Comprehensive Appraisals of Contemporary Documents

Rob Kooper and Peter Bajcsy
National Center for Supercomputing Applications,
University of Illinois at Urbana-Champaign,
Urbana, Illinois, USA
{kooper,pbajcsy}@ncsa.illinois.edu

Abstract

This paper describes problems related to contemporary document analyses. Contemporary documents contain multiple digital objects of different type. These digital objects have to be extracted from document containers, represented as data structures, and described by features suitable for comparing digital objects. In many archival and machine learning applications, documents are compared by using multiple metrics, checked for integrity and authenticity, and grouped based on similarity. The objective of our work is to design methodologies for contemporary document processing, visual exploration, grouping and integrity verification.

1. Introduction

The objective of our work is to design a methodology, algorithms and a framework for document appraisal by (a) enabling exploratory document analyses and integrity/authenticity verification, (b) supporting automation of some analyses and (c) evaluating computational and storage requirements for archival purposes. In order to address the aforementioned criteria, our approach has been to decompose the series of appraisal criteria into a set of focused analyses, such as (a) find groups of records with similar content, (b) rank records according to their creation/last modification time and digital volume, (c) detect inconsistency between ranking and content within a group of records, and (d) compare sampling strategies for preservation of records.

In this work, we had chosen a specific class of electronic documents that (a) correspond to information content found in scientific publications about medical topics, (b) have an incremental nature of their content in time, and (c) contain the types of information representation that are prevalent in contemporary medical environments. We selected to work with PDF documents since PDF is an open file format, and the open nature of the file format is critical for automated electronic document appraisal and long term preservation.

In order to address the appraisal criteria [2], we adopted some of the text comparison metrics used in [9], image comparison metrics used in [11] and lessons learnt stated in [8]. Then, we designed a new methodology for grouping electronic documents based on their content similarity (text, image and vector graphics), and prototyped a solution supporting grouping, ranking and integrity verification of any PDF files and HTML files [6]. First, text based, vector based and multi-image based comparisons are performed separately. The current prototype is based on color histogram comparison, line count in vector graphics and word frequency comparison. The image colors and word/integers/floating numbers can be analyzed visually to support exploratory analyses. Subsets of the undesirable text and image primitives could be filtered out from document comparisons (e.g., omitting conjunctions, or background colors). The results of text, image and vector based comparisons are fused to create a pair-wise document similarity score. The matrix of pair-wise document similarity scores are used for grouping. The other appraisal criteria are approached by ranking documents within a group of documents based either on time stamps or on file name indicating the version number. The inconsistency between ranking and content within a group of records is based on frequency tracking, where the frequency of text, image and vector primitives is monitored over the time/version dimension of the grouped documents.

Currently, we hypothesized that the correct temporal ranking correlates with the content (images, vector and text) in such a way that the content is being modified without sharp discontinuities. Sharp content discontinuities are perceived as significant changes of document descriptors that would correspond, for instance, to large text/image deletions followed by large text/image additions or large text/image additions followed by large text/image deletions. We have experimented with real PDF documents of journal papers to validate the above hypothesis.

The novelty of our work is in designing a methodology for computer-assisted appraisal, in developing and pro-

totyping a mathematical framework for automation of appraisals based on image, vector graphics and text types of information representation.

2. Previous work

Related work to the proposed framework: Our work is related to the past work of authors in the area of digital libraries [9], content-based image retrieval [11] and appraisal studies [8]. For example, the authors of [7] analyze PDF document by examining the appearance and geometric position of text and image blocks distributed over an entire document. However, they did not use the actual image and vector graphics information in their analyses. Similarly, the focus in [1] is only on the logical structure of PDF documents but not the content. The work in [3] and [10] is based on analyses of vector graphics objects only since it is focused on diagrams represented by a set of statistics, e.g., the number of horizontal lines and vertical lines. Other authors also focused only on chart images using a model-based approach [5]. There is currently no method that would provide a comprehensive content-based PDF comparison and appraisal strategy according to our knowledge. In order to prototype a solution for comprehensive document comparisons and clustering, the difficulties lie in dealing with vector graphics objects, fusion of similarities of multiple digital objects, and in providing a scalable solution with the increasing number of documents.

Related work on comparing vector graphics objects: Vector graphics objects are described by the most primitive feature of the graphics object, lines, are practically useless in meaningful comparisons. Due to this, it is necessary to compare objects at a slightly more sophisticated level by comparing groups of lines and their relationship to each other. However, the manner in which this can be done varies, and many techniques for comparison of simple geometric shapes exist, making it not trivial to choose which graphic comparison to use. Veltkamp [14] showed ten distinct methods for the comparison of polygons and open paths, and it is assumed that more may exist. However, the principle difficulty in implementing almost all of these methods is that they rely on a direct, side-by-side computationally expensive comparison between two graphical objects resulting in a real number comparison value in the range between 0 and 1. In addition, the problem of formulating a metric measuring approximate similarity between visual objects has also been known as an open problem [12, 4]. Finally, there is the problem of the sequential arrangement of the line segments since they could be encoded in the document arbitrarily. This introduces a plethora of problems because there is no guarantee that a visually similar object will be encoded in a document such as an Adobe PDF file in anything like the same way.

3. Methodology

This section presents the methodology and theoretical framework for addressing grouping, ranking and integrity verification problems (see figure 1).

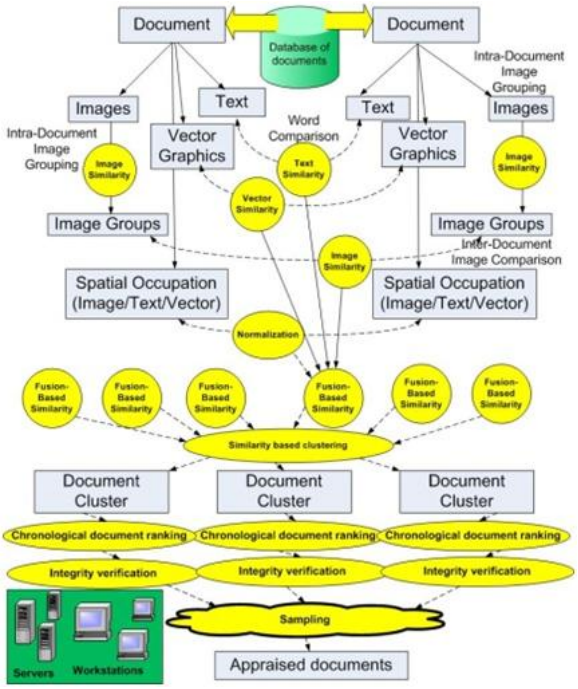

Figure 1. An overview of the document appraisal framework based on comprehensive comparisons of document's internal digital objects.

3.1. Overview

The designed methodology consists of the following main steps: (1) Extract components and properties stored in PDF files/containers. (2) Define text, image and vector graphics primitives, and extract their characteristic features. (3) Group images within each document into clusters based on a pair-wise similarity of image primitives and a clustering similarity threshold. (4) Compute a pair-wise similarity of image clusters across two documents based on their corresponding features. (5) Compute a pair-wise similarity of text & vector graphics primitives across two documents. (6) Calculate fusion coefficients per document to weight the contribution of text-based, image-based and vector-based similarities to the final pair-wise document similarity score. (7) Repeat steps (4-6) for all pairs of documents. (8) Group documents into clusters based on the pair-wise document

similarity score and a selected similarity threshold. (9) Assign ranks to all documents based on their time stamps and storage file size. (10) Calculate the second difference of the document characteristic features over time and file size dimensions. Report those documents for which the second difference exceeds a given threshold defining allowed discontinuities in content.

3.2. Theoretical Framework

Document grouping problem. Given a set of documents, $\{D_i\}; i = 1, 2, , N$ compute pair-wise similarity of documents $sim(D_i, D_j)$ and aggregate them into clusters based on the similarity values for further ranking within each cluster. The similarity of documents is understood as the combined similarity of document components. In our case, it would be the similarity of text, vector and raster (image) graphics components. The three components are decomposed into multiple images I_{ik} and their image primitives e_m^I, vector graphics and their primitives e_m^V, and text primitives e_m^T in textual portions $T_{ik} = T_i$ of a document D_i. The similarity for each component type is derived either directly using the features of its primitives (the case of text) or average features of multiple components of the same type and their primitives (the case of images and vector graphics).

Calculations of Statistical Features. The text feature for the word primitives is the frequency of occurrence of each unique word. The image feature for the color primitive is the frequency of occurrence of each unique color (also denoted a one-dimensional color histogram). The vector graphics feature is the frequency of occurrence of lines forming each vector graphics. The frequency of occurrence provides a statistical estimate of the probability distribution of primitives.

Calculation of Document Similarity. Given two PDF documents D_i, D_j the similarity is defined as a linear combination of the similarities of the document components. In our case, the formula contains only the text and raster graphics components.

$$\begin{gathered} w_T * sim(T_i, T_j) + \\ sim(D_i, D_j) = w_I * sim(\{I_{ik}\}_{k=1}^{K}, \{I_{jl}\}_{l=1}^{L}) + \\ w_V * sim(V_i, V_j) \end{gathered} \quad (1)$$

where w_T, w_I, w_V are the weighting coefficients.

We have derived the weighting coefficients from the spatial coverage ratio of images, vector graphics and text in two compared documents. The weight assignment could be viewed as the relevance (the weight) of each PDF component according to the amount of space it occupies in a document. The motivation is based on a typical construction of documents where the space devoted to a textual description or an illustration reflects its importance and hence should be considered in the similarity calculation. For example the weights for images are calculated as

$$w_I(D_i, D_j) = \frac{R_I(D_i) + R_I(D_j)}{2}$$

$$w_I(D_i, D_j) + w_V(D_i, D_j) + w_T(D_i, D_j) = 1$$

where

$$R_I(D_i) = \frac{Area_I(D_i)}{Area_I(D_i) + Area_V(D_i) + Area_T(D_i)}$$

$$R_I(D_i) + R_V(D_i) + R_T(D_i) = 1$$

Calculation of Text Similarity. The similarity of text components from two documents is computed using the text features and similarity metric defined according to [9]. The equation is provided below.

$$sim(T_i, T_j) = \sum\nolimits_{k1,k2} \omega_{i,k1}\omega_{j,k2} \quad (2)$$

where $k1, k2$ are those indices of text primitives that occur in both documents (in other words, there exist $e_{i,k1} = e_{j,k2}; e_{i,k1} \in T_i; e_{j,k2} \in T_j$. The ω terms are the weights of text primitives computed according to the equation below.

$$\omega_{ik} = \frac{f_{ik} \log(\frac{N}{n_k})}{\sqrt{\sum_{l=1}^{L} (f_{il})^2 (\log(\frac{N}{n_l}))^2}} \quad (3)$$

where f_{ik} is the frequency of occurrence of a word e_k in D_i, N is the number of documents being evaluated, L is the number of all unique text primitives (words) in both documents, and n_k is the number of documents in that contain the word $e_k (n_k = 1 or 2)$.

Calculation of Raster Graphics (Image) Similarity. In contrary to text that is viewed as one whole component, there are multiple instances of raster graphics components (images) in one document. Thus, the similarity of image components in two documents is really a similarity of two sets of images.

Due to the fact that many documents contain images that are sub-areas or slightly enhanced versions of other images in the same document, we have observed biases in image-based document similarity if all possible image pairs from two documents are evaluated individually and then the average similarity would be computed. In order to avoid such biases, we approached the similarity calculation by first computing a pair-wise similarity of all images within each document and clustering them. Next, the pair-wise similarity of clusters of images from each document is computed using the average features of clusters.

A. Intra-document image similarity: The similarity of two raster graphics (image) components from one document $sim(I_{ik} \in D_i, I_{il} \in D_i)$ is computed using the one-dimensional color histogram feature and the same similarity metric as defined before for text according to [9].

B. Inter-document image similarity: The similarity of two sets of raster graphics (image) components, one from each document is computed similarly to the intra-document similarity.

Calculation of Vector Graphics Similarity. The methodology used in the text and raster image comparison is based on a statistical comparison technique for comparing large numbers of small information units of precisely defined structure, such as words or colors, which are reused multiple times to construct the document. In the case of vector graphics however, the primitive features of the graphics object, lines, cannot be simply compared and a more sophisticated level by comparing groups of lines and their relationship to each other is needed.

Several comparison techniques [14, 4, 13] exist for the comparison of polygons and open paths. However, all of these methods rely on a direct, side-by-side computationally expensive comparison between two graphical objects resulting in a value between 0 and 1. Our methodology relies on counting exactly reproducible informational units and these techniques don't fit in our methodology. Instead we are looking for a characterization of the shape of the graphic, which can then be easily compared with other shape characterizations.

We developed an algorithm which binned the angle of each line in any series of connected lines in the document and recorded the binned angles. The angles could be grouped into unordered histograms or stored in the order they are rendered. The angular bin size is also variable. Tests presented in section 4.1 were used to find an optimal method and number of bins. The chosen methodology was unordered grouping with 9 angular bins. Using a textual encoding of the angular description, the algorithm used for text comparison can be used on the vector graphic data.

The angle used in the similarity metric can be defined as the angle and length of the line with respect to the page in which it is drawn. This does not allow for rotational and scale invariance of a similarity metric for comparing vector graphics objects. To ensure rotational invariance, there must be a common definition of the basis line against which all the angles can be computed. Since the main feature of the shape is the location of the vertices, we use the average of all the vertices in the shape and find a center point about which all reference lines can be drawn. The reference line for each angle computation is then drawn through the center point and through the center of the line whose angle is being computed. These two points will always be in the same relation to the line whose angle is being measured despite rotational variance (see figure 2 the red point is the center, the blue point is the center of the line and ? is the rotational invariate angle used)

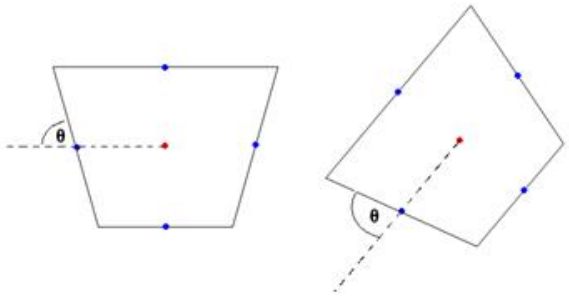

Figure 2. Rotational invariance of the similarity metric designed for comparing vector graphics objects.

4. Experimental Results

We report experimental results that illustrate the choice of vector graphics comparison metrics, the accuracies of comprehensive document comparisons based on all contained digital objects, and the prototype approach to document integrity verification.

4.1. Evaluations of Vector Graphics Comparisons

In the comparison of images it is standard to ensure that the comparison metric is invariant under translation, rotation, and rescaling. Because the method is based on the angle between lines it stands to reason that there should be no variation under translational variation. Also, as long as rescaling does not distort the image's aspect ratio it should also not make a difference in comparison. To test this, the three documents appearing in figure 3 were compared.

Note that the image of the computer in the document on the left has been both translated and rescaled in the center document. The document on the right contains an unrelated image of a CD-ROM drive. The result of a comparison of these three images is 100% similarity between documents 1 and 2 and 0% similarity between documents 1 and 3 and documents 2 and 3. Therefore the comparison is both translation and scale invariant.

We performed a similar experiment where we rotated the images. Using the unordered implementation with 9 bins without the rotationally invariant modification, the similarity between the rotated documents is only 0.18, while the similarity between unrelated documents is 0. Using the modification to allow rotational invariance brings the similarity between the rotated documents to 1, while the similarity between the unrelated documents remains at 0. This process was repeated for rotations of $25°$, $50°$, $75°$, and $180°$. The $180°$ rotation also showed a similarity of 1, but when the graphic was rotated by the intermediate angles, the sim-

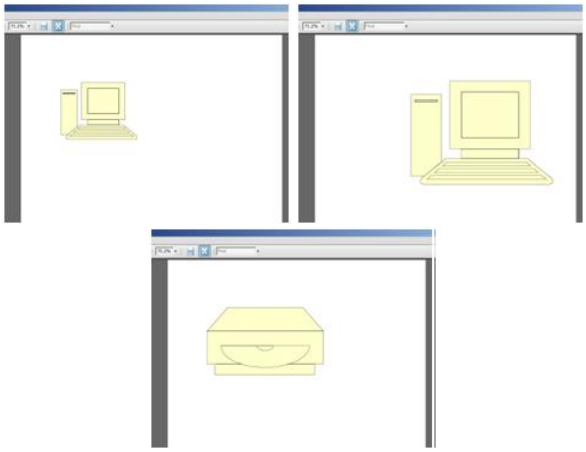

Figure 3. Test images to illustrate invariance properties of vector graphics comparisons.

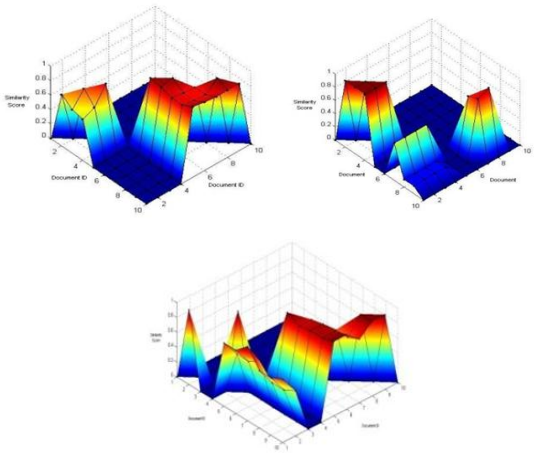

Figure 4. Showing the three similarity comparisons (image, vector graphs and words).

ilarity dropped to 0.72. However, the rotations at intermediate angles all showed similarity 1 between other rotations at intermediate angles. This indicates that during partial rotations some line segments may become distorted in the PDF rendering, but this type of problem cannot be rectified because the actual shape is being distorted and there is no way to recover it.

4.2. Evaluations of Comprehensive Document Comparisons and Clustering Results

Using the presented framework, we appraised sets of PDF documents generated during scientific medical journal preparations. The document sample set consisted of 10 documents 4 of which were modified versions of one article and 6 were of a different though related article. The pair-wise comparisons of the features are presented in the following graphs.

The z values of the graphs in figure 4 represent the similarities between the documents identified with the numbers by the x and y axes. The word similarity comparison clearly shows that the documents 5 through 10 score highly in comparison with each other while they score poorly in comparison to the documents 1 through 4. Likewise, documents 1 through 4 show higher scores for comparison within the group. The vector graphics of documents 1 through 4 are nearly identical while in the second subgroup only documents 7 through 9 are conclusively linked. The raster images within the two subgroups of the documents shows high similarities for the documents 5 through 10 score but the rest are not obvious how they are related.

Throughout the documents the relative apportionment of visible space of the three document features varies. Figure 5 shows the fraction of each document covered by words, images and vector graphics.

Combining the three comparison techniques with weights allotted by the proportion of coverage of the feature represented by that comparison allows for a final similarity score to be established. Figure 6 displays the final comparison matrix, which clearly distinguishes the two subgroups of the original document set.

4.3. Results of Document Integrity Verification

With the documents adequately grouped and ordered by PDF timestamp, the verification process looks for conspicuous editing habits from one document to the next. Successful document order verification relies on multiple failures of the tests displayed in figure 7. The Integrity Tests are aligned from top to bottom, currently we support the following tests: (1) appearance or disappearance of document images, (2) appearance and disappearance of dates appearing in documents, (3) file size, (4) image count, (5) number of sentence, and (6) average value of dates found in document.

5. Conclusions

We have described a framework for addressing document appraisal criteria. The framework consists of feature extraction from multiple digital objects contained in contemporary documents, pair-wise similarity metrics for comparing digital objects, a comprehensive content-based grouping of documents, ranking based on temporal or file size attributes

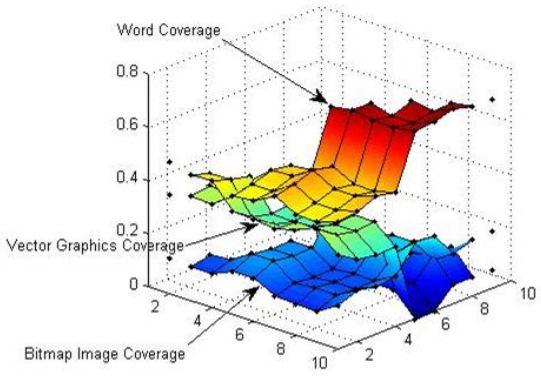

Figure 5. Area covered by words, images and vector graphics per document.

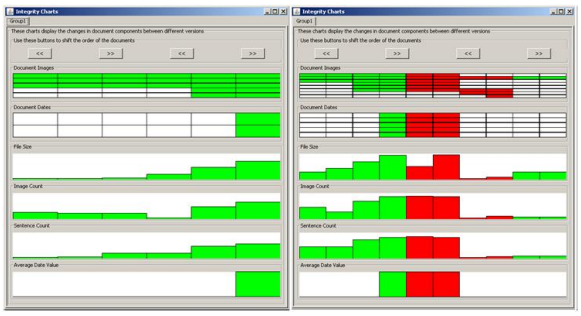

Figure 7. Verification of the Document Ordering Based Upon the Time Stamps of PDF Documents. Left image shows all integrity tests passing, and on the right it shows tests are failing.

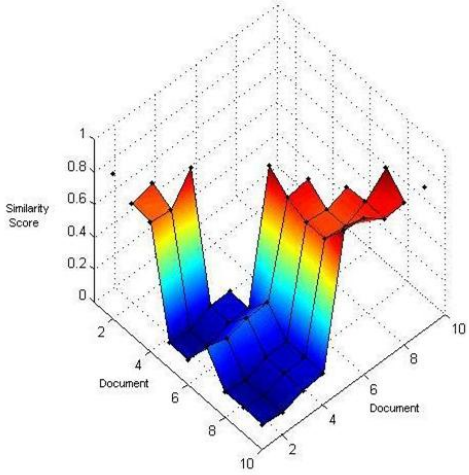

Figure 6. Final similarity score.

and verification of document integrity, as well as the parallel algorithms supporting applications with large volumes of documents and computationally intensive processing. Although we selected to work with documents in PDF format, the framework is applicable to any file format as long as the information can be loaded from any proprietary file format. In future, we will be exploring other hypotheses to increase the likelihood of detecting inconsistencies and understanding the high-performance computing requirements of computer-assisted appraisal of electronic records.

6. Acknowledgment

This research was partially supported by a National Archive and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archive and Records Administration, or the U.S. government.

References

[1] A. Anjewierden. AIDAS: Incremental logical structure discovery in PDF documents. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 374–378, 2001.

[2] P. Bajcsy and S.-C. Lee. Computer assisted appraisal of contemporary pdf documents. In W. V. Oz and M. Yannakakis, editors, *ARCHIVES 2008: Archival R/Evolution & Identities, 72nd Annual Meeting Pre-conference Programs*, San Francisco, CA, 2008.

[3] R. Futrelle, M. Shao, C. Cieslik, and A. Grimes. Extraction, layout analysis and classification of diagrams in PDF documents. In *ICDAR 2003 (Intl. Conf. Document Analysis & Recognition*, pages 1007–1014, 2003.

[4] L. Hess and B. Mayoh. Graphics and the Understanding of Perceptual Mechanisms: Analogies and Similarities. *Geometric Modeling and Imaging–New Trends, 2006*, pages 107–112, 2006.

[5] W. Huang, C. Tan, and W. Leow. Model-based chart image recognition. *Lecture notes in computer science*, pages 87–99, 2004.

[6] S. Lee, W. McFadden, and P. Bajcsy. Text, image and vector graphics based appraisal of contemporary documents. In *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications-Volume 00*, pages 729–734. IEEE Computer Society Washington, DC, USA, 2008.

[7] W. Lovegrove and D. Brailsford. Document analysis of PDF files: methods, results and implications. *Elecctronic Publishing*, 8:207–220, 1995.

[8] J. Marshall. *Accounting for disposition: A comparative case study of appraisal documentation at the National Archives and Records Administration in the United States, Library and Archives Canada, and the National Archives of Australia*. PhD thesis, University of Pittsburgh, 2006.

[9] G. Salton, J. Allan, and C. Buckley. Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2):97–108, 1994.

[10] M. Shao and R. Futrelle. Recognition and classification of figures in PDF documents. *Lecture Notes in Computer Science*, 3926:231, 2006.

[11] D. Squire, W. Müller, H. Müller, and T. Pun. Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters*, 21(13-14):1193–1198, 2000.

[12] M. Tanase and R. Veltkamp. Part-based shape retrieval. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 543–546. ACM New York, NY, USA, 2005.

[13] R. Veltkamp. Shape matching: similarity measures and algorithms. In *Shape Modeling and Applications, SMI 2001 International Conference on.*, pages 188–197, 2001.

[14] R. Veltkamp and L. Latecki. Properties and performance of shape similarity measures. In *Proceedings of IFCS 2006 Conference: Data Science and Classification*. Springer, 2006.