer Science*, pages 143–166. Springer, 2003.

[12] G. Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *ISMIR*, 2007.

[13] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629 –639, July 1990.

[14] P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag New York, Inc., 2003.