J Vis
https://doi.org/10.1007/s12650-021-00780-0

REGULAR PAPER

Yingchaojie Feng · Jiazhou Chen · Keyu Huang · Jason K. Wong · Hui Ye · Wei Zhang · Rongchen Zhu · Xiaonan Luo · Wei Chen

iPoet: interactive painting poetry creation with visual multimodal analysis

Received: 8 July 2021 / Accepted: 18 July 2021

Abstract Chinese painting poetry is an extraordinary aesthetic phenomenon in world art history. It is not only part of the paintings but also helps us to better understand the spiritual conception that the artists express. In this paper, we present an interactive visual system to enable ordinary users to compose customized painting poetry for ancient Chinese paintings, which contain three properties: (1) We employ object detection and image captioning to describe the scenery depicted in the painting. (2) We extend the modern color theory to analyze the underlying emotions of each painting. (3) We propose an interactive poetry generation method that takes the content description and the emotional expression to add the diversity of the poetry creation. Several visual components are carefully designed to visualize and contextualize the features in the painting. They effectively guide users to steer the creation of personalized painting poems. We conduct efficient case studies and user interviews to demonstrate the effectiveness of our system.

Keywords Poetry creation · Chinese painting · Visual analysis · Multimodal analysis

1 Introduction

Ancient Chinese paintings are usually drawn on silk papers with ink brushes dipped in water, ink, and pigments. Their elements consist mainly of landscapes, figures, flowers, and birds, reflecting ancient people's conception of nature, society, politics, philosophy, and religion. Without a comprehensive understanding of this spiritual motif, appreciation of ancient Chinese painting is strongly limited to visible contents, thus being impossible to immerse oneself into the historical context and the painter's spiritual conception behind the painting (Fig. 1).

Traditionally, talented painters are highly educated and are often great poets. They nourished an extraordinary aesthetic phenomenon in world art history called Chinese painting poetry that increases the literary value of paintings. Poems are written in the negative space (empty space) by the painters themselves

Y. Feng · H. Ye · W. Zhang · R. Zhu · W. Chen
State Key Lab of CAD&CG, Zhejiang University, Zhejiang, China

J. Chen (✉) · K. Huang
College of Computer Science and Technology, Zhejiang University of Technology, Zhejiang, China
E-mail: cjz@zjut.edu.cn

J. K. Wong
Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

X. Luo
Guilin University of Electronic Technology, Guilin, China

Published online: 19 November 2021

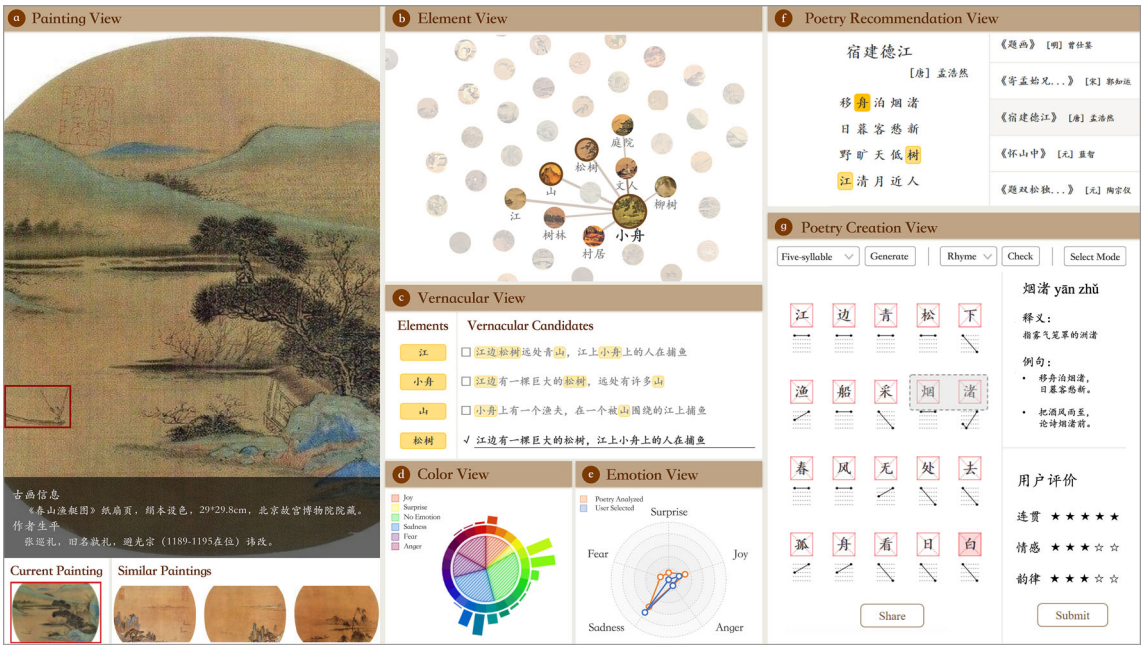

Fig. 1 Interface of *iPoet* consists of seven components: **a** Painting View explores ancient Chinese paintings; **b** Element View shows the scenery and object elements detected by algorithms; **c** Vernacular View offers various candidates for element-related vernacular description; **d** Color View presents the color distribution of the painting; **e** Emotion View supports emotion recommendation and configuration; **f** Poetry Recommendation View lists related poems for reference; **g** Poetry Creation View supports poem creation with multiple auxiliary functions

or other poets. These poems extend the painting in a literary manner to express their feelings and artistic ideas. Therefore, painting poetry is an important way to interpret and appreciate ancient paintings, especially to understand the spiritual world of painters. For instance, the Song Dynasty painter Hui Chong painted "The picture of playing ducks on the spring river," and the poet Su Shi wrote a poem on this painting titled "Spring River and Dawn Scenery," which describes its content and conceptions. This poem reveals the abstract feeling that ducks are the first to know the warmth of the water when the spring arrives. This conception and personification can only be expressed by the poem, as they are too complicated to convey with the painting itself. Unfortunately, Hui Chong's painting can no longer be found, but Su Shi's painting poem has transcended through times and enjoys great popularity nowadays.

Despite the importance of the painting poetry in appreciation of ancient paintings, very few ancient paintings have corresponding poems. Writing poems for paintings are much more challenging than creating paintings or poetry alone. It requires expertise in ancient Chinese art and literature simultaneously, which is hardly possible for ordinary people. If poems can be created for each ancient painting, it will help us appreciate them better and increase their literary values. This has motivated us to study how to allow more people to appreciate and engage in ancient Chinese arts through composing painting poetry.

The rapid development of natural language processing and computer vision technologies provides the possibility to solve the aforementioned problems. A naive idea is to directly train a neural network model for the painting poetry creation. Nevertheless, the difficulties lie in establishing an accurate one-to-one correspondence between paintings and poems, as there is no large-scale and high-quality dataset available for the end-to-end training at present. An indirect way is to identify elements in the paintings through object detection methods and then generate poems using the element-related keywords. However, the generated poems are often monotonous, indistinctive, and inconsistent with the paintings in mood and composition.

In this paper, we present a visual analysis system, *iPoet*, to enable ordinary users to compose customized painting poetry for ancient Chinese paintings. We extract both element-related description and emotions from the painting through object detection, image captioning, and statistical color analysis. These are visualized to help users intuitively understand the multimodal information behind the painting. With these features, we further propose a novel poetry creation method to aid users in generating poems consistent with the painting's contents and emotions. We carefully design a visual interface to facilitate human-in-the-loop in the poetry creation process to deal with the imperfect machine-generated poems. *iPoet* enables users to

add their personal understanding and feelings about the paintings to steer the creation process and iteratively refine the results. The main contributions of this paper are summarized as follows:

- A novel automatic poetry generation method that preserves consistency with visual elements and emotional feelings of the ancient Chinese paintings.
- A visual analysis system that provides a set of intuitive visual designs to support interactive exploratory analysis of ancient Chinese paintings and flexible customization of poetry creation.
- Two case studies and a user study that demonstrate the usability and effectiveness of the proposed method on composing painting poetry for ordinary users.

2 Related work

This section reviews analytical techniques that are most relevant to our work, including computer vision algorithms for painting analysis, poem generation models, and literary visualizations.

2.1 Object detection and captioning for painting analysis

According to the "Six principles of Chinese painting" summarized by Xie He, a fundamental Chinese painting principle is the correspondence of drawn shapes and lines with the shape of real elements. Therefore, the analysis of elements is conducive to the appreciation of ancient Chinese paintings. It can be divided into two steps: (1) determine whether an element exists in the painting, and (2) make sense of the element composition. Object detection has been conventionally used to determine object existence in images. Huang et al. (2017) compare popular algorithms with detailed experiments and open-sourced implementation and conclude the relative speed-accuracy trade-offs to serve as model selection criteria. These methods can identify and locate the elements in paintings for further analysis.

An important approach is to transform images into text through automatic image annotation and captioning. The early development in image captioning is based on traditional machine learning models, which extract visual concepts, such as targets and attributes, to generate rule-based sentences (Kulkarni et al. 2013). These rule-based methods have limited generalizability because they require enormous human efforts and a clear definition of the application to be summarized. Many effective approaches (Vinyals et al. 2015; Karpathy and Li 2015) adopt the Encoder–Decoder architecture (Cho et al. 2014) to generate image-specific captions. Xu et al. (2015) first propose to merge the attention mechanism into image captioning, while Lu et al. (2017) later improve by adding self-adaptability to the attention mechanism. Relationship detection plays a crucial role in describing the content of an image (Lu et al. 2016; Newell and Deng 2017), and the use of scene graph can reflect the properties and relationships of objects (Xu et al. 2017). There is a lot of work around scene diagrams, such as generating scene graphs (Anderson et al. 2016), using scene graphs to achieve image retrieval (Johnson et al. 2015; Schuster et al. 2015), and generating captions by auto-encoding scene graphs (Yang et al. 2019). However, for art appreciation, each person has different regions of interest, leading to high demand for customization. This paper proposes first detecting the elements in the paintings and contextualizing the element space to help users select elements of interest. Then, these elements are piped into the image captioning algorithm to obtain a description for content analysis.

2.2 Poem generation

In the long course of history, ancient literati have written countless beautiful poems that are invaluable treasures in Chinese culture. Classical poetry has to be rich in content expression, compliant in strict poetic forms, and articulate in pronunciation. These constraints pose significant challenges to poets, but this does not stop them from innovating. Modern literati have also studied numerous ways to generate poems automatically with machine learning methods (Zhang and Lapata 2014). Inspired by the human creation process, the quality of generated poems can be improved by iterative polishing (Yan 2016) and sub-topics planning (Wang et al. 2016). Chen et al. (2019) extend the semi-supervised VAE with a temporal module to control sentiment at the sentence level. Yi et al. (2020) model the poetic style as the combination of several controllable factors and apply the adjusted style in poem generation.

In image-inspired poetry generation, the poem's theme is extracted from visual images. Since there is no high-quality large-scale dataset of image-poem pairs, most works are mainly restricted to two different

strategies to train the poem generation models. The first uses the keywords extracted from the image to generate poems (Cheng et al. 2018; Wu et al. 2020). The extraction and generation models are trained separately on their own datasets. The second one matches the keywords extracted from images with poems for end-to-end training (Xu et al. 2018; Liu et al. 2018). Both ways share the same idea of taking keywords as information bridges to connect paintings with poems. However, they cannot explicitly express the internal relationship of painting, such as relative positions between elements, leading to a significant decline in the consistency between paintings and poems. Our work differentiates itself from previous works in that our information bridges are the caption-like descriptors that contain more spatial and semantic information than keywords. Moreover, we provide visual analysis to facilitate user engagement, as in many fields (Li et al. 2020; Leite et al. 2020; Zhao et al. 2019). As a result, our work reduces the information loss to provide higher painting-to-poem consistency and allows users to use their natural language to steer the creation process.

2.3 Literary visualization

Visualization effectively helps the users gain insight from the data by synergizing human and machine intelligence (Zhou et al. 2019).

Art appreciation has inspired many literary visualization designs to enable users to explore collections of art pieces. Yang et al. (Yang et al. 2006) apply semantic image analysis techniques to accelerate user retrieval, visualize content similarities, and support annotation process evaluation. Kang et al. (Kang et al. 2018) analyze paintings' emotions by matching the primary colors with emotions to define a similarity metric for recommendations. Kaneko et al. (2020) propose a content-based image browser to visualize image features such as color temperature, composition, and author information. However, these works only serve for content retrieval and exploration purpose. Our system focuses on the in-depth analysis of a single painting and engages users in the art creation process to deepen their understanding of the painting.

Text visualization is conducive to many applications (Shu et al. 2021; Wang et al. 2020, 2018). Our work mainly concerns poetry, in particular Chinese classical poetry. McCurdy et al. (2015) propose *Poemage* to visualize the sonic topology of poetry. Meneses and Furuta (2015) present a set of interactive tools to visualize poetry, including a graph and a wheel to reveal emotions expressed in individual words and their context. Hu (2018) present several visual designs to show the style evolution of Songci and trace information of poets. Zhang et al. (2019) explore the potential associations between poets' life, time background, and similar creations, which provide literary researchers with a new perspective. These works focus on several aspects of poetry that are astonishing for exploration but cannot be directly applied to facilitate poetry creation. We consider achieving the goal by combining automatic methods and visualization to analyze and contribute to poetry creation.

3 Background

This section describes the datasets used for interactive painting poetry creation and the task analysis that guides the design of our system.

3.1 Painting poetry data and preprocessing

We focus on the "small paintings" of the Song Dynasty, of which canvas size spans about 30 cm×30 cm, and image resolution varies from 750×750 to 1200×1200. These paintings are relatively small but their content diversity and richness are not compromised. They depict various elements and express multiple themes, which provide heterogeneous topics for users to focus on during the painting poetry creation. Also, the paintings convey profuse artistic moods and underlying emotions such that the viewers can freely stretch their imaginations when reading them. We collect 450 pieces of "small paintings" from LTFC[1] online art collection and several digital copies of Song Dynasty art books. To provide ground truth for the paintings, we manually annotated a part of paintings in the collection, which includes 1k labeled captions, with three schemes. First, we look for object elements in paintings and label their bounding boxes because existing object detection methods might not perform well on the ancient art forms with different elements. We find

[1] http://www.ltfc.net/.

78 different kinds of object elements from the paintings. Second, we identify actions or states of elements and write at most five sentences for each painting to describe the spatial and semantic relationships between elements. Third, based on the background information and the content interpretation of the paintings, we improve the sentences with appropriate adjectives, adverbs, and other descriptive languages to make them more "poetic."

For poems, we target at "*Jueju*" or Chinese quatrain that is characterized by having four verses and has two forms: five- and seven-syllable. Based on the open-source data of Gushiwen[2], we build a corpus on *Jueju* that contains 10k labeled poems and 130k unlabeled poems. Each of the labeled poems contains the title, author, dynasty, content, and corresponding vernacular translation. Then, we invite poetry experts (none of them are co-authors) to annotate the sentiment of the poems classified into five classes: joy, anger, sadness, surprise, and fear. We also extract the element-related keywords from both labeled and unlabeled poems to prepare them for the poetry recommendation.

3.2 Task analysis

Our system is mainly designed for ordinary users, thus requiring us to lower the standard of poetry creation with reference to experts' domain knowledge and research experience. Through investigating the literature and interviewing painting and poetry researchers, we characterize the problems and identify the following tasks in poetry creation:

- **T1** **Extract scenery and object elements from the painting** Object detection accuracy is a critical bottleneck that affects the consistency of painting and poetry. Since some elements in the painting may not be detected by the algorithm, the system should help users confirm or prioritize specific element candidates with high co-occurrence probabilities. Users should be able to remove the non-existent objects falsely detected by the algorithm.
- **T2** **Describe the content and composition of the painting in the vernacular** Regarding the user-selected elements, the system should assist in describing the contents of the painting in the vernacular, which ordinary users could easily comprehend. The description may relate to the spatial relationship of the scenery and the demeanor and actions of the object elements.
- **T3** **Identify the emotions that users want to express** Emotion expression is an important consideration and one of the purposes of painting poetry. The system should analyze the painting and recommend its underlying emotions to users. They can then refine and clarify the emotions they feel and want to express in the poem.
- **T4** **Create a poem for the painting** Based on the multimodal features obtained from the previous tasks, the system should help users create a painting poem by automatically generating a skeleton poem as a basis. It should also recommend classics of poetry based on the user-selected elements to provide reference and inspiration. It should offer word interpretation or a dictionary to help users with little poetry expertise understand the poem. Finally, it should provide a feedback interface allowing the users to rate and evaluate the creation.

4 The framework of *iPoet*

The purpose of *iPoet* aims to help users analyze the different perspectives of a painting and use them to create a painting poem. From the content perspective, we first use an object detection algorithm to identify the elements in the painting, on which users confirm (**T1**). Then, we use an enhanced image captioning algorithm to generate descriptive statements that match the contents (**T2**). From the emotional perspective, we observe the colors in the painting and adopt a color-emotion transformation method to model the intensity of five emotion types (**T3**). These emotions are recommended to users and support free adjustments. Based on the two features, users can start creating a poem. The system provides auxiliary functions during the process (**T4**), which include poetry recommendation, description- and emotion-based poem creation, word interpretation, fluency and rhyme checks, and a user feedback interface to evaluate the poem (Fig. 2).

[2] https://www.gushiwen.org/.

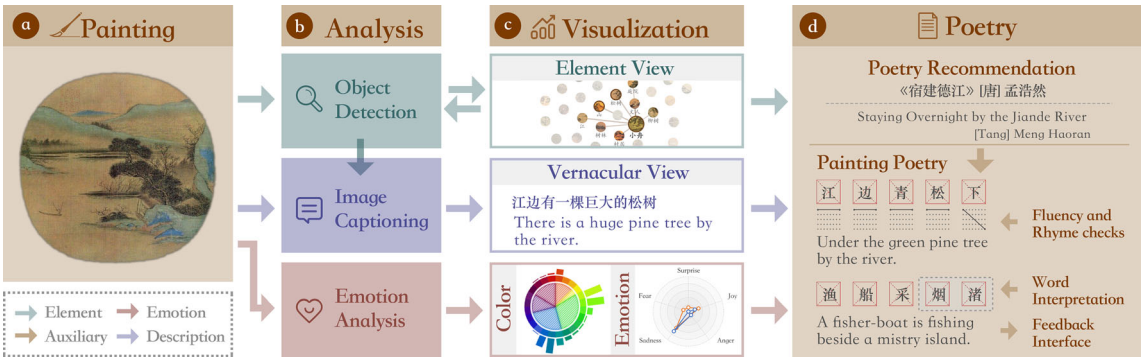

Fig. 2 *iPoet* framework. From the painting input **a** to the poetry output **d**, *iPoet* help users to seamlessly and interactively navigate through analysis and visualization. In the analysis process **b**, object detection and image captioning detect elements in the painting and describe their composition to facilitate understanding. The emotion analysis transforms the color distribution of the painting to emotions and models the intensity of five emotion types. The poetry creation **d** produces painting poetry that is consistent with the content and the emotion of the painting. The visual interface **c** corresponds to visualize the interim results for users to manipulate and steer the poem creation process

4.1 Painting content analysis

The inspiration of painting poetry comes from the poet's perception and cognition of the painting. Every user has their individual interpretation when reading the paintings. This difference will affect the result of composing poems. Those one-to-one poem generation methods go against the spiritual purpose of painting poems. To enable users to choose elements independently and create poems, particularly for the elements of interest, *iPoet* adopts a Faster R-CNN (Ren et al. 2016) implementation with dense output Inception Resnet models from TensorFlow Object Detection API that attain the best possible accuracy among all the models provided (Huang et al. 2017). We refine the models by our labeled data to detect the elements in paintings (**T1**). The motivation of training automated methods is to improve the generalizability of the system toward other types of ancient Chinese paintings. The method can scan through the dataset to discover elements with high co-occurrence probabilities and recommend them as candidates in case the algorithm malfunctions.

The image captioning algorithm takes in an image and outputs a sequence of words. In this paper, we use InceptionV3 (Szegedy et al. 2016) to extract image features because its number of parameters and model complexity are far less than VGGnet (Simonyan and Zisserman 2014). We fine-tune the pretrained model by the annotated dataset so that it can better adapt to the detection of ancient Chinese paintings. It facilitates a faster training speed and a higher classification accuracy of the system. We combine the word embedding vector and the painting feature vector to feed into a recurrent neural network which transforms the variable-length combined vectors into a fixed-dimension vector. The transformed vector is then fed into the decoder to obtain the output sentence. The probability of the correct description given the image is defined as follows:

$$\theta^* = \arg\max_{\theta} \sum_{(I,V)} \log p(V|I;\theta) \tag{1}$$

where θ denotes the model parameters, I is the image vector, $V = \{ v_1, v_2, \ldots, v_n \}$ is the collection of captions, n is the number of description statements. In this paper, we select $n = 2$ because a smaller number allows greater flexibility on dataset selection and wider variety in poem creation when it comes to abstract expressions. It can be easily extended $n \leq 4$, where four sentences are the straight limit of *Jueju*.

The captioning algorithm employs LSTM as the decoder. All recurrent connections are transformed to feed-forward connections, the unrolling procedure reads:

$$p(w_t|w_{0:N-1}, I) = LSTM(e(w_t)), t \in \{ 0, \ldots, N-1 \}, \tag{2}$$

where w_t is the word in the sentence, $e(w_t)$ is embedding vector of w_t. Before the LSTM, the initial vector is influenced by the image and the objects, such that $e(w_{-1}) = CNN(I) + e(O)$, where $CNN(I)$ is the image feature vector, and $e(O)$ is the embedding vector of the chosen objects. Each word in the sentence can be predicted according to the image feature vector and word embedding vector.

4.2 Painting emotion analysis

The desire to express emotions has been an important motivation for poetry creation; therefore, neglecting emotions may lead to inconsistency between painting and poetry. Previous studies attempt to derive a correlation between colors and emotions by surveying users' subjective feelings about different colors. We refer to the study of Takahashi and Kawabata (2018) to establish a color–emotion correspondence and model the intensity of five emotion types. The derived emotions are estimators to users' emotional feelings when they read the paintings and recommendations in emotion perspective for poetry creation.

We divide the color of the Hue–Saturation-Value(HSV) spectrum into 24 equal intervals. The median hue value of each interval is considered representative. Then we classify each color interval into six main color descriptor types and eventually their corresponding emotion (Takahashi and Kawabata 2018). Neutral gray-scale colors and the background color resulted from the aging silk papers are removed by saturation and value. The correspondence between color interval and emotion is shown in Fig. 3. To analyze the mood and the underlying emotion in a painting, we obtain its color distribution by extracting the HSV of each pixel of the painting and map them onto the color intervals. The intensity of each conveyed emotion is then modeled by the proportion of colored pixels of the painting.

4.3 Poem creation

Painting poetry creation centers on our requirements (**T2, T3**) and the multimodal analysis. We build upon the idea that separates the sentiment with polarity and other attributes in capturing the latent representation (Hu et al. 2017) to improve the controllability of text generation. We encode the vernacular text and the polarity of the sentiment separately. For vernacular text $v = \{x_1, \ldots, x_t\}$, GRU is employed to encode v by taking the state at the last time step t as the representation. For sentiment labels, we directly use its embedding vector. Then we get the mixed representation through concatenation and linear projection as $z = f([s_t, e_y])$, where f is the nonlinear layer. Similar to previous works (Wang et al. 2016), we generate the poem line by line to achieve sentence-level semantic consistency and poem-level topic conformity. The probability of each character is calculated as:

$$p(x_{i,j}|x_{i,1:j-1}, x_{1:i-1}, v, y) = softmax(f(s_{i,j})) \tag{3}$$

$$s_{i,j} = GRU(s_{i,j-1}, [e(x_{i,j-1}), r_t, c_i]), \tag{4}$$

where $s_{i,j}$ is the state of GRU at line i and time step j, r_t is the genre embedding (Yi et al. (2018)) to encode the rhyme and position information, c_i is the content representation which is initialized before each poem sentence generated through $c_i = f([z, e_p])$ where e_p is embedding vector of line index. However, a static embedding vector may sacrifice grammatical correctness (Zhou et al. 2018). We adopt a read gate g_t^r and a write gate g_t^w to dynamically update the expression of poem sentences $m_{i,j}^e$ varying with time, $m_{i,j}^e$ is initialized with c_i and read as

$$g_t^r = f([e(x_{i,j-1}), s_{i,j-1}, m_{i,j}^e]) \tag{5}$$

$$m_{i,j}^r = g_t^r * m_{i,j}^e, \tag{6}$$

where f is the nonlinear layer with *sigmoid*. Therefore, Eq. (4) then becomes

$$s_{i,j} = GRU(s_{i,j-1}, [e(x_{i,j-1}), r_t, m_{i,j}^r]) \tag{7}$$

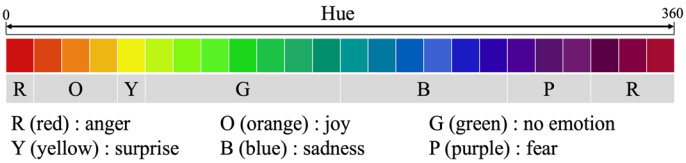

Fig. 3 Color–emotion transformation is acquired by dividing the hue color space into 24 intervals and assigning a color descriptor to each of them. The median hue value is the representative color of each interval

After generating each character, we update $m^e_{i,j}$ as $m^e_{i,j+1} = g^w_t * m^e_{i,j}$ and $g^w_t = f([s_{i,j}, m^e_{i,j}])$. Here f is also the nonlinear layer with *sigmoid*. The loss function is defined as $Loss = -log(P(X|v, y)) + sum_i(||m^e_{i,-1}||)$, where the first term is MLE loss and the second is L2 regularization to ensure a complete expression.

5 The visual interface of *iPoet*

Based on the framework above, we design and implement the visual interface of *iPoet* to support painting multimodal analysis and poetry creation. As shown in Fig. 1, *iPoet* contains four modules and seven views: (1) Element Exploration Module to explore ancient Chinese s their elements detected by algorithm; (2) Content Analysis Module to offer users various vernacular description candidates; (3) Emotion Analysis Module to present the color distribution of the and support emotion recommendation and configuration; Poetry Creation Module to list several related poetry classics and support poem creation with multiple auxiliary functions.

5.1 Element exploration module

view (Fig. 1a) is a browser that allows users to explore the collection of ancient Chinese s. Users can select a in the dataset in a selection list or upload their s of interest to the system. Apart from merely browsing, the thumbnails at the bottom depict the current and its top-three similar s from the dataset as a recommendation. Each is annotated with a list of elements it contains. For uploaded s, the object detection algorithm is responsible to generate such lists. similarity ranking is acquired by comparing the element list of the current with all other s in the dataset and positioning them by the number of overlapping elements in descending order. Users can click on the thumbnail to switch the displayed . Browsing multiple s and viewing related ones helps users explore the dataset and have a better understanding of a particular type of paintings. The black translucent area at the bottom of the painting introduces the background information of the painting and the painter for the basic need of art appreciation. Interactions such as zooming and dragging are supported for users to observe the different levels of detail.

Element view (Fig. 1b) provides an overview of the landscapes and object elements (**T1**). It employs a node-link diagram, which is widely used and studied (Zhao et al. 2020; Giovannangeli et al. 2020; Shi et al. 2020; Pinaud et al. 2020), to show not only the elements detected in the current painting, but also all those appear in the data set. We use NetV.js (Han et al. 2021) to draw the graph. In the graph, each element type is represented as a node. The detected elements are distinguished from the others by making undetected elements transparent. The links between nodes indicate that the nodes have appeared together in any painting. The more co-occurrence in paintings, the wider the width of links. Element View provides users an interface to interact with the object detection algorithm by letting users confirm the detection results. Users can manually add the undetected elements or delete the non-existent elements by simply clicking the corresponding nodes. To support the identification of element candidates with high co-occurrence probabilities, users can follow the links that connect between the detected elements and other undetected elements. These connected elements are more likely to appear in the current painting because they have appeared in other paintings. Hovering over the nodes will show their name and highlight the connected nodes and links.

5.2 Content analysis module

Vernacular view (Fig. 1c) produces the description of the painting's composition that serves as an interim result of poem creation (**T2**). The left panel of Vernacular View shows the list of elements which are detected by the algorithm or selected by users. The right panel provides several description candidates in vernacular which are generated according to the painting and the selected elements. The candidates may differ from each other in descriptive emphasis or sentence structure. Users can select and refine one candidate whom they find most appropriate. They can also enter their own descriptive sentences in the input box option. Keywords in the description will be highlighted if they are related to selected elements; meanwhile, the corresponding area of the elements will be distinguished by bounding boxes, which provide convenience for users to check the consistency between the painting and the description.

5.3 Emotion analysis module

Color view (Fig. 1d) is designed to present the color distribution and extract the primary color of the painting, which are the basis of emotion analysis. Inspired by the color wheel, we form color intervals shown in Fig. 1b into a "chromatography ring." Outside the ring is the histogram that shows the count of each color interval of the current painting. The inner part of the ring is a pie chart that shows the emotion types corresponding to each color interval. Generally, the color descriptors we use span several color intervals of the chromatography ring. Therefore, these continuous intervals correspond to an emotion. When the user hovers the mouse over a part of the pie chart, the corresponding color intervals and the statistical histogram will be highlighted by colors. **Emotion View** (Fig. 1e) aims to help users to analyze the emotions of the painting (**T3**). It provides a radar chart to present the intensity of five emotions: joy, surprise, sadness, fear, and anger. Based on the color–emotion transformation method introduced in Sect. 4.2, the system recommends an intensity for each emotion according to the color distribution of the painting, which is depicted by the blue line. Users can adjust the intensity of each emotion to indicate their preference for poem creation. The red line represents the emotion analyzed from the created poem and thus is non-adjustable. The two lines on the radar chart support the comparison of the emotion conveyed between the painting and the poem, which may provide an improvement direction for users to refine iteratively.

5.4 Painting poetry creation module

Poetry recommendation view (Fig. 1f) is an exhibition hall of the poetry corpus, displaying the excellent works of ancient poetry. Users may benefit from referring to these poems to enhance their knowledge of poetry. The right panel of the view shows the list of recommended poems. Similar to the painting recommendation mentioned in Sect. 5.1, we use the overlap with the scenery of the current painting to rank the poems in the corpus and recommend the top ten related ones. Each item in the list shows the title, author, and the Dynasty of the poem. Users can click the item to switch between poems. The left panel shows the poem line by line. The characters will be highlighted with a yellow background if they are related to the selected elements. If users hover the mouse over one of these characters, the corresponding elements in Element View and Painting View will also be highlighted. The coordinated visualizations are designed to demonstrate the explicit connection between the painting and the poem.

Poetry creation view (Fig. 1g) provides a stage for inspirations and creativity (**T4**). Based on the vernacular description and emotion analysis obtained from the previous views, the system provides a machine-generated poem as a skeleton that users can freely modify, enhance and refine. The view provides word interpretation, fluency and rhyme checks, and a user feedback interface to evaluate the poem. The left panel is the main creation interface. Each box in this panel is decorated with a Tianzige background and contains a Chinese character. Under the box visualizes the tone of the character. We use the widely adopted five-degree marking method (Zhang et al. 2019) to encode four tones: *yinping*, *yangping*, *qu*, and *ru*. At the top of the creation panel is the toolbar, providing various configurations such as creating five- or seven-syllable *Jueju*, selection mode for word interpretation, and revision mode for fluency, tone, and rhyme check. Users can select any words and characters in the poem under the selection mode, which will trigger the top right panel to present their definition and usage examples to support word interpretation. In the revision mode, the system identifies words in the poem as unfluent if they barely appear in the poetry corpus and warns users about words that do not rhyme or follow the tonal rule of Pingze. These words will be highlighted with a light red background, suggesting that they require revisions. The bottom right panel provides a feedback interface, which allows users to evaluate the poem from three dimensions: fluency, sentiment and rhythm. After the creation, users can click the share button, and the system will automatically generate a picture to mimic real painting poetry.

6 Case study

We demonstrate how the system supports users to create poems for paintings interactively through two case studies.

6.1 Case 1: "spring mountain and fishing boat"

The user selects a painting named "Spring Mountain and Fishing Boat" as the starting point to create a poem. Figure 1 provides an overview of the creation process. Under the guidance of the system, the user first analyzes the content of the painting in Element View (Fig. 1b). Elements such as boats, mountains, and pine trees are automatically detected by the system and shown in opaque colors. Hovering over the "boat" node, the user notices that the "river" element frequently co-occurs with the "boat" indicated by a thick line connecting the two elements. The "river" node, however, is in opaque color, meaning that no river is detected in the current painting. He revisits the painting and finds that the boat is actually floating on a river which is not drawn explicitly (without colors for water or lines for waves), so the algorithm did not detect the "river" element. The user considers the river as an important element for creation, so he clicks on the "river" node in Element View to add it to the current collection. Then, he goes through the descriptions in Vernacular View (Fig. 1c) and finds that they are appropriate regarding the content. The first one appears to be the best but contains too many elements in its first sentence. With the help of other descriptions, he revises the first one as "A huge pine tree stands along the river, and fishermen in the boat are fishing on the river."

The user then navigates to the emotion analysis part of the system. Color View (Fig. 1d) shows the color distribution of the painting. The primary colors of this painting are green and blue, which contribute to a feeling of sadness. The radar chart to the right (Fig. 1e) shows the prediction of the emotions, which is acceptable to the user.

After clarifying the content and emotion of the painting, the user begins to create the poem (Fig. 1g). As the user is inexperienced in poetry, he decides to create the poem based on the one generated by the system. The generated poem (Fig. 4a) has five syllables as selected by the user. After reading through the poem, the user selects the word "烟渚" (misty island) to get some explanation using the word interpretation function. Based on the painting and the descriptions in vernacular, the user evaluates the first draft of the poem (Fluency-5, Emotion-3, Rhyme-3). He thinks that the poem is relatively consistent with the vernacular description, but the "boat" in the last sentence repeats the previous text, and the last sentence ("白–white") does not rhyme with the second sentence ("渚–island"), which is also highlighted by the system under the "Rhyme Check" mode. The user intends to revise the last sentence. He notices that the scene of the painting is open and sparsely populated, which gives him a feeling of loneliness and coldness, so he modifies the last sentence to "I'm so lonely without anyone to talk to" The revised sentence not only rhymes with the second sentence but also increases the mournful mood. After that, the user makes some modifications to other parts of the poem by referring to the recommended poems (Fig. 1f). Then, the user is satisfied with the created poem (Fluency-5, Emotion-5, Rhyme-5) and finishes the creation process by saving the results (Fig. 4b).

6.2 Case 2: "fairy mountain pavilion"

In this case, the user intends to create poems from another painting named "Pavilion of Fairy Hill." After checking the basic information displayed at the bottom of the painting (Fig. 6a), the user explores Element View (Fig. $6b_1$). The view shows the detected elements, including mountains, pavilion, and clouds. The vernacular (Fig. $6b_2$) describes the painting as "There are pavilions on the high mountain, surrounded by

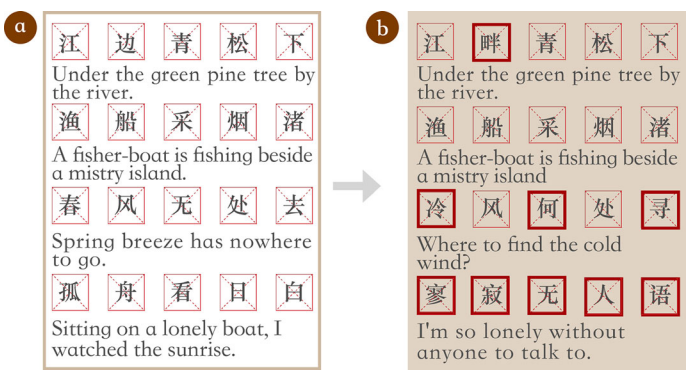

Fig. 4 Poems of "Spring Mountain and Fishing Boat" before revision (**a**) and after revision (**b**)

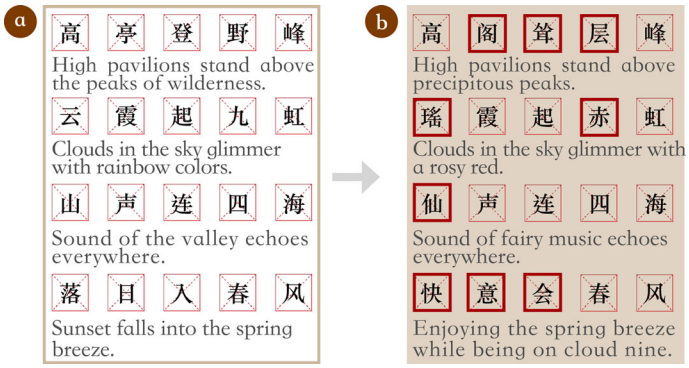

Fig. 5 Joyful poems of "Fairy Mountain Pavilion" before revision (**a**) and after revision (**b**)

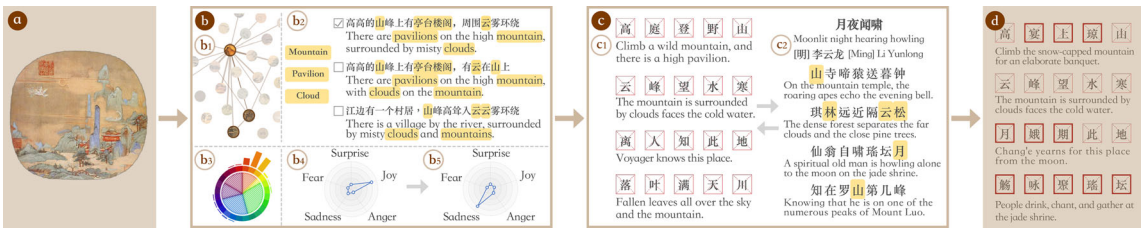

Fig. 6 Process of creating a poem for the "Fairy Mountain Pavilion." **a** Browse the painting; (b1) Identify the scenery and object elements contained in the painting; (b2) describe the content of the painting according to the elements; (b3) Analyze the color distribution of the painting; (b4) Consider the emotion recommended by the system; (b5) Adjust emotion based on the painting and background knowledge; **c** Adopt the poem generated by the system as the basis and revise it by referring to the recommended poem; **d** Get the final result of the creation

clouds" by considering these elements. Color View (Fig. 6b_3) depicts orange and red as the primary colors of this painting. Therefore, the system recommends joy as the main emotion in the radar chart (Fig. 6b_4). The user agrees with the result and reads through the poem generated by the system (Fig. 5a). He appreciates the implicit expression of warmth and leisure, and then makes some revisions to the poem to further enhance the joy and fluency of the poem (Fig. 5b).

After composing such a joyful poem, the user is also curious about trying other emotions, especially sadness, because the fairy riding a crane in the painting reminds him of the fairy tale of Chang'e flying to the moon. Therefore, the user sets sadness as the dominant emotion and reduces the intensity of joy in the radar chart of Emotion View (Fig. 6b_5). Then, the user instructs the system to generate a five syllables quatrain (Fig. 6c_1). Then, he considers that the generated poem rhymes well, the first two lines are consistent with the description of the painting, and the words such as "cold" and "departed people" in the poem express the sadness. The only defect is that the last sentence, "The sky is full of falling leaves." is not reflected in the painting, so there is a lack of consistency in the content. Therefore, the user tries to modify the last sentence and its context for better consistency.

He browses through some recommended poems (Fig. 6c_2), one of which titled "Hearing the whistling in the moonlit night" depicts mountains, clouds, the moon, pine trees, and immortals. The user thinks that the poem fits perfectly with the content of the current painting. Referring to the description of the moon in this poem and the story of Chang'e, the user modifies the last two lines of the poem as "Chang'e looks at this place from a distance, people drinking, chanting, and gathering at the altar made of jade," making the poem more consistent with the content of the painting and his feelings of sorrow and longing. The user then revises the rest of the poem to further strengthen the poem's consistency with the painting (Fig. 6d).

7 Evaluation

To evaluate the practicality and effectiveness of the system, we conducted user interviews with 11 users who used the system.

7.1 Participants and procedure

Eleven university students (age mean: 23.6; female: 7) participated in our interviews voluntarily. Most of them (8/11) majored in computer science in which two have more than two years of visualization research experience, and the rest (3/11) majored in chemistry or education. None of them had painting or poetry-related research experience.

The user interviews were structured in three parts. First, the participants watched a ten-minute pre-recorded video explaining the workflow and the functions of each view through the two case studies. The participants could pause the video at any time to ask the experimenter questions. After the video, the participants used the system to create a poem from a painting of their selection independently. No restrictions were set regarding the content of the creation. The participants could finish the task once they were satisfied with their poems. The creation process lasted for 5-30 minutes. Finally, the participants completed a semi-structured questionnaire for system evaluation, which took around 20 minutes. Each question in the questionnaire contained a five-point Likert scale assessment (from 1 = strongly disagree to 5 = strongly agree) and a comment space for the participants to explain the reasons behind their ratings. On average, the entire experiment took about 45 minutes. Figure 7 has listed the questions in the questionnaire.

7.2 Results

Usability. In terms of the system usability, the participants were most satisfied with the function of vernacular generation for paintings (4.54/5). The vernacular "provided a good description of the elements in the painting" and "fitted the scene in the painting." The object detection algorithm detected the elements correctly for the participants (4.45/5). "The algorithm could identify all obvious objects in the painting and even catch those that human eyes could not quickly distinguish." The participants evaluated the emotion recommended by the system as "relatively accurate" (4.27/5) and "able to capture my subjective feeling of sadness." The generated poems could reflect the content and emotional features of the paintings (4.27/5). Participants expressed their fondness for the generated poems: "The poem matched the atmosphere of the painting. Although you could not see that people were drinking in the painting, you could imagine the scene of two literati drinking and having fun, as rendered by the poem." Color View provided the participants with cues for emotion analysis (4.18/5), which was "helpful," but "the interface was a little bit complicated." In

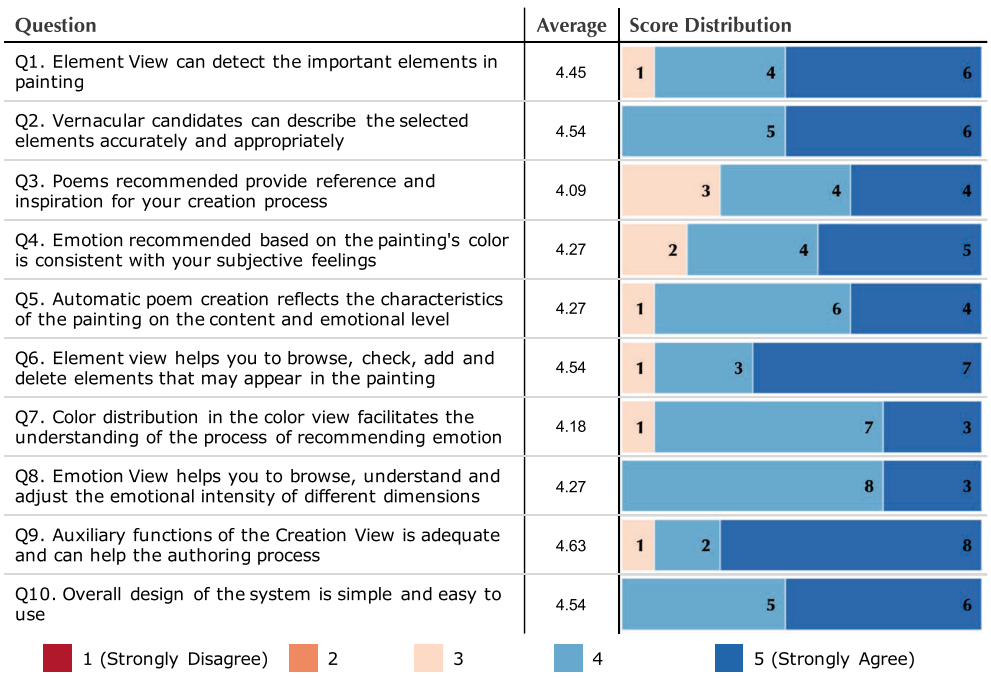

Fig. 7 Questions in our questionnaire and their corresponding score average and distribution

terms of the overall usability of the system, our system was highly rated by the participants (4.54/5). "The overall process of the system was clear and fun, and it might be helpful in education."

Effectiveness In terms of the effectiveness of the system, Poetry Creation View (4.63/5) "had powerful functions, which played a great role in the process of creation, and the generated poems basically met my expectations." "Through checking the rhythm and coherence, suggestions for modification were provided to improve the poem, which was a very practical function." Element View was helpful in browsing, checking, adding and deleting scenes (4.54/5). "Element View was intuitive and clear, and the operations of adding and deleting were very convenient." "Comparing Element View with the painting could help improve my imagery, which was useful for the creation of subsequent poems." Emotion View (4.27/5) was "intuitive" and "ingenious in design." However, some participants said that "the interaction was a little bit complicated." The function of poetry recommendation was a little weak compared with other views (4.09/5), but several participants benefited from it. "There was no doubt that the view was helpful" and "insightful." A participant considered that the poem and the painting "were not necessarily consistent" in emotion, but he also stated that "different perspectives provided more ideas for the creative process."

Suggestions for improvement Several participants hoped to have more emotion types(poetry style) to choose from, such as the ones related to wars, nostalgia, and patriotism. The elements in Element View could be further refined, such as the characters could be further distinguished by gender, age, etc. Participants also suggested considering emotions in recommending poems so that the recommended poems could be consistent with the paintings not only on the content level but also on the emotional level. These invaluable suggestions help us improve the functionality of the system and provide future direction.

8 Discussion

Although the user interview validated the usability and effectiveness of our system, it inevitably has some limitations.

Emotion analysis The color–emotion transformation we adopted may not apply to all kinds of ancient Chinese paintings, as the ancient use of color in emotional expression may have differed from nowadays. However, modern color theory suggests that colors influence viewers' emotions. We think from the viewer's perspective to speculate on their subjective feelings, which is consistent with the targets in color–emotion association studies. Moreover, it is quite common that people may come up with different interpretations by combining their background knowledge, life experiences, and different concerns about the painting. Therefore, we can use a modern interpretation of ancient paintings without diminishing the artistic value of the paintings.

Generalizability The framework of *iPoet* is generalizable to other similar tasks that create specific texts for images. The Greeks have a similar art type, Ekphrasis, that describes visual arts with verbal descriptions. Our framework can be adapted to create such art. However, finding the datasets for such creation is difficult. Our poetry corpus also suffers from uneven data distribution such as a small amount of fear and surprise samples. It limits the performance of these emotions on the training-based poetry generation algorithm, compared to the other types that have ample amount training data. This uneven distribution can be solved by finding more annotated poetry data for these two emotions and attempting to use data augmentation techniques.

Evaluation We have received very positive comments regarding *iPoet* in the user interview. This has demonstrated the usability and effectiveness of the overall approach and implementation of *iPoet* in creating painting poetry. Yet, the interviewees are all recruited from universities and mostly post-graduate students. They might not represent the population of our target ordinary users, thus resulting in a potential bias that they could have composed such poems by themselves. However, all of our interviewees do not have an advanced academic background and expertise in poetry creation. In this sense, the difference between our interviewees and ordinary users is small. Our interviewees are also able to create painting poems within the interview process (5-30 minutes) with the help of *iPoet*, illustrating an increased efficiency for poetry creation.

9 Conclusion and future work

In this work, we present an interactive visualization system for painting poetry creation with multimodal analysis. Both element-based description and emotional feelings are extracted from ancient Chinese paintings, and used to generate a poem that is consistent with the content and the emotion of the painting. Based on this framework, our system allows ordinary users to interactively explore the insight in the painting, and inject their own ideas and interpretation in the poetry creation process. Two in-depth case studies and a user interview demonstrate the usability and effectiveness of *iPoet* in facilitating user creation. The feedback from user interview also provides us with several valuable suggestions regarding future work.

In future work, we intend to expand the range of object detection and boost the emotional expression in poetry generation to a more granular level. Our system can broaden its support to different styles of ancient Chinese paintings and poems, such as Songci which has variable-length verses that poses interesting research challenges. Supporting richer emotional expressions and poetry styles can further improve the consistency of paintings and poems. Moreover, we plan to involve more modalities related to the ancient paintings and poems into account, such as the historical background and the experience of the artists/poets.

Acknowledgements This work is supported by National Natural Science Foundation of China (61972122, 61772456).

References

Anderson P, Fernando B, Johnson M, Gould S (2016) Spice: Semantic propositional image caption evaluation. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision - ECCV 2016. Springer, pp 382–398

Chen H, Yi X, Sun M, Li W, Yang C, Guo Z (2019) Sentiment-controllable chinese poetry generation. pp 4925–4931

Cheng W-F, Wu C-C, Song R, Fu J, Xie X, Nie J-Y (2018) Image inspired poetry generation in xiaoice. *arXiv preprint*arXiv:1808.03090

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint*arXiv:1406.1078

Giovannangeli L, Bourqui R, Giot R, Auber D (2020) Toward automatic comparison of visualization techniques: application to graph visualization. Vis Inform 4(2):86–98

Han D, Pan J, Zhao X, Chen W (2021) Netv. js: a web-based library for high-efficiency visualization of large-scale graphs and networks. Vis Inform 5(1):61–66

Hu H (2018) Visualization design and research of the style and sects change of song ci. Harbin Institute Of Technology (**Master's thesis**)

Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Toward controlled generation of text. In: Proceedings of ICML, pp 1587–1596

Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S, Murphy K (2017) Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of CVPR, pp 3296–3297

Johnson J, Krishna R, Stark M, Li L-J, Shamma DA, Bernstein MS, Fei-Fei L (2015) Image retrieval using scene graphs. In: Proceedings of CVPR, pp 3668–3678

Kaneko A, Komatsu A, Itoh T, Wang FY (2020) Painting image browser applying an associate-rule-aware multidimensional data visualization technique. Vis Comput Ind Biomed Art 3(1):1–13

Kang D, Shim H, Yoon K (2018) A method for extracting emotion using colors comprise the painting image. Multimed Tools Appl 77(4):4985–5002

Karpathy A, Li F (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of CVPR, pp 3128–3137

Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL (2013) Babytalk: understanding and generating simple image descriptions. TPAMI 35(12):2891–2903

Leite RA, Arleo A, Sorger J, Gschwandtner T, Miksch S (2020) Hermes: guidance-enriched visual analytics for economic network exploration. Vis Inform 4(4):11–22

Li Y, Fujiwara T, Choi YK, Kim KK, Ma K-L (2020) A visual analytics system for multi-model comparison on clinical data predictions. Vis Inform 4(2):122–131

Liu L, Wan X, Guo Z (2018) Images2poem: Generating Chinese poetry from image streams. In: Proceedings of ACMMM, pp 1967–1975

Lu C, Krishna R, Bernstein M, Fei-Fei L (2016) Visual relationship detection with language priors, vol 9905, pp 852–869

Lu J, Xiong C, Parikh D, Socher R (2017) Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of CVPR, pp 3242–3250

McCurdy N, Lein J, Coles K, Meyer M (2015) Poemage: visualizing the sonic topology of a poem. TVCG 22(1):439–448

Meneses L, Furuta R (2015) Visualizing poetry: Tools for critical analysis. paj: J Init Digit Hum Med Cult 3:1

Newell A, Deng J (2017) Pixels to graphs by associative embedding, vol NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp 2168–2177

Pinaud B, Vallet J, Melançon G (2020) On visualization techniques comparison for large social networks overview: a user experiment. Vis Inform 4(4):23–34

Ren S, He K, Girshick R, Sun J (2016) Faster r-cnn: towards real-time object detection with region proposal networks. TPAMI 39(6):1137–1149

Schuster S, Krishna R, Chang A, Fei-Fei L, Manning C (2015) Generating semantically precise scene graphs from textual descriptions for improved image retrieval. pp 70–80
Shi L, Liao Q, Tong H, Hu Y, Wang C, Lin C, Qian W (2020) Oniongraph: Hierarchical topology+ attribute multivariate network visualization. Vis Inform 4(1):43–57
Shu X, Wu J, Wu X, Liang H, Cui W, Wu Y, Qu H (2021) Dancingwords: exploring animated word clouds to tell stories. J Vis 24(1):85–100
Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint:1409.1556*
Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of CVPR, pp 2818–2826
Takahashi F, Kawabata Y (2018) The association between colors and emotions for emotional words and facial expressions. Color Res Appl 43(2):247–257
Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: Proceedings of CVPR, pp 3156–3164
Wang X, Zeng H, Wang Y, Wu A, Sun Z, Ma X, Qu H (2020) Voicecoach: Interactive evidence-based training for voice modulation skills in public speaking. In: Proceedings of CHI, pp 1–12. ACM
Wang Y, Haleem H, Shi C, Wu Y, Zhao X, Fu S, Qu H (2018) Towards easy comparison of local businesses using online reviews. Comput Gr Forum 37(3):63–74
Wang Z, He W, Wu H, Wu H, Li W, Wang H, Chen E (2016) Chinese poetry generation with planning based neural network. arXiv preprint arXiv:1610.09889
Wu L, Xu M, Qian S, Cui J (2020) Image to modern chinese poetry creation via a constrained topic-aware model. TOMM 16(2):1–21
Xu D, Zhu Y, Choy C, Fei-Fei L (2017) Scene graph generation by iterative message passing. pp 3097–3106
Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of ICML, pp 2048–2057
Xu L, Jiang L, Qin C, Wang Z, Du D (2018) How images inspire poems: Generating classical chinese poetry from images with memory networks. In: Proceedings of AAAI, vol 32
Yan R (2016) i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. pp 2238–2244
Yang J, Fan J, Hubball D, Gao Y, Luo H, Ribarsky W, Ward M (2006) Semantic image browser: bridging information visualization with automated intelligent image analysis, pp 191–198
Yang X, Tang K, Zhang H, Cai J (2019) Auto-encoding scene graphs for image captioning. In: Proceedings of CVPR, pp 10677–10686
Yi X, Li R, Yang C, Li W, Sun M (2020) Mixpoet: diverse poetry generation via learning controllable mixed latent space. Proc AAAI 34:9450–9457
Yi X, Sun M, Li R, Yang Z (2018) Chinese poetry generation with a working memory model. arXiv preprint arXiv:1809.04306
Zhang W, Siwei T, Liu K, Lei S, Chen S, Chen W (2019) A new perspective on the study of literature (songci): text correlation and spatio-temporal visual analytics. J Comput-Aided Des Comput Gr 31(10):1687–1697
Zhang X, Lapata M (2014) Chinese poetry generation with recurrent neural networks. In: Proceedings of EMNLP, pp 670–680
Zhao Y, Jiang H, Qin Y, Xie H, Wu Y, Liu S, Zhou Z, Xia J, Zhou F et al (2020) Preserving minority structures in graph sampling. IEEE Trans Vis Comput Gr 27(2):1698–1708
Zhao Y, Luo X, Lin X, Wang H, Kui X, Zhou F, Wang J, Chen Y, Chen W (2019) Visual analytics for electromagnetic situation awareness in radio monitoring and management. IEEE Trans Vis Comput Gr 26(1):590–600
Zhou F, Lin X, Liu C, Zhao Y, Xu P, Ren L, Xue T, Ren L (2019) A survey of visualization for smart manufacturing. J Vis 22(2):419–435
Zhou H, Huang M, Zhang T, Zhu X, Liu B (2018) Emotional chatting machine: emotional conversation generation with internal and external memory. In: Proceedings of AAAI, vol 32