TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/12 pp1-16
Volume 17, Number 4, August 2012

Sequential Document Visualization Based on Hierarchical Parametric Histogram Curves*

Haidong Chen, Guizhen Wang, Dichao Peng**, Wuheng Zuo†, Wei Chen

State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China;
† **Zhijiang College of Zhejiang University of Technology, Hangzhou 310024, China**

Abstract: Recently, sequential document visualization has attracted much attention for its superior capability in depicting the sequential semantic progression in a single document. However, existing methods commonly take abstractive visual forms such as texts, numbers, and glyphs, and require much user expertise for document exploration. In this paper we propose a sequential visualization to represent a single document with a two-dimensional picture-based storyline, which semantically enhances the comprehension of textual information. We introduce a new parametric modeling approach called the Hierarchical Parametric Histogram Curve (HPHC), which encodes the statistical progression locally and adaptively. By transforming an HPHC into the two-dimensional space with a new locality-preserving embedding algorithm, we create a mapping from points along the curve to descriptive pictures and generate the visualization result. The new representation expresses the primary content with a graphical form, and allows for efficient multi-resolution and focus+context exploration in a long document. Our approach compares favorably with previous work in that it is more intuitive and requires less user expertise. Informal evaluation shows that it is useful in quick document browsing, communication, and understanding, especially for people with low literacy skills.

Key words: document visualization; sequential document visualization; pictorial communication; histogram curve

Introduction

Document visualization has been an important research topic in information visualization[1−6]. While much of the effort has been devoted to visualizing large document corpus, it is of equal, if not more, importance to present visual summaries of a single (large) document. Different from most multi-document visualization approaches which focus on building and exploring document manifolds with various inter-document similarity metrics, the within-document visualization requires a detailed representation and reformulation to the underlying document components. State-of-the-art approaches[5,7,8] represent text units such as a word, a sentence, and a paragraph with visual abstractions such as a glyph, a knot in a graph, and a curve, respectively. Studies have shown the effectiveness of these methods towards revealing structural and hierarchical information for a document.

Inspired by the fact that a traditional document is written and read orderly like an evolving story, the

Received: 2012-05-01 ; Accepted: 2012-06-13
∗ Supported by the National High-Tech Research and Development (863) Program of China (No. 2012AA120903); the National Natural Science Foundation of China (No. 61003193); Commonweal Project of Science and Technology Department of Zhejiang Province (No. 2011C21058)
∗∗ To whom correspondence should be addressed.
E-mail: pengdichao@cad.zju.edu.cn Tel: 86-15157180966

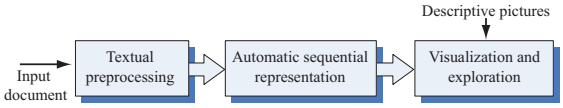


Fig. 1 Three main components of our approach

technique of sequential document visualization[9] has been developed to exploit the inherent sequentiality of a document. It can greatly facilitate document segmentation, topic extraction, and information retrieval. For tasks that demand efficient semantic presence like quick document browsing and understanding, however, current solutions with limited basic representations such as numbers, texts, and glyphs may become inefficient. This is mainly because the capability to capture and visualize the document semantics is limited by the textual or glyph-based visual forms. Users have to be very careful to align the visualization curve and the texts, and to interact with the digital numbers in the text panel that indicate the positions corresponding to the points on the curve. This may be extremely undesirable for specific users such as children or one with less literacy skills. Furthermore, mastering those tools typically requires a fair amount of expertise on the background knowledge behind the visualization, such as the PCA analysis, and thus involves a steep learning curve.

The difficulty of understanding the information in a large amount of texts may be addressed by taking advantage of the power of human perception. An intuitive method is to exploit the visual expressiveness of pictures to reveal thematic patterns and topic transitions in a document. With this method, the sequential process of reading a document can be enhanced by a parallel perceptual process that benefits greatly from complex human visual system[6,10].

In this paper, we propose a new multi-resolution and sequential document modeling approach, which is equipped with a picture-based visual representation for generating effective visualization of a single (long) document. Our approach consists of three components: a textual preprocessing module; a sequential, parametric, and multi-resolution representation; visualization and interactive exploration. Figure 1 shows the flowchart of the visual exploration system. The novelty of our work is a new sequential document representation and its reformulation to a two-dimensional representation in conjunction with the employment of a pictorial communication mode.

The resulting visualization favorably lowers the obstacles to the understanding of the users and shortens the user exploration time by a great extent. A snapshot of the visualization of a section in the book *My Life*[11] is shown in Figure 2.

To summarize the contribution of this paper, we overcome the limitation of the previous sequential document visualization scheme[9] with a more informative and perceivable two-dimensional picture-based representation. The suite of techniques presented in this paper, including the multi-scale content summarization and focus+context visualization, can also be seamlessly incorporated with a new parametric document representation. And finally, our system allows users to navigate the document in a set of creative modes with intuitive controls. Informal evaluation shows that our system is useful in quick document browsing, communication, and understanding, especially for people with low literacy skills.

1 Related Work

Document visualization. A rough classification for document visualization techniques would be the ones that visualize a document corpus and the ones that handle a single document. For the visualization of a document collection, researchers[1,2,4] seek to map the collection into a multivariate space, explore, and present the connections and relevance among documents. Often, a text unit (a sentence, a paragraph, or a document) is encoded with a knot in a graph or a glyph. Representative techniques and systems include Gist Icons[7], DocuBurst[8], Phrase Nets[5], Topic Island[12], IN-SPIRE[13], Omniviz and Refviz[14].

On the other hand, visualizing a single document reduces the scope to the texts in a document. Concerning the visualization, many efforts[15] have been made in text pattern mining, presentation, and spatialization. To enable the sequential analysis, the similarity among different segments of a document is judged. The approach presented in Ref. [9] represents a categorical sequence as a smooth curve in the histogram space, and makes use of a set of differential geometry and smooth analysis to detect thematic transition. Due to the involvement of several local statistical modeling techniques, the designed user interface is not well suited for people with low literacy skills.

Pictorial communication. Communication through pictures eases user recognition by avoiding lexical analysis and text understanding[10]. The idea that converts

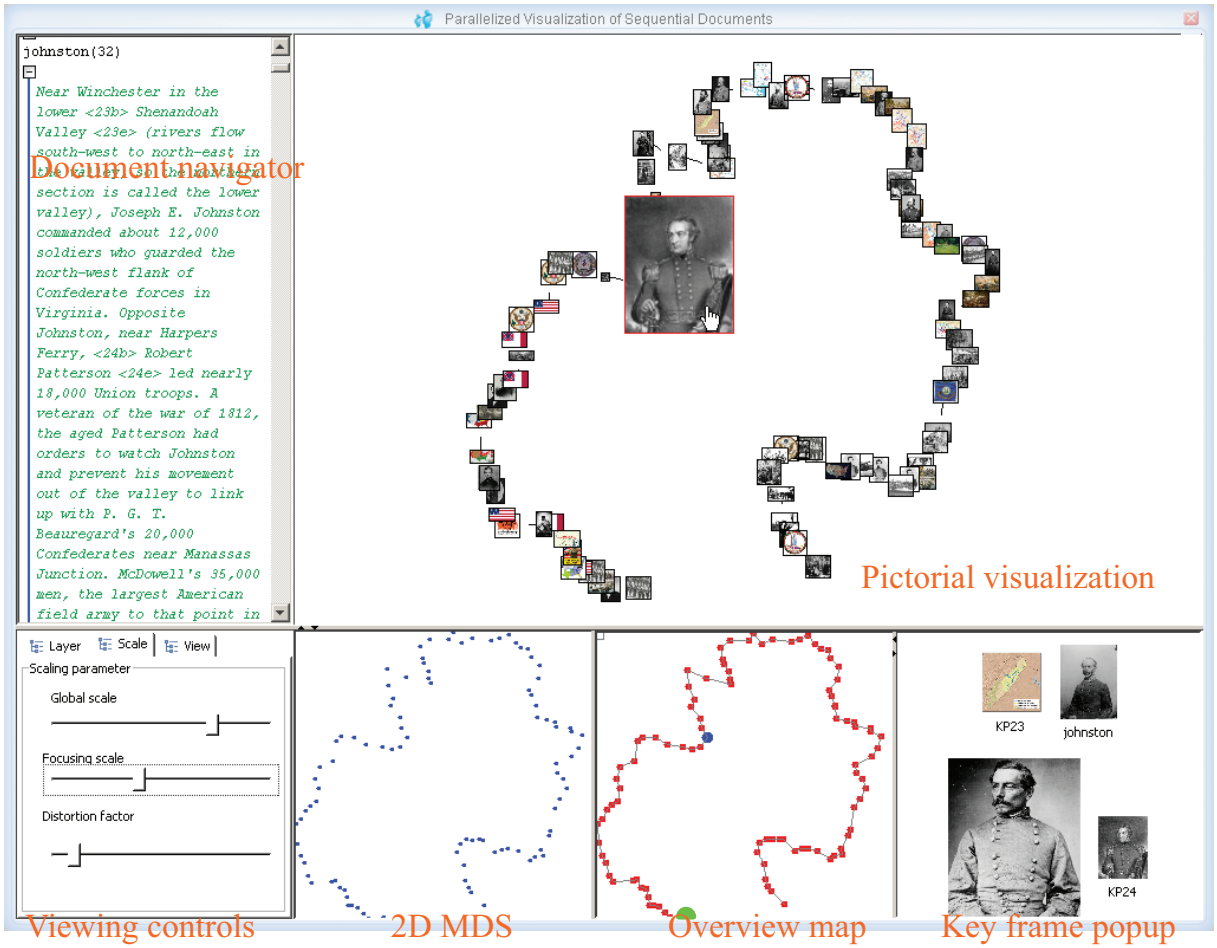


Fig. 2 Visualization of a single long document with the start point highlighted in green and active point in blue. The layout of the user interface consists of a pictorial visualization window, a text navigator, a two-dimensional MDS view, an overview map, a key frame popup window, and a set of interactive viewing controls.

text to a perceivable scene is prevailing in many fields. For example, the WordsEye system[16] assembles a set of pre-defined models from token sets to build a three-dimensional world. Obviously, its focus is the spatial organization of the built scene. By leveraging content-based image analysis techniques, an intelligent text illustration system[17] was generated. However, this algorithm does not address the spatialization issue. A nicely operated system[18] can automatically convert descriptive texts to two-dimensional traffic accidents. Aiming at communicating general texts, a text-to-picture framework was proposed in Refs. [19-21]. However, the proposed approach tends to be feasible for a short segment, like a sentence. The goal of this paper is to design an effective visualization scheme for communicating a single long document.

2 Hierarchical Parametric Histogram Curve

The representation of a document is a key component in visualizing its contents. We expect the representation to have the following properties:

- Considers the locality of words and captures medium or long range sequential dependencies
- Reflects the semantic variation within a document under different levels of user interest

Both properties should be met in the context of sequential document visualization. To model a very long document, it is appropriate to assume that the portion of the document that is far away from the location of interest should be considered irrelevant, thus does not contribute to the modeling of local content. The second property results from the fact that document content can be captured under multiple resolutions, and the lower the level, the more the detail.

Traditional representations such as Bag Of Words (BOW) or *n*-gram either ignore the word ordering or are only able to capture short range sequential dependency. The recently introduced Locally Weighted Bag Of Words (LOWBOW) representation[9,22] is able to capture sequential progression within a document. It is a non-parametric document modeling approach. By varying the bandwidth of the smoothing kernel, it can capture the document content at various resolutions. However, at a fixed resolution level, it simply uses a uniform sampling scheme to discretize the original continuous representation, thus no adaptiveness has been considered there.

In the following, we introduce a parametric document modeling approach called Hierarchical Parametric Histogram Curve (HPHC). The steps to construct an HPHC representation include building a list of local word histograms, approximating the underlying continuous process by a piecewise linear interpolation, and constructing a hierarchical description by selecting feature points based on local curvature information. Figure 3 shows a simple example. A word list D is discretized as a set of local histograms and linearly interpolated as a continuous curve in a high dimensional space (here 3-D). To visualize this curve, each point can be considered as a barycentric coordinate and is drawn as a point in a triangle.

First, we will review some basic notions related to BOW that constitute the basis of our representation. Given a document D which contains a sequence of words $\{w_1,\ldots,w_N\}$ where $w_i \in V, \forall i$ for some vocabulary set $V = \{v_1,\ldots,v_{|V|}\}$, the BOW representation counts the number of times each word from V appears in the document, and uses this information to estimate the underlying distribution. This results in a $|V|$-entry normalized vector, whose k-th item is computed as $\frac{1}{N}\sum_{i=1}^{N}\delta_{i,k}$ with $\delta_{i,k}$ being defined as

$$\delta_{i,k} = \begin{cases} 1, & w_i = v_k, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Similar to LOWBOW, we assume that the distribution that generates the document varies with the position in the document. More specifically, for a document $D = \{w_1,\ldots,w_N\}$ and some position $j \in \{1,\ldots,N\}$, we use the BOW assumption to estimate the distribution at j over a restricted portion of the document determined by a window of size $2s+1$, which yields

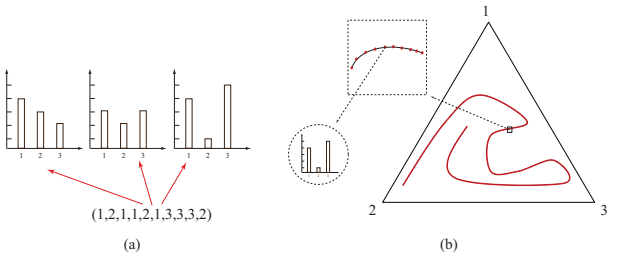


Fig. 3 (a) Given a word list $D =< 1,2,1,1,2,1,3,3,3,2 >$, and a vocabulary $V = \{1,2,3\}$, the set of local histograms is discrete; (b) A piecewise linear histogram curve for a long document.

$$\boldsymbol{h}_j := \frac{1}{2s+1}\langle \Sigma_{i=-s}^{s}\delta_{j+i,1},\ \ldots,\ \Sigma_{i=-s}^{s}\delta_{j+i,|V|}\rangle \tag{2}$$

where $\delta_{i,k}$ is defined in Eq. (1). To preserve the sequentiality within a document, s is set to be bigger than the length between each two consecutive positions such that every tow consecutive portions of document have overlaps.

For an arbitrary set of positions $\{t_1,\ldots,t_M\} \subset \{1,\ldots,N\}$ where $t_1 = 1$, $t_M = N$ and $t_1 < \ldots < t_M$, and the corresponding $\{\boldsymbol{h}_{j_1},\ldots,\boldsymbol{h}_{j_M}\}$, we approximate the word distribution at any position $t \in [1,N]$ with the following linear interpolation

$$\boldsymbol{C}_M(t) = \sum_{i=1}^{M-1}\left(\frac{t_{i+1}-t}{t_{i+1}-t_i}\boldsymbol{h}_i + \frac{t-t_i}{t_{i+1}-t_i}\boldsymbol{h}_{i+1}\right)\mathbf{1}_{B_i}(t) \tag{3}$$

where $B_i = [t_i, t_{i+1})$ and $\mathbf{1}_A(t)$ is the indicator function which is 1 if $t \in A$, and 0 otherwise.

This linear approximation is critical for the second property discussed above. The value of M as well as the realization of M positions may vary to reflect the change in document content. The larger the M, the finer resolution we get for the representation. In addition, if majority of t_is are concentrated at regions where semantic meanings varies a lot, we make use of the adaptive features provided by our technique. In the following, we focus on the construction of HPHC from an initial position configuration where $M_0 = \{t_1,\cdots,t_{M_0}\}$. Note the slight abuse of using M_0 once as a set and once as an integer will simplify the notation below.

The hierarchical parametric histogram curve is essentially a set $\{\boldsymbol{C}_{M_0}(t),\boldsymbol{C}_{M_1}(t),\ldots,\boldsymbol{C}_{M_l}(t)\}$ with the additional requirement $M_0 \supset M_1 \supset \ldots \supset M_l$. $\boldsymbol{C}_{M_{i+1}}$ is constructed from $\boldsymbol{C}_{M_i}$ by using $\{t_1,\ldots,t_{M_i}\}$ as the feature points for $\boldsymbol{C}_{M_i}$ through the following steps:

- Choose a constant m that serves as a smoothing factor influencing the level of detail.
- Compute the following measure of local curvature for position $t_j \in M_i$,

$$\kappa_{jk} = \frac{(\boldsymbol{h}_j - \boldsymbol{h}_{j+k}) \cdot (\boldsymbol{h}_j - \boldsymbol{h}_{j-k})}{|\boldsymbol{h}_j - \boldsymbol{h}_{j+k}||\boldsymbol{h}_j - \boldsymbol{h}_{j-k}|}, \quad k = 1, 2, ..., m,$$

  where κ_{jk} varies from 1 for the sharpest angle to -1 for a straight line. The support r_j and the corresponding curvature κ_{j,r_j} for position t_j are then determined by the following property

$$\kappa_{j,m} < \kappa_{j,m-1} < \cdots < \kappa_{j,r_j} \geq \kappa_{j,r_j-1}.$$

- Perform a nonmaxima suppression[23] to detect feature points t_j where $\kappa_{j,r_j} \geq \kappa_{s,r_s}$ for all s with no more than r_j distance from j. The set M_{i+1} then contains all those detected feature points.
- Update $\boldsymbol{h}_j$ for $t_j \in M_{i+1}$ using Eq. (2) with an asymmetric window centered at t_j, ranging from $t_j - s - \lfloor \frac{t_j - t_{j-1}}{2} \rfloor$ to $t_j + s + \lfloor \frac{t_{j+1} - t_j}{2} \rfloor$. Estimate $\boldsymbol{C}_{M_{i+1}}(t)$ using Eq. (3).

We repeat the above procedure and each time generate a new curve under the controllable threshold m. It is clear from the construction that $\boldsymbol{C}_{M_i}(t)$ contains more details than $\boldsymbol{C}_{M_{i+1}}(t)$. By gradually enlarging the smoothing factor m at each iteration, the entire process is able to discover features in multiple scales, making it possible to capture the document characteristics such as the semantic variation and thematic patterns. Figure 4 depicts an HPHC representation for a 2-D curve in 4 levels.

It is worth mentioning that the last step of updating $\boldsymbol{h}_j$ is essentially an adaptive kernel bandwidth selection scheme that depends on semantic variation within the document. Since fewer feature points are selected for regions of low semantic variation, the effective window size is then enlarged due to relatively large distances between adjacent point locations which results in more smoothing, and vice versa.

3 Visualization and Exploration

The scatterplot[24] is a popular statistical graphics technique that plots a collection of points. It gives a good visual interpretation to the relationship among points. In this section we describe a different way by linking a set of scattered points with a parametric and piecewise linear curve and representing the points such that each one indicates a portion of the document with the most relevant descriptive picture manually collected from the internet. In case of no pictures, a simple point is drawn instead. This new form is able to explicitly express the document structures with a sequential curve. To generate a multi-resolution visualization for sequential documents, two central problems need to be addressed:

(1) How can a sequential document be represented in a hierarchical visual form without the loss of locality and sequentiality?

(2) How can as many flexibilities as possible be provided to enable intuitive document exploration?

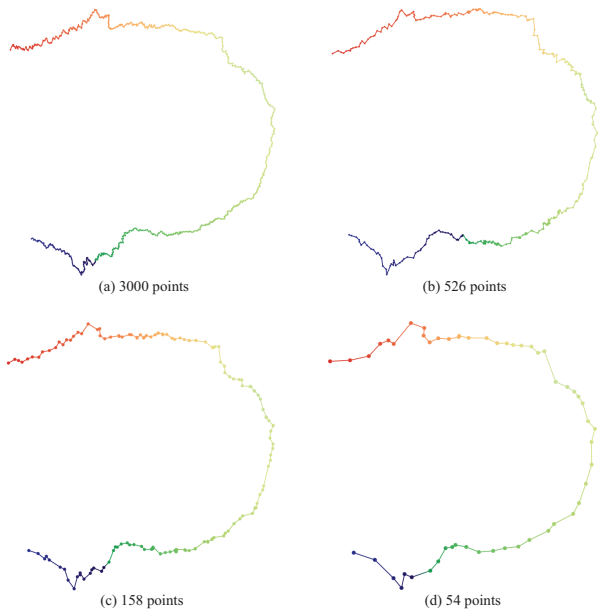


Fig. 4 Four consecutive hierarchies of a hierarchical 2-D curve representation. The rainbow color scheme from blue to red is adopted to encode the sequentiality.

3.1 Locality-preserving two-dimensional reformulation

An HPHC is a high-dimensional curve, whose basic elements are the statistical word frequencies. Many well-studied low-dimensional embedding techniques can reduce the dimensionality for effective visualization. As the dissimilarity among document pieces is the main consideration for exploring the thematic patterns and the topic changes, we employ the Multidimensional Scaling[25] (MDS).

An MDS algorithm tries to assign locations in low-dimensional space, typically 2-D or 3-D space, while preserving the similarities of data points measured in high dimensional space. Suppose that we expect to compute a set of two-dimensional points $\boldsymbol{P} = \{p_i = (x_i, y_i) \in \mathbb{R}^2, i = 1, 2, 3, ..., N\}$, where N is the item number in the high-dimensional space. The optimization of

$\boldsymbol{P}$ is fulfilled by minimizing stress energy

$$\sigma(\boldsymbol{P}) = \sum_{i<j} (d_{ij} - \rho_{ij})^2 \quad (4)$$

where d_{ij} and ρ_{ij} denote the Euclidean distance between p_i and p_j, and the dissimilarity between the i-th and j-th items, respectively.

Along the curve $C(t)$, the local histogram list $< \boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_{|D|} >$ represents the textual information at each word with a local statistical measure. The document sequentiality is actually encoded in the generation mode of the local histogram list so that two consecutive histograms share the most similar information. Other important features are the curvature κ and torsion τ based on the arc-length parametrization of $C(t)$ in $\mathbb{R}^{|V|}$[26]. Therefore, we define the dissimilarity as:

$$\xi_{ij} = \|\boldsymbol{h}_i - \mathbf{h}_j\| + \alpha\|\kappa_i - \kappa_j\| + \beta\|\tau_i - \tau_j\| \quad (5)$$

where α and β are two adjustable variants. They are 0.4 and 0.2 respectively in our system.

In this paper, we assume that the robustness and accuracy issues can be addressed by advanced MDS techniques[25] and focus on its use for curve layout.

When the point positions are obtained, we can compute a smooth two-dimensional curve with local smoothing kernels other than the uniform kernel presented in Eq. (3), like the Gaussian, Tricube, Epanechnikov, and Cauchy kernels[24].

3.2 Multi-resolution visualization

For a substantially long document, the visual exploration should be made at different levels of detail. For example, an abstraction at the coarsest level with about one hundred words may roughly outline the content, but a more detailed summarization at a finer level provides an overview for each document section. However, the computational complexity of MDS becomes a major concern. To reduce the computational complexity of MDS configuration, distributional MDS algorithm[27] is adopted.

First, a sufficiently fine level $\boldsymbol{C}_k(t)$ of the *HPHC* list $< \boldsymbol{C}_0(t), \boldsymbol{C}_1(t), \boldsymbol{C}_2(t), ... >$ is determined so that the distance between $\boldsymbol{C}_0(t)$ and $\boldsymbol{C}_k(t)$ is below a given threshold ε. The distance between a fine level $\boldsymbol{C}_j(t)$ and a coarse level $\boldsymbol{C}_k(t)$ is as follows:

$$d_c(\boldsymbol{C}_j(t), \boldsymbol{C}_k(t)) = \sum_{i=1}^{i=N_j} d_p(\mathbf{h}_{i_j}, \overrightarrow{\mathbf{h}_{i_k}\boldsymbol{h}_{i_{k+1}}}) \quad (6)$$

$$d_p(\mathbf{p}_3, \overrightarrow{\mathbf{p}_1\mathbf{p}_2}) := \frac{(\mathbf{p}_3 - \mathbf{p}_1)\cdot(\mathbf{p}_2 - \mathbf{p}_1)}{\|\mathbf{p}_2 - \mathbf{p}_1\|^2} \quad (7)$$

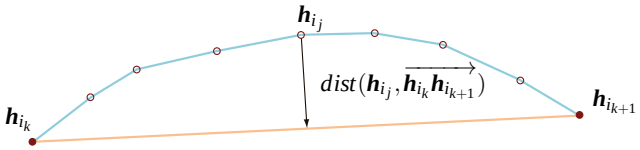


Fig. 5 The distance from a point in a finer level to a line segment in a coarser level.

where N_j is the point number in $\boldsymbol{C}_j(t)$. $\boldsymbol{h}_{i_k}$ and $\boldsymbol{h}_{i_{k+1}}$ are two neighboring points in $\boldsymbol{C}_k(t)$, and $\boldsymbol{h}_{i_j}$ is between $\boldsymbol{h}_{i_k}$ and $\boldsymbol{h}_{i_{k+1}}$ in $\boldsymbol{C}_j(t)$. Figure 5 illustrates the concept of the distance between a point and a line segment with a 2-D example.

Suppose that the selected level is k. We choose its consecutive levels, of which each coarser level is a clustered result built upon its previous level. The distributional MDS algorithm[27] can then be employed to get a multi-level two-dimensional MDS configuration $< \mathbf{P}_k, \mathbf{P}_{k+1}, \mathbf{P}_{k+2}, ... >$. By displaying this sequence orderly, a progressive insight of the underlying document structure is gained. Figure 6 shows three levels of a section selected from the book *My Life*[11]. The configuration of this visualization is listed in the first row of Table 1.

3.3 Focus+Context visualization

For rapid understanding of a long document, both the overview (context) and detailed depiction (focus) are needed to investigate places with semantic transitions. A well-structured layout would allow the users to browse the entire content and focus on interesting regions. Based on the built hierarchical curve structure, we can easily construct a focux+context visualization.

Suppose, for example, the drawn curve under the current viewing is $\boldsymbol{C}_j(t)$. When a focusing operation is specified by the user, the curve segment that intersects with the focus region and the rearranged layout are determined. Subsequently, their projections in $\boldsymbol{C}_{j-1}(t)$ are extracted, yielding a finer curve segment. Each picture in the segment is scaled in a way that is proportional to its distance to the focus center, as depicted in Figure 7.

3.4 Interactive document exploration with linked views

We have developed a user interface (see Figure 2) that integrates multiple exploration windows. The left side of the window is designed to navigate the textual information of the underlying document which is organized as a hierarchical tree through key words or chapter structures. The bottom left window provides some basic interaction settings such as selecting levels. The main window in the top right portion of

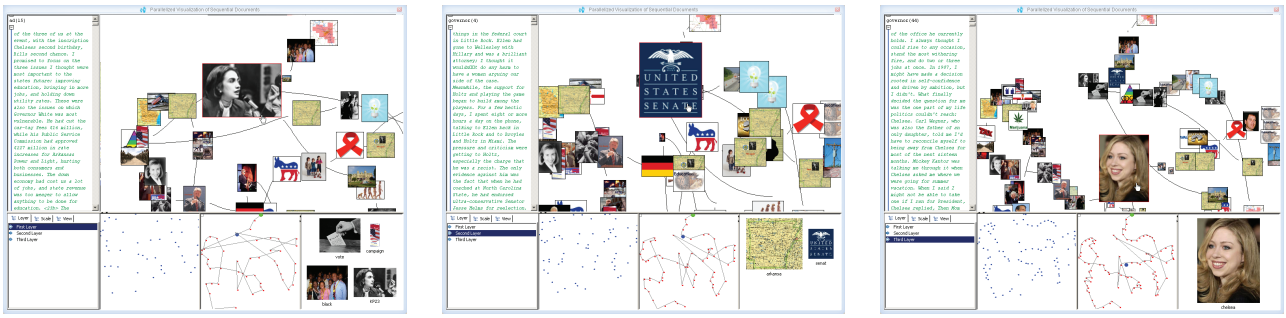


Fig. 6 Multi-resolution views of a section (Chapter 20-Chapter 28) in the book *My Life*[11].

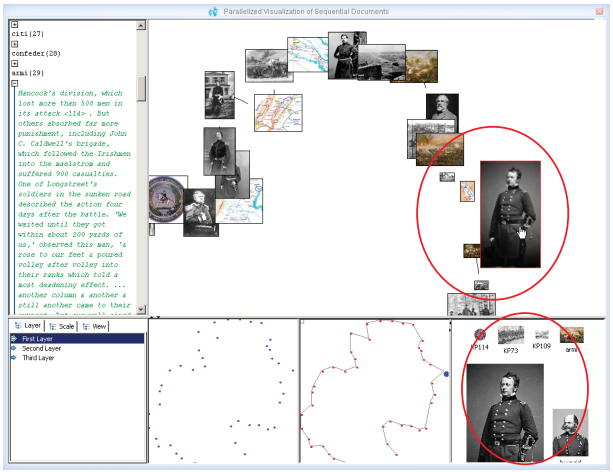


Fig. 7 Focus+context visualization of a section of *The American Civil War*. Pictures within selected segment are also presented in a key frame pop-up window.

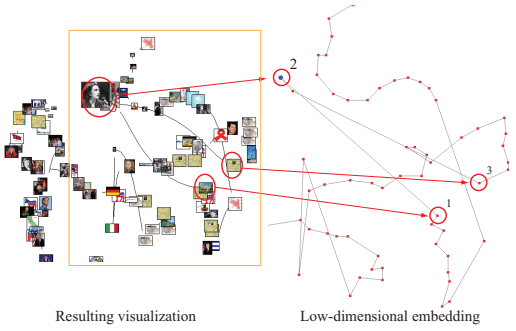
the interface displays an HPHC curve and descriptive pictures located at the two-dimensional projection of each curve point. Selection, zooming, and browsing operations are supported in this window. The key frame popup window and the overview map are shown in the bottom of the interface. In the key frame popup window, multiple picture candidates that relate to the underlying curve point are shown. These windows provide the user a details-on-demand display of the document in conjunction with the main window. To some extent, the two-dimensional MDS view and the Overview Map are redundant. However, the two-dimensional MDS view can help the user discover the point density of the low-dimensional embedding. The Overview Map can provide indications of the sequentiality of document and the active text segment (The blue circle).

In particular, our system features the following operations:

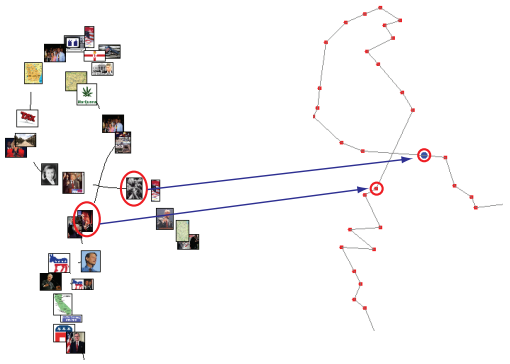
- **Rapid document browsing.** The descriptive pictures and the cuve lower the user's workload, making the browsing suitable for anyone, especially persons with low-literacy skills.
- **Document segmentation.** The most coarse levels in the *HPHC* representation suggest potential segmentations of a document. Users can also rapidly justify the analysis by checking the textual information where the curve variations are large, and generate a pictorial summarization based on the ones that pass the examination. Figure 8a shows an interesting example. The numbers denote three important events during the contest for the Governor of Arkansas in 1982: (1) Bill decided to contest for the Governor. Thereafter he was busy in other affairs; (2) Hillary showed her support for Bill by taking the Clinton family name during the contest; (3) Bill was elected as the Governor for the second time. Here, the dominant points in the HPHC clearly show sharp topic transitions, and lead to a potential document segmentation.
- **Document pattern discovery.** The sequentiality of the *HPHC* representation leads to a long evolving picture sequence, which is extremely suitable for documents with clear temporal or spatial relationships like the biography, geography, and historical books. If there are some similar or period thematic patterns or concepts in a document, our multi-resolution visualization scheme effectively preserves them while making a reasonable two-dimensional layout. An interesting example is shown in Figure 8b, where the curve intersects with itself indicating a repetitive event was detected.

4 Experiments and Discussion

We have implemented the proposed approach with Java. For efficiency, the computation for the dissimilarity matrix and MDS configuration is accelerated by using CUDA. All experiments are conducted on a PC equipped with a P4 2.4 GHz CPU, 3.5GB host memory, and a nVidia GTX 280 video card.



(a)



(b)

Fig. 8 Visualizing a section of *My Life*. (a) A piece of HPHC that shows the process of Clinton's second run for the Governor of Arkansas in 1982. (b) A repeating pattern in the *hphc* reveals the entire procedure of the presidential election preparation of Bill Clinton in 1991.

4.1 Textual Preprocessing

Before the visualization, sequential processing of the input document is needed. First, all words are converted to lower case and non-letter symbols are removed. The Porter stemming algorithm[28] is then employed to stem the word list. To get a smoothly progressed list of local histograms along the document, a small numeric value (e.g., 0.05) is added to zero-valued items.

Three books, namely, *My Life*[11] (abbr. ML), *Robinson Crusoe*[29] (abbr. RC), and *The American Civil War*[30] (abbr. AW), were tested with our proposed method in this paper.

The first and second columns of Table 1 list the item numbers of the initial input documents and their stemmed versions. For each document, three levels were constructed to build the textual hierarchy (keywords) and the hierarchical levels, as reported in the third and fourth columns. The time (in minutes) and memory (in megabyte) consumed for preprocessing are listed in the last two columns.

Table 1 Statistics on the tested documents

	Input document	Stemmed document	Key words	LOD	Time	Mem
ML	101526	44947	164	48/67/97	10	150
RC	104318	36785	482	55/93/146	10	120
AW	26237	13655	186	41/63/102	6	30

4.2 User Evaluation

The results for the document *My Life* have been demonstrated throughout this paper. We show the results for another two documents in Figure 9 and Figure 10.

To evaluate the efficiency of our approach, we conducted a preliminary informal user evaluation that consisted of 10 volunteers who had no experience in the field of text visualization. All participants were graduate students and three of them were female. Shortly before they used our toolkit to explore the three documents, a 10-minute introduction to the user study was given. Three document exploration modes were then provided, namely, the text-based navigation, the curve view, and the visualization with linked views. The first two modes are provided by disabling related features within our system. For each mode, an individual document was used as the testing data. The users were then asked to answer the three questions listed below. Five categorical values ranging from 1 (very negative) to 5 (very positive) were given for each question and each mode.

- How long do you expect to roughly understand what this book is about with the help of this mode?
- Do you think that this mode can help you understand this document?
- Will you choose this mode when you are asked to present this document to your grandmother or grandfather?

The total user study took 40 minutes. The scores for each mode are listed in the order of the questions: the text-based navigation (2.4/3.6/2.0); the curve view (2.5/2.8/2.3), and the visualization with linked views (3.0/4.0/3.8). These statistics show a strong favorability for the linked views visualization. All users found that the visualization with linked views mode very interesting and novel. They also agreed that the system was very useful for rapid document understanding and were always willing to read the ones with pictures. Some users commented that our tool "clearly provided a useful complement to the hierarchical document structures". This evaluation also confirmed the interactive performance and plausible interactivity of our system

thanks to the combination of the sequential document visualization and the pictorial communication techniques. Furthermore, in certain situations, it provides a rapid method for identifying important document structures.

4.3 Discussions

Our HPHC representation shares many common components as the LOWBOW representation. For instance, both are built upon the local histogram list, and maintain sequential and position information. However, our visualization approach extensively considers the two-dimensional layout of a sequential curve, and is different from the LOWBOW-based document visualization[9] in several aspects:

- LOWBOW is a sophisticated framework for representing sequential or categorical time series data. Our approach takes a much simpler form, and is flexible to undergo both the frequency-based or geometric operations like simplification, query, and frequency analysis.
- The HPHC representation is parametric, while the LOWBOW representation is non-parametric and globally smoothed by using a smoothing filter scheme. Although the *hphc* is formed by fitting the discrete local histograms with a uniform kernel (Eq. (3)), all other smoothing kernels suitable for the LOWBOW framework like the Gaussian, Epanechnikov, and Cauchy kernels, can also be employed.
- The HPHC takes a hierarchical structure whose finest level encodes every word, and whose coarsest level gives a high abstraction to the document. In contrast, the LOWBOW representation uses a uniform sampling and achieves a multi-resolution view by changing the size of the smoothing kernel. Thus, it can hardly afford adaptive exploration.

Furthermore, the selected images have an effect on the resulting visualization. But automatically generating pictures from texts is not the main focus of this paper. All pictures used in this paper are manually collected from the internet and labeled by five volunteers in one week who are otherwise uninvolved to this work. However, more advanced techniques can be seamlessly incorporated into our approach.

5 Conclusions

Document visualization is useful in browsing, searching, and analyzing hidden themes of textual archives. In this paper we describe an integrated approach that supports hierarchical visualization, interactive navigation, comparative analysis, as well as querying capabilities. With the combination of the one-dimensional document sequentiality and the two-dimensional graphical forms, we create an effective visual exploration system that is capable of providing users (e.g., a 12 years old child) with a powerful tool that is easy to use and requires little expertise by leveraging their abilities to understand the visual forms.

For future directions, we would like to test the proposed scheme with other categorical time series data such as video sequences and financial data. We also expect to extend the pictorial communication mode for the visualization of document corpus. We believe that our approach can help develop a new document browsing technology, potentially reshaping the way in which users comprehend textual information.

References

[1] Fortuna B, Grobelnik M, Mladenic D. Visualization of text document corpus. *Informatica Journal*, 2005, **29**(4): 497-502.

[2] Havre S, Hetzler E, Whitney P, et al. ThemeRiver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 2002, **8**(1): 9-20.

[3] Rohrer R, Sibert J, Ebert D S. The shape of shakespeare: Visualizing text using implicit surfaces. In: IEEE Symposium on Information Visualization. Washington, D.C., USA, 1998.

[4] Cui W, Liu S, Tan L, et al. TextFlow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 2011, **17**(12): 2412-2421.

[5] Ham F V, Wattenberg M, Viegas F B. Mapping text with phrase nets. *IEEE Transactions on Visualization and Computer Graphics*, 2009, **15**(6): 1169-1176.

[6] Wise J, Thomas J, Pennock K, et al. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In: Information Visualization. Washington, D.C., USA, 1995: 51-58.

[7] DeCamp P, Frid-Jimenez A, Guiness J, et al. Gist Icons: Seeing meaning in large bodies of literature. In: IEEE Symposium on Information Visualization, Poster Compendium, IEEE CS Press, 2005.

[8] Collins C, Carpendale S, Penn G. DocuBurst: Visualizing document content using language structure. *Computer Graphics Forum*, 2009, **28**(3): 1039-1046.

[9] Mao Y, Dillon J, Lebanon G. Sequential document visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2007, **13**(6):1208-1215.

[10] Tufte E. Visual Explanations: Images and quantities, evidence and narrative. Graphics Press, 1997.

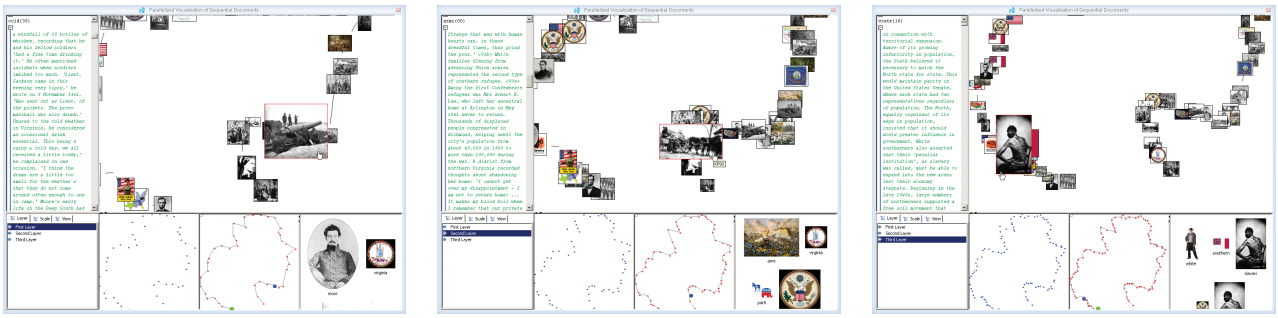


Fig. 9 Visualizing three levels for the book *The American Civil War*[29]. The third row of Table 1 lists the configuration of this experiment.

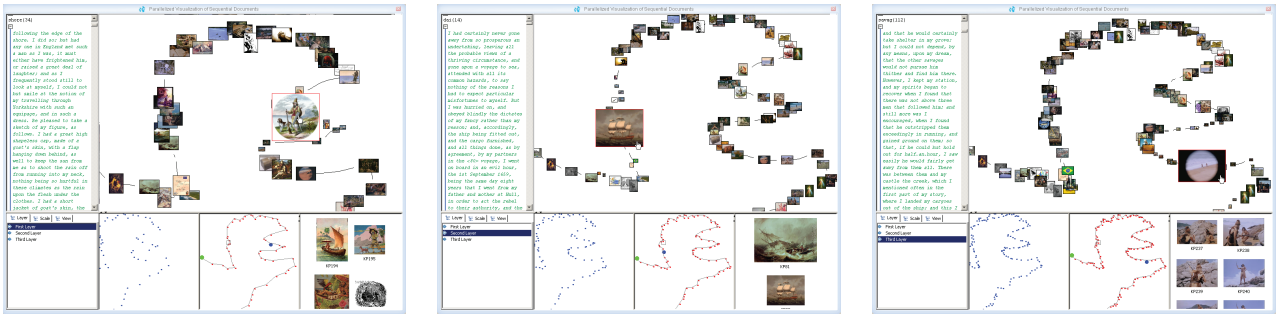


Fig. 10 Visualizing three levels for the book *Robinson Crusoe*[30]. The second row of Table 1 lists the configuration of this experiment.

[11] Clinton B. My Life. Vintage , 2005.

[12] Miller N E, Wong P C, M. Brewster, et al. Topic islands-a wavelet-based text visualization system. In: IEEE Visualization. North Carolina, United States, 1998.

[13] Pacific Northwest National Laboratory. IN-SPIRE Visual Document Analysis. [Online]. Available: http://in-spire.pnnl.gov/, 2012.

[14] BioWisdom. Omniviz and Refviz. [Online]. Available: http://www.biowisdom.com/tag/omniviz/, 2012.

[15] Wong P C, Cowley W, Foote H, et al. Visualizing sequential patterns for text mining. In: IEEE Symposium on Information Visualization, Washington, DC, USA, 2000.

[16] Coyne B, Sproat R. WordsEye: An automatic text-to-scene conversion system. In: ACM SIGGRAPH. Los Angeles, USA, 2001: 487-496.

[17] Joshi D, Wang Z, Li J. The story picturing engine: A system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2006, **2**(1): 1356-1363.

[18] Johansson R, Berglund A, Danielsson M, et al. Automatic text-to-scene conversion in the traffic accident domain. In: 19th IJCAI. Edinburgh, Scotland, 2005: 1073-1078.

[19] Goldberg A B, Rosin J, Zhu X, et al. Toward text-to-picture synthesis. In: NIPS 2009 Symposium on Assistive Machine Learning for People with Disabilities, 2009.

[20] Goldberg A, Zhu X, Dyer C R, et al. Easy as ABC? facilitating pictorial communication via semantically enhanced layout. In: Computational Natural Language Learning. Manchester, United Kingdom, 2008.

[21] Zhu X, Goldberg A, Eldawy M, et al. A text-to-picture synthesis system for augmenting communication. In: The Integrated Intelligence Track of the Twenty-Second AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2007.

[22] Lebanon G. Sequential document representations and simplicial curves. In: Uncertainty in Artificial Intelligence (UAI), Arlington, Virginia, USA, 2006.

[23] Teh C H, Chin R T. On the detection of dominant points on digital curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989, **11**(8): 859-872.

[24] Wilkinson L. The Grammar of Graphics. Springer, 1999.

[25] Borg I, Groenen P. Modern multidimensional Scaling: theory and applications. Springer, 1997.

[26] Spivak M. A Comprehensive Introduction to Differential Geometry. Perish Press, 1999.

[27] Quist M, Yona G. Distributional scaling: An algorithm for structure preserving embedding of metric and nonmetric spaces. *Journal of Machine Learning Research*, 2004, **5**: 399-420.

[28] Porter M F. An algorithm for suffix stripping. *Readings in Information Retrieval*, 1980, **14**(3): 130-137.

[29] Glatthaar J, Krick R. The American Civil War. Osprey Publishing, 2001.

[30] Defoe D. Robinson Crusoe. W. Taylor, 1719.