Multimedia Systems (2016) 22:5–16
DOI 10.1007/s00530-014-0371-3

SPECIAL ISSUE PAPER

Chat with illustration

Yu Jiang · Jing Liu · Hanqing Lu

Published online: 11 April 2014

Abstract Instant messaging service is an important aspect of social media and sprung up in last decades. Traditional instant messaging service transfers information mainly based on textual message, while the visual message is ignored to a great extent. Such instant messaging service is thus far from satisfactory in all-around information communication. In this paper, we propose a novel visual assisted instant messaging scheme named Chat with illustration (CWI), which presents users visual messages associated with textual message automatically. When users start their chat, the system first identifies meaningful keywords from dialogue content and analyzes grammatical and logical relations. Then CWI explores keyword-based image search on a hierarchically clustering image database which is built offline. Finally, according to grammatical and logical relations, CWI assembles these images properly and presents an optimal visual message. With the combination of textual and visual message, users could get a more interesting and vivid communication experience. Especially for different native language speakers, CWI can help them cross language barrier to some degree. In addition, a visual dialogue summarization is also proposed, which help users recall the past dialogue. The in-depth user studies demonstrate the effectiveness of our visual assisted instant messaging scheme.

Y. Jiang · J. Liu (✉) · H. Lu
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences,
Beijing, China
e-mail: jliu@nlpr.ia.ac.cn

Y. Jiang
e-mail: yjiang@nlpr.ia.ac.cn

H. Lu
e-mail: luhq@nlpr.ia.ac.cn

Keywords Instant messaging service · Text-to-picture · Layout

1 Introduction

Social media has been showing a status of rapid development as network technique and computer technique moves forward, since Web2.0 era began. Different from traditional internet applications, social media emphasizes interaction. People are connected together by such interaction, so that they could exchange ideas and share all kinds of information through social media.

Instant messaging service like Tencent QQ, Google talk and Skype, is an important aspect of social media and also sprung up in last decades. As of 20 March 2013, there are 798.2 million active QQ accounts, with a peak of 176.4 million simultaneous online QQ users.[1] Information interaction through such kinds of traditional instant messaging service (TIMS) is usually confined on textual message. As a result, users can only obtain abstract information rather than all-around information. Thus many problems cannot be avoided, just like in the following aspects:

1. *Talk is just a talk.* The media between the users of TIMS is text. Limitations of text, abstractness and monotone, decide that such a kind of communication lacks of interesting qualities. TIMS also tries to do something to make the communication more funny. For example, the user of QQ also could sent images, but manually finding such appropriate images is time consuming.
2. *Talk may be around a mistake.* Due to different ages, geographical positions, and educational levels, the gap

[1] http://en.wikipedia.org/wiki/Tencent_QQ.

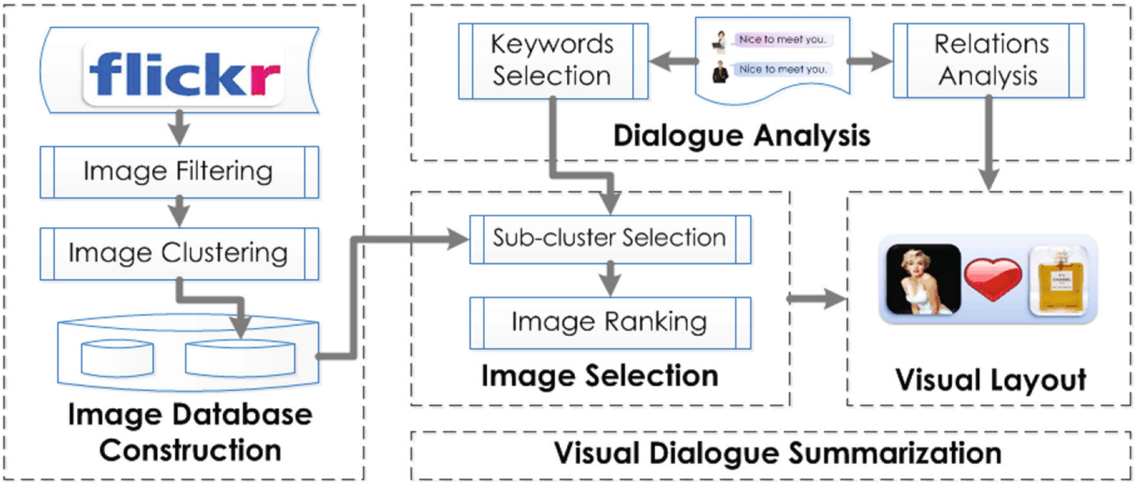


Fig. 1 The framework of CWI. There are five components: (1) image database construction, (2) dialogue analysis, (3) image selection and (4) visual layout and (5) visual dialogue summarization

on culture and understanding is widespread in population. A simple example, a talk on TIMS around "football" between a Chinese and an American, may become a totally confusing conversation. The American is talking about rugby football, while in Chinese's opinions, football is just soccer.

3. *Talk may be impossible.* Non-native speakers are usual poor in vocabularies or even have no ideas about a second language. In such a case, talk through TIMS between different language speakers is impossible. Machine translation may be resorted to. But confusion brought by machine translation still cannot be avoided. Therefore, a simple combination of machine translation and TIMS still cannot help different native language speakers carry out smooth conversations.

Early research efforts in education and psychology support the assertion that carefully constructed text illustrations generally enhance learners' performance on a variety of text-dependent cognitive outcomes [4]. Experiments comparing learning from illustrated text with learning from text alone in [15], also strongly reveal that illustrations aid learning of text material. These researches prove an old saying intuitively: "one image is worth of thousands of words". Motivated by the above observations, we argue that visual information could help users specify the others' true intents better, and an introduction of related visual message is a more natural way for instant messaging service than only passing textual message.

In this paper, we propose a novel instant messaging service named Chat with illustration (CWI). Different from TIMS, the CWI system presents visual message (i.e., illustration related to chat content) as well as textual message automatically. Thus, CWI overcomes the drawbacks of TIMS to a great extent. Due to the vividness and intuitiveness of visual message, a more interesting experience would be brought by CWI for users. CWI also would help to cross the gap on culture and understanding. When a picture of "rugby football" is shown to a Chinese, he or she would never think it as an association football. What is more, with the combination of illustration and results of machine translation, it is much easier to understand each other for different native language speakers, since the visual message compensates for the ambiguity brought by fallacious machine translation. In conclusion, the textual message transfers abstract concepts, while visual message transfers vivid concepts. Both of them are complementary, and the combination of them brings user a smooth, intelligible and vivid conversation experience.

As shown in Fig. 1, CWI comprises of five main components: (1) image database construction, (2) dialogue analysis, (3) image selection, (4) visual layout and (5) visual dialogue summarization. In order to support the CWI system, a large image database with semantic index is built offline firstly, whose construction falls back on social media, specifically, the image sharing service Flickr.[2] Image filtering and image clustering are performed to filtering junk images and build cluster-based semantic index, respectively. When users start their chat, CWI identifies meaningful keywords from their dialogue, and analyzes grammatical relations between these keywords and some logical relations between clauses. For each keyword, CWI discovers its most representative image from the image database. A two-step approach, including sub-cluster selection and image ranking, is developed to discover the

[2] http://www.flickr.com/.

representative image of a picturable[3] keyword. Afterward, image layout are performed with well-designed templates and different logical parts are assembled into an illustration. Finally, visual message is presented, as well as textual message. In addition, in order to help user recall what they have talked a few days ago, CWI would form a visual dialogue summarization so that they need not check textual chat history logs one by one, but just flash a look at the visual dialogue summarization. Visual message will refresh their memory.

The main contributions of this paper can be summarized as follows:

1. We propose a new visual assisted instant messaging scheme, named CWI, which provides users a more effective and interesting interaction experience through transferring both textual and visual message. Especially, CWI is very useful for different language speakers.
2. Some intuitive templates for image layout are designed based on grammatical relations between words and logical relations between clauses. This template-based method provides an easy and simple way to generate logical and integral visual messages of sentences in the dialogue.
3. A visual dialogue summarization scheme is proposed. The main picturable conceptions appearing in the dialogue are integrated into one illustration which can help users recall their memory of the past dialogue quickly.

The rest of this paper is organized as follows. Section 2 reviews related research. In Sect. 3, we describe the components of the scheme in detail. User study is presented in Sect. 4. Finally, we conclude the paper in Sect. 5.

2 Related work

Words-to-picture conversion has a close relation with our work. Image search engine, such as Google images,[4] just perform such a task. The user submits some keywords, and the image search engine returns related pictures. However the mechanism adopted by image search engine is a unimodal approach which mainly depends on text meta-data. A representative project of cross-model is Word2Image [16], which attempts to leverage both image tags and visual feature to translate a concept into its visual counterpart with sets of high quality, precise, diverse and representative images. Wang et al. [20] proposed a diverse relevance ranking scheme during tag-based image search that is able to take relevance and diversity into account by exploring the content of images and their associated tags. Another work related to us is text-to-scene conversion, the classical projects including NALIG [1], SPRINT [21], WordsEye [7] and CarSim [2, 8, 12]. SPRINT and NALIG are two early work to produce spatial reasoning visualization of the simple descriptive sentences. SPRINT operates with the Japanese language, while NALIG operates with the English language. WordsEye is a natural language understanding system for automatically converting text into representative 3D scenes by adding objects identified in the text. CarSim specially generates 3D graphics from traffic accident reports. It is intended to be a helpful tool that can enable people to imagine a traffic situation and understand the course of events properly [12]. Words-to-picture conversion cannot delivery meaning of a sentence, and text-to-scene conversion works only for descriptive sentences of collated objects. While our proposed CWI is a chat system, which is under a more complex and general natural language condition and delivery meaning of dialogue.

The work most closely related to ours is text-to-picture conversion. Zhu et al. [10, 22] investigated a text-to-picture system that synthesizes a picture from general, unrestricted natural language text. Dhiraj et al. [13] proposed a story picture engine, which depicts the events, happenings, and ideas conveyed by a piece of text in the form of a few representative pictures. Rada et al. [18] described a system for the automatic construction of pictorial representations for simple sentences. Ustalov [19] developed a text-to-picture system for Russian Language especially. Duy et al. [3] developed a system to create pictures to illustrate patient instructions. But all of the above text-to-picture work lack reasonable and intuitive visual layout, and some of them [18, 19] depends on a third-party image college. Beside, different from CWI, they are all not specially designed as a chat system.

3 The framework of CWI

In this section, we elaborate the implementation of CWI. First, we introduce the off-line process i.e., the construction of image database. And then the on-line processes including dialogue analysis, image selection and visual layout are explained. At last, visual dialogue summarization is interpreted. The notation used in the paper is defined in Table 1.

3.1 Image database construction

In order to meet the real-time characteristic of CWI, an image database facilitating to the chat visualization should

[3] Picturability is an attribute of word, which measures the probability of finding a good image to represent the word [22]. Some words with higher picturablily are more picturable, that means it is easier to find an image to represent this conception. Some other words are unpicturable, which usually are abstract concepts.

[4] http://images.google.com/.

Table 1 Notation

Symbol	Description
q	A keyword
i	An image of q
t	A tag of q
$\mathcal{T}$	Original tag set of q
$tagweight\ (t)$	The importance weight of t to q
$V^s(i)$	The semantic representation of i
$correlate(i)$	The semantic correlation of i to q
$saliency^v(i)$	The visual saliency of i
$saliency^s(i)$	The semantic saliency of i to q
$saliency(i)$	The saliency of i to q
$V^v(i)$	The visual representation of i
$similarity^s(i,j)$	The semantic similarity between i and j
$similarity^v(i,j)$	The visual similarity between i and j
$similarity\ (i, j)$	The similarity between i and j

be well collected and indexed in advance. The image database is divided into two parts, in light of keywords' picturability. One of two sub-databases corresponding to unpicturable keywords is built artificially. It is hard to automatically find proper images to represent these concepts, such as "large", "small", because they are abstract. But they convey important information of chat content. As a result, a few unpicturable concepts, such as some verbs, adjectives, fixed phrases and interrogatives etc., are considered and labeled to images manually. Some examples are showed in Fig. 2. In this sub-section, we elaborate the other sub-database, which is corresponding to the picturable keywords and built automatically. The data source is Flickr, a popular photo-sharing service, with plentiful resource of tagged images. It is reported in March 2013 that more than 3.5 million new images uploaded to Flickr daily.[5]

During the automated process of building the sub-database, there are several challenging issues. One problem we have to consider is that the quality of images on Flickr is not ensured. Some images may be irrelevant to the keyword, i.e., they even do not contain the concept of the keyword, while some images are indistinctive, though they include the concept of the keyword. Such images will lead to misunderstanding. Thus, we should remove these images firstly according to the tag semantic correlation to keywords and the saliency measure. The other problem we face is polysemy, which is very common. To take a simple example, the word of "*pitcher*" has two completely different semantic aspects: in baseball, pitcher is the player who throws the baseball, while it's also the meaning of ewer, an open vessel with a handle. To this end, we adopt a hierarchical clustering method by exploring both semantic and visual information, so as to obtain some sub-clusters of images with specific semantic aspect.

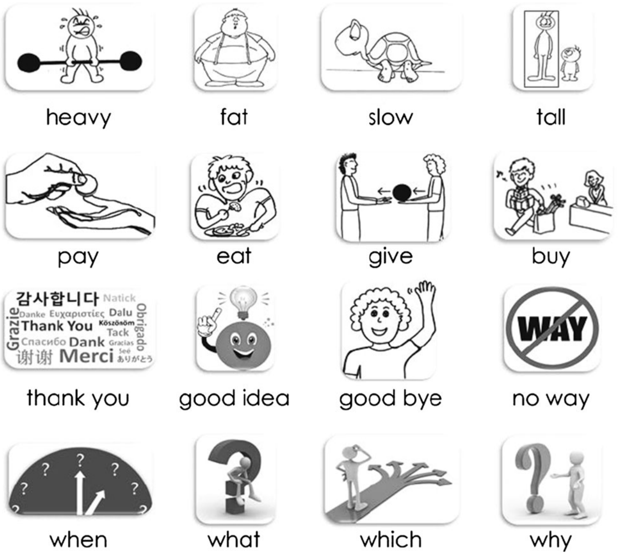


Fig. 2 Some example images of unpicturable keywords. In the first row are adjectives; second are verbs; third are fixed phrases and four are interrogatives

3.1.1 Image filtering

For each picturable keyword q, we collect the top P images with their tags from Flickr with the API.[6] As a result, a tag set $\mathcal{T}$ are formed. Many tags maybe appear more than once in $\mathcal{T}$. In order to ensure that images are precise and do not cause ambiguity, a two-level image filtering strategy is performed.

Firstly, we calculate the weight of each tag in $\mathcal{T}$ which measures the importance of tag to the keyword. The top M tags is kept to represent each image with an M-dimensional vector in the semantic space of keyword q. And then images could be ranked and filtered according to the semantic correlation score to q. Specifically, Two methods, the normalized Google Distance (NGD) [6] and the modified TFIDF scheme, are used to measure the weight of each tag. NGD is a theory of similarity between words and phrases based on information distance and Kolmogorov complexity. The smaller NGD is, the larger weight tag t has to keyword q. NGD(t, q) is defined as:

$$\mathrm{NGD}(t,q) = \frac{\max(\log N(q), \log N(t)) - \log N(t,q)}{\log N - \min(\log N(q), \log N(t))} \qquad (1)$$

where N(q), N(t) and N(q,t) denotes the number of pages reported by Google containing q, t and both q and t,

[5] http://en.wikipedia.org/wiki/Flickr.

[6] http://www.flickr.com/groups/api.

separately. N is the total number of web pages of search engine. Intuitively, more frequent tag will be more important, however, tag with higher document frequency (DF) will be too general and less informative. Similar to traditional TFIDF, the modified TFIDF is defined as follow:

$$\mathrm{TFIDF}(t) = freq(t) \times \log\left(\frac{N}{N(t)}\right) \tag{2}$$

where $freq(t)$ is the frequency of tag t in $\mathcal{T}$, the tag collection of query q, instead of word frequency in a document. N is the total number of images in Flickr, and $\mathrm{N}(t)$ is the number of images with tag t in Flickr. The weight for tag t to keyword q is estimated by a linearly combination of NGD and TFIDF.

$$tagweight(t) = \alpha f(\mathrm{NGD}(t,q)) + (1-\alpha)\mathrm{TFIDF}(t) \tag{3}$$

$f(.)$ is a certain monotonically decreasing function. Given the weight of each tag in $\mathcal{T}$, the top M tag $\{t_1, \ldots t_M\}$ are kept, and image i could be represented in semantic space by $V^s(i) = (V^s_1(i), \ldots V^s_M(i))$, where $V^s_m(i)$ is defined as:

$$V^s_m(i) = \begin{cases} 1, & \text{if } t_m \text{ is one tag of } i; \\ 0, & \text{otherwies.} \end{cases} \tag{4}$$

The semantic correlation of i to q is calculate by:

$$correlate(i) = \frac{1}{N}\sum_{m=1}^{M} tagweiht(t_m) \times V^s_m(i) \tag{5}$$

where N is the number of tags for image i. So far, images can be ranked according to the semantic correlation, and certain proportion images are filtered.

We do not just expected that images are precise, that means the images should include the concept of the keyword, but also wish that images are salient, i.e., the object corresponding to the keyword should be dominant in the image. The filter strategy by the semantic correlation removes most junk images. In the second-level filtering, we try to solve the issue of saliency through considering both visual and semantic aspects. Intuitively, the larger the saliency region, the more salient the image is. Recently, a new image saliency method is proposed in [5], which we adopt to detect image salient region. And visual saliency of image i is simply measured by

$$saliency^v(i) = \frac{salientarea(i)}{area(i)} \tag{6}$$

where *salientarea(i)* is the area of salient region of i and *area(i)* is the total image area. However, the visual saliency is not enough to ensure the image is salient for keyword. An example is show in Fig. 3, $saliency^v(i)$ would be high, but obviously it is not a good representative image of apple (fruit). Thus, we further make use of tags to solve this problem. In this example, besides “apple” and “fruit”, other tags have no relations with apple. Thus, we argue that the tags more semantic consistent to query keyword, the more salient the image is. the semantic saliency of image i to keyword q is defined as

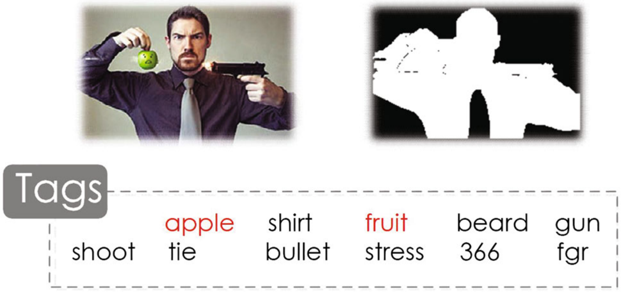


Fig. 3 The *left image* is not a good representative image of “apple”, though saliency region (the *white area* in the *right image*) is large. Only two tag is about “apple”. Tag information can be used to judge the saliency of image

$$saliency^s(i) = f\left(\frac{\sum_{n=1}^{N}\mathrm{NGD}(t_n,q)}{N}\right) \tag{7}$$

where $f(.)$ is a certain monotonically decreasing function, N is the number of tags for image i, and t_n is the nth tag of i. Finally, the saliency score for image i can be evaluated by

$$saliency(i) = \beta saliency^v(i) + (1-\beta)saliency^s(i) \tag{8}$$

and the images are further filtered according to the saliency score.

3.1.2 Image clustering

As mentioned above, polysemy is a serious issue which we should pay attention to. In this subsection, our objective is to divide these images into some sub-clusters with definite semantic aspects. Therefore, both semantic similarity and visual similarity of images are exploited, and an advanced clustering algorithm is applied to hierarchically cluster the images to semantically and visually consistent groups.

The semantic similarity between images i and j is defined as

$$simlarity^s(i,j) = \frac{\sum_{m=1}^{M} tagweight(t_m) \times V^s_m(i)V^s_m(j)}{|V^s(i)| \times |V^s(j)|} \tag{9}$$

where

$$|V^s(i)| = \sqrt{\sum_{m=1}^{M} tagweight(t_m) \times V^s_m(i)V^s_m(i)} \tag{10}$$

To construct visual feature space, we consider both the global and local features. 225-dimensional grid color moments (divide the image into 5 × 5 grids), 500-dimensional bag of words (SIFT) and 75-dimensional edge distribution histogram are extracted and normalized, respectively. Then, $V^v(i)$ the visual representation of image i is composed of these above feature. Visual similarity between each images i and j is calculated by

$$simlarity^v(i,j) = \exp\left(-\left\|\frac{V^v(i) - V^v(j)}{\sigma}\right\|^2\right) \tag{11}$$

The pair-wise similarity between each pair images i and j could be defined as a linear combination of semantic and visual similarities as

$$simlarity(i,j) = \gamma simlarity^s(i,j) + (1-\gamma) simlarity^v(i,j) \tag{12}$$

Given the pair-wise similarity, we apply a well-known clustering algorithm (i.e., affinity propagation [9]) to hierarchically cluster and generate sub-clusters with specific meaning. The clustering algorithm is selected since it is flexible on clustering numbers and has been extensively proved to be effective. So far, the cluster-based semantic index for the picturable keyword q is built.

3.2 Dialogue analysis

Once users start chat with CWI, the module of dialogue analysis start work, which is responsible of two tasks, meaningful keywords detection and grammatical and logical relations analysis. Meaningful keywords reflect users' intent in chat and are used as query words in image selection, while both of these two relations are used as foundation for visual layout.

For simplicity, pronouns, nouns, adjectives and verbs are considered as meaningful keywords, in that these words are informative and usually play important roles in sentences.

Grammatical relations represent dependency relations between words. For instance, which words are the subject or object of a verb, or which noun is served to by an adjectival modifier, etc. There are 52 grammatical relations proposed by [17]. We only pick up eight most important relations including adjectival modifier, conjunct, direct object, indirect object, negation modifier, nominal subject, possession modifier and prepositional modifier. All of them involve relations between pronoun, nouns, adjectives and verbs.

The Stanford Parser,[7] a Java implementation of probabilistic natural language parsers, is adopted as the tool of meaning keywords detection and grammatical relation analysis.

Logical relations represent relationship between clauses. Here, we only consider six logical relations, i.e., causal relationship, assuming relationship, turning relationship, progressive relationship and parallel relationship. Usually, there are some feature words standing for these relations in the sentences. For example, "*due to*" represents causal relationship, while "*but*" represents turning relationship. We detect these feature words and judge the logical relation.

3.3 Image selection

After obtaining the meaningful keywords of dialogue, representative images for them should be searched and selected from the image database, which are most consistent with the context of dialogue. For unpicturable keywords, the representative images are searched in the manual sub-database directly, while for picturable keywords, a two-step approach is developed to discover representative images in the other sub-database built automatically. In the first step, the most proper sub-cluster is selected from all sub-clusters which are clustered with specific semantic aspects. In the second step, with the help of visual and tag information, images in the selected sub-cluster is ranked and the most representative image for keyword in the specific dialogue is selected.

3.3.1 Sub-cluster selection

Only when the representative image for keyword meets with the dialogue context, the visual information help to improve chat experience. Thus, the most proper sub-cluster is selected according to so-called context clues which are other keywords in the same sentence or last a few sentences. First of all, the semantic representation of context clues is constructed. Since context clues usually are composed by just a few words, context clues are extended into a group of tags by Flickr related tag.[8] A M-dimensional semantic representation for context clues $V^s(context)$ is obtained through the most salient tag frequency statistics in the extended tag set. Similarly, for each sub-cluster, the semantic representation $V^s(subcluster)$ is also obtained through tag statistics in the sub-cluster. The distance between the context clues and each sub-cluster could be calculate with Eq. 9, and the nearest sub-cluster is deem as the candidate sub-cluster.

3.3.2 Image ranking

In order to obtain the most representative image, image ranking is performed within the candidate sub-cluster

[7] http://nlp.stanford.edu/software/lex-parser.shtml.

[8] Flickr API provides related tag inquiry. An interesting application of Flickr related tag: http://airtightinteractive.com/projects/related_tag_browser/app/.

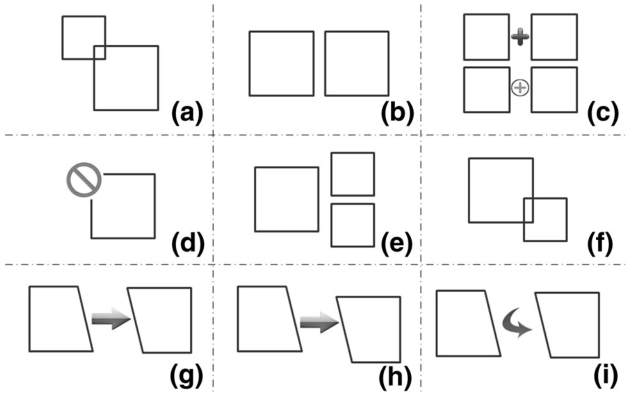


Fig. 4 Some most used layout templates. **a** adjectival modifier and possession modifier; **b** nominal subject and direct object; **c** conjunct, the upper one is for "and" conjunction while the lower one is for "or" conjunction; **d** negation modifier; **e** indirect object; **f** prepositional modifier; **g** causal relationship; **h** assuming relationship; **i** turning relationship

based on both semantic and visual aspects. On semantic aspect, semantic similarity between context clues and image $i\,similarity^{s}(context, i)$ is calculated as Eq. 9. On visual aspect, as in [14], we rank the images within the candidate cluster according to how well they represent the cluster. The intra cluster distance is used to estimate the visual representativeness of a image and defined as

$$intra(i) = \frac{\sum_{n=1}^{N} similarity^{v}(i, i_n)}{N} \tag{13}$$

where N is the image number and i_n is the nth image in the candidate sub-cluster. The final representativeness score of i is the linear combination of the above two terms.

$$represcore(i) = \eta similarity^{s}(context, i) + (1 - \eta)intra(i) \tag{14}$$

Given the representativeness score of each image in the candidate sub-cluster, images could be ranked, and the image with highest score is selected as the most representative one.

3.4 Visual layout

So far, we have obtained the proper representative images for meaningful keywords in dialogue. These images can only represent the certain concepts in isolation, but not an integral and logical meaning. In this section, we focus on a reasonable visual layout of images in the range of once reply, and present these images as a logical illustration.

A good layout scheme is expected to exhibit the following qualities:

1. It is intuitive to humans;
2. It is easy to generate by computers.

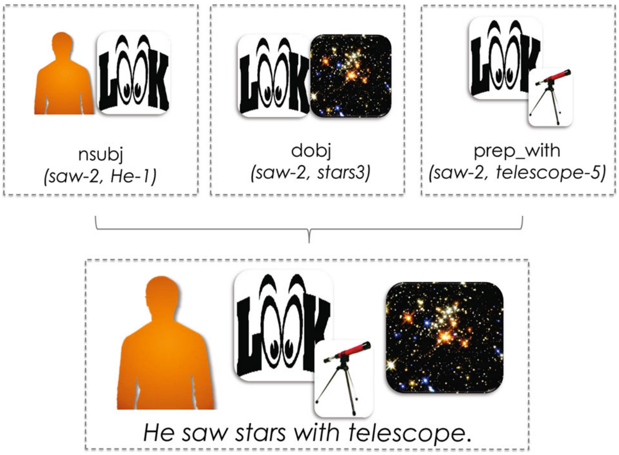


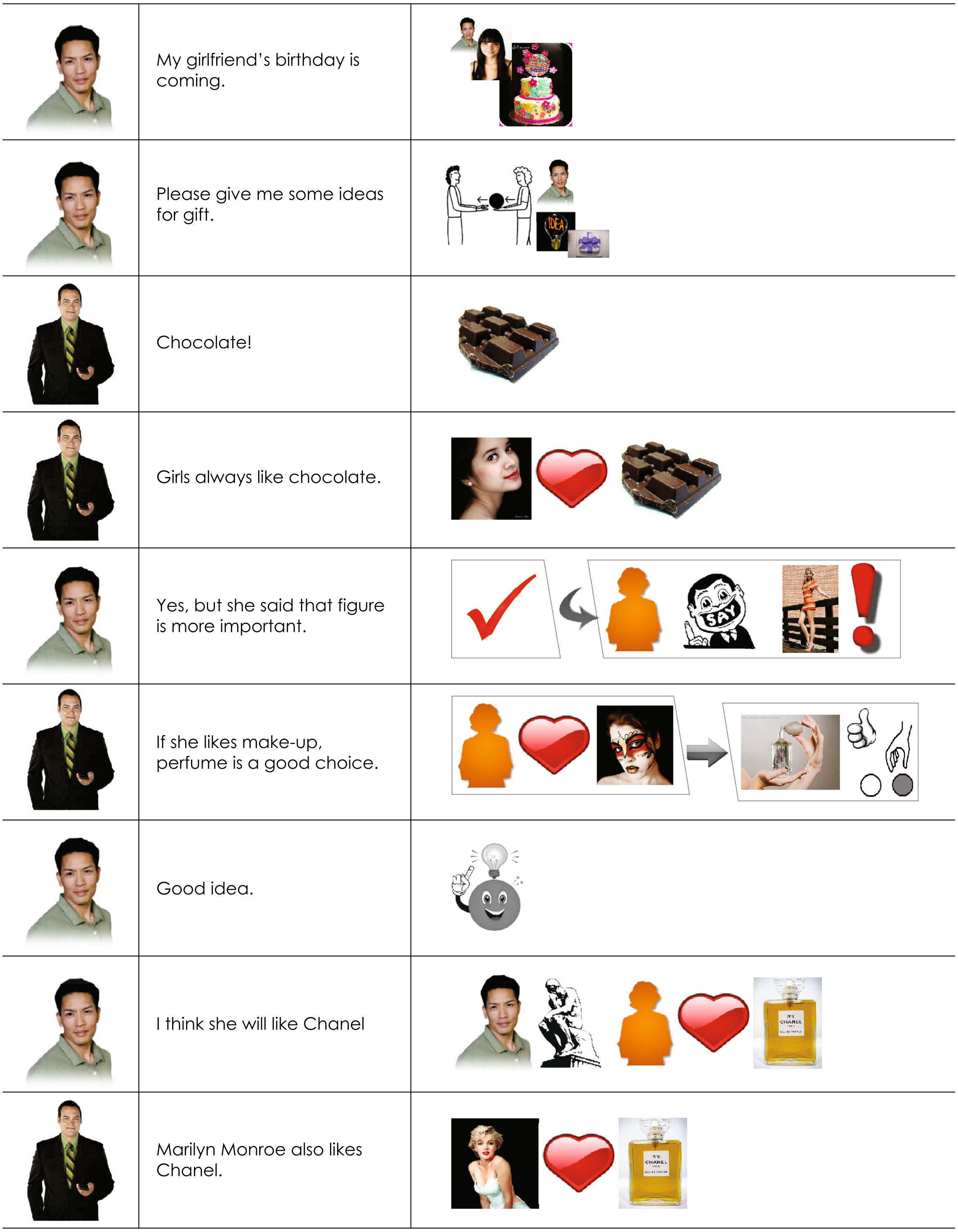
Fig. 5 An example of visual layout

We propose a template-based visual layout scheme. Some image layout templates have been designed based on grammatical relations between words and logical relations between clauses. Each template stands for a certain semantic unit. Parts of most common used layout templates are shown in Fig. 4. Templates (a)–(f) are semantic units for grammatical relations between words, while the other three are semantic units for logical relations between clauses. Templates are simple but intuitive. It is very easy to understand them. For example, the template (a) is stand for adjectival modifier or possession modifier. The big image is the object, and the small image located in the upper left corner is the adjectival phrase or somebody who owns the object. What need system to do is just to arrange images with these templates according to grammatical and logical relations. The grammatical and logical relations have been detected during dialogue analysis.

Generally, one reply in chat dialogue includes a few semantic units. Therefore, it is needed to synthesis different templates together. The synthetic step follows the below two criterion:

1. Different semantic units are connected by the co-ownership parts, if there are co-ownership parts among them;
2. The order of semantic unites are in accordance with order of keywords.

A simple visual layout example is demonstrate in Fig. 5. In this example, "*nsubj(saw-2, He-1)*" stands for "*He*" is the nominal subject of "*saw*"; "*dobj(saw-2, stars-3)*" means the direct object of "saw" is "*stars*", and "*prep_with(saw-2, telescope-5)*" reflects a propositional modifier serves to the meaning of verb. The co-ownership part is "*saw*", by which three parts are attached together.



My girlfriend's birthday is coming.

Please give me some ideas for gift.

Chocolate!

Girls always like chocolate.

Yes, but she said that figure is more important.

If she likes make-up, perfume is a good choice.

Good idea.

I think she will like Chanel

Marilyn Monroe also likes Chanel.

Fig. 6 A dialogue by CWI on the topic of "recommend gift"

3.5 Visual dialogue summarization

Complex pictorial information can be represented and retrieved from memory as mental visual images [11]. Visual information are not only helpful to remember something, but also to recall people's memory. Users often forget what they talked a few days ago. We propose a visual dialogue summarization scheme to help users recall their memory of the past dialogue quickly. Visual dialogue summarization, which comes into being when a chat is finished, is an illustration which including main picturable conceptions in the dialogue.

We only pick up picturable nouns and their corresponding images from the dialogues to ensure simplicity of the summarization. The challenge is how to integrate these images together as an illustration. We consider three properties of each image:

1. Size: A larger size of image stands for a more confident concept. The confidence of image i is measured by semantic similarity to context clues [i.e., $similarity^s(context, i)$]. In other words, a larger image is more consistent with context. It is worth to mention that the context clues here includes all the keywords in the dialogue.
2. Centrality: Each image only appear once in our visual summary, no matter how many times the corresponding conception appear. But the image of high-frequency concept which may be important, should be in the center of the illustration.
3. Distance: The distance of images measure the closeness of conceptions. If conceptions appear in the same sentence, their images should be more closer.

To sum up, size of image i is proportional to image semantic similarity to context clues; while centrality and distance determine the location of images in the illustration. The locations of all images are formulated as an optimization problem to minimize the objective:

$$\min \lambda \sum_{n=1}^{N} freq(i_n) dis(i_n, center) + (1-\lambda) \sum_{n=1}^{N-1} \sum_{m>n}^{N} q(i_n, i_m) dis(i_n, i_m) \tag{15}$$

where $freq(i_n)$ is frequency of the keyword corresponding to the nth image i_n. $dis(i_n, center)$ is the distance of the center of i_n to the center of illustration, while $dis(i_n, i_m)$ is the distance between the nth and mth images. $q(i_n, i_m)$ is an indicator function defined as

$$q(i_n, i_m) = \begin{cases} 1, & \text{if keywords in the same sentence;} \\ 0, & \text{otherwies.} \end{cases} \tag{16}$$

To solve this highly non-convex optimization problem, inspired by [10], we use a Monte Carlo randomized algorithm to construct multiple candidate illustration and then pick the one that minimizes the objective function.

During the process of constructing a candidate illustration, images are selected one by one and their locations are determined. First of all, the image of most high-frequency concept is placed at the center of the illustration. To select the next image to add to the illustration, we make a random decision between selecting an image based on word frequency or based on obeying closeness constraints.

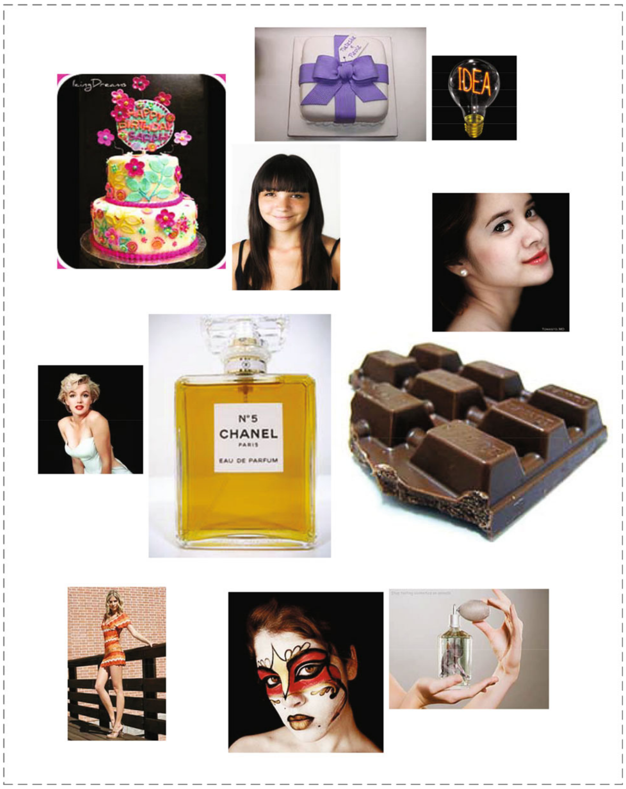


Fig. 7 An example of visual dialogue summarization which recalls users they once talked about "recommend gift". The high-frequency concepts chocolate and Chanel are in the center of illustration. If two concepts appear in the same sentence, two images would be close to each other, such as Marilyn Monroe and Chanel

The process of creating a candidate illustration is repeated many times (e.g. 500), and the best illustration (with the lowest objective function) is selected as the final result. Figure 6 is a dialogue example with CWI, and Fig. 7 shows the visual summarization of the dialogue.

4 Experiments

In this section, we conduct the user study to evaluate the effectiveness of the proposed scheme.

4.1 Image database and methodologies

We set 1,286 initial keywords totally on 15 different topics as shown in Table 2. To every keyword, top 1,000 images and their tags are downloaded from Flickr. The semantic feature space of one keyword is represented by 50 most related tags, i.e., $M = 50$. The function $f(.)$ in Eqs. 3 and 7 is a negative exponential function. As for

Table 2 Dialogue topics

Ask the way	Book hotel	Borrow book
Buy air ticket	Buy car	Recommend gift
Buy phone	Electrical repair	Money change
Movie	Order pizza	Rent house
Reserve sickness	Selected courses	Travel

the manual image database, we collect images for 132 common unpicturable concepts, most of which are line drawings or clip arts.

It is worth to mention that the image database is only built in English, i.e., sematic feature space of each query is in English. The reason is that the available resource of other language is not so extensive as English, e.g., in our system, images, tags and related tags of keyword from Flickr are applied extensively, however, Flickr API is not support other language, but English. Thus, our system work when the users chat in English, however, our proposed scheme is pervasive for other language.

4.2 User study

We invited 20 volunteers (16 male and 4 female, ages vary from 22 to 47) to participate into the evaluation of CWI. They are grouped into ten pairwise combinations randomly. All of the volunteers are skillful in both English and Chinese. As mention above, image can be presented only in English, so volunteers are always required express themselves in English. Before the study, volunteers are informed some priori knowledge, such as what is the meaning of different layout templates.

To sufficiently investigate the effectiveness of CWI, we first measure how much advantage our proposed scheme is able to gain for common users and different native language users, and then we further evaluate the components. Every pairwise combination are required to make five dialogues in the following strategies:

1. SL + TIMS: Chat in the same language on traditional instant messaging service, only textual message are transferred. And both of the users chat in English.
2. SL + CWI-L: Chat in the same language on Chat with Illustration without layout. Similar to CWI, but the images are arranged according to the words' order rather than proposed layout scheme. And both of the users chat in English.
3. SL + CWI: Chat in the same language on Chat with Illustration. Both textual and visual messages are transferred. And both of the users chat in English.
4. DL + TIMS + M: Chat in different languages on traditional instant messaging service with machine translation. A module of Google translation is integrated. The input text is in English, while the output text is in Chinese.
5. DL + CWI: Chat in different languages on chat with Illustration. Both textual and visual messages are transferred. The input text is in English, while the output text is in Chinese.

The last two strategies simulate international interaction, volunteers are always thought as Chinese speaker when they read the messages. In order to avoid the same chat content, before the dialogue, they should pick up a different topic from topic list as show in Table 2.

4.2.1 *Evaluation of full scheme*

We evaluate the full scheme of CWI from two aspects. In the first aspect, we compare CWI with TIMS, while in the other aspect, we evaluate usefulness of CWI during an international interaction.

The comparison between CWI and TIMS is based on SL + CWI and SL + TIMS. And the volunteers are required to provide the following evaluations:

- Enjoyment. It measures the extent to which user feel the process of interaction is enjoyable.
- Comprehension. It measures the extent to which user can understand each other easily.
- Recallment. It measures the extent to which the visual dialogue summarization can recall the memory of user easily, after a few days.

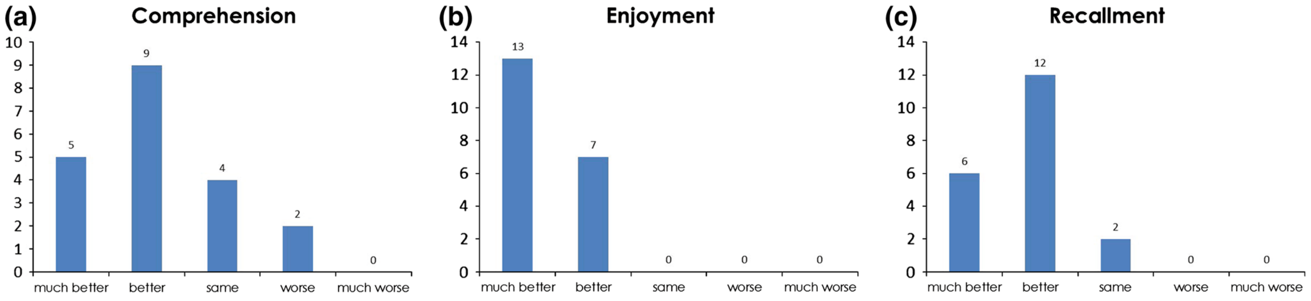


Fig. 8 a–c Comparisons between CWI and TIMS from three aspects: enjoyment, comprehension and recallment

Volunteers are asked to make a decision whether CWI performs "much better", "better", "same", "worse", or "much worse" than TIMS in the indicators of comprehension and enjoyment, respectively. We also ask the volunteers "do you think the visual dialogue summarization of CWI would help you recall the past dialogue 'much better', 'better', 'same', 'worse', or 'much worse' than traditional instant message tools?" Fig. 8 shows the results of individual indicators. Obviously, on all of the three indicators, CWI outperforms TIMS remarkably. Especially, all of the volunteers choose "much better" or "better" on the indictor of enjoyment. And almost all of the volunteers think the visual dialogue summarization would more helpful when they want to recall the dialogues after a few days. As for comprehension, 30 % volunteers choose "same" or "worse". Via communication with volunteers, it is found that they do not suit to the form of interaction.

To evaluate the usefulness of CWI system during an international communication, volunteers were invited to answer the question "Is CWI useful for expressing your intents and understanding the other's intents?" They were asked to choose one from three options: "very useful", "somewhat useful", and "unuseful". Figure 9 shows the evaluation results. The CWI system was regarded to be very useful by 35 % users and be useful by the remaining 65 % users. As for another question, "which form of communication will you choose during an international communication, DL + CWI or DL + TIMS + M", 18 volunteers choose our proposed scheme, while the other two have no preference between these two way.

From the above user studies, it can be found that CWI, transferring information through both textual and visual messages, brings users a better experience of communication than TIMS. What is more, CWI is very useful during an international communication when users are different native language speakers.

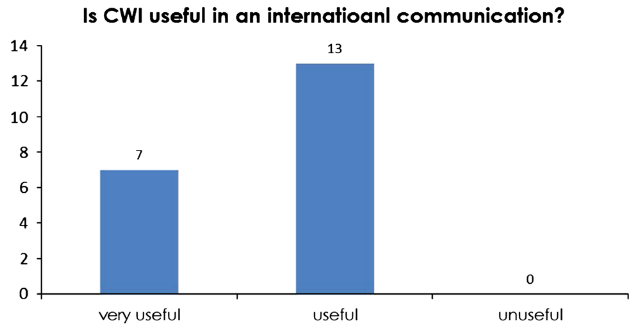


Fig. 9 Evaluation whether CWI is useful in an international communication

4.2.2 *Evaluation of components*

Now we further evaluate two components in the proposed scheme: image selection and image layout. We propose four criteria:

- Precision of selected image. It measures whether the selected images are correct and suitable for chat context.
- Saliency of selected image. It measures whether the object in the dialogue is salient in the selected image.
- Helpfulness of image layout. It measures whether the images layout is helpful to understand the content.
- Naturalness of image layout. It measures whether the images layout is consistent with human cognition.

The former two criteria measure the effectiveness of image selection. If the selected image includes the keyword, it is thought to be precise. And if the area corresponding to keyword is larger than quarter of the image, it is thought to be salient. We evaluate 448 images of picturable keywords in the 30 dialogues (10 pairwise, each pairwise performed SL + CWI-L, SL + CWI and DL + CWI). The precision is 93.97 %, and the salience rate in the precise images is 91.92 %. Thus, the image selection obtain a good performance both on the precision and saliency.

The latter two criteria measure the effectiveness of image layout. The user study is based on volunteers' experience on SL + CWI and SL + CWI-L. Volunteers are asked to choose "very helpful/natural", "helpful/natural" or "unhelpful/unnatural".

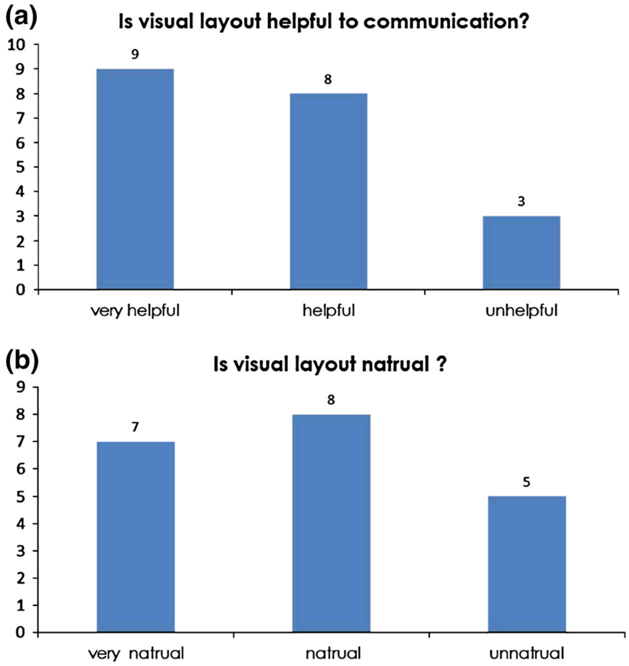


Fig. 10 Comparisons between SL + CWI and SL + CWI-L from two aspects: helpfulness and naturalness

Figure 10 shows the result of comparison between SL + CWI and SL + CWI-L. From the result, we find that majority of volunteers think SL + CWI performs much better than SL + CWI-L on both helpfulness and naturalness. That means our proposed visual layout scheme is effective.

All of the experiments above, clearly demonstrate the effectiveness of image selection and image layout.

5 Conclusion

This paper has demonstrated a novel visual assisted instant messaging scheme, named Chat with Illustration. Different from traditional instant messaging service, CWI not only provides textual message but also vivid visual message. Visual message is definite, vivid and intuitive. To common users, CWI is thus able to offer a more interesting, vivid and all-round information interaction. What is more, to different native language speakers, CWI can help them overcome language barrier to some degree. Besides, we also propose a visual dialogue summarization scheme, which integrates main concepts of dialogue into an illustration and help to recall people's memory. Extensive experiments have shown that CWI outperforms TIMS.

Our future investigations may include: (1) applying CWI on more different real-world chat scenes; (2) integrating other automatic dialogue topic detection techniques into the proposed system; and (3) introduce other multimedia information, such as flash, music into the system.

Acknowledgments This work was supported by the 973 Program (2010CB327905) and National Natural Science Foundation of China (61272329 and 61273034).

References

1. Adorni, G., Manzo, M.D., Giunchiglia, F.: Natural language driven image generation. In: Proceedings of the 10th International Conference on Computational Linguistics, pp. 495–500 (1984)
2. Akerberg, O., Svensson, H., Schulz, B., Nugues, P.: Carsim: An automatic 3d text-to-scene conversion system applied to road accident reports. In: Conference of the European Chapter of the Association for, Computational Linguistics, pp. 191–194 (2003)
3. Bui, D., Nakamura, C., Bray, B.E., Zeng-Treitler, Q.: Automated illustration of patients instructions. In: AMIA Annual Symposium Proc., pp. 1158–1167 (2012)
4. Carney, R.N., Levin, J.R.: Pictorial illustrations still improve students' learning from text. Educ. Psychol. Rev. **14**(1), 5–26 (2002)
5. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–416 (2011)
6. Cilibrasi, R., Vitnyi, P.M.B.: The google similarity distance. IEEE Trans. Knowl. Data Eng. **19**(3), 370–383 (2007)
7. Coyne, B., Sproat, R.: Wordseye: an automatic text-to-scene conversion system. In: Annual Conference on Computer Graphics, pp. 487–496 (2001)
8. Dupuy, S., Egges, A., Legendre, V., Nugues, P.: Generating a 3d simulation of a car accident from a written description in natural language: the carsim system. Computing Research Repository cs.CL/0105 (2001)
9. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
10. Goldberg, A.B., Zhu, X., Dyer, C.R., Eldawy, M., Heng, L.: Easy as abc?: Facilitating pictorial communication via semantically enhanced layout. In: Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL '08, pp. 119–126 (2008)
11. Ishai, A., Haxby, J.V., Ungerleider, L.G.: Visual imagery of famous faces: effects of memory and attention revealed by fmri. Neuroimage **17**, 1729–1741 (2002)
12. Johansson, R., Berglund, A., Danielsson, M., Nugues, P.: Automatic text-to-scene conversion in the traffic accident domain. In: International Joint Conference on Artificial Intelligence, pp. 1073–1078 (2005)
13. Joshi, D., Wang, J.Z., Li, J.: The story picturing engine - a system for automatic text illustration. ACM Trans. Multimed. Comput. Commun. Appl. **2**, 68–89 (2006)
14. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: Proceedings of the 17th international conference on World Wide Web, pp. 297–306 (2008)
15. Levie, W.H., Lentz, R.: Effects of text illustrations: A review of research. Educ. Technol. Res. Dev. **30**(4), 195–232 (1982)
16. Li, H., Tang, J., Li, G., Chua, T-S.: Word2image: towards visual interpreting of words. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 813–816. ACM, New York (2008)
17. de Marnee, M.C., Manning, C.D.: Stanford typed dependencies manual. Stanford University (2008)
18. Mihalcea, R., Leong, C.W.: Toward communicating simple sentences using pictorial representations. Mach. Transl. **22**, 153–173 (2008)
19. Ustalov, D.: A text-to-picture system for russian language. In: Proceedings of the Sixth Russian Young Scientists Conference in, Information Retrieval, pp. 35–44 (2012)
20. Wang, M., Yang, K., Hua, X.S., Zhang, H.J.: Towards a relevant and diverse search of social images. Trans. Multi. **12**(8), 829–842 (2010)
21. Yamada, A., Yamamoto, T., Ikeda, H., Nishida, T., Doshita, S.: Reconstructing spatial image from natural language texts. In: Proceedings of the 14th Conference on Computational Linguistics, pp. 1279–1283 (1992)
22. Zhu, X., Goldberg, A.B., Eldawy, M., Dyer, C.R., Strock, B.: A text-to-picture synthesis system for augmenting communication. In: Proceedings of the 22nd national conference on Artificial intelligence, pp. 1590–1595 (2007)