International Journal on Digital Libraries (2006) 6(1): 18–29
DOI 10.1007/s00799-005-0121-4

REGULAR PAPER

James Z. Wang · Kurt Grieb · Ya Zhang · Ching-chih Chen · Yixin Chen · Jia Li

Machine annotation and retrieval for digital imagery of historical materials

Received: 20 May 2004 / Published online: 23 February 2006

Abstract Annotating digital imagery of historical materials for the purpose of computer-based retrieval is a labor-intensive task for many historians and digital collection managers. We have explored the possibilities of automated annotation and retrieval of images from collections of art and cultural images. In this paper, we introduce the application of the ALIP (Automatic Linguistic Indexing of Pictures) system, developed at Penn State, to the problem of machine-assisted annotation of images of historical materials. The ALIP system learns the expertise of a human annotator on the basis of a small collection of annotated representative images. The learned knowledge about the domain-specific concepts is stored as a dictionary of statistical models in a computer-based knowledge base. When an un-annotated image is presented to ALIP, the system computes the statistical likelihood of the image resembling each of the learned statistical models and the best concept is selected to annotate the image. Experimental results, obtained using the Emperor image collection of the *Chinese Memory Net* project, are reported and discussed. The system has been trained using subsets of images and metadata from the Emperor collection. Finally, we introduce an integration of wavelet-based annotation and wavelet-based progressive displaying of very high resolution copyright-protected images.

Keywords Content-based image retrieval · Statistical modeling · Hidden Markov models · Image annotation · Machine learning

A preliminary version of this work has been presented at the *DELOS-NSF Workshop on Multimedia in Digital Libraries*, Crete, Greece, June 2003. The work was completed when Kurt Grieb and Ya Zhang were students of The Pennsylvania State University. James Z. Wang and Jia Li are also affiliated with Department of Computer Science and Engineering, The Pennsylvania State University. Yixin Chen is also with the Research Institute for Children, Children's Hospital, New Orleans.

J. Z. Wang(✉)
School of Information Sciences and Technology,
The Pennsylvania State University, University Park, PA 16802, USA
E-mail: jwang@ist.psu.edu (http://wang.ist.psu.edu)

K. Grieb
Lockheed Martin Corporation, Philadelphia, PA, USA

Y. Zhang
Department of Electrical Engineering and Computer Science,
The University of Kansas, Lawrence, KS 66045, USA

C.-c. Chen
Graduate School of Library and Information Science,
Simmons College, Boston, MA 02115, USA

Y. Chen
Department of Computer Science, University of New Orleans,
New Orleans, LA 70148, USA

J. Li
Department of Statistics, The Pennsylvania State University,
University Park, PA 16802, USA

Present address:
Information Sciences and Technology Building, The Pennsylvania State University, University Park, PA 16802, USA

1 Introduction

Annotating digital imagery of historical materials is labor-intensive. Typically, a well-trained human annotator must go through individual images and type in keywords or linguistic descriptions. As these image databases grow larger and larger, it is becoming prohibitively expensive to annotate these images manually. In our work, we attempt to study whether it is possible for computers to learn the expertise from historians and use the learned knowledge to annotate collections of images automatically and linguistically.

Digital images of the same type of historical objects often share similar appearances. We hypothesize that a computer program can learn some of the human expertise based on the image data and some sample annotations. For instance, the First Emperor of China, Qin Shi Huang Di (259–210 BC), has his mausoleums surrounded by more than 7,000 life-sized terra-cotta warriors and horses. Each warrior is unique, being modeled after an actual guard-of-honor. It should be possible to train computers with the concept of "Terracotta Warriors and Horses" with a few sample images and let computers annotate other such images in the collection automatically. An human annotator can then review

Fig. 1 Sample images from the Emperor collection and their manually-created image titles (copyright of the original images belongs to C.-c. Chen)

the machine-generated annotations and make modifications if necessary. C.-c. Chen of Simmon College, a member of our project team, created extensive documentary using the historical materials related to the First Emperor of China [1] and created *The First Emperor of China* image collection. The extensive image collection was further expanded as a part of this *Chinese Memory Net*[1] image databases since 2000 [2]. Figure 1 shows some examples of this image collection and the manually-prepared image titles. Chen's project staff took thousands of such photos and manually created extensive metadata including keywords and descriptive annotations. Their effort has enabled computer scientists to study the suitability of applying state-of-the-art machine learning techniques to images of historical materials.

1.1 Content-based image retrieval

Content-based image retrieval techniques allow users to search for visually similar images within an image collection by comparing and analyzing numerical features extracted from images [3, 4]. In our prior work, we successfully applied the SIMPLIcity content-based image retrieval system [5] to the Emperor collection so that users can search for images based not only on annotations but also on visual similarity (Fig. 2) [6]. We have also shown that text-based retrieval techniques can help users to find images of historical materials if these images are annotated (Fig. 3).

Machine annotation of images is generally considered impossible because of the great difficulties in converting the structure and the content of a digital image into linguistic terms. In our work, we explore the use of machine learning and statistical modeling techniques in automated image annotation.

The field of content-based image retrieval has been evolving rapidly. Readers are referred to some recent review articles [3, 7], monographs [4, 8], and additional references [5, 9–19] for more information.

1.2 Automatic linguistic indexing of pictures

Since year 2000, the Penn State research team has been developing machine learning and statistical modeling-based techniques for image annotation and image retrieval [8, 19]. The Automatic Linguistic Indexing of Pictures (ALIP) system has been developed to learn objects and concepts through image-based training. The system was inspired by the fact that a human being can recognize objects and concepts through matching a visual scene with the knowledge structure stored in the brain. For instance, even a 3-year old child is typically capable of recognizing a number of concepts or objects.

The ALIP system builds a knowledge base about different concepts automatically from training images. Statistical models are created about individual concepts by analyzing a set of features extracted from training images using wavelet transforms [4]. A *dictionary* of these models is stored in the memory of the computer system and used in the recognition process. The team has conducted large-scale learning experiments using general-purpose photographic images representing 600 different concepts. In the published work, it has been demonstrated that the ALIP system with 2-D multiresolution hidden Markov models (2-D MHMM) [20] is capable of annotating new images with keywords after being trained with these concepts [19].

Since late 2002, researchers at Penn State have been collaborating with C.-c. Chen of Simmons College on the application of ALIP to the problem of annotating digital imagery of historical materials. We attempt to determine if ALIP can learn about domain-specific knowledge, i.e., the basis of the expert annotations of images. The Emperor image collection is suitable for this task because of both the high quality of the images and the comprehensiveness of the metadata descriptions.

[1] The project is now named as *Global Memory Net*. URL: `http://www.memorynet.org`

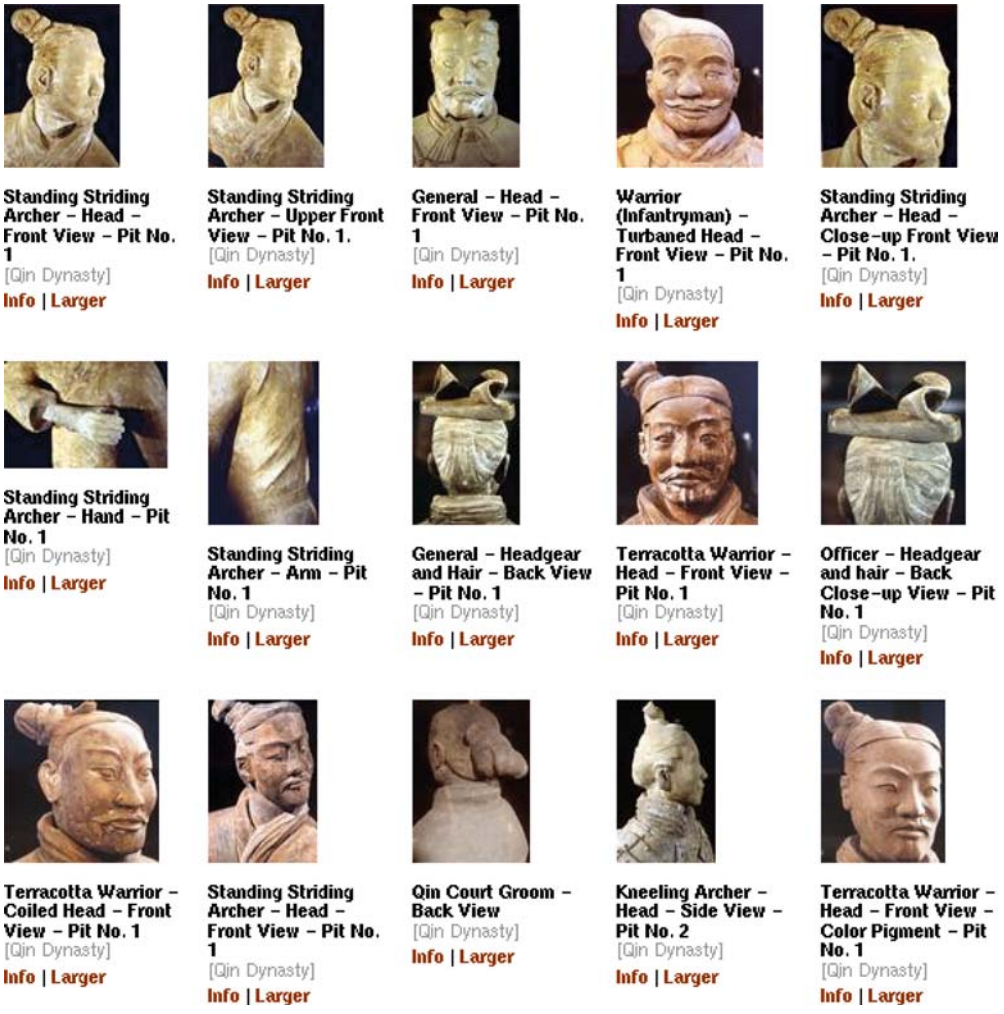

Fig. 2 Textual annotations can be integrated with content-based image retrieval engines. Similarity search results using the SIMPLIcity system are shown. The *top left* corner image is the query image. The images are ordered according to their visual similarity to the query image

1.3 Outline of the paper

In the remainder of this paper, we present our machine annotation approach, our integrated progressive displaying technique, and the experiments we have conducted. Specifically, Sect. 2 describes the training and annotation process. In Sect. 3, we introduce the integrated wavelet-based progressive image displaying technique for very high resolution copyright-protected images. In Sect. 4, we present the results of our extended experiments. Discussions on the limitations of the approach are included in Sect. 5. We present our conclusions and suggest future research directions in Sect. 6.

2 Training the ALIP system to annotate images of historical materials

We now review the ALIP (Automatic Linguistic Indexing of Pictures) system and introduce its application to the problem of annotating images of historical materials. The ALIP system [19] has three major components: wavelet-based feature extraction, model-based training, and statistical likelihood estimation for annotation.

2.1 The ALIP technology

The first process of the ALIP system is the model-based learning process. Before we can use the ALIP system to annotate any images of historical materials, we must train the system about the domain. For each concept, we need to prepare a set of training images. Ideally, these training images should be representative to the concept. For example, if we would like to train the concept "Roof Tile-End", we need to use images of different roof tile-ends rather than different images of the same roof tile-end.

For each training image, we extract localized features using wavelet transforms. An image is first partitioned into small pixel blocks. The block size can vary depending on the resolution of images in the collection and the subject of the collection. The block size is chosen to be 4×4 in our

Fig. 3 Textual annotations of images can be used in keyword-based search. Top 10 search results for the search phrase (`+"terracotta warrior" +side -front`) are shown

experiments as a compromise between the texture detail and the computation time. Other similar block sizes may also be used. The system extracts a feature vector of six dimensions from each block. Three of these features are the average color components of pixels in the block. The other three are texture features representing energy in high frequency bands of wavelet transforms. Specifically, each of the three features is the square root of the second-order moment of wavelet coefficients in one of the three high frequency bands. The features are extracted using the LUV color space, where L encodes luminance, and U and V encode color information (chrominance). The LUV color space is chosen because of its good perception correlation properties.

We manually prepare a training database of concepts, each with a small collection of images representing the concept. The system is capable of handling different number of training images per concept. The images are typically stored in JPEG format, while our system can process virtually any standard-image format. It is not required that the training images for a concept must all be visually similar. However, intuitively, the more diverse a concept is, the more training images can be required to obtain a reasonable training of the system. If it takes many examples to train a human about a concept, it would take even more images to train a computer system.

Figure 4 illustrates the training process of the system. For each concept, we estimate a 2-D MHMM [20] based on the wavelet-based features extracted from the training images. These models, stored in computer memory, are used during the annotation process. A 2-D MHMM captures both the inter-scale and intra-scale statistical dependence among training images. The inter-scale dependence is modeled by the Markov chain over resolutions. The intra-scale dependence is modeled by the HMM. At the coarsest resolution, feature vectors are assumed to be generated by a 2-D HMM. At all the higher resolutions, feature vectors of sibling blocks are also assumed to be generated by 2-D HMMs. The HMMs vary according to the states of parent blocks. The 2-D MHMM can be estimated by the maximum likelihood criterion using the EM algorithm.

In the annotation process, we first extract a collection of feature vectors at multiple resolutions from the image. The technique for extracting the features is the same as the technique used in the training process. We regard the features of an image as an instance of a stochastic process defined on a multiresolution grid. The similarity between the image and a concept of images in the database is assessed by the log likelihood of this instance under the model trained from images in the concept. A recursive algorithm [20] is used to compute the log likelihood. After determining the log likelihood of the image belonging to any concept, we sort the log likelihoods to find a concept with the highest likelihood.

2.2 Initial experiment

To verify the feasibility of training technique for images of historical materials, we first present an initial experiment with five concepts from the Emperor collection. The five concepts we have chosen are: (1) Terracotta Warriors and Horses, (2) The Great Wall, (3) Roof Tile-End, (4) Terracotta Warrior—Head, and (5) Afang Palace—Painting.

Figure 5 shows all the images used to train two of the five concepts. For the concept of "Terracotta Warriors and Horses," a total of only eight images are used to train the ALIP system. For the concept of "Roof Tile-End," a total of

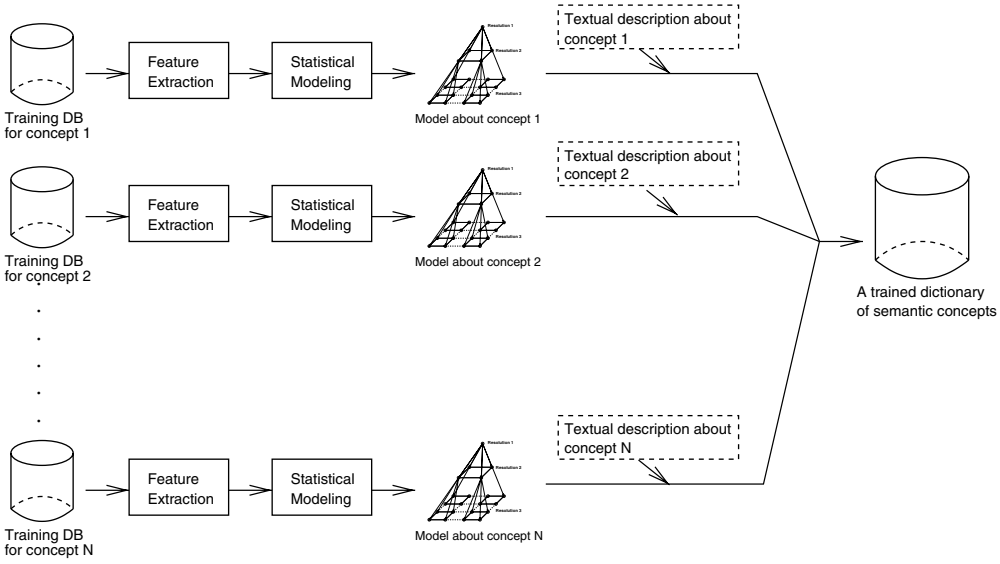

Fig. 4 The training process of the ALIP system

four images are used. We purposely use a minimum set of training images for each concept because the system can be useful in practice only if it does not require labor-intensive preparation of training images. One of the key advantages of the ALIP approach is that training images for a concept are not required to be all visually similar. The system is highly scalable because the training of one concept does not involve images related to other concepts. If we need to add another concept to the collection of concepts, we simply need to train a new model based on the training images to this new concept. If more images are added to the training collection for a given concept, only that concept needs to be retrained.

Because of the small number of training images per concept, only a couple of minutes of CPU time are required to train a concept on a Pentium PC running LINUX operating system. The training process is parallelizable because the training of one concept requires only images related to that concept.

To validate the effectiveness of the ALIP training, we tested ALIP with other Emperor images related to the five trained concepts. In another word, we give the ALIP system an examination on its learning progress.

It takes the computer only a few seconds to compute the statistical likelihoods of an image to all five learned concepts and sort the results. The concept with the highest likelihood is used to annotate the image. The experimental results are summarized as follows:

1. *Terracotta Warriors and Horses*: A total of 52 images were tested. We tested all non-training images of this concept available in the Emperor collection. Only one image was mistakenly annotated as "The Great Wall". The accuracy for this concept is 98%.
2. *The Great Wall*: A total of 65 non-training images were tested. Again, we used all available images. Only one image was mistakenly annotated as "Terracotta Warriors and Horses". The accuracy for this concept is 98%.
3. *Roof Tile-End*: A total of 28 available non-training images were tested. Three images were mistakenly annotated as "The Great Wall". Two images were marked as "Terracotta Warriors and Horses". The accuracy for this concept is 82%.
4. *Terracotta Warrior—Head*: A total of 57 available non-training images were tested. Two images were mistakenly annotated as "The Great Wall". The accuracy for this concept is 96%.
5. *Afang Palace—Painting*: A total of 33 available non-training images were tested. Six images were mistakenly annotated as "The Great Wall". The accuracy for this concept is 82%.

Fig. 5 All the images used to train the ALIP system to recognize the concepts "Terracotta Warriors and Horses" and "Roof Tile-End"

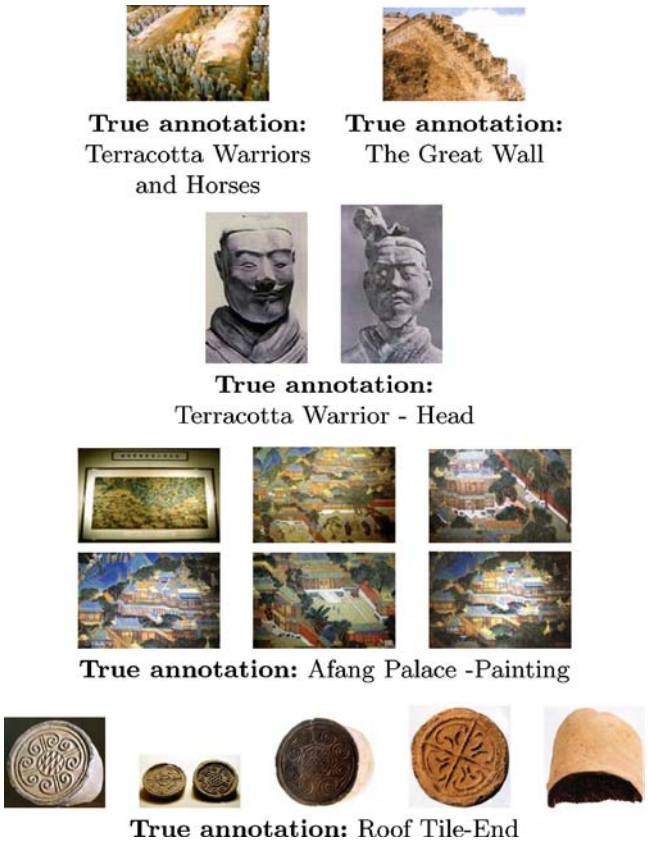

Fig. 6 ALIP mistakenly marked these images. The manually-created annotations are shown in the figure

This result is remarkable considering the small number of images we have provided to the system as training examples. Figure 6 shows all those images mistakenly marked by the ALIP system. In this initial experiment, we used only 4–8 images to train each concept. We expect the performance to improve if more images are used for training or if the training images are selected more carefully.

3 Wavelet-based progressive displaying and copyright protection

Copyright concerns are significant obstacles that have prevented the wide use of the Internet for distributing high-valued and high-resolution images of historical materials. We developed a technique on the basis of wavelet transforms to enable both progressive displaying and copyright protection. In this section, we briefly introduce the coding and decoding algorithms.

Some commonly-used existing methods such as user authentication and whole-image watermarking [21] are not sufficient. For instance, invisible watermarking techniques are not robust enough to prevent people from illegal copying. Typical whole-image visible watermarking techniques are ineffective or distractive for images of very high resolutions. Our approach aims at alleviating these problems.

We developed a wavelet-based progressive displaying method with dynamic watermarking for viewing very high resolution copyright-protected images. The encoder, which *dynamically* determines levels of transform and partition of coefficients, is based on a modified Haar wavelet transform. The decoder retrieves the necessary data and reconstructs the requested region-of-interest at a scale specified by the user. The system enables virtually any size of images to be displayed progressively. The system has low computational complexity for both the encoding and the decoding processes.

The core algorithm of the encoder is the modified Haar wavelet transform, which allows all computation to be done with integer arithmetic. Integer computation is desired because it saves storage space. The level of the transform and partition of coefficients are dynamically determined on the basis of the spatial resolution of the image to be processed.

The modified Haar wavelet transform is constructed by ignoring the scaling factor of Haar wavelet transform. Thus the coefficients are sums and differences of adjacent pixels. The encoder performs encoding only once when an image is added to the collection. We assume that the images are in the RGB color space, consisting of the red component, green component, and the blue component. Each color component image is encoded by the modified Haar wavelet transform and stored separately. The level of transform, s, can be calculated with $s = \log_2(\max(m, n)/k)$, where m is the height of the image and n is the width of the image, both expressed in numbers of pixels, and k is a parameter selected manually. By setting the parameter k, we enforce that coefficient files are smaller than a certain size for fast online access of large images. In our experiments, we set k to 100. Depending on the speed of the computer server systems and the desired processing speed, the parameter can vary.

After determining the transform level s, the right and bottom borders of the image need to be padded according to the transform level so that the size of the image is suitable to the level of the transform applied. The way that these borders are padded is not important because the padded borders are not to be displayed to the users. After the padding, we apply the s-level transform to each of the three gray-scale images of a color image.

The transform algorithm is lossless in nature. In order to save the storage when fully lossless is not critical, the resulting transform coefficients are quantized in the highest frequency. To make the whole process to be of low loss, we only quantize the highest frequency to fit in the scale of -128 to 128, which causes a loss of 2 out of a range of 512. Because the coefficient files for higher frequencies can be too large for fast retrieval of the region-of-interest, we partition these coefficient files spatially and store them in separate files. The partition of the coefficient files is dynamically determined by $d_c = n/(2^{(s-k)}t)$ and $d_r = m/(2^{(s-k)}t)$, where d_c and d_r are the numbers of division on the original coefficients by column and row, respectively, s is the total level of the transforms, k is the current transform level, and t is an adjustable parameter. t determines the size of

Fig. 7 A user can progressively zoom in the region of interest. The requested region at the requested resolution is dynamically generated from wavelet coefficients. A light visible watermark is added at the lower right corner of the image before transmission

individual coefficient files. We constrain that each partitioned coefficient file is no larger than a certain size. In our experiments, we set t to be 1,000.

Wavelet transform decomposes an image into sums and differences of adjacent pixels. For smooth areas, the difference elements are near zero. Huffman coding is employed to further compress the coefficient files and then the compressed files are stored. When a user request comes in, the decoder retrieves all necessary coefficient files and reconstructs the image on the fly.

Wavelet transform preserves the localization of data, which means that if a user queries a small contiguous portion of a whole image, the decoder only need to retrieve the coefficients corresponding to the queried region. Given a query with x-location, y-location, and scale variable, the decoder searches the stored representation files of the image for the related data files, decompresses them and finds the necessary coefficients to reconstruct the block of pixels at the specified scale. The decoding process is the exact reverse of the encoding process. Because of the quantization of the highest-level coefficients, the reconstructed image has a small loss. We have implemented a Web-based user interface that allows users to magnify any portion of the images in different resolution scales. Figure 7 shows the some screen shots. The reconstructed image is converted to a Web-image format, such as JPEG or PNG, and sent to the client/user through the Internet.

There are several advantages of this wavelet-based approach: (1) the image of the region-of-interest is dynamically generated to accurately answer to the user selection; (2) on the basis of different levels of controlled usage, different user groups can be given permissions to different resolutions; (3) users cannot easily reconstruct the original high-resolution images; and (4) non-reversible visible watermarks can be added to all the retrieved regions.

4 Extensive annotation experiments

Because of the promising results from the initial five-concept annotation experiment described above, we conducted additional tests on this database of images. We would like to find out the performance of ALIP when additional variables were introduced to the equation. In this section, we describe the set up of our experiments and report the results.

4.1 Design of the experiments

The Emperor database has about 2,000 images. It has about 100 different major concepts, though the textual description of images in the same major concept may vary significantly. That is, on an average, there are 20 images in a major concept. Because we need to select many sets of training and testing images from each concept, we need to choose only those concepts with larger number of images. We have only about 20–30 concepts to use.

We ran two additional sets of tests on the Emperor image collection. The first set contained nine general training concepts which included: (1) Black and White Sketches, (2) The Great Wall, (3) Horses, (4) Pots and Plates, (5) Roof Tile-Ends, (6) Color Paintings, (7) Maps, (8) Terracotta Warriors Faces, and (9) Soldiers. It is important to note that for the Terracotta Warrior Faces concept, both color and gray-scale faces were used. The reason is because we wanted this set of concepts to have more general image concepts such as Soldiers, Horses, and Faces. Figure 8 shows some sample training images from the nine concepts.

In the second set of experiments, we broke apart these concepts into more specific concepts to determine how ALIP performed with overlapping image concepts, and to see how well ALIP could identify specific features of a more general concept. The second set of testing included all of the above concepts, with certain concepts divided into two separate concepts. The additional concepts were: (10) Stones, (11) Text, (12) Hands, (8) Terracotta Warrior Faces being broken down into (13) Black and White Faces and (14) Color Faces, (15) Upper Bodies, (16) Full Bodies, (17) Two Heads, (18) Feet, (3) Horses being broken down into (19) Horse

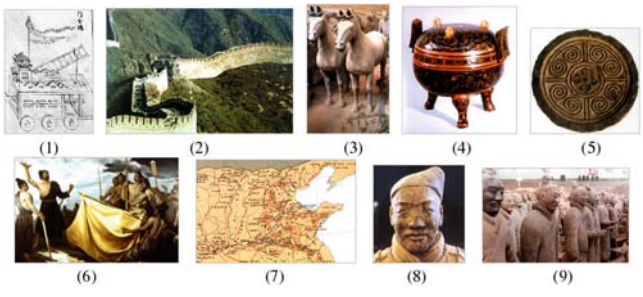

Fig. 8 Sample training images of the first nine concepts

Fig. 9 Sample training images of the second set of 11 concepts

Heads and (20) Horse Bodies. There are slightly less desirable concepts in this set because certain concepts had a more limited amount of total pictures in the original Emperor image collection. Figure 9 shows some sample training images from this set of 11 concepts.

Both sets were broken down into four subsets. Each subset had different sets of training images. The first subset's training images were chosen by our lab. We attempted to give ALIP the best possible representation of each concept with the pictures we chose. We did this by providing ALIP a diverse set of pictures within each concept. The second subset was also chosen by the lab. In this subset, we attempted to give ALIP the worst possible set of training images. Here we chose very similar pictures with little diversity, hence giving a rather centered and poor representation of the concept. The final two subset's training images were randomly chosen.

We further broke down the subsets into four different training image dataset sizes. We began with three images, and for each additional size, we simply added three more images to the training set. Therefore, we had training image sizes of 3, 6, 9, and 12 for each of the four subsets of the two sets of experiments. In other words, we ran a total of 32 tests on the Emperor image collection.

4.2 Goals and reasoning

The above experiments were designed specifically to discover certain characteristics of ALIP. We chose two sets of training concepts to determine how ALIP's performance is affected by more concepts and also by more specific concepts. Here, we expected the performance of ALIP to decrease with a larger number of more complicated concepts.

We chose four subsets of concepts so that we could track the performance of ALIP when given different training images. We expected that different training images would lead to different levels of performance. Additionally, we expected that our hand-picked representative training images would give better results than our hand-picked poorly representative training images. We also expected that the two randomly chosen training image subsets would give results somewhere between the two hand-picked results.

Finally, we wanted to determine what a reasonable training image size would be. We expected that more training images would lead to better ALIP performance. However, we did not want to use too many training images, because that would defeat the purpose of the ALIP work. Our goal is to automate as much of the identification process as possible, which in turn requires minimizing human input.

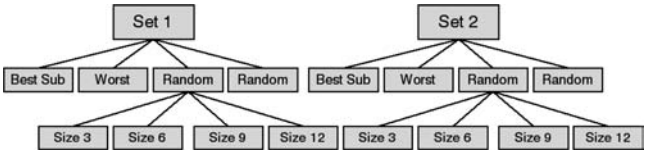

Fig. 10 Breakdown of 32 tests we ran on the Emperor database

4.3 Results

Overall, ALIP's success rate was promising. It ranged from as low as 36% in the absolute worst case scenario, to close to 75% in the best case scenarios. Most runs had over a 50% success rate. Since there have been no previous precedents set with computer-based image recognition for images of historical materials, we compare our results with a random-based classification scheme.

4.3.1 Results of the first set

Results from the first set of nine concepts were very good. Success rates here were from a low of about 42% to a high of 74%. Figure 11 shows the general trends of the testing.

As expected, the hand-picked best case test did perform, on average, better than any of the other subsets. Also, the hand-picked worst case test performed worse than any other subsets. The two random subsets were generally between the best and worst case scenarios.

In terms of training image size, we see that there is an improvement of roughly 5% per three training images added in all the four subsets. This is not surprising because of the small number of training images we have used. We would not expect the trend to last as we add more and more images.

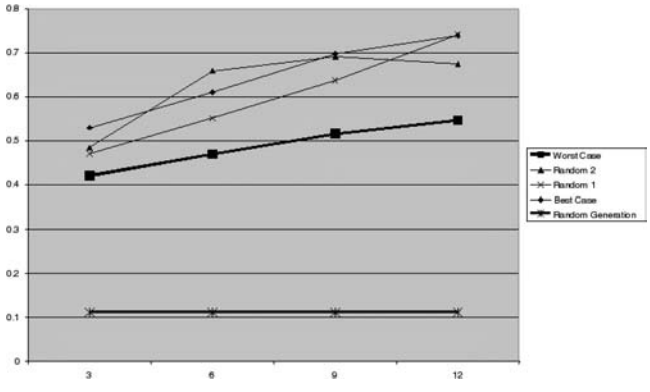

Fig. 11 Accuracy for the runs in the first set. *X*: sample size, *Y*: accuracy

We could not add more images to the training due to the size of the Emperor database.

We also see that all results are far better than the unintelligent random picking of any concept a picture may fall in. Here, since we have nine concepts of equal size, each image has an 11% chance of falling into any of the concepts.

4.3.2 Results of the second set

Results were also promising on the more complicated set. Subset accuracy ranged from 35 to 65%, slightly lower than the first set but also much better than the random selection success rate of 5%. Figure 12 shows the testing results with this set.

As with the first test, the hand-picked best case and worst case scenarios generally set the upper and lower boundaries for the percentage correct. Also, the percentage of images correctly identified again rose roughly 5% on an average for every three images added to the training pool. Again, we would not expect the trend to last as even more images are added.

4.3.3 Comparisons between the two sets

As shown above, the initial accuracy of the second set is roughly 10% lower than the initial accuracy of the first set. We hypothesized earlier that this would be the case, due to the increase of the number of concepts in addition to the added complexity of the concepts themselves. However, the accuracy increase with each of the sample sizes is roughly the same, at 5% per three pictures for both sets. This suggests that additional training can overcome the reduced accuracy of more complex image concepts.

4.4 Experimental findings

There are certain trends and features that became apparent during the testing, and we believe they are worthwhile to be noted. These features include extremely successful concepts, general trends, and reasoning why ALIP incorrectly identified certain images.

4.4.1 Black-and-white sketches

ALIP had the most success in identifying black-and-white sketches. The success rate for classifying these images was over 99% for all tests. This is remarkably high, compared to most other image concepts, especially for the smaller training samples. This is expected because black-and-white sketches look very different from photographic images.

4.4.2 Training size vs. accuracy

It appears that a larger set of training images leads to better accuracy. However, in certain cases, the accuracy occasionally went down after adding training images to each of the concepts. It is difficult to find out the actual reason for this behavior as there are so many factors. It can be due to the limitation of the ALIP approach. It could also be that a particular image added to the training group was not a good choice for training ALIP. Figure 13 shows one example of this behavior.

4.4.3 Training quality vs. accuracy

It appears when images are trained with the worst-case set of images, certain concepts perform much better than before. This is shown in Fig. 14 for the horses concept from the first set of experiments. The worst-case scenario performed better than all other runs in the horses concept, though it performed worse than the others in nearly every other concept. These results suggest that there can be performance trade-offs among different concepts in the training process.

4.4.4 Black-and-white images vs. textual images

Black-and-white graphical images and textual images (i.e., images with only text) are very similar visually. Both have a large amount of white space and very distinct boundaries where the black ink starts and the white paper stops. Also, in many images there is text describing the image, leading to a

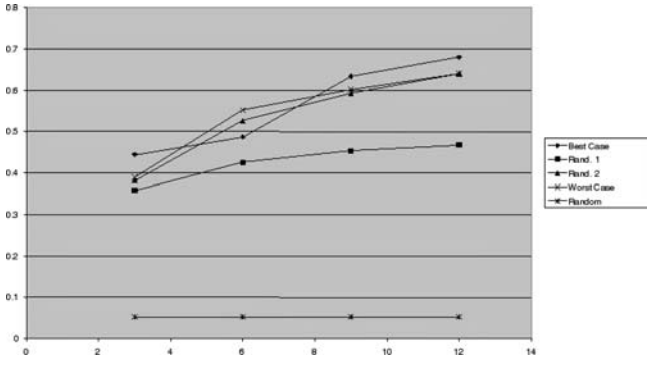

Fig. 12 Accuracy for the runs in the second set. *X*: sample size, *Y*: accuracy

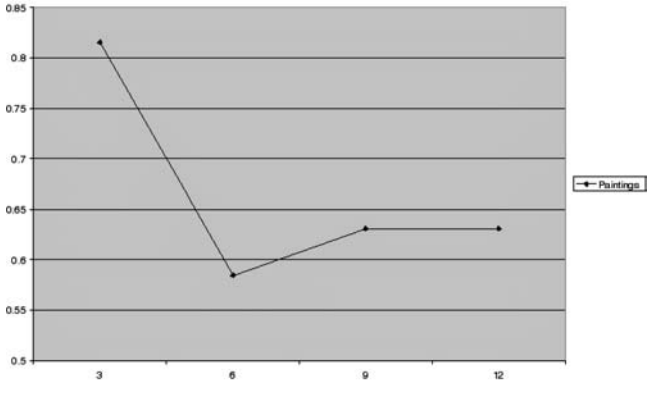

Fig. 13 Accuracy of classifying paintings in a subset. *X*: number of training images, *Y*: accuracy

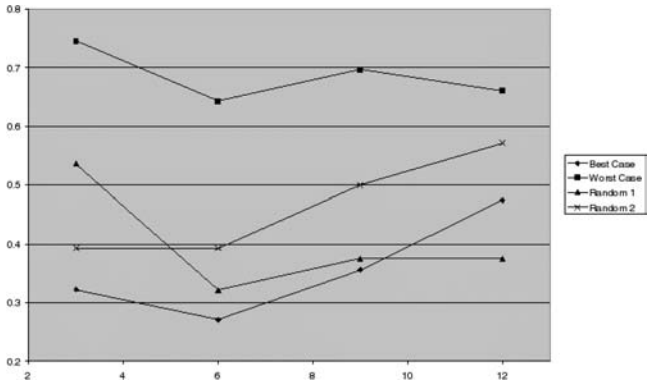

Fig. 14 Accuracy of classifying horses. *X*: number of training images, *Y*: accuracy

further meshing of these two concepts. Though ALIP did a fairly good job at correctly identifying these two concepts, most of the misclassified black-and-white images were incorrectly classified as textual images. These same images also had text in the foreground.

4.4.5 All faces vs. color or black-and-white faces

Our results suggest that combining both types of faces into one face concept requires less training for similar results. Admittedly, our experiment cannot conclusively determine this due to other changing variables, but the combined success of the second set's black-and-white and color faces are roughly equal to the success of the first set at most points. Other misclassifications also affected the success of the second set. Some of these stem from the blending of concepts as described below.

4.4.6 Overlapping semantics

Faces are predominantly featured in the upper body images of the Emperor image database. It is possible that ALIP was again thrown off by this blending of concepts. It appears that roughly half of all errors in the Color Faces, Black and White Faces, Upper Bodies, and Two Faces concepts were misclassified as one of the other three concepts. Again, depending on the perspective of the observer, these errors could in fact be deemed correct because they are returned in part what was asked for.

5 Discussions

If a trained human was asked to classify all the pictures of the Emperor database, the success rate of the human would probably be close to perfect. Humans still have many advantages over the training-based computer annotation systems such as ALIP. For instance, pictures of a certain object, a horse as an example, are taken from a variety of different angles (Fig. 15). Humans have developed a 3-D representation of a horse. If we see a picture of a back horse, we could still be able to correctly identify the image. It would be difficult for computer systems based only on 2-D image training to do this because they do not have 3-D representations of the horse. A computer program needs to first see a back of a horse and be told that the image it sees is the back of a horse, for it to correctly identify that image as a horse. Simply showing the head of a horse to ALIP is not enough to allow it to correctly identify other parts of a horse.

Currently, work is being done on creating a system that can analyze 3-D images [22]. The techniques to deal with them will be even more sophisticated and computationally intensive than the current 2-D techniques. However, this is the natural extension of the current 2-D systems and algorithms.

Additionally, ALIP needs to be given the flexibility to deal with image concepts that overlap. If there is a picture of a soldier with a horse (Fig. 16), should ALIP classify that picture as a horse or a soldier? What we are searching for will determine whether this picture is of relevance. If ALIP classifies this as a horse, but we are searching for soldiers, we may never see this picture when we query the database of images.

ALIP currently ranks all the concepts by their likelihood of being the home concept of the image it currently is analyzing. Developing some technique to correctly determine a group of concepts that may be related to a certain image would be useful. The system would have to be flexible enough to give the user a reasonable amount of good images without flooding the user with images that have nothing to do with the search.

The success rate of ALIP must be improved for it to be useful in more difficult situations. Here, we are dealing with rather simple images. However, an example of the eventual goal is to separate soldiers from different cultures into their respective cultures. Of course, ALIP would probably not perform well if given this task now. However, our results are

Fig. 15 Different aspects of the Terracotta Horses

Cavalryman and Cavalry Horse - Pit No. 2

Fig. 16 A terracotta soldier with a horse

promising, in that we have shown that content-based image annotation is possible with certain images of art or historical materials. Future systems, with more computational power and much better algorithms will be able to do what ALIP cannot currently do. Even though our success rates may not be as good as a human's success rate, they are high enough to suggest that further research in this area is warranted. ALIP is just one step in the realization of a highly successful and flexible content-based image annotation system for our ever growing digital image collections.

6 Conclusions and future work

In this paper, we demonstrated the application of the ALIP system to the problem of automatic annotation of digital imagery of historical materials. We used categorized images to train computers with semantic concepts. Wavelet-based features are used to describe local color and texture in the images. We have shown a wavelet-based progressive displaying and copyright protection techniques for very high-resolution and high-valued images. We conducted extensive experiments with the Emperor image collection. Promising results have been obtained and reported.

There are many possible future directions. Automatic linguistic indexing of pictures, as a research field, is just at its beginning. We envision that the annotation accuracy can be improved by integrating model-based learning with a rule-based system so that some human expertise can be incorporated. It can be interesting to study ways to improve the statistical modeling process to include human assistance or feedback. Three-dimensional image-based statistical modeling can be very challenging but useful. Finally, some intuitive user interfaces can be developed so that machine learning and statistical modeling-based annotation can be used in practice.

Acknowledgements This work is supported by the US National Science Foundation (NSF) under Grant nos. IIS-0219272 and ANI-0202007, The Pennsylvania State University, the PNC Foundation, and SUN Microsystems under grant EDUD-7824-010456-US. Emperor images collection provided by C.-c. Chen is a part of *PROJECT EMPEROR-I* supported by the Humanities in Libraries Program of the US National Endowment for the Humanities (NEH). This collection and the associated metadata is a part of her *Chinese Memory Net* (now expanded to *Global Memory Net*) project support by NSF/IDLP under Grant No. IIS-9905833. Conversations with Michael Lesk, Stephen Griffin, and members of the DELOS-NSF Working Group on Digital Imagery for Significant Cultural and Historical Materials have been very helpful.

References

1. Chen, C.-c., *The First Emperor of China*: interactive videodisc, the Voyager Company, 1991. Multimedia CD-ROM published in 1993. Result of PROJECT EMPEROR-I, supported by the US National Endowment for the Humanities
2. Chen, C.-c.: Chinese Memory Net (CMNet): A model for collaborative global digital library development. In: Chen, C.-c. (ed.) Global Digital Library in the New Millennium: Fertile Ground for Distributed Cross-Disciplinary Collaboration, pp. 21–32. Tsinghua University Press, Beijing (2001)
3. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
4. Wang, J.Z.: Integrated Region-Based Image Retrieval. Kluwer Academic Publishers, Dordrecht (2001)
5. Wang, J.Z., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001)
6. Wang, J.Z., Li, J., Chen, C.-c.: Interdisciplinary research to advance digital imagery indexing and retrieval technologies for Asian art and cultural heritages. In: Proceeding of the 4th International Workshop on Multimedia Information Retrieval, in conjunction with ACM Multimedia, Juan Les Pins, France, ACM, 6 pp. (2002)
7. Chen, C.-c., Wactlar, H., Wang, J.Z., Kiernan, K.: Digital imagery for significant cultural and historical materials: an emerging research field bridging people, culture, and technologies. Int. J. Digital Libr. Special Issue: Towards the New Generation Digital Libraries: Recommendations of the US-NSF/EU-DELOS Working Groups (2005)
8. Chen, Y., Li, J., Wang, J.Z.: Machine Learning and Statistical Modeling Approaches to Image Retrieval. Kluwer Academic Publishers, Dordrecht (2004)
9. Kriegman, D., Ponce, J.: On recognizing and positioning curved 3D objects from image contours. IEEE Trans. Pattern Anal. Mach. Intell. **12**(12), 1127–1137 (1990)
10. Dickinson, S., Pentland, A., Rosenfeld, A.: 3-D shape recovery using distributed aspect matching. IEEE Trans. Pattern Anal. Mach. Intell. **14**(2), 174–198 (1992)
11. Wactlar, H.D., Kanade, T., Smith, M.A., Stevens, S.M.: Intelligent access to digital video: Informedia project. IEEE Comp. **29**(3), 46–52 (1996)
12. Chandrasekaran, S., Manjunath, B.S., Wang, Y.F., Winkler, J., Zhang, H.: An eigenspace update algorithm for image analysis. Graph. Models Image Process. **59**(5), 321–332 (1997)
13. Chu, W.W., Hsu, C.C., Cardenas, A.F., Taira, R.K.: A knowledge-based image retrieval with spatial and temporal constructs. IEEE Trans. Knowledge Data Eng. **10**(6), 872–888 (1998)
14. Sheikholeslami, G., Chatterjee, S., Zhang, A.: WaveCluster: A multi-resolution clustering approach for very large spatial databases. In: Proceeding of the VLDB Conference, New York City, pp. 428–439 (1998)
15. Wang, J.Z., Wiederhold, G., Firschein, O., Sha, X.W.: Content-based image indexing and searching using Daubechies' wavelets. Int. J. Digital Libr.(IJODL) **1**(4), 311–328 (1998)
16. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.J.: Content-based hierarchical classification of vacation images. In: Proceeding of the IEEE International Conference on Multimedia Computing and Systems, ICMCS, Amsterdam, The Netherlands (1999)
17. Chen, Y., Wang, J.Z.: A region-based fuzzy feature matching approach to content-based image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **24**(9), 1252–1267 (2002)
18. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D.A., Blei, D.M., Jordan, M.I.: Matching words and pictures. J. Mach. Learn. Res. **3**, 1107–1135 (2003)

19. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1075–1088 (2003)
20. Li, J., Gray, R.M.: Image Segmentation and Compression Using Hidden Markov Models. Kluwer Academic Publishers, Dordrecht (2000)
21. Cox, I.J., Miller, M.L., Bloom, J.A.: Digital Watermarking. Morgan Kaufmann, San Francisco, CA (2002)
22. Li, J., Joshi, D., Wang, J.Z.: Stochastic modeling of volume images with a 3-D hidden Markov model. In: Proceeding of the IEEE International Conference on Image Processing, Singapore, IEEE, pp. 2359–2362 (2004)