A Dataset for Research of Traditional Chinese Painting and Calligraphy Images

Hong Bao¹,², Ye Liang³, Hong-zhe Liu¹,² and De Xu¹

¹: School of Computer and Information Technology, Beijing Jiaotong University
²: Institute of Information Technology, Beijing Union University, Beijing, China

Abstract—Traditional Chinese painting and calligraphy is a unique form of art and highly regarded for its theory, expression and techniques throughout the world. However, popular image processing techniques have little been used in this specific domain. Especially, there is no standardized image dataset up to now. Authors often use their own images or do not specify the source of their datasets in their papers. Naturally this makes comparison of results somewhat difficult. In this paper we introduce a new dataset, TCCPID -- a dataset of traditional Chinese painting and calligraphy images which tries to bridge this gap. The TCCPID dataset currently consists of 6400 images. In addition, the process of creating this dataset is dissertated in detail, including how to model traditional Chinese calligraphy and painting images based on ontology and how to use the system of the dataset. Our work is very meaningful to create a standardization image databases in specific Chinese art image domain.

Keywords- traditional Chinese painting and calligraphy; ontology; image dataset

1. Introduction

Chinese civilization has been around for thousands of years. Traditional Chinese calligraphy and painting is the gem of Chinese traditional arts and date back to the Neolithic Age, about 5000 years ago. At present, with the advances of multimedia and computing technology, the special old art form of traditional Chinese calligraphy and painting is spreading and getting more popular by these modern techniques. Many invaluable antique paintings have been digitized at a much higher resolution. On one hand, this makes the original copy being well preserved, on the other hand, allows a lot of people being able get access to view them. Various museums and more and more artists are constructing digital archives of art calligraphy and painting and preserve the original artifacts. Effective indexing, browsing and retrieving digitized art images become an important and imperative task to be solved. However, digital archives belonged to each museum are different greatly and there is not an integrated database which includes all digital archives which are approbated by authoritative institution. As a result, it is very difficult to access and exploit the integrated digital archives for researchers and artistic fans. Ontology is good avenues to integrate these distributed and heterogeneous systems. Especially, authors often use their own images or do not specify the source of their datasets in their papers. Naturally this makes comparison of results somewhat difficult. In this paper we introduce a new dataset, TCCPID -- an image dataset of traditional Chinese calligraphy and painting which tries to bridge this gap.

The rest of paper is organized as follows. The related work is given in section 2. In section 3, we first analyze the characteristics of traditional Chinese painting and calligraphy. Then, the process of modeling traditional Chinese painting and calligraphy based on ontology is dissertated, including how to select appropriate terms in CRM ontology and how to extend them in specific domain. In section 3, the process of creating this dataset and the functions of the system is discussed in detail. At last, conclusions are given.
2. Related Work

2.1 Related Research on Traditional Chinese calligraphy and Painting Images

Although there are many research work on image retrieval, classification and automatic semantic tagging, but little attention has been paid to the domain of traditional Chinese painting images. Existing research work in the field of for Chinese Calligraphy and Painting is introduced as follows.

James. Wang et al.(2003) [4] used wavelet transform extract features from images of calligraphy and painting, then proposed a classifier model which using the 2-D multi resolution hidden Markov model(MHMM) to achieve on the automatic classification of Shen Zhou, Tang Yin, Zhang Daqian and others painters’ images. Their experiment dataset included 276 painting images which were from 5 renowned artists: Shen Zhou, Dong Qichang, Gao Fenghan, Wu Changshuo, Zhang Daqian.

Danqing Zhang et al. (2004) [5]present a framework for modeling traditional Chinese paintings, and examine various existing CBIR proposals and algorithms for their suitability for traditional Chinese paintings, and also present a research agenda to study the problems of Chinese paintings classification and retrieval based on the framework. Their experiment dataset was a collection of over 1000 high quality digital Chinese paintings over 2000 years.

Shuqiang Jiang and Qingming Huang et al. (2006) [6] used autocorrelation texture features to describe the complexity of the digital images of Chinese Painting and Calligraphy, and then propose a scheme for detecting Traditional Chinese Painting from general images and categorize them into Gongbi and Xieyi schools. Low-level features such as color histogram, color coherence vectors, autocorrelation texture features and the newly proposed edge-size histogram are used to achieve the high-level classification. The used database included 3688 traditional Chinese painting images.

Shuqiang Jiang and Qingming Huang et al. (2005) [7] propose a framework for retrieving art images using an ontology-based method. The proposed ontology describes images in various aspects. Non-objectionable semantics are first introduced, and also give the expression of these semantics. Concepts in the ontology could be automatically derived. They used 1254 traditional Chinese painting images to test the ideas.

We can conclude that little attention of the research has been paid to the field of traditional Chinese Painting. Especially, the authors of these papers used different datasets to do the experiments. Because of no uniform datasets, later researchers have many difficulties in comparing their experimental results to the former’s.

2.2 Ontology Technique

Ontology is a formal shared conceptualization of domains of interest providing common vocabularies used for asserting and searching concepts [1]. Since sharing information among disparate organizations is often hampered due to difference in terminologies, an ontological approach to provide a “semantic” glue is of use. Ontology facilitate the interoperability between heterogeneous systems by providing a shared understanding of domain problems. In addition, an ontology can make domain knowledge explicit and as such it enables non-domain experts to understand which concepts are of interest and value for a given application.

Significant efforts have been made to provide standard data representations within cultural heritage domains have been invested in order to enable disparate organisations to share their information across different locations and information systems. These efforts produced reference models which can be used as a basis for creating specific applications, or for classifying concepts and thesauruses. A reference model is a useful representation to help the understanding, controlling and modelling of specific domains of interest. Vernadat [2] defines the reference model as a model which can be used as a basis for a dedicated model for specific aspects of a given model in order to develop or evaluate a particular model. It is argued that the reference model can enhance reusability, correctness and development efforts as well as support the compatibility between participants.

The CIDOC Object-oriented Conceptual Reference Model (CRM), originates from earlier standers proposals produced by ICOM/CIDOC Documentation Standards Group. Since September of 2000, the CRM has been processing as an ISO standard in a joint effort of the CIDOC CRM SIG and ISO/TC46/SC4. It
represents ontology for culture heritage information. The primary role of the CRM is to serve as the semantic ‘glue’ needed to transform disparate, localized information sources into a coherent and valuable global resource. CIDOC CRM is more than a core standard of the most basic entities and relationships. It attempts to adequately capture the semantics behind the most common data structures of the culture heritage domain and related domains. The collection of the CIDOC CRM, which is due for completion in fall 2002, comprises 84 classes and 130 properties [3].

To facilitate search, ontology is used as a structuring device for an information repository (e.g., documents, web pages, names of experts); this supports the organization and classification of repositories of information at a higher level of abstraction than is commonly used today. They can be used as a sophisticated indexing mechanism into such repositories. The use of ontology to overcome the limitations of keyword-based search has been put forward as one of the motivations of the Semantic Web since its emergence in the late 90’s. There have been contributions in this direction in the last few years. Most achievements so far make use of the expressive power of an ontology-based knowledge representation.

3. Modeling for Traditional Chinese calligraphy and Painting images Based on Ontology

We should analyze relative properties of entities before creating database system for our introduced dataset. The properties of entities should be described from two aspects. One aspect is to comply with global metadata standards in culture heritage; the other is considering the characteristics of traditional Chinese calligraphy and painting.

3.1 Characteristics Analysis of Traditional Chinese calligraphy and Painting

As an important part of the East Asian cultural heritage, traditional Chinese painting is highly regarded for its theory, expression, and techniques throughout the world. Chinese painting is a unique form of art, creating dramatic visual effects by applying water-soluble ink and/or colors on paper or silk using Chinese brushes.

Traditional Chinese painting is distinguished from Western art in many aspects. The structure of most traditional Chinese painting doesn’t conform to golden section which western paintings commonly conform to. The exhibiting formats include shaft format, volume format, page format, fan format, book format, screen format and so on. The art of traditional Chinese painting promotes capturing the essence of objects, rather than their shapes. The perspective of traditional Chinese painting is disperse. However, the perspective of western painting is concentrated. The strokes of traditional Chinese painting vary greatly. They are thick or thin, fast or slow, squared or circled. Traditional Chinese painting is distinguished from Western art in that it is executed with the Chinese brush, Chinese ink, Chinese mineral and vegetable pigments.

According to the creation of artists, TCP may appear various art styles. The types of traditional Chinese painting are many from different perspectives. The main body of interest in traditional Chinese paintings includes animals, flowers and grasses, rocks and trees, figures, buildings and mountains. For each object type, there are usually a small number of sub varieties. From perspective of used technique, traditional Chinese painting is generally classified into two styles: Xieyi (freehand strokes) and Gongbi (skilled brush). The Xieyi is marked by exaggerated forms and freehand brushwork. In contrast, the brushwork in Gongbi paintings is fine and visually complex. The main used colors include water-soluble ink, golden, green, deep red and purple.

Another distinct feature of transitional Chinese paintings is that they nearly all contain scripts part and seals part, mostly in black ink and read colors respectively. They are integral parts of traditional Chinese paintings, not only for noting the title, name and date of a piece of art work, but also as a part of the visual art to enhance and reveal the artistic conception of a painting. A piece of calligraphy is typically followed by one or several pieces of seals, applied by the creator and sometimes owners of a painting. Scripts part and seals part on paintings can be sometimes used for identification and authentication purposes.

Figure 1 shows three typical Chinese paintings.
Chinese calligraphy itself is an art with a history of thousands of years and is unique to Asian cultures. Calligraphy is the art of making beautiful or elegant handwriting. Chinese Calligraphy, which is written in water-soluble ink using Chinese brush, is the art of writing Chinese characters and an art unique to Asian cultures. To write Chinese characters, a brush, ink, paper and ink stone are needed, which are commonly referred to as the four treasures of the study. Over the thousands of years, the style of Chinese calligraphy has evolved continually. From that time to the present, there are five major styles of calligraphy: Zhuanshu (the Seal Script), Lishu (the Official /Clerical Script), Kaishu (the Formal/Regular Script), Xingshu (the Semi-cursive script), Caoshu (the Cursive script).

The Seal Script is the formal script of the Qin system of writing, the informal script of which was precursor to the Clerical Script. Most people today cannot read the seal script, so it is generally not used outside the fields of calligraphy and carved seals. The Clerical Script developed from the Seal Script. In general, characters are often "flat" in appearance, being wider than they are tall. The strokes may appear curved and with variations in width. Most noticeable is the dramatically flared tail of one dominant horizontal or downward-diagonal stroke, especially that to the lower right. Clerical Script of the Chinese Warring States period to Qin Dynasty and early Han Dynasty can often be difficult to read for a modern East Asian person, but the mature Clerical Script of the middle to late Han dynasty is generally legible. Modern works in the Clerical Script tend to use the mature, late Hán style, and may also use modernized character structures, resulting in a form as transparent and legible as Regular Script. The Semi-cursive Script approximates normal handwriting in which strokes and, more rarely, characters are allowed to run into one another. In writing in the Semi-cursive Script, the brush leaves the paper less often than in the Regular Script. Characters appear less angular and rounder. The Cursive Script is a fully cursive script, and a person who can read the Semi-cursive Script cannot be expected to read the Grass Script without training. Entire characters may be written without lifting the brush from the paper at all, and characters frequently flow into one another. Strokes are modified or eliminated completely to facilitate smooth writing and to create a beautiful, abstract appearance. Characters are highly rounded and soft in appearance, with a noticeable lack of angular lines. The Regular Script is one of the last major calligraphic styles to develop, emerging between the Chinese Han dynasty and Three Kingdoms period, gaining dominance in the Southern and Northern Dynasties, and maturing in the Tang Dynasty. It emerged from a neatly written, early period semi-cursive form of clerical script. As the name suggests, the Regular Script is "regular", with each of the strokes placed slowly and carefully, the brush lifted from the paper and all the strokes distinct from each other.

Figure 2 shows three typical Chinese Calligraphy images.

3.2 Modeling Traditional Chinese calligraphy and Painting Images based on Ontology

We can annotate a traditional Chinese painting and calligraphy from two aspects. On one hand, we should describe its creation information, including creation time, creation place, the author, its size, its material, its exhibition formats, main body of the painting, used techniques and its collected information. On the other
hand, the seals part and scripts part are also very important which note the information of the authors and owners and should be annotated. Image semantic classification is a form of semantic image understanding. So the classification of traditional Chinese painting and calligraphy in annotation process is a key problem to assist image retrieval and related process.

Based on analysis above, we adopt the existing classes and properties in the CRM ontology to describe the common characteristics of a painting. In addition, we appropriately extend ontology to suit to the specific characteristics of traditional Chinese painting and calligraphy, including seals, the scripts part and the classification. The designed ontology model is as figure 3.

In the figure 3, the new added classes include Exhibiting formats, Types of painting, Types of calligraphy, Types of painting. Figure, Mountain /river, Flower/bird, Types of main body, Types of technique, Types of main color, Seal, Prescript /postscript. The new added properties includes Has exhibiting formats, Has subclassification, Has author. The words in the ellipses denote instance values of the corresponding classes.
4. Creating database system for TCCPID

4.1 The Database of Traditional Chinese Calligraphy and Painting Images

At present, traditional Chinese paintings are preserved in different distributed museums. The typical avenue is to scan art books using high performance scanner. These books also have copyright. We first signed the agreement with State Administration of Cultural Heritage and guaranteed that the use of the dataset was only for research. The used scanners are Epson V940 scanner and Ever Smart Supreme Flatbed Scanner. The resolution of the images is between 100 and 1200dpi. The color of images is 24bit true color. The format of the images is TIF.

The traditional Chinese painting images in TCCPID dataset are scanned from Chinese Painting Collected Edition, which is a authoritative edition in traditional Chinese painting domain and published by Cultural Heritage Press of China. There are 6581 painting images which belong to different classification and were created by 487 artists in different dynasties. The table 1 shows the number of painting images in different dynasties created by different artists.

<table>
<thead>
<tr>
<th></th>
<th>Chinese Warring States period – Tang Dynasty</th>
<th>Wudai Dynasty</th>
<th>Song Dynasty</th>
<th>Liao Dynasty</th>
<th>Jin Dynasty</th>
<th>Ming Dynasty</th>
<th>Qing Dynasty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of</td>
<td>157</td>
<td>875</td>
<td>467</td>
<td>2018</td>
<td>3064</td>
<td></td>
<td></td>
</tr>
<tr>
<td>painting images</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of</td>
<td>7</td>
<td>82</td>
<td>67</td>
<td>156</td>
<td>175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>artists</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The traditional Chinese calligraphy images in TCCPID database are scanned from Chinese Calligraphy Collected Edition, which is a authoritative edition in traditional Chinese calligraphy domain and published by Cultural Heritage Press of China. There are 4800 calligraphy images which belong to different classification and were created by 263 artists in different dynasties. The table 2 shows the number of calligraphy images in different dynasties created by different artists.

<table>
<thead>
<tr>
<th></th>
<th>Pre-Qing Dynasty</th>
<th>Wei Dynasty</th>
<th>Su-tang Dynasty</th>
<th>Ming Dynasty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Jin Dynasty</td>
<td>Wudai Dynasty</td>
<td>Wudai Dynasty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nan-bei Dynasty</td>
<td>Wudai Dynasty</td>
<td>Wudai Dynasty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tang Dynasty</td>
<td>Sui Dynasty</td>
<td>Sui Dynasty</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tang Dynasty</td>
<td>Tang Dynasty</td>
</tr>
<tr>
<td>Number of</td>
<td>200</td>
<td>400</td>
<td>1800</td>
<td>2400</td>
</tr>
<tr>
<td>painting images</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of</td>
<td>0</td>
<td>11</td>
<td>36</td>
<td>216</td>
</tr>
<tr>
<td>artists</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In all, there TCCPID database currently consists of 7565 images, which include 2000 calligraphy images and 5565 painting images. We are still working on expanding the image set.

4.2 The Functions of the System

The flat of system is based on Java EE. WEB Server and App Server are adopted to satisfy retractility and extensibility of abundant data of traditional Chinese painting and calligraphy images. The developing language is Java. The database is Oracle 9i. The main interface is as Figure 4. We can search a painting according a key word or many key words. The search interface is as Figure 5. Then a result interface can be
got as Figure 6. Similarly, we can search a artist, a museums, a collector and so on. At the time, we can further search relative painting images when we get a result. In addition, the result dataset can be downloaded.

5. Conclusions
OVER THE LAST FEW YEARS, MORE AND MORE ATTENTION HAS paid on the domain of traditional Chinese art images. Unfortunately, many authors used their own images sets to do the experiments. This makes comparisons of various algorithms virtually impossible. In the paper, we introduce TCCPID -- a dataset of traditional Chinese calligraphy and painting images. The aim of the dataset is to provide a standard dataset for researchers. TCCPID database currently consists of 7565 images. We are still working on expanding the image set. We believe our work is very meaningful for domain of traditional Chinese art. In addition, we will add some functions to facilitate the research work. One function is standard classification. This function is to let volunteers to classify work images and the researchers can compare their results with the volunteers’ annotation results. The other function is standard observing. This function is to let volunteers to annotate work images’ salient regions and the researchers can compare their salient results with the volunteers’ annotation results. At last, we can add the function that stochastic sample datasets can be created according to researchers’ demands.

6. Acknowledgement
This work is partly supported by National Nature Science Foundation of China (No. 60972145).

7. Reference