The Story Picturing Engine: Finding Elite Images to Illustrate a Story Using Mutual Reinforcement

Dhiraj Joshi
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802
djoshi@cse.psu.edu

James Z. Wang
School of Information Sciences and Technology and the Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802
jwang@ist.psu.edu

Jia Li
Departments of Statistics and Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802
jiali@stat.psu.edu

ABSTRACT

In this paper, we present an approach towards automated story picturing based on mutual reinforcement principle. Story picturing refers to the process of illustrating a story with suitable pictures. In our approach, semantic keywords are extracted from the story text and an annotated image database is searched to form an initial picture pool. Thereafter, a novel image ranking scheme *automatically* determines the importance of each image. Both lexical annotations and visual content of an image play a role in determining its rank. Annotations are processed using the Wordnet to derive a lexical signature for each image. An integrated region based similarity is also calculated between each pair of images. An overall similarity measure is formed using lexical and visual features. In the end, a mutual reinforcement based rank is calculated for each image using the image similarity matrix. We also present a human behavior model based on a discrete state Markov process which captures the intuition for our technique. Experimental results have demonstrated the effectiveness of our scheme.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: retrieval models; search process; selection process.

General Terms: Algorithms, Design, Experimentation, Human Factors.

Keywords: Story picturing, lexical referencing, image retrieval, mutual reinforcement, Markov chain.

1. INTRODUCTION

Stories are often accompanied with pictures. Classic tales of folklore were illustrated with beautiful sketches before the advent of the camera. We always associate historic characters we have read about with pictures we have seen in books. Stories written for children ought to be decorated by pictures to sustain a child's interest, for *a picture is worth a thousand words*. In modern times, pictures are used everywhere, from newspapers, magazines and Websites to movie banners in order to punctuate the effect of stories surrounding the pictures. Often the pictures themselves become more important than the accompanying text. In such scenarios, a key task is to choose the best pictures to display.

Story Picturing, aptly called, denotes the process of depicting the events, happenings and ideas conveyed by a piece of text in the form of few representative pictures. This work is performed by news writers and speakers who select a few images from their repertoire to complement their news stories. In our work, we attempt to automate the process of story picturing.

Choosing a few representative images from a collection of candidate pictures is challenging and highly subjective because of the lack of any defined criteria. People may choose to display one particular image over another entirely on merit or pure prejudice. Similarly a photographer's personal disposition will be reflected in his or her photo archives. For example, on a visit to Europe, a history lover will click more pictures of historical places while a nature lover will capture the Alps in his camera.

1.1 Related work in image retrieval

Many efficient content-based image retrieval systems have come up in the last decade [23, 19, 7, 24]. Most of the work has been focused upon quantifying image similarity and retrieving images similar to a query image. In recent years, linguistic indexing of pictures using learned statistical models has been explored [15]. Statistical associations between image regions and words have been studied [3]. Spectral graph clustering has been found to be effective for image retrieval [9]. Machine learning approaches are being applied to study ancient art [16]. Concerns have been expressed to archive all ancient historical and cultural materials in digital form for posterity [8]. As the sizes of digital image libraries grow, a constant need for machine intervention in all aspects of search is pressing.

Let us try to analyze how the research issues mentioned in this section are related to story picturing. Linguistic index-

MIR'04, October 15–16, 2004, New York, New York, USA.

ing addresses the problem of finding the most suitable text to describe a given image. Story picturing, on the other hand, attempts to find the best set of pictures to describe a given piece of text. Thus, the two problems can be regarded as inverse of each other. Notice that story picturing is a content-based image retrieval problem as here, the aim is to retrieve images similar, in content, to a story.

1.2 Related work using mutual reinforcement

Methods based on mutual reinforcement principle have been widely reported in literature especially in the domains of journal evaluation and more recently on Web search [4, 14, 17, 21, 13].

All such methods fall under the category of link analysis techniques wherein the underlying structure is a graph, the entities can be represented as the nodes and the edges represent endorsements that entities give each other. Endorsements can be in the form of *citations* (as in journal evaluation systems) as well as hyperlinks (in Web analysis). The aim is to associate a numeric measure of importance (also referred to as *standing*) with each entity on the basis of the link structure. Kleinberg's HITS algorithm finds *hubs* (Web-pages pointing to many important sites) and *authorities* (important Websites pointed to by many other pages) [14]. Google's *pagerank* of a particular Web-page is a measure of its standing based on its link structure [4]. In [17], modification of HITS by assigning a weight to each link based on textual similarities between pages has been found to perform better than the original HITS.

1.3 Our approach

In this paper, we present an interesting scheme for image ranking and selection based on mutual reinforcement principle. This ranking process is at the heart of our Story Picturing Engine. The set of candidate images is assumed to form a graph with the images acting as nodes and image similarities forming the weights in the edges. Under a special condition, the image selection process can be modeled as a random walk in the graph.

Automatic text to scene conversion by computer graphics techniques has been studied for several years [18, 22, 10, 11, 5]. The WordsEye system developed at AT&T Labs [11] is a natural language understanding system which converts English text into three-dimensional scenes that represent the text. Our Story Picturing Engine, on the other hand, attempts to choose the best set of images from an image database to illustrate a piece of text. The goals of the former and the latter are similar. However, even with an annotated image database available, it is not simple to choose a few images which best represent the text. As pointed out earlier, the problem is subjective and a human being would use his knowledge acquired over the years to judge the importance of each image. Quantifying image importance in a pool of images is the focus of our work. Authors of [2] introduced the idea of auto-illustration as an inverse problem of auto-annotation. Statistical associations between images and text were used to find images with high likelihood. However, the discussion was brief. In this paper, we present a new approach to this problem.

1.4 Outline of the paper

The rest of the paper is organized as follows. In section 2, we describe the story picturing engine and present the mathematical formalism of reinforcement principle. In section 3, we discuss some results and in section 4, we list some future research directions.

2. THE STORY PICTURING ENGINE

Our Story Picturing Engine consists of four components, (1) the story processing, (2) the image selection process, (3) estimation of similarity between pairs of images based on their visual and lexical features, and (4) mutual reinforcement based rank estimation process. In this section, we describe these individual components and show how they work together to form a robust image ranking system. Figure 1 shows a schematic flow diagram of our system.

2.1 Story processing

The purpose of the story processing component of the system is to identify certain descriptor keywords and proper nouns in the story and estimate a lexical similarity between keywords using the Wordnet. We first perform stopword elimination from the story. Stopwords are common words (i.e., a, of, the, on) that have little or no meaning by themselves [1]. A list of stopwords has been manually prepared which serves the purpose. Wordnet, developed by the Cognitive Science Laboratory at Princeton University, is a lexical reference system, the design of which is inspired by psycholinguistic theories of human lexical memory [20, 12]. Nouns are arranged in topical hierarchies and the hierarchy induces a transitive relation *hypernymy* among concepts. Consider the following wordnet hierarchy *oak* $\rightarrow$ *tree* $\rightarrow$ *plant* $\rightarrow$ *organism*, here the arrow represents the *hypernymic* relation. A *polysemy count* is defined for each word based on the number of senses it has (i.e., the number of different contexts it can be used in). For example, according to Wordnet, the noun *case* has 18 different senses. *Case* is synonymous with a *lawsuit*, an *instance*, a *container* among others and can be used in place of any of them without changing the meaning of the sentence. On the other hand, the polysemy count of *kangaroo* is 1.

For our purpose, we choose to eliminate nouns with very high polysemy count because such words offer little weight to the meaning conveyed by the story (as they could mean several things). Besides this, we also eliminate adjectives, verbs and adverbs, with high polysemy counts, from the story. Hence, in our system, common nouns with *not-so-high* polysemy counts and verbs, adjectives and adverbs with *low* polysemy counts are descriptor keywords of a piece of text. In addition to this, we also identify a set of proper nouns from the text which are deemed more important than the rest of the keywords as they denote the place and people the story is about. At the end of this process, we denote the keywords by k_1, k_2, ..., k_n and the set of proper nouns by N_1,N_2,...,N_d.

We made use of the semantic organization of Wordnet to derive *rough* semantic similarities between keywords. The following approach is adopted for the present implementation. We now list a few Wordnet definitions before proceeding to explain the mathematics:

1. k_i and k_j are *synonyms* if they can be used interchangeably within the same context.

2. k_i is a *hypernym* of k_j if k_i occurs in k_j's topical hierarchy tree (*hyponymy* is the inverse of *hypernymy*).

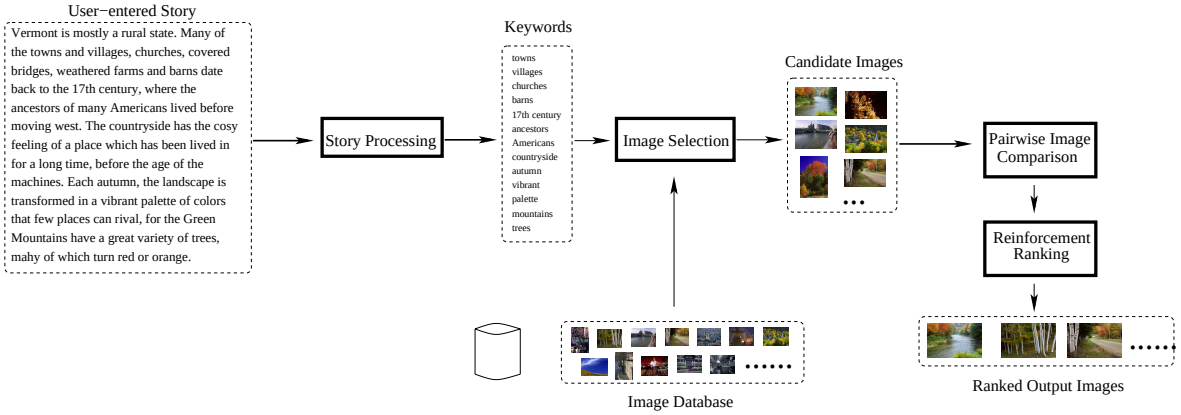


Figure 1: A flow diagram of our Story Picturing Engine.

3. k_i is a *meronym* of k_j if k_i is a part of k_j (e.g., beak is a meronym of bird).

4. k_i and k_j are *co-ordinate terms* if they have the same *hypernym* (e.g., mountain and ridge).

If k_i and k_j are two keywords, we define a similarity measure $\mathcal{K}_{ij}$ between them as follows

$$\mathcal{K}_{ij} = \begin{cases} 1 & \text{if } k_i \text{ and } k_j \text{ are identical or synonyms.} \\ \mu & \text{if } k_i \text{ and } k_j \text{ are related by meronymy or are co-ordinate terms.} \\ \mu^t & \text{if } k_i \text{ and } k_j \text{ are related by hypernymy and } k_i \text{ appears in } k_j\text{'s topical hierarchy tree, } t \text{ edges away or vice versa.} \\ 0 & \text{if } k_i \text{ and } k_j \text{ are unrelated.} \end{cases}$$

In the previous expression, μ is a real number between 0 and 1. In our experiments, we have arbitrarily fixed μ as 0.5. As an example, if k_i is *kangaroo* and k_j is *fauna* then according to our definition, $\mathcal{K}_{ij}$ is 0.015625 as *fauna* appears in *kangaroo's* topical hierarchy tree 6 edges away from *kangaroo*.

In literature, there exist more elaborate semantic distance calculation mechanisms based on Wordnet hierarchy which have been used in computational linguistics [6]. However, for our present implementation, this was not the prime focus and so we kept the relations simple. It would be an interesting future research to incorporate more sophisticated lexical similarity measures into our system.

2.2 The image selection process

In this section, an initial picture pool is formed composed of images whose annotation text has terms matching the story text. For clarity we restate our notations. The keywords extracted from the story are denoted by k_1, k_2, ..., k_n and a set of proper nouns denoted by N_1,N_2,...,N_d. In order to select an initial pool of images upon which the ranking algorithm will be applied, we select images which are annotated with at least one of the N_r's and one of the k_s'.

2.3 Estimation of similarity

Once a pool of images has been identified, we assign a numeric similarity to each pair of images $\mathcal{I}_i$ and $\mathcal{I}_j$ based on their visual and lexical features.

2.3.1 Similarity assignment

The following steps elaborate our approach.

1. An Integrated Region Matching (IRM) distance is calculated between images. IRM is an image matching mechanism which identifies regions in images and calculates an overall region-based distance using visual features. Details of IRM have been skipped here due to lack of space and can be found in [24].

2. IRM distances d_{ij} are converted into percentile IRM similarities η_{ij}, such that η_{ij} is the fraction of all d_{st} such that $d_{st} \geq d_{ij}$.

3. An annotation based similarity is also calculated between pairs of images as follows

$$\zeta_{ij} = \sum_{k_l \in A_i} \sum_{k_m \in A_j} \mathcal{K}_{lm}. \quad (1)$$

   In the previous expression, A_i and A_j denote the set of words that form the annotations of images I_i and I_j respectively. In the present implementation, we only consider words in annotation sets A_i and A_j which are among k_1,...,k_n. ζ_{ij}'s are also converted to respective percentile similarities.

4. The two forms of similarities are combined to form a unified similarity measure between pairs of images. If $\alpha \in [0, 1]$, s_{ij} is defined as follows

$$s_{ij} = \alpha\eta_{ij} + (1 - \alpha)\zeta_{ij}. \quad (2)$$

Note that $s_{ij} > 0\ \forall i, j$ and $s_{ij} = s_{ji}$ by the previous definition. The parameter α balances the effects of lexical and visual features in determining s_{ij}. It is desirable to have an intermediate value of α. The manual annotations associated with an image are reflections of human intelligence applied to interpret the content of an image whereas visual features have pronounced effects on integrated region based similarity. It is important to perform visual processing to identify visually good images from among those which bear similar manual annotations. Therefore, we combine the two to form a similarity measure which awards visually similar images as well as images judged similar by annotations.

2.4 Reinforcement-based rank

Mutual reinforcement refers to the process where each entity contributes towards the rank of other based on some similarity between them. This kind of ranking has been successfully used in several domains over the years. In our system, we use this idea to assign a measure of importance to each image based on its similarity with other images. An iterative mechanism is used to estimate the rank of each image. Finally the most highly ranked images are selected. Figure 2 illustrates the idea of mutual reinforcement. We now discuss the mathematical details of this procedure.

Let $\mathcal{I}_1$,$\mathcal{I}_2$,...,$\mathcal{I}_N$ represent the working set of images obtained as explained in section 2.2. We define the rank of image $\mathcal{I}_i$ as r_i which is the solution to the equation:

$$r_i = \sum_{j=1}^{N} s_{ij} r_j. \tag{3}$$

Let us first discuss the consequence of rank assignment based on Eq 3 using a standard result from linear algebra [14].

If $\mathcal{S}$ is a symmetric $M \times M$ matrix and $\vec{u}$ is a vector which is not orthogonal to the principal eigenvector of $\mathcal{S}$, then the unit vector in the direction of $\mathcal{S}^t u$ converges to the principal eigenvector $\vec{v}$ of $\mathcal{S}$ as t increases without bound. Further, if the entries of $\mathcal{S}$ are non-negative, then the principal eigenvector $\vec{v}$ has only non-negative entries.

Eq. 3 is non-linear but can be solved iteratively by the previous result. The following algorithm, commonly referred to as *power method* finds the principal eigenvector of a symmetric matrix with non-negative entries.

1. Initialize $\overrightarrow{r^0}=(r_1^0, r_2^0, \ldots, r_N^0)$ randomly such that $\sum_{i=1}^{N} r_i^0 = 1$ and $r_i^0 > 0\ \forall i$.
2. $t \leftarrow 1$.
3. $r_i^t = \sum_{j=1}^{N} s_{ij} r_j^{t-1}\ \forall i \in 1, \ldots, n$.
4. $r_i^t \leftarrow \frac{r_i^t}{\|\overrightarrow{r^t}\|_1}$, $\|\overrightarrow{r^t}\|_1 = \sum_{i=1}^{N} r_i^t$.
5. $t \leftarrow t + 1$.
6. Repeat steps 3 to 5 till convergence (i.e., $\overrightarrow{r^t} = \overrightarrow{r^{t-1}}$).

Since $\mathcal{S}$ has only non-negative entries, its principal eigenvector $\vec{v}$ also has only non-negative entries, hence the constraints on our choice of the initial vector $\overrightarrow{r^0}$ in step 1 ensure that it is not orthogonal to $\vec{v}$.

Consider a graph $\mathcal{G}$ such that $\mathcal{I}_1$,$\mathcal{I}_2$,...,$\mathcal{I}_N$ constitute the nodes of $\mathcal{G}$ and s_{ij} is the weight of edge from image I_j to image I_i, then finding high ranked images from among $\mathcal{I}_1$,$\mathcal{I}_2$,...,$\mathcal{I}_N$ using mutual reinforcement becomes akin to finding *hubs*, *authorities* or pages with high *pagerank* in the World-Wide-Web graph. Pictures possessing a high rank are expected to be *authoritative* in terms of their content.

2.5 Discrete state Markov chain model

We will now present an interesting model of human behavior under the condition that $\sum_{j=1}^{N} s_{ij} = 1\ \forall i$ which can be achieved by normalizing each row of the image similarity matrix $\mathcal{S}$. Imagine a human operator trying to select a few representative images from a pool of images that capture a particular concept or concepts conveyed by a piece of text. He begins by looking at a random image and with an intent to look for a better image, proceeds to another. However, it is natural to assume that the selection of the next image would be influenced by its similarity to its precursor. This model captures the real scenario to a certain extent as human beings do have a fickle mind and scan through a number of things before deciding upon the *best*. Consider the case when the operator is looking at image $\mathcal{I}_j$, we claim that s_{ij} now represents the probability that he or she will proceed to image $\mathcal{I}_i$ next. Note that $\sum_{j=1}^{N} s_{ij} = 1$ and so $s_{ij}\ \forall i$ is a probability mass function.

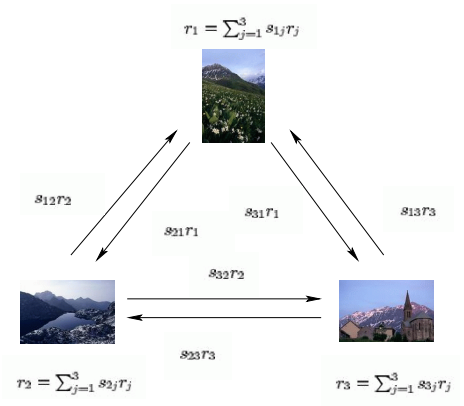


Figure 2: Mutual reinforcement is depicted by arrows for three images. Every image reinforces the rank of every other image based on how similar they are.

This pattern can be aptly described by a discrete state Markov chain in which each image forms a discrete state and $\mathcal{S}$=$[s_{ij}]$ represents the transition probability matrix. Thus, the solution $\vec{r}$ to the equation $\vec{r} = \mathcal{S}\vec{r}$, is the stationary distribution of this stochastic model. In other words, r_i represents the probability that the operator will reach image I_i, while performing the random walk.

This model of human behavior has been inspired from Brin and Page's explanation of Google's pagerank [4]. In their words, the pagerank represents the probability that a surfer beginning from a random Web-page arrives at a particular page following only forward *hyperlinks*.

A small variation to the previous algorithm is that the similarity matrix $\mathcal{S}$ does not remain symmetric under the constraint $\sum_{j=1}^{N} s_{ij} = 1\ \forall i$. However, the rank vector $\vec{r}$ can still be calculated using the following set of standard results for Markov chains.

1. A transition matrix which is element wise positive is irreducible, aperiodic and positive recurrent.
2. If the transition probability matrix $\mathcal{S}$ is irreducible, aperiodic and positive recurrent then $\mathcal{S}^k$ converges to a matrix in which each column is the unique stationary distribution of $\mathcal{S}$.

3. EXPERIMENTS

For our first set of experiments to test this procedure, we used the *Terragalleria* [1] and the *AMICO* [2] image databases. Terragalleria is the collection of personal photographs of

[1] `http://www.terragalleria.com/`
[2] `http://www.amico.org/`

Quang-Tuan Luong who is famous for his scientific work on camera self-calibration and his devotion to the large-format photography community. He took these pictures during his numerous treks and visits to various parts of the world. The number of pictures in the *Terragalleria* database is around 7200 and each image bears a manual annotation provided by the photographer. Besides this, in his Website, he has included small descriptions of his travels and experiences in different places, which we adopted as our test stories.

The Art Museum Image Consortium (*AMICO*) has been formed by around 40 museums from all over the world. Images, paintings, and sculptures have been archived in digital form to enable educational use of multimedia data in the museums. The number of digital images in this database is around 118, 300. Important metadata about the images including their name, their artist, their country, their genre, the material used and the museum were treated as manual annotations. We obtained some sample test stories from *ARTKids* [3]. ARTKids is a nonprofit website offering art resources for the classroom and independent students of art. An important assumption about our test stories is that they revolve around *a few* central ideas and talk about places, people, times (e.g., Roman).

3.1 Experiments on Terragalleria database

The two example stories and their results discussed here are the photographer's descriptions of *Paris and Vermont* (a province in the US). These stories were written by the photographer and provided on his Website. We took these stories verbatim. The effect of story picturing applied to his textual descriptions can be reminiscent of a scenario when Quang-Tuan Luong is trying to choose some representative pictures from his visits to Paris or Vermont in order to show them to a friend or to a visitor of the Website. Naturally, we would expect that the results will be influenced by his photographic patterns. By patterns, we refer to the way in which a human being has internalized concepts. Abstract ideas like *nature, spring* could stand for a number of things in real life. It is subjective what concrete things a person prefers to best associate with them. The two stories used in our experiments are listed below.

- *Paris.* The cradle of Paris is Ile de la Cite, a small island in the middle of the Seine river. Paris then first developed along the river, and from the Louvre to the Eiffel tower or the Place de la Concorde to Notre Dame, its evolution and its history can be seen from the river. The central portion of the banks form a very homogeneous ensemble and makes for a fine walk. The banks of the Seine are UNESCO World heritage site. The right bank of the Seine is dominated by the large perspectives due to the avenues designed by Haussman in the 19th century. The most prominent of them is the one extending from the Louvre to the Arc de Triomphe, through the Champs Elysees, France's most famous avenue. (by:Q-T. Luong)

- *Vermont.* Vermont is mostly a rural state. Many of the towns and villages, churches, covered bridges, weathered farms and barns date back to the 17th century, where the ancestors of many Americans lived before moving west. The countryside has the cozy feeling of a place which has been lived in for a long time, before the age of the machines. Each autumn, the landscape is transformed in a vibrant palette of colors that few places can rival, for the Green Mountains have a great variety of trees, many of which turn red or orange. (by:Q-T. Luong)

[3] `http://www.artfaces.com/artkids`

3.2 Experiments on AMICO database

As mentioned earlier, AMICO is a growing digital library which at present contains over 100, 000 paintings, sculptures, drawings and watercolors, prints, photographs, textiles, costumes and jewelery, works of decorative art, books and manuscripts in digital form. Works date from prehistoric (around 2000 B.C.) to contemporary times. A diverse range of cultures from Native American to ancient and medieval Greek, Roman, Egyptian, Chinese and Indian civilizations have been evenly represented. There is a good collection of modern works too.

Sample stories for this set of experiments were obtained verbatim from ARTKids which is an educational Website (as mentioned earlier). Here, we have included results for 3 short-stories which are listed below. Each story is a small description of art or sculpture of a particular civilization.

- *Roman Art.* While there is evidence that the Romans also painted on portable panels, the surviving paintings that we can see today were painted directly on the walls of their rooms. Domestic interiors were claustrophobic, windowless and dark, so the Romans used painted decoration to visually open up and lighten their living spaces. Technical elements of Roman painting include the fresco technique; brightly colored backgrounds; division of the wall into multiple rectangular areas, tic tac toe design; multipoint perspective, and trompel oeil effects.

- *Egyptian Sculpture.* As far back as 5,000 years ago Egypt had introduced a style that, with surprisingly little change, continued for almost 3,000 years. Rules for the making of statues were rigidly prescribed, as were social and religious customs. Religion was the dominant force in life on earth and it required certain preparations for the life beyond. Sculpture was entirely associated with the needs of religion and the gods or with the earthly rulers who were regarded as their representatives. To symbolize the Godlike role of the kings, they were represented as half human, half animal. The great Sphinx at Gizeh is the best known example. to express their power and eternal life they were carved in the hardest stone and in colossal proportions. The statues of Rameses 2 at Abu Simbel are examples.

- *Greek Sculpture.* The glory of Greece was its sculpture. The roots of Greek sculpture reach into the earlier cultures of Crete, Mycenae, and even Egypt. The figures of the 7th and 6th centuries B.C. lack life and movement; their faces wear the frozen smile peculiar to archaic sculpture. even so, these early craftsmen, whose names are lost with the temples they decorated, show sensitivity to the qualities of marble and a superb sense of design. As if to make up for the lack of life in their statues, archaic sculptors sought naturalism by painting them. Greek sculpture rose to its highest achievement in the 5th century B.C, when the spirit of Greece itself was at its height.

3.3 Results

Figures 3 and 4 show the results of story picturing applied to the stories on Paris and Vermont respectively. Figure 5 shows the results for stories on Egyptian and Greek sculptures. We show only 9 images in Figures 3 and 4 and 8 images in Figure 5 in order to save space.

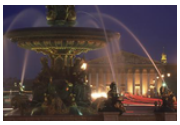
We believe that a photographer's archives are governed by his tastes. In a discussion of Paris, if pictures of *Arc de Triomphe* get higher reinforcement compared to pictures of *Eiffel tower*, it might reflect that the photographer was more enamored by the former than the latter or alternately, the pictures of *Eiffel tower* were taken in the night or were of low quality. In Figures 3 and 4, we notice that the low

H_1. Small park on the Ile de la Cite Paris, France.

H_2. Basin in Louvre, Notre Dame cathedral and the Ile de la Cite

H_3. Notre Dame cathedral and the Ile de la Cite seen from Ile Saint Louis. Paris, France.

H_5. Notre Dame seen from the Montparnasse Tower, sunset. Paris, France.



H_4. Fountain on Place de la Concorde and Assemblee Nationale by night. Paris, France.

H_6. Walking on the banks of the Seine on the Saint-Louis island. Paris, France.

L_1. Ecclesiastic clock inside the Notre Dame cathedral. Strasbourg, Alsace, France.

L_2. Backpackers being picked up by a floatplane at Circle lake. Gates of the Arctic National Park, Alaska, USA.

L_3. Astrological clock inside the Notre Dame cathedral. Strasbourg, Alsace, France.

Figure 3: Pictures provided by the Story Picturing Engine to illustrate the story on Paris. H_1-H_6 are the 6 top ranked images while L_1-L_3 are the 3 lowest ranked images. The parameter α is 0.5.

ranked images, which received low reinforcement are either odd, irrelevant, or of poor quality.

Notice that in Figure 5, the picture (E_2) is irrelevant to the story which was included in the pool by our scheme as its painter is Egyptian. Figure 6 shows the two top ranked images for different values of parameter α for the story on Roman art. The parameter α, as described in Section 2.3.1, balances the effects of lexical and visual features in determining the reinforcement ranks. If the value of α is 0, visual features play no role towards the rank of an image while if the value of α is 1, lexical features are unimportant. An intermediate value of α such as 0.5 controls the effects of both. We can see from Figure 6 that in the case when only lexical features are dominant (i.e., A_1, A_2), images, which are visually not so appealing, may receive high ranks. However, on the other hand, when only visual features are used, image ranks are completely decided by color, texture and shape. An image which is distantly relevant to a story but has an appealing visual content may receive a high rank. By inspection, the readers may see that the top ranked images are mostly relevant to the stories and have a fairly diverse visual content. More results can be viewed at our Website [4].

For all the results shown, we used the constraint $\sum_{j=1}^{N} s_{ij} = 1\ \forall i$. The CPU times taken to perform different parts of the experiments are listed in Table 1. All the experiments were performed on a 2.6 GHz Xeon processor running Linux.

Story	T_1	M	T_2	T_3
Paris	103s	136	16s	11s
Vermont	11s	38	5s	0.08s
Roman Art	199s	29	0.48s	0.05s
Egyptian Sculpture	362s	202	3s	62s
Greek Sculpture	2031s	556	10s	3415s

Table 1: Timing study for experiments on AMICO image database. T_1 = Time for Story Processing, image selection and estimation of lexical similarity, M = Image pool size as obtained by selection described in section 2.2, T_2 = Time for IRM distances calculation, T_3 = Time for calculation of ranks.

[4] `http://wang.ist.psu.edu/~djoshi/storypicturing`

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a scheme for automated story picturing. Stopword elimination was performed and the text of the story was processed using the Wordnet to form a list of keywords. A set of proper nouns was also identified. An initial image pool was formed by searching an annotated image database using the lexical descriptors of the passage (keywords and proper nouns). An image ranking algorithm was applied to determine mutual reinforcement based rank for each image. The few highest ranked images are believed to be strong representatives of the ideas conveyed by the story. An interesting human behavior model was presented according to which image selection process becomes a stochastic random walk over a graph of images as nodes with the normalized image similarities acting as edges. Performance of the algorithm on sample stories was shown in this paper.

In the present implementation, the mutual reinforcement ranks are calculated in real time. We would like to build a system in which ranks are pre-calculated using the entire image database. Image retrieval would then amount to searching images with high ranks with annotation text matching the text in lexical descriptors of the story. As a result, the time for real time story picturing is expected to be drastically reduced.

One of the reviewers pointed out that many images could receive high ranks in the event they are all very similar to each other. In this case, mutual reinforcement ranking could be unfair if some important concepts in a story are not well represented in pictures. A post ranking image clustering could be performed and few representatives from every cluster could be selected in order to break the monopoly of many high ranked images. In the future, sequence of taken time of photographs can be aligned with the story text as well.

We are exploring benchmarks and preparing a user survey to evaluate story picturing results. A Turing test based evaluation system will be very useful and we wish to take the help of human evaluators in this respect. We plan to integrate several image databases together and build an online system which can accept stories given by users. Once the integration is achieved and an offline rank calculation system is built, our Story Picturing Engine can potentially be used by teachers and students alike.

H_1. River with trees in autumn color. Vermont, New England, USA.

H_2. Birch trees. Vermont, New England, USA.

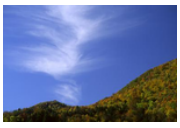
H_3. Birch trees and sugar house in Reading. Vermont, New England, USA.

H_4. Village surrounded by trees in brilliant autumn foliage. Vermont, New England, USA.

H_5. Autumn reflexions, Green Mountains. Vermont, New England, USA.

H_6. East orange village in autumn. Vermont, New England, USA.

L_1. Bouquet of trees in fall foliage. Vermont, New England, USA.



L_2. Hills in fall colors and cloud, Green Mountains. Vermont, New England, USA.

L_3. Hills and cloud, Green Mountains. Vermont, New England, USA

Figure 4: Pictures provided by the Story Picturing Engine to illustrate the story on Vermont. H_1-H_6 are the 6 top ranked images while L_1-L_3 are the 3 lowest ranked images. The parameter α is 0.5.

E_1. Unfinished statue of a King, carved limestone, Artist not recorded, Egyptian, The Detroit Institute of Arts

E_2. Woman holding and poking a small stuffed animal, autochromes, Arnold Genthe, Egyptian, Library of Congress Prints and Photographs Division.

E_3. Statue of a seated woman, limestone, Artist not recorded, Egyptian, Museum of Fine Arts, Boston.

E_4. Seated statue of Akhmeretnesut, limestone, Artist not recorded, Egyptian, Museum of Fine Arts, Boston.

G_1. Head, marble, Amorgos, 5th century B.C., Greek, The Cleveland Museum of Art.

G_2. Gravestone in the form of a lekythos, marble, 4th century B.C., Greek; Athenian, The Cleveland Museum of Art.

G_3. Piping and dancing Satyr, bronze, 3rd-2nd centuries B.C., Greek, The Cleveland Museum of Art.

G_4. Figurine of standing woman, terracotta, 3rd century B.C., Greek, The Cleveland Museum of Art.

Figure 5: Pictures provided by the Story Picturing Engine to illustrate the stories on Egyptian and Greek sculptures. E_1-E_4 are results for story on Egyptian sculpture and G_1-G_4 are results for story on Greek sculpture. The parameter α is 0.5.

A_1. Fresco panel, fresco, Roman Empire, Museum of Fine Arts, Boston.

A_2. Fresco panel, fresco, Roman Empire, Museum of Fine Arts, Boston.

B_1. Allegory of Birth , after Giulio Romano's design for the fresco decoration of the Loggia Della Grotta in the Palazzo Te, Mantua, Giorgio Ghisi, Italian, Fine Arts Museums of San Francisco.

B_2. The triumph of two Roman Emperors, pl. 4 from a series of engravings after paintings by Polidor Da Caravaggio, Cherubino Alberti, Italian, Fine Arts Museums of San Francisco.

C_1. Plate with relief decoration, silver with gilding, Roman Empire, J. Paul Getty Museum.

C_2. The Snake Charmer, after the drawing by Giulio Romano for a fresco in the Sala Dei Venti, Italian, Fine Arts Museums of San Francisco.

Figure 6: The two top ranked images for the story on Roman art for $\alpha = 0$, 0.5 and 1 are shown here. For A_1-A_2, $\alpha = 0$, for B_1-B_2, $\alpha = 0.5$ and for C_1-C_2, $\alpha = 1$.

Acknowledgments

This work is supported by the US National Science Foundation under Grant Nos. IIS-0219272, IIS-0347148, and ANI-0202007, The Pennsylvania State University, the PNC Foundation, and SUN Microsystems under grants EDUD-7824-010456-US and EDUD-7824-030469-US. Discussions with Ruqian Lu have been helpful. The authors would like to thank Q.-T. Luong and J. Trant for providing the image databases used in the project.

5. REFERENCES

[1] M. Agosti, F. Crestani and G. Pasi, *Lectures on Information Retrieval,* Lecture Notes in Computer Science, vol. 1980, Springer-Verlag, Germany, 2000.

[2] K. Barnard and D. Forsyth, "Learning the Semantics of Words and Pictures," *Proc. Int. Conf. on Computer Vision*, pp. 408–415, July 2001.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. -de. Freitas, D. M. Blei, and M. I. Jordan, "Matching Words and Pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[4] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. Seventh Int. World Wide Web Conf.*, pp 107–117, April 1998.

[5] D. C. Brown and B. Chandrasekaran, "Design Considerations for Picture Production in a Natural Language Graphics System," *ACM SIGGRAPH Computer Graphics.*, vol. 15, no. 2, pp. 174–207, 1981.

[6] A. Budanitsky and G. Hirst, "Semantic Distance in Wordnet: An Experimental, Application-Oriented Evaluation of Five Measures," *Workshop on WordNet and Other Lexical Resources, NAACL*, June 2001.

[7] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Color and Texture-Based Image Segmentation using EM and its Application to Image Querying and Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026–1038, 2002.

[8] C.-c. Chen, H. Wactlar, J. Z. Wang, and K. Kiernan, "Digital Imagery for Significant Cultural and Historical Materials - An Emerging Research Field Bridging People, Culture, and Technologies," *International Journal on Digital Libraries*, 2004, accepted.

[9] Y. Chen, J. Z. Wang, and R. Krovetz, "CLUE: Cluster-based Retrieval of Images by Unsupervised Learning," *IEEE Transactions on Image Processing*, vol. 13, no. 15, pp. 2004, accepted.

[10] S. R. Clay and J. Wilhelms, "Put: Language-Based Interactive Manipulation of Objects,", *IEEE Computer Graphics and Applications*, vol. 16, no. 2, pp 31–39, 1996.

[11] B. Coyne and R. Sproat, "WordsEye: An Automatic Text-to-Scene Conversion System,", *Proc. 28th Annual Conf. on Computer Graphics and Interactive Techniques*, pp 487–496, August 2001.

[12] C. Fellbaum, *WordNet - An electronic lexical database*, MIT Press, Cambridge, Massachusetts and London, England, 1998.

[13] E. Garfield, "Citation Analysis as a Tool in Journal Evaluation," *Science*, vol. 178, pp. 471–479, 1972.

[14] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[15] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.

[16] J. Li and J. Z. Wang, "Studying Digital Imagery of Ancient Paintings by Mixtures of Stochastic Models," *IEEE Transactions on Image Processing*, vol. 13, no. 3, pp. 340–353, 2004.

[17] L. Li, Y. Shang, and W. Zhang, "Improvement of HITS-based Algorithms on Web Documents," *Proc. Eleventh Int. World Wide Web Conf.*, pp 527–535, 2002.

[18] R. Lu and S. Zhang, *Automatic Generation of Computer Animation*, Lecture Notes in Artificial Intelligence, vol. 2160, Springer-Verlag, Germany, 2002.

[19] W. Y. Ma and B. S. Manjunath, "NeTra: A Toolbox for Navigating Large Image Databases," *Multimedia Systems*, vol. 7, no. 3, pp. 184–198, 1999.

[20] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: An On-line Lexical Database," *Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[21] G. Pinski and F. Narin, "Citation Influence for Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics," *Information Processing and Management*, vol. 12, pp. 297–312, 1976.

[22] R. Simmons and G. Novak, "Semantically Analyzing an English Subset for the CLOWNS Microworld," *American Journal of Computational Linguistics*, Microfiche 18. 1975.

[23] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[24] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947–963, 2001.