Image annotation tactics: transitions, strategies and efficiency

Jing Chen\(^a\), Dan Wang\(^a\), Iris Xie\(^b\), Quan Lu\(^c\)*

\(^a\) School of Information Management, Central China Normal University, Wuhan 430079, China
\(^b\) School of Information Studies, University of Wisconsin-Milwaukee, WI 53201, USA
\(^c\) Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China

**ARTICLE INFO**

**Keywords:**
- Image semantics
- Image annotation
- Human-computer interaction
- Annotation behavior

**ABSTRACT**

Human interpretation of images during image annotation is complicated, but most existing interactive image annotation systems are generally operated based on social tagging, while ignoring that tags are insufficient to convey image semantics. Hence, it is critical to study the nature of image annotation behaviors and process. This study investigated annotation tactics, transitions, strategies and their efficiency during the image annotation process. A total of 90 participants were recruited to annotate nine pictures in three emotional dimensions with three interactive annotation methods. Data collected from annotation logs and verbal protocols were analyzed by applying both qualitative and quantitative methods. The findings of this study show that the cognitive process of human interpretation of images is rather complex, which reveals a probable bias in research involving image relevance feedback. Participants preferred applying scroll bar (Scr) and image comparison (Cim) tactics comparing with rating tactic (Val), and they did fewer fine tuning activities, which reflects the influence of perceptual level and users’ cognitive load during image annotation. Annotation tactic transition analysis showed that Cim was more likely to be adopted at the beginning of each phase, and the most remarkable transition was from Cim to Scr. By applying sequence analysis, the authors found 10 most commonly used sequences representing four types of annotation strategies, including Single tactic strategy, Tactic combination strategy, Fix mode strategy and Shift mode strategy. Furthermore, two patterns, “quarter decreasing” and “transition cost,” were identified based on time data, and both multiple tactics (e.g., the combination of Cim and Scr) and fine tuning activities were recognized as efficient tactic applications. Annotation patterns found in this study suggest more research needs to be done considering the need for multi-interactive methods and their influence. The findings of this study generated detailed and useful guidance for the interactive design in image annotation systems, including recommending efficient tactic applications in different phases, highlighting the most frequently applied tactics and transitions, and avoiding unnecessary transitions.

1. Introduction

Millions of images become accessible to people with the new development of Internet technology. In order to retrieve the digital images that exactly satisfy users’ needs, annotating the content with semantically meaningful labels is of great necessity. Prior image annotation approaches, regardless of interactive ones, such as relevance feedback techniques (Nezamabadi-Pour & Kabir, 2009; Rui, Huang, Ortega, & Mehrotra, 1998), or non-interactive ones, such as multiple feature spaces (Sun, 2013; Xu, Tao, & Xu, 2015) and other statistical models (Feng & Lapata, 2010; Qi & Han, 2007), fail to tackle the challenge of semantic gap fundamentally (Ke, Li, &
on one hand, users of current human-computer interactive image annotation tools are supposed to annotate an image by tagging, while ignoring whether the description expressed by tags matches image semantics. On the other hand, human interpretation of images is complicated, while its process and mechanisms remain uncertain (Ivasic-Kos, Ipsic, & Ribaric, 2015). At the same time, how users perceive images in semantic ways is not very well understood. Consequently, more insight into human-computer interactive image annotation behavior is needed, so that users' interpretation process in annotation of images can be understood, narrowing the semantic gap, ultimately resulting in better image annotation experiences for all.

At present, there are three major interactive methods usually used to investigate an individual's image interpretation: (a) normative rating method (Scherer, 2005), (i.e. typically features a rating scale from 1-5, 1-9, etc.); (b) semantic visualization based scroll bar method (Schmidt & Stock, 2009), i.e. dragging the scroll bar to express the image's semantic meaning; (c) ranking visualization based comparison method (Hardoon & Pasupa, 2010), i.e. comparing with other images when evaluating the current image. Different methods represent users' diverse image cognition processes, and their own strengths and weaknesses. Rating method is characterized by quantifying evocation into specific value(s). The scroll bar and image comparison methods are relatively perceptual. In the scroll bar method, users are supposed to adjust the scroll bar to express perceived semantic intensity. When performing a comparison method, users' image interpretation may be influenced by reference image(s) since they would estimate differences in generating an appropriate annotating result. The existing studies on image semantic interpretation are all conducted under one single method, such as rating (Grühn & Scheibe, 2008), scroll bar (Schmidt & Stock, 2009) or image comparison (Junge & Reisenzein, 2015), which may produce a bias in collecting image relevance feedback. Users may not express the image interpretation accurately by just rating, the most widely used annotation method. In addition, this kind of single annotation method application seems to limit the comprehensive investigation of users' sophisticated annotation behavior. Hence, the combination of multiple interactive methods helps researchers understand the complicated process of image interpretation.

Moreover, in the field of information retrieval (IR), systems usually offer users multiple interactive activities, such as formulating queries, scanning websites, and so on (Baskaya, Keskustalo, & Järvelin, 2013; Niu & Kelly, 2014). Searching behavior patterns (e.g., frequently applied actions, transition of these actions) generated by selecting or combining these operations enable researchers to study general users' searching behaviors and strategies (Joho, Jatowt, & Blanco, 2015; Xie, 2011). Therefore, this approach of user behavior investigation in search fields has provided reference for image annotation. In this study, the authors intend to create an interactive annotation environment that provides various methods to explore users' image annotation behavior.

With respect to the essential components of the annotation process, the authors offered an explanation of annotation moves, tactics, and strategies for this study after merging definitions proposed by Bates (1979, 1990) and Marchionini (1995) for search moves, tactics, and strategies. “Move” is a common alternative word for “tactic”. Annotation moves are basic thoughts or actions in the annotation process. Tactics indicate a move or moves, including annotation choices and actions that users apply to advance their annotation. Annotation strategies represent patterns of sequential tactics which imply users' plans for the annotation. In the annotation process, annotation tactics can be identified on the basis of annotation moves (e.g., dragging the slider, considering the position of the scroll bar, evaluating the value, and comparing with the reference image.) (Schmidt & Stock, 2009; Zhang, Fu, Liang, Chi, & Feng, 2010), which is similar to the search process (Shiri & Revie, 2003). Since the objective of this study is to figure out users’ image interpretation processes, annotation moves discussed here are mainly related to the user's cognitive involvement. In addition to these typical moves, tactics, such as fine tuning (Cuzzola, Jovanović, Bagheri, & Gašević, 2015), are critical to the success of the annotation process. In that most works of the search strategies are in relation to topic refinement (Chen & Dhar, 1991; Xie & Joo, 2010), the act of fine tuning plays the same role as refinement, which is the basis to classify annotation strategies into two types in this study. The first type represents the common patterns of tactic selection without fine tuning, while the second type reveals the tactic chains including fine tuning. Regardless of whether a user refines the annotation result or not, the strategy he or she employs implies his or her plan for the annotation.

The annotation process is a complicated and dynamic one, in which tactics adopted by users may change during the process. To understand the process, merely identifying tactics is not enough. It is also critical to look into the transitions among annotation tactics. Each transition focuses on detecting changes between two annotation actions. In addition, a chain of tactic transitions within a strategy illustrates the annotation process. Unfortunately, few researchers have investigated transitions in annotation tactics. Parallel to search tactic research, where transitions have been conducted via discourse or log analysis (Rieh & Xie, 2006; Xie & Joo, 2010), the occurrences of transitions in tactics could also be derived from utterances or log data to highlight a user's successive decision-making process when accomplishing an annotation task.

This study focuses on exploring users' annotation tactics, transitions, strategies and efficiencies by investigating users' behaviors in dealing with annotation tasks. The analysis of annotation tactics helps researchers acquire a better understanding of the nature of human's interpretation process to images. The study of annotation tactics, transitions, strategies and efficiencies offers an opportunity to recognize users' image annotation behavior, thus provide insights into annotation behavior research, and finally improve existing annotation systems to assist users optimize their annotation behaviors and enhance their image annotation experience.

2. Related Work

2.1. Image Annotation

The semantic gap, in most of the previous research, has been defined as the difference between the subjective users' understanding of an image and the objective computer's interpretation of the users' annotation (Bahmanyar, Murillo Montes De Oca, & Datcu, 2015). Early solutions attempting to fill the semantic gap is to use relevance feedbacks from users (Nezamabadi-Pour & Kabir, 2011).
However, the need of significant amount of user intervention makes it an impractical approach. Therefore, automatic image annotation (AIA) was emerging to label images with semantic keywords through machine learning tools so as to facilitate textual queries. Many machine learning algorithms and statistical models have been proposed in recent years (Bahrami & Abadeh, 2014; Bahrololoum & Nezamabadi-Pour, 2017). Even though some of these techniques have shown impressive results, it is still difficult to contemplate an automatic procedure for users to recognize objects and understand images (Song, Wang, & Zhang, 2003). Enser, Sandom, Hare, and Lewis (2007) pointed out two limitations of automatic annotation involving generic features and non-visible features including time, space, events and significance, abstract and emotive concepts, and unwanted features. These phenomena can only be inferred by high-level interpretation. Meanwhile, Ivasic-Kos, Ipsic, and Ribaric (2015) offered their explanation for the complexity of the interpretation process since it requires additional reasoning with general and domain-specific knowledge which is often incomplete, imprecise, uncertain and ambiguous in nature. In addition, the understanding of users’ image interpretation process is still limited. Therefore, the challenge of automatically annotating images by semantics and the scarcity of annotation behavior research calls for the need to investigate human-computer interactive annotation processes.

Most existing interactive image annotation systems are generally operated on the basis of social tagging, while few have functioned from the dimension of semantic interpretation. Representative examples include the ESP system (von Ahn & Dabbish, 2004), LabelMe (LabelMe, 2012) and Flickr (Flickr, 2018), where users are allowed to attach appropriate labels or tags to each image. However, it is inadvisable to neglect associating tags with semantic levels. Shatford (1986) confirmed that tags do not separate the different semantic levels of ofness and aboutness. Kipp (2009) observed that linguistic tags, such as cool or fun, do not appear to add anything to the subject classification of an item and seem to be poor candidates for search terms for information retrieval. Similarly, Huron (2000) emphasized that in music information retrieval, the standard reference tags have very limited applicability in most music-related queries.

In addition, image semantics can be divided into several levels. Eakins (1996) identified three levels of image semantics in order of increasing abstraction. Level 1, the lowest level, comprises primitive features such as color, texture, shape or the spatial location of image elements (e.g. “Find all images containing yellow or blue stars arranged in a ring”); Level 2 consists of derived attributes involving some degree of logical inference about the identity of the objects depicted in the image (e.g. “Find images of a passenger train crossing a bridge”); Level 3 includes abstract attributes involving a high degree of abstraction, and possibly subjective reasoning about the meaning and purpose of the objects or scenes depicted (e.g. “Find images illustrating pageantry”). This level can be subdivided into events depicted in the image and emotional or symbolic significance of a picture. At this level, individuals’ emotional response, comparing to tagging, plays a more critical role in understanding their complicated image interpretation process.

In previous works, three fundamental interactive methods were presented for users to perceive image semantics: (a) normative rating method; (b) semantic visualization-based scroll bar method; (c) ranking visualization-based comparison method. Rating method features numeric emotional evocation is by far the most common one. Lang and colleagues used the Self-Assessment Manikin (SAM) (Lang, 1980), a nonverbal affective rating system, to assess the dimensions of valence, arousal and dominance of pictures derived from the International Affective Picture System (IAPS) (Lang, Bradley, & Cuthbert, 1997). In this rating system, a series of graphical figures is presented for each dimension, and each series forms a 9-point rating. Instead of using the SAM rating format, Libkuman, Otani, Kern, Viger, and Novak (2007) applied a 9-point Likert scale for the purposes of uniformity to obtain ratings on 14 dimensions, including the previous norms in the IAPS, the additional 5 dimensions, and the 6 emotions. Scroll bar and image comparison methods are comparatively less numeric. Scroll bar method reflects the extent of emotion has been experienced. Lee and Neal (2007) developed the Glass Engine, which allows users to browse several emotional facets of Glass’ music via sliding bars. Following Lee and Neal’s approach, Schmidt and Stock’s (2009) applied a scroll bar to determine the intensity of five basic emotions upon viewing images. On a scale from 0 (no evocation of the given emotion quality) up to 10 (very evident evocation), the test subjects moved the scroll bar to the point of their perceived emotion intensity. As for the image comparison method, it enabled users to discriminate emotional stimulus differences when viewing different pictures simultaneously. Hardoon and Pasupa (2010) explored an image ranking experiment with implicit feedback according to eye movements, where users were shown 10 images on a single page and were asked to rank the top five images in order of relevance to one topic. Zhang, Fu, Liang, Chi, and Feng (2010) evaluated the possibility of inferring the relevance of images based on eye movement data as well. In their experiment, each participant was asked to use their eyes to locate the positive image on each searching stimulus with 20 candidate images. Rather than studying only image relevance, Junge and Reisenzein (2015) compared two nonmetric probabilistic scaling methods (MLDS and ODS) for the measurement of emotion intensity. Participants were asked to indicate the size of the difference between two stimuli (graded pair comparisons, ODS) and between pairs of stimuli (quadruple comparisons, MLDS). And ODS was found to perform at least as well as MLDS in the reliability of the estimated scale values and their correlations to direct ratings of emotion intensity. Although the three interactive methods have been applied by users to interpret images, cognitive involvement of each method has rarely been researched.

2.2. User annotation behavior

Existing literature on interactive image annotation behavior emphasizes annotation motivations, results or providing insight into image retrieval, while few studies have examined users’ image annotation behaviors from the perspective of cognitive processes. For instance, Marlow, Naaman, Boyd, and Davis (2006) described a detailed set of motivations in social tagging systems comprising of: future retrieval, contribution and sharing, attention seeking, play and competition, self-presentation and opinion expression. Ames and Naaman (2007) proposed a taxonomy of motivations for tagging in Flickr and ZoneTag in relation to two dimensions: (a) function, and (b) sociality. Apart from research on annotation motivations, Grünl and Scheibe (2008) provided rating result of
valence and arousal for 504 IAPS pictures by young and older adults to verify the age-related difference in the perception of these pictures. Other research investigates the relationship between image annotation and image retrieval. Han, Xu, Li, Guo, and Liu (2014) developed an interactive system for users to efficiently retrieve and annotate image objects on an iPad. This system allows users to manually annotate the query objects and their annotations can be exported to the retrieval results. Evaluations on three publicly available benchmark image databases and comparisons with previous approaches have demonstrated the effectiveness and efficiency in image retrieval of the proposed system. Im and Park (2015) also proposed a tag-based image annotation system (LinkedTag) that exploited Linked Data, such as DBPedia to insert semantic relationships between tags, and the results show the system is promising for image search.

In the field of users’ interactive behavior, search behaviors have been studied extensively. Apart from the existing literature on image annotation behavior, relevant literature on search behaviors incorporated in this paper, in particular related to tactic, transition, efficiency, and system.

First, annotation behaviors can be discussed on two levels (tactic/moves and strategies) based on their units of analysis like search behaviors. Annotation tactics can be formulated based on annotation moves to represent the process. The most commonly occurring interactive move in image annotation is rating. In a significant portion of relevant studies, a Likert type of rating format (e.g., a 9-point scale) is used (Scherer, 2005), and individuals are instructed to assign a certain integer value for pictures derived from the IAPS on various dimensions based on the research purposes. However, assigning exact numbers, e.g., such as “I like the picture of mountain 2 times better than sea” is difficult for users (Pommeranz et al., 2012). Instead of evaluating pictures by accurate numbers in the range, scroll bar enables users to coherently annotate images (Schmidt & Stock, 2009). Users’ image perceptions are expressed by means of moving the slider to a point within the range. The adjusted position of the graphically presented scroll bar is then converted into numerical values, which serves as an indicator for the accuracy of the tag. A more expressive and intuitive way, image comparison has also been used in some research to obtain users’ difference judgments through image ranking (Hardoon & Pasupa, 2010). Cognition is less demanding since it does not require numerical input for images from users. Image recognition needs to be compared with reference images but are not associated with a numerical weight. Moves mentioned are all essential components in conducting annotation tasks. Apart from these, acts like fine tuning also facilitate users to make the annotation process success. It functions as a mechanism to refine in search query reformulation (Shute & Smith, 1993), and divides an annotation process into different phases. All the annotation moves mentioned above are the basis for annotation tactics identified for this study. Additionally, annotation strategies, which determine the overall plan for a whole annotation session, are classified based on whether the act of refinement occurs. The first type represents the common patterns of tactic selection without fine tuning, which enables the authors to explore the usage of tactics, such as single or combinative. The second type reveals the tactic chains including fine tuning, which assists the investigation of whether tactics employed change after refinement.

Second, few studies on transitions in annotation moves have been conducted, compared with studies on moves. However, tactics are not supposed to be regarded as discrete activities, the study of tactic transition shows great importance since it better reflects how users interpreting images. In search studies, Bates (1989) proposed a searching model called “berrypicking”, which argues that searches are evolving and dynamic. A person constantly changes their search tactics in response to the results returned from the IR system. The same holds true for image annotation. The act of annotating generates feedback which may lead users to change their cognition to images, and thus the annotation evolves. In this study, transitions between annotation tactics are investigated to reflect a user’s successive decision-making process, determine how users annotate images, and capture the nature of interactive image annotation.

Third, annotation efficiency is quite important for successful image annotation, since it reduces annotation time during the annotation process. As Debowski (2001) pointed out search activities may lead to redundant search processes in the form of repetitive strategies. These redundant activities caused distractions and challenges the achievement of search quality. He examined the search process of novices and found that much time was spent in attempting to manage the search history, exploring the search space and reading the records. Hence, it is also important to discover productive annotation tactics in order to reduce redundant annotation behaviors. Also, annotation efficiency can be improved by supporting these efficient annotation tactics.

Finally, previous studies on evaluating image interpretation are mainly conducted under one single annotation method (Hardoon & Pasupa, 2010; Scherer, 2005; Schmidt & Stock, 2009). All of these methods are suitable for image semantic evaluation, whereas cognitive involvement in each method differs. The probable bias generated in these works limits the investigation on users’ sophisticated annotation behaviors. On the contrary, various options are available for users in IR systems, including scrolling up and down, checking search history, and so on. For instance, Xie and Joo (2010) examined real users’ transitions in search tactics. Their results highlighted users’ multiple search tactics, the most frequent and most probable tactic transitions, and the most common search strategies occurring at different phases within one search session.

Unlike research on search behaviors, the existing studies on image annotation have not explored the nature of human image interpretation behaviors and process. The unanswered questions are in relation to whether users’ diverse image interpretation processes exist, whether the annotation process evolves during the searching process, whether annotation behavior patterns exist or efficient annotation tactics exist. There is a need to explore annotation tactics, transitions, strategies, and efficiency in the human image annotation process.

3. Research Questions

Even though previous research has identified several interactive annotation moves, little is known about interactive annotation patterns during the image annotation process, and in particular, patterns of annotation tactic transitions. This study addresses the following four research questions:
RQ1: What are the frequently applied annotation tactics in the image annotation process?
RQ2: What is the probability of annotation tactic transitions in the image annotation process?
RQ3: What are the most frequently applied annotation strategies representing the common patterns of annotation tactic transitions?
RQ4: What is the time distribution in annotation tactics in different phases of image annotation process?

4. Methodology

To address these questions, the researchers designed a user study, and systematically collected and analyzed the data. Multiple methods were applied to collect data; correspondingly, these data were analyzed through quantitative and qualitative methods.

4.1. Sampling

A total of 90 participants were recruited from various schools of Wuhan University to complete the designed image annotation tasks. The sample was stratified by educational level. Few participants were experienced in image annotation. Characteristics of the participants are presented in Table 1.

4.2. Data Collection

In the previous work, researchers have developed an Image Semantic Annotation Research Platform (ISARP) (Lu, Liu, & Chen, 2014; Fig. 1), where the experiment was conducted. In Figure 1, the rating number is shown under the annotated image on the left. Two reference images were provided for comparison on the right with the scroll bar displayed below. The lower scroll bar was a larger version of the upper one and was offered for the current annotation. There were three sliders with different colors within each scroll bar: yellow represents annotated image, red represents the left reference image, and blue represents the right reference image. The participants were asked to express the intensity of nine images from three emotional dimensions (pleasure, arousal, dominance) on ISARP. The nine images and seventy-eight reference images (26 * 3 dimension) were from IAPS provided by University of Florida. According to IAPS, every image is rated with one of three different affective dimensions: pleasure, arousal and dominance. Pleasure refers to affective valence (ranging from 1 for unpleasant to 9 for pleasant); arousal marks the intensity of an activated motivational system (ranging from 1 for calm to 9 for excited); and dominance represents the control state of an individual on situations or others (ranging from 1 for in-control to 9 for controlled). Ratings for the nine annotated images were evenly distributed in each affective dimension. Ratings for the twenty-six reference images in each dimension approximately divided the slider into 27 equal segments. Table 2 presents the annotated images’ affective ratings determined by IAPS based on normative rating by college students. For example, the image of Aimed Gun on the affective scale (2.37, 7.35, 2.15) represents a low rating on pleasure (2.37), a high rating on arousal (7.35), and a low rating on dominance (2.15). To eliminate the influence of the order of presented images, three annotation tasks in this experiment corresponding to the three dimensions assigned to each participant through rotation. The participants were assigned randomly to one of three groups depending on the order of task completion: group A performs the tasks by the sequence of one-two-three; group B is two-three-one; and group C is three-two-one.

At first, participants were asked to fill out a pre-questionnaire requesting their demographic information and their experiences in image annotation. Then, they were invited to a university usability lab to visually annotate each dimension of an image for the above nine images by applying three interactive methods. Participants could drag and drop a slider on the scroll bar and were allowed two chances of further refinement. The three viewing options could be freely combined with each other, including the slider’s position on

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics of participants (N = 90).</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Demographic characteristics</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Education level (current status)</td>
</tr>
<tr>
<td>Undergraduate</td>
</tr>
<tr>
<td>Graduate</td>
</tr>
<tr>
<td>Major</td>
</tr>
<tr>
<td>Liberal arts</td>
</tr>
<tr>
<td>Social Science</td>
</tr>
<tr>
<td>Science</td>
</tr>
<tr>
<td>Engineering</td>
</tr>
<tr>
<td>Information Science</td>
</tr>
<tr>
<td>Image annotation experience</td>
</tr>
<tr>
<td>Never</td>
</tr>
<tr>
<td>Rarely</td>
</tr>
<tr>
<td>Sometimes</td>
</tr>
<tr>
<td>All the time</td>
</tr>
</tbody>
</table>
the scroll bar, a numerical value converted from the slider's position, and two reference images from IAPS to divide the rating range into three almost equal parts. During the annotation process, the participants were asked to "think aloud" to express their feelings, thoughts and intentions regarding their annotations. A usability testing software, Morae, was installed to the computer for recording
the entire process of the experiment, including timestamp, participants’ movements and their thinking aloud protocols. The process of annotating one image in a dimension was defined as an annotation session. In total, 2430 (3*9*90) log files (namely, annotation sessions) were recorded by Morae.

4.3. Data Analysis

The unit of analysis is each annotation tactic. As defined in the Introduction, annotation tactics refer to a move or moves, including annotation choice and actions that users apply to advance their annotation process. A coding scheme (Table 3), consisting of key annotation tactics, was created through content analysis, a highly flexible research method that has been widely used in library and information science studies for user behavior research (White & Marsh, 2006). Based on previous research and data generated from this study, the authors classified all tactics into three types of codes: (a) image selection tactic, (b) fine tuning tactics, and (c) annotation tactics. “Fine tuning tactics” was further divided into “First fine tuning” and “Second fine tuning” that participants employed to optimize their annotation and reach more accurate results. Upon clicking the fine tuning button, the reference images changed, and the slider interval range narrowed to 1/3 size of the original one, reducing measurement accuracy to 0.1 after refining results twice. “Annotation method” consists of the three interactive methods proposed earlier (i.e. rating, scroll bar, and image comparison). Rating number was generally set in integer (e.g. Likert scale); in this study, it was set in real number which could more accurately assess an emotional intensity value, for example, the pleasure rating value on the image of Aimed Gun was 3.18. The definition and example of each code are presented in Table 3. Notably, “comparing to pre-image (PIM)”, a thinking activity observed during the annotation process, which does not belong to the original interactive annotation methods, was regarded as a subclass.

In the end, seven types of annotation tactics were defined in this coding scheme. An abbreviation is used to represent each tactic for simplicity. Descriptive analysis was conducted to obtain the frequency of each type of tactic applied in the annotation process.

Apart from coding for each annotation tactic, transitions of tactics were coded for all annotations as well. A tactic was coded according to the order of the first selection. For example, if the move of “considering the position of scroll bar (Scr)” was selected first, and occurred constantly along with “assess the value (Val)”, then the transition Scr to Val would be coded only once. In addition, reoccurring tactics were coded once as well. The transitions between tactics were co-determined by click-through log analysis and think aloud protocol analysis. The first two annotation types (namely Sel, NfF, and NfS) of the coding scheme were identified based on click data, think aloud data was used to identify the third annotation type (namely Scr, Val, Cim, and Pim). Each move was coded and linked with its previous and following moves to see whether a participant modified their annotation tactics. In addition, verbal protocols corresponding with each move was analyzed to figure out participants’ mental activities, thus assisting in coding and seeking out the reasons for changing tactics. The transition of annotation tactics were recorded when a change of annotation tactic was identified.

To test the intercoder reliability of each annotation tactic, two researchers independently coded 30 samples randomly selected from 90 experimental samples. The result of intercoder reliability for coding each tactic has a high consistency, at 0.91 according to Holsti’s (1969) reliability formula. Reliability = 2M/(N1 + N2), where M refers to the number of coding decisions on which two coders agree, N1 and N2 are the total number of coding decisions by the first and second coder respectively.

To explore transitions in annotation tactics, a matrix was generated to show the overall relationship between annotation tactics, and Markov chain was employed to calculate the probabilities of transitions between all possible annotation-tactic pairs. In general, Markov chain is widely used for giving the probabilities of moving from one state to another (Chen & Cooper, 2002). In this study, a first-order Markov chain was applied to examine the probabilities of transitions between two consecutive tactics. In addition, in order to make a comparison of tactic application before and after refinements, a table was made to display the tactic selection with high frequency at three phases: (a) Low precision phase (from Sel to NfF), (b) Medium precision phase (from NfF to NfS), and (c) High precision phase (after NfS). The three phases were divided and identified based on the level of precision requirement in the image annotation process.

Finally, the most applied annotation strategies were identified as the sequences of tactics that frequently occurred. Probabilities of tactic transitions were further calculated by applying different orders of Markov chain according to the number of tactics involved in the strategy. Time was recorded for annotation tactic application in the three phases to identify efficient tactics. The authors selected the top 10 strategies that occurred more than 20 times to represent common transition patterns of annotation tactics. The top 10 tactics include:

- Select an image
- Tune first time
- Tune second time
- Consider the position of scroll bar
- Assess the value
- Compare to the current reference image
- Compare to the pre-image
- Fine tuning tactics
- Tuning first time
- Tuning second time
- Assess the value
- Compare to the pre-image
- Compare to the current reference image
- Compare to the last reference image

Table 3

<table>
<thead>
<tr>
<th>Types</th>
<th>Definition</th>
<th>Code</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image selection tactic</td>
<td>Select an image</td>
<td>Sel</td>
<td>Click “enter” or “next” to start an annotation task</td>
</tr>
<tr>
<td>Fine tuning tactics</td>
<td>Tune first time</td>
<td>NfF</td>
<td>Click the “Fine tuning” button for the first time</td>
</tr>
<tr>
<td></td>
<td>Tune second time</td>
<td>NfS</td>
<td>Click the “Fine tuning” button for the second time</td>
</tr>
<tr>
<td>Annotation tactics</td>
<td>Consider the position of scroll bar</td>
<td>Scr</td>
<td>“I suppose pleasure for the image is the lowest.”</td>
</tr>
<tr>
<td></td>
<td>Assess the value</td>
<td>Val</td>
<td>“(The score of) this image is about 1.5.”</td>
</tr>
<tr>
<td></td>
<td>Compare to the current reference image</td>
<td>Cim</td>
<td>“Pleasure for this image is to the left of this reference image.”</td>
</tr>
<tr>
<td></td>
<td>Compare to the pre-image</td>
<td>Pim</td>
<td>“Score for this image's pleasure is not very high, but a little higher than the last one to be annotated for pleasure.”</td>
</tr>
</tbody>
</table>

4.3. Data Analysis

The unit of analysis is each annotation tactic. As defined in the Introduction, annotation tactics refer to a move or moves, including annotation choice and actions that users apply to advance their annotation process. A coding scheme (Table 3), consisting of key annotation tactics, was created through content analysis, a highly flexible research method that has been widely used in library and information science studies for user behavior research (White & Marsh, 2006). Based on previous research and data generated from this study, the authors classified all tactics into three types of codes: (a) image selection tactic, (b) fine tuning tactics, and (c) annotation tactics. “Fine tuning tactics” was further divided into “First fine tuning” and “Second fine tuning” that participants employed to optimize their annotation and reach more accurate results. Upon clicking the fine tuning button, the reference images changed, and the slider interval range narrowed to 1/3 size of the original one, reducing measurement accuracy to 0.1 after refining results twice. “Annotation method” consists of the three interactive methods proposed earlier (i.e. rating, scroll bar, and image comparison). Rating number was generally set in integer (e.g. Likert scale); in this study, it was set in real number which could more accurately assess an emotional intensity value, for example, the pleasure rating value on the image of Aimed Gun was 3.18. The definition and example of each code are presented in Table 3. Notably, “comparing to pre-image (PIM)”, a thinking activity observed during the annotation process, which does not belong to the original interactive annotation methods, was regarded as a subclass.

In the end, seven types of annotation tactics were defined in this coding scheme. An abbreviation is used to represent each tactic for simplicity. Descriptive analysis was conducted to obtain the frequency of each type of tactic applied in the annotation process.

Apart from coding for each annotation tactic, transitions of tactics were coded for all annotations as well. A tactic was coded according to the order of the first selection. For example, if the move of “considering the position of scroll bar (Scr)” was selected first, and occurred constantly along with “assess the value (Val)”, then the transition Scr to Val would be coded only once. In addition, reoccurring tactics were coded once as well. The transitions between tactics were co-determined by click-through log analysis and think aloud protocol analysis. The first two annotation types (namely Sel, NfF, and NfS) of the coding scheme were identified based on click data, think aloud data was used to identify the third annotation type (namely Scr, Val, Cim, and Pim). Each move was coded and linked with its previous and following moves to see whether a participant modified their annotation tactics. In addition, verbal protocols corresponding with each move was analyzed to figure out participants’ mental activities, thus assisting in coding and seeking out the reasons for changing tactics. The transition of annotation tactics were recorded when a change of annotation tactic was identified.

To test the intercoder reliability of each annotation tactic, two researchers independently coded 30 samples randomly selected from 90 experimental samples. The result of intercoder reliability for coding each tactic has a high consistency, at 0.91 according to Holsti’s (1969) reliability formula. Reliability = 2M/(N1 + N2), where M refers to the number of coding decisions on which two coders agree, N1 and N2 are the total number of coding decisions by the first and second coder respectively.

To explore transitions in annotation tactics, a matrix was generated to show the overall relationship between annotation tactics, and Markov chain was employed to calculate the probabilities of transitions between all possible annotation-tactic pairs. In general, Markov chain is widely used for giving the probabilities of moving from one state to another (Chen & Cooper, 2002). In this study, a first-order Markov chain was applied to examine the probabilities of transitions between two consecutive tactics. In addition, in order to make a comparison of tactic application before and after refinements, a table was made to display the tactic selection with high frequency at three phases: (a) Low precision phase (from Sel to NfF), (b) Medium precision phase (from NfF to NfS), and (c) High precision phase (after NfS). The three phases were divided and identified based on the level of precision requirement in the image annotation process.

Finally, the most applied annotation strategies were identified as the sequences of tactics that frequently occurred. Probabilities of tactic transitions were further calculated by applying different orders of Markov chain according to the number of tactics involved in the strategy. Time was recorded for annotation tactic application in the three phases to identify efficient tactics. The authors selected the top 10 strategies that occurred more than 20 times to represent common transition patterns of annotation tactics. The top 10
strategies, which account for 83.4% of all observed tactic sequences, are considered to be appropriate in identifying common patterns in this study. These annotation strategies are divided into two types from the perspectives of refinement. Then, based on the change of tactic usage, more specific strategies in each type were identified. In particular, the authors explored the usage of the single tactic method or tactic combination method before refinement and examined whether participants modified their tactic selection after refinement. Since tactics, applied in a mixed and repeated way during a short period, were coded only once, time for each single tactic or tactic combination cannot be accurately recorded. However, time for tactic application in a certain phase was more easily captured, namely the duration of annotating a single image at the three phases. Hence, the mean time of a specific tactic or applied combination was computed, and it was used to examine efficient tactic applications in the process of annotation. In order to minimize the influence of the use of think aloud method on time data analysis, we cleaned the irrelevant data and only kept image annotation related time for analysis, excluding the time before and after a participant’s image annotation move. A table (Table 4) was created to summarize research questions, data collection method and data analysis methods.

5. Results

The findings of this study offer answers to the proposed research questions, namely the frequently applied annotation tactics, the probability of tactic transitions, the most frequently applied annotation strategies, and the time distribution on annotation tactics in different phases of the image annotation process.

5.1. Types of frequently applied annotation tactics

A total of 6151 tactics were observed in performing nine annotation tasks by 90 participants. A minimum value of applied tactics was 64, and a maximum value was 2211. An overall average of annotation tactics was 1025.2 (SD = 903.7). Average time for each session was 34.3s (SD = 24.4s), and 3.5 tactics (SD = 1.9) employed in each session on average. Table 5 presents each tactic’s frequency and proportion.

Among all the tactics, the position of scroll bar (Scr) was the most frequently applied tactic, at 35.9%. Adjusting the position of the scroll bar to express image recognition served as the major tactic in the image annotation process. The second most frequently applied tactic was comparing to the current reference image (Cim), which accounts for about 32.6% of total tactics. More than half of the observed tactics were constituted by Scr and Cim. Assessing value (Val) and first fine tuning (NrF) ranked third and fourth, representing 15.7% and 11.5% respectively. Additionally, fine tuning (NrS) and comparing to the pre-image (Pim) were relatively insignificant, with the percentage of 3.3% and 1.0% correspondingly.

The proportion of Scr and Cim occurred at double the rate than Val, which implies that compared with assessing values for images, participants preferred to use the scroll bar and image comparison during the image annotation process. Although Val was observed to be used at first by some participants, its proportion was much smaller than that of Scr or Cim. The tactics in relation to fine tuning (NrF and NrS) only attributes to 14.8%, with the trend for gradual optimization decreasing from 11.5% to 3.3%.
5.2. Probability of Tactic Transitions

A directed matrix of annotation-tactic transitions was generated to explore transitions between tactics. Table 6 presents the transition matrix which tabulates total transitions from one tactic to another for all participants. The matrix includes 6151 transitions of tactics analyzed from 2430 log files. Each cell’s value indicates the frequency of transitions from row tactic to column tactic. In this matrix, there were 20 pairs of tactics showing no transition between 49 pairs of tactics (the same as the number of total cells). The average value of transition between two tactics was 125.53 (SD = 307.28). Cim, the most active tactic that shifts to others, occurred 1927 times. While Scr occurred most frequently, accounting for 2211 cases, and the majority of transitions initiated from Cim (N = 1514). The frequency of transfer to Cim was also relatively high (N = 2008), where nearly all the transitions occurred at the start of the annotation task, such as “Sel→Cim”, “NrF→Cim”, and “NrS→Cim”. As the fine tuning button modifies reference images and the slider range, the act of refinement can be regarded as a new start to a task, to a certain extent. Meanwhile, the number of transitions from Sel to Cim (N = 1249) doubled compared to transitions from Sel to Scr (N = 603) and Val (N = 503). The disparity widened in the medium and high precision phase, as the frequency of transitions from NrF to Cim (N = 561) and from NrS to Cim (N = 172) occurred more frequently than transitions from NrF to Scr (N = 49), Val (N = 37) and from NrS to Scr (N = 12), Val (N = 2). In comparison with Scr and Val, Cim is the most perceptual method without requirement for precision and quantification. This indicates that participants might prefer using cognitive visualization methods requiring low accuracy. Transitions to Pim were also notable in the frequency matrix. Compared with Scr (N = 2) and Cim (N = 11), a larger number of transitions to Pim initiated from Sel (N = 51). This suggests that participants were inclined to compare the image with the previous one at the beginning while the frequency of employing Pim decreased during the task process. Interestingly, there were eight transitions shifting from Sel directly to NrF, and 28 from NrF to NrS, without any other tactic occurring within the phase. This implies that few participants accepted the initial annotation setting, and then directly entered into a fine-grained phase.

A first-order Markov chain was applied to scrutinize the probabilities of tactic transitions that were presented in Table 7. Since probability is also affected by the number of options available after employing a specific annotation tactic, it may not consistent with frequency of transitions. The most probable transitions were from Sel tactic to Cim (51.7%), Scr (25.0%) and Val (20.8%) respectively. As participants refined the annotation results, the probabilities of Cim selected increased to 83.1% and 92.5%. By contrast, both Scr and Val were less likely applied, especially for the Val tactic, where the percentage plummeted to 1.1% after NrS. In addition, the Val tactic likely led to NrF, showing 87.1% of probability, while the likelihood reduced dramatically to 6.1% when shifting to NrS. Although the Scr tactic exhibited the same declining pattern in the second fine tuning, its proportion (20.6%) was still higher than that of Val. Since Scr is more perceptual compared with Val, it can be inferred that the perception level of annotation methods would influence whether users choose to optimize their annotation results. Participants were most likely to apply the Scr (78.6%) after Cim, otherwise, they preferred to choose Val (19.6%).
Table 8 presents high-frequency tactic application in three phases (see the definition of the three phases in Data Analysis). It is evident that the participants tended to employ Cim at first no matter at which phase. It corresponds to the finding in the probability matrix that the most likely transition from Sel, NrF and NrS was Cim. The Cim tactic appears in combination with other tactics, in that it was usually used together with Scr or Val, and less frequently applied alone. For example, participant 75 stressed, “the kitten is much more adorable than the two reference images, the pleasure rating is 9.” “I don’t have strong feelings for the lamp compared with the two reference images, I would drag the slider here.” said participant 64. Cim was mainly associated with Scr, and the frequency of (Cim, Scr) combination was applied most frequently among the three phases. Scr and Val occurred frequently in the low precision phase, while it occurred less frequently in the later phases, which is consistent with the result in the frequency matrix that both Scr and Val were less likely applied as participants optimized their annotation results.

5.3. Most Frequently Applied Annotation Strategies

To analyze the most common patterns of tactics transition, the researchers selected the top 10 annotation sequences that occurred more than 20 times (Table 9). Overall, there were 2430 observed sequences. The top 10 strategies occurred 2027 times, which accounted for 83.4% of all tactic sequences. These most frequently applied sequences revealed two main types of annotation strategies: (a) non-refinement strategy, and (b) refinement strategy. Non-refinement strategy refers to a strategy without employing fine tuning in the annotation process. Refinement strategy refers to a strategy using fine tuning (NrF or NrS) to enhance the annotation process.

To explore patterns of tactic usage changes, specific subtypes of strategies were further identified. Particularly, non-refinement strategies were divided into: (a) Single tactic strategy, and (b) Tactic combination strategy. Refinement strategy was classified into: (c) Fix mode strategy, and (d) Shift mode strategy.

Single tactic strategy includes only one kind of tactic (excluding Sel) to accomplish the annotation task. It mainly consists of Sel and one of the interactive methods (Scr, Val or Cim). The frequency of single tactic strategy is nearly half of total sequences. In Table 9, of 10 frequently observed sequences, three can be viewed as part of the single tactic strategy.

Tactic combination strategy consists of more than one tactic (excluding Sel) in a task. For example, as the annotation pattern “Sel→Cim→Scr” shows that the participant used a tactic combination of Cim and Scr. Of 10 frequently observed sequences, two were related to tactic combination strategy.

Fix mode strategy refers to a strategy that consists of only one kind of tactic combination before and after refinement. For instance, the sequence “Sel→Cim→Scr→NrF→Cim→Scr→NrS→Cim→Scr” identifies iterative pattern by the combination of Cim and Scr. Of 10 frequently observed sequences, two of them fall into fix mode strategy.

Shift mode strategy involves different tactic selections before and after refinement. For instance, the pattern of “Sel→Cim→Scr→NrF→Cim→Val” contains two different tactic combinations including (Cim, Scr) and (Cim, Val). Of 10 frequently observed sequences, three are considered as shift mode strategy.

To better understand the entire annotation process, the most frequently occurring eight strategies were selected for analysis based on the probabilities of tactic transitions calculated by Markov chain (Fig. 2). The order of Markov chain was applied based on the

### Table 8
High-frequency tactic application in three phases.

<table>
<thead>
<tr>
<th>Tactic application</th>
<th>Frequency</th>
<th>Tactic application</th>
<th>Frequency</th>
<th>Tactic application</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low precision phase Sel → NrF</td>
<td></td>
<td>Medium precision phase NrF → NrS</td>
<td></td>
<td>High precision phase After NrS</td>
<td></td>
</tr>
<tr>
<td>Sel, Scr</td>
<td>952</td>
<td>(Cim, Val)</td>
<td>432</td>
<td>(Cim, Scr)</td>
<td>130</td>
</tr>
<tr>
<td>Scr</td>
<td>587</td>
<td>(Cim, Scr)</td>
<td>88</td>
<td>(Cim, Val)</td>
<td>30</td>
</tr>
<tr>
<td>Val</td>
<td>495</td>
<td>Cim</td>
<td>47</td>
<td>Cim</td>
<td>16</td>
</tr>
<tr>
<td>(Cim, Val)</td>
<td>260</td>
<td>Scr</td>
<td>47</td>
<td>Scr</td>
<td>12</td>
</tr>
<tr>
<td>Cim</td>
<td>31</td>
<td>Val</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Pim, Scr)</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 9
Most frequently applied annotation strategies.

<table>
<thead>
<tr>
<th>Strategy type</th>
<th>Case</th>
<th>Tactic sequence</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-refinement strategy</td>
<td>Single tactic strategy</td>
<td>SA1</td>
<td>Sel Scr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SA2</td>
<td>Sel Val</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SA3</td>
<td>Sel Cim</td>
</tr>
<tr>
<td>Tactic combination strategy</td>
<td>TC1</td>
<td>Sel Cim Scr</td>
<td>520</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TC2</td>
<td>Sel Cim Val</td>
</tr>
<tr>
<td>Refinement strategy</td>
<td>Fix mode strategy</td>
<td>FM1</td>
<td>Sel Cim Scr NrF Cim Scr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FM2</td>
<td>Sel Cim Scr NrF Cim Scr NrS Cim Scr</td>
</tr>
<tr>
<td>Shift mode strategy</td>
<td>SM1</td>
<td>Sel Cim Scr NrF Cim Val</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SM2</td>
<td>Sel Scr NrF Cim Scr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SM3</td>
<td>Sel Val NrF Cim Scr</td>
</tr>
</tbody>
</table>
number of tactics involved in the strategy. For example, for SA1, SA2 and SA3, the first-order Markov chain was applied; for TC1 and TC2, the second-order Markov chain was employed; and for FM1, the fifth-order Markov chain was used. In strategy TC1, 1249 tactics related to Cim were chosen in the first order (2430 total cases), so the probability of shifting from Sel to Cim was 51.4%. In the second order, 259 Val tactics were selected on the basis of 1249 total cases, then the probability of transitions from Cim to Val was 20.7%. The same approach applies to calculate the remaining probabilities. Additionally, the self-circulation of each final tactic means the annotation task was completed by that tactic.

In Figure 2, participants always started with one of the three interactive annotation methods, with Cim tactic being used more often (51.4%). The Scr (24.8%) and Val (20.7%) tactic is often applied as single method, while Cim was prone to sequentially shift to Scr (76.0%) and Val (20.7%) in the tactic combination strategy. In fix mode strategy, the most common applied tactic combination is (Cim and Scr). It was iteratively applied until participants completed their annotation tasks. In shift mode strategy, the single occurrences of Scr and Val only happened prior to the act of refinement. The Cim tactic was more likely selected after fine tuning. In the strategy without Cim at the beginning phase (SM2), Cim (64.2%) was selected immediately after entering the next phase. In the strategy beginning with Cim (SM1), it was more likely to shift to Cim (89.1%) after refinement.

5.4. Time distribution on annotation tactics in three phases

As mentioned in Data Analysis, time spent for a specific tactic cannot be accurately recorded. In order to further explore efficient tactics or combinations, the mean time for tactic application in each phase was computed. Since tactics like ‘Sel’, ‘NrF’ or ‘NrS’ are actions of clicking a button, time used for these tactics was ignored. Specific time devoted for tactic application in each phase was listed in Table 10. It is evident less time was spent with less tactics being applied, as participants spent the least amount of time in applying single tactic strategies to finish annotation tasks. However, Val is the inefficient tactic which results in an increase of time. In each type of strategy, time spent for strategies with Val occurred more than strategies without Val. The most obvious example is in the shift mode strategies, in which SM2 took the least time among the three.

The time data reveal a “quarter decreasing” pattern; time distributed for a specific tactic or tactic combination would decrease approximately 25% as annotation proceeds. On one hand, the decreasing pattern in non-refinement strategy appeared to be quite distinctive, with almost 25% of the total time saved when two tactics were applied in combination, compared to these tactics applied separately. For example, the mean time for the tactic combination (Cim, Scr) in strategy TC1 was 30.8s, nearly 75% of the sum of time
Table 10
Time spent for annotation tactics in three phrases.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Tactic/Tactic combination (Mean time)</th>
<th>Low precision phase</th>
<th>Medium precision phase</th>
<th>High precision phase</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA1</td>
<td>Scr</td>
<td>20.4</td>
<td></td>
<td></td>
<td>20.4</td>
</tr>
<tr>
<td>SA2</td>
<td>Val</td>
<td>24.6</td>
<td></td>
<td></td>
<td>24.6</td>
</tr>
<tr>
<td>SA3</td>
<td>Cim</td>
<td>20.1</td>
<td></td>
<td></td>
<td>20.1</td>
</tr>
<tr>
<td>TC1</td>
<td>(Cim, Scr)</td>
<td>30.8</td>
<td></td>
<td></td>
<td>30.8</td>
</tr>
<tr>
<td>TC2</td>
<td>(Cim, Val)</td>
<td>34.5</td>
<td></td>
<td></td>
<td>34.5</td>
</tr>
<tr>
<td>FM1</td>
<td>(Cim, Scr)</td>
<td>24.6</td>
<td>(Cim, Scr)</td>
<td>26.9</td>
<td>51.5</td>
</tr>
<tr>
<td>FM2</td>
<td>(Cim, Scr)</td>
<td>19.5</td>
<td>(Cim, Scr)</td>
<td>17.3</td>
<td>36.8</td>
</tr>
<tr>
<td>SM1</td>
<td>(Cim, Scr)</td>
<td>29.9</td>
<td>(Cim, Val)</td>
<td>28.5</td>
<td>58.4</td>
</tr>
<tr>
<td>SM2</td>
<td>Scr</td>
<td>15.9</td>
<td>(Cim, Scr)</td>
<td>25.8</td>
<td>41.7</td>
</tr>
<tr>
<td>SM3</td>
<td>Val</td>
<td>27.2</td>
<td>(Cim, Scr)</td>
<td>22.7</td>
<td>49.9</td>
</tr>
</tbody>
</table>

that Cim and Scr being applied separately (40.5s). And 34.5s was consumed by (Cim, Val), which was also about 75% of the sum of time that Cim and Val being selected separately (44.7s). In other words, the “quarter decreasing” pattern demonstrates the efficiency of combined tactics would improve by 25%. On the other hand, in refinement strategy, time spent for the same tactic application decreased by 20%-30% when refinement occurred. For example, the mean time spent on (Cim, Scr) of the three phases in FM2 (18.7s) was nearly 70% of that in FM1 (25.8s), which was almost 80% of that in TC1 (30.8s). Similarly, time used for the tactic combination (Cim, Val) in SM1 (28.5s) also reduced by approximately 25% compared with that for TC2 (34.5s). In conclusion, both multiple tactics and refinement activities represent efficient tactic applications based on the above time analysis. Moreover, since Val always costs more time than Cim and Scr, (Cim, Scr) was found to be the efficient tactic combination.

However, there were two tactic applications that did not show the decreasing pattern even if the act of refinement occurred. Unlike strategy FM1 and FM2, time spent on the tactic combination (Cim, Scr) in SM1 (29.9s) was similar to that in the non-refinement strategy TC1 (30.8s). Furthermore, the tactic Val in SM3 (27.2s) consumed more time compared with Val in SA2 (24.6s). At the same time, in SM1 and SM3 strategies, participants shifted annotation tactics between numerical and perceptual methods as refinement occurred, which was different from strategy SM2 where participants applied tactics between the two perceptual methods.

The tactic shifting process possibly contributed to the non-decreasing time consumption, which could be labelled as “transition cost”.

6. Discussion

6.1. Patterns of annotation tactics and strategies

This study revealed the complex cognition process of users in image annotation. Specifically, the findings did uncover patterns of tactics and strategies ranging from the most and least applied annotation tactics, most applied tactic transitions, and different types of annotation strategies, as well as time spent on the application of a variety of annotation tactics in different phases. Here are several main contributions of this study:

First, the adoption of annotation tactics mainly depends on the perceptual level of annotation methods. Val is the tactic which requires users to transform the image evocation into particular value; Scr requires participants to use a visual slider to make a judgement; Cim requires participants to consider reference images for comparison. Even though Val is more familiar to participants, it entails more cognitive resources than Scr and Cim in image semantic interpretation. Also, the frequency of using Val is far less than Scr and Cim reaffirms Pommeranz et al’s (2012) viewpoint that it is difficult for individuals to assign a certain value for images through individual perception.

The main contributing factors for the decreasing optimization echoes the law of least effort introduced by Zipf (1949), in which people tend to choose the way requiring the least amount of effort to finish tasks. Most participants in this study were not willing to put additional effort in further optimizing their annotation results. Instead, they acted in a manner which is not to maximize gain but rather to minimize loss in terms of effort. Besides, the perception level of annotation methods also influenced users’ act of gradual optimization. Transition probability from Scr to NrS was much higher than that from Val to NrS, because Scr offers a more visual and intuitive way (scroll bar) for users to match the value with the image compared with Val. Therefore, the fact that participants applying Scr were more willing to continue refining their annotation results suggests that methods with less accuracy requirement and more visual helpful could enhance users’ desire to refine.

As Borgman, Hirsh, and Hiller (1996) noted, searching is a complex process that cannot be understood simply by comparing the output of a search session to a query stated in advance. Evaluating the search process requires multiple methods. Image interpreting, a complicated process, also requires multiple evaluation methods to achieve a full characterization of the behavioral processes that occur. However, previous research on interactive image annotation generally adopts a single rating method and simplifies the human image interpretation process. This study demonstrates that participants need to apply multiple tactics and strategies to achieve their diverse image interpretation process.

Second, the analysis of tactic transitions shows that annotation-tactic transitions have their own patterns. Participants favored using perceptual methods at the beginning of each phase rather than the method demanding high accuracy, with Cim always initiating from Sel, NrF, and NrS. At the same time, since the annotation range for the task image initially is relatively wide (from 1-9),
the task did not place much burden on users at first, so the probability for transitions from Sel to Scr and Val were not low as well. In the information searching environment, task complexity is a critical factor that influences users' search behavior (Aula, Khan, & Guan, 2010; Liu, Zhang, & Huang, 2016). As the task becomes more complex, the number of steps in the search, and the time taken to search increases. In addition, the decision to use various search strategies is an adaptive response to the demands of tasks (Payne, Bettman, & Johnson, 1993). In this study, as the act of fine tuning narrows the scrolling range and modifies the former reference images, it is no doubt that the annotation task becomes more difficult. With limited capability to judge image semantics from the perspective of numerical precision, perceptual visualization method (Cim) was relied on more heavily than the numeric method (Val), which also indicates that Cim enables users to more easily judge an image than Val under a tiny scale environment. Here is a typical example from a participant (P72) who clicked fine-tuning twice, “The value of arousal for this image is not easy to judge. I need to fine tune again. Now I can judge the arousal for the broken glass easily with the help of reference images.”

In spite of the less cognitive demanding feature of Cim, it does not provide a mechanism for participants to express annotation result clearly, thus it was rarely used alone. The most remarkable transition (Cim, Scr) shows that participants tend to determine the interval for task image by Cim first, and then use perceptual method Scr to make concrete annotations. Simultaneously, the combination with Scr also helps Scr become the most frequently applied tactic, while the Cim tactic, the most perceptual method, turns out to be the second. In other words, Scr and Val often appeared following Cim, which can be considered as auxiliary methods to present the result annotated by Cim.

Third, a significant contribution of this study lies in the findings that discover typical tactic patterns to illustrate participants' sequential movements in detail. These findings offer not only opportunities for researchers to understand the nature of the image annotation process but also suggestions for annotation systems designed to support efficient tactic and strategy application.

In this study, non-refinement sequences constitute 80.9% of the total frequently applied annotation strategies, which means participants were confident about the results they annotated, or they did not pursue high accuracy of the annotation result. The following two participants expressed their thoughts behind their annotations: “For me, the tornado deserves a high rating on arousal, dragging its slider to the right of this reference one is ok (P84);” “It is disgusting to see the naked body, I will give a lowest number regarding pleasure (P15).”

In terms of image annotation involving refinement, on one hand, tactic applications in the fix mode strategy show unique patterns. Only one tactic combination (Cim, Scr) was used iteratively within the annotation process, which validates Luchins' (1942) theory of Einstellung effect that an idea coming immediately to mind in a familiar context prevents alternatives being considered. Once participants applied the tactic combination of (Cim, Scr) at the beginning of a task, the Einstellung effect makes them not willing to make changes to employ other tactics to solve similar problems. Moreover, rather than other tactic combinations, (Cim, Scr) becomes the fix mode strategy that does not impose a high cognitive load on participants. Therefore, fix mode strategy was preferred after refinements when participants performed a familiar annotation task.

On the other hand, patterns reflected in shift mode strategy are diverse. It is worth noting that participants shifted their annotation tactics between numerical and perceptual methods in strategy SM1 and SM3. Participants who applied SM1 or SM3 expressed more uncertain thoughts when annotating image in their low precision phase. Here is a typical statement from Participant 16, “I think the value is higher than 5, maybe between 5 to 7, actually between 5 to 6 or 5 to 7 are both okay.” Perhaps it is the uncertainty that makes participants turn to other annotation methods after refinement. According to Participant 63, “It is hard to assign the dominance value, maybe it is between the two reference images. I need to click fine tuning. Now the value is 3.09, it is pretty close.” And Participant 20 also said “It is difficult to judge the value of arousal, after refining, the arousal is higher than the two reference images obviously.” Interestingly, different tactics applied in shift mode strategy in different phases also correspond to the shift in search tactics in different phases in the field of information retrieval (Xie & Joo, 2010). Participants in the study modified their tactic as they entered the next phase. It suggests that the existence of “berry picking” tactics (Bates, 1989) occurs in the image annotation process. The changes of tactics in different phases call for the need for image annotation system to provide different support mechanisms at various stages.

Fourth, the analysis for the time of each strategy uncovers efficiency levels of annotation tactic usages. The numeric character of Val accounts for its inefficiency. The requirement for transforming the image evocation into a specific value took participants more time and cognitive effort compared with the other two perceptual methods. For example, participant 77 stated his hesitation in assigning an accurate number, “I feel I can control the kitten, its dominance should be higher than the two reference images. Number 8 is fine, no, between 8 and 9, actually it could even be higher than 9.” Additionally, although less tactics consume less time, it is more important that user interpret image semantics by refinement, as the “quarter decreasing” pattern of time indicates the high marginal efficiency of applying refinement activities. Human information-processing has been commonly considered as consisting of distinct stages according to the history of cognitive psychology (Rayner, Pollatsek, Ashby, & Clifton, 2012). For instance, reading was thought of involving two sets of processes, one that makes visual information available (perceptual processes) and the other that makes use of that information in support of the language processes involved in reading (cognitive processes) (Mcconkie, Reddix, & Zola, 1992). In the image interpretation process, users also experience the perceptual process which forms an initial recognition to the image to be annotated. When annotating the same image with provided methods, users do not need to spend time to reconstitute the recognition, which helps explain the high marginal efficiency. In addition, the two exceptions in SM1 and SM3 that are caused by “transition cost” can be attributed to participants' uncertain annotation behaviors.

6.2. Implications for image annotation system design

The analysis of annotation tactics, transitions and strategies sets up a foundation for image annotation system design. In Figure 3, key findings of this study and their relationships are highlighted, and most importantly, corresponding design recommendations are presented as well.
First, the occurrence of combined tactics implies individuals have the need to use various interactive methods to annotate images. Ivasic-Kos, Ipsic, and Ribaric (2015) pointed out that knowledge in context of image interpretation is often incomplete, imprecise, uncertain and ambiguous in nature. However, previous image interpretation system, such as standardized digitized image corpus, generally adopts the single rating method to collect this kind of fuzzy knowledge, which makes the results inaccurate. It validates the existence of bias in collecting image relevance feedback. For example, a user marks any retrieved images as highly relevant, relevant or irrelevant to his query to express image relevance, but this method ignores the complex cognitive process it may involve. In addition, relatively perceptual methods (Cim and Scr) show more efficiency than the numeric method (Val) in this study. Therefore, apart from providing a single rating method, which is the most common practice, multiple annotation options, especially perceptual methods, should be offered to users to facilitate their image interpretation. Specifically, providing intuitive reference images and a scroll bar along with a rating method should be accessible in the annotation interface. It is important to enable users to select their preferred annotation method and thus generate more reliable image relevance feedback.

Second, Debowski (2001) argued that sequencing search moves into search tactics can be viewed as a component of search quality, which also yields insights into the significance of investigating annotation tactic sequence. The typical annotation tactic sequences could be used to guide users’ annotation behavior and offer smoother transitions. Relevant reference images need to be highlighted in the annotation interface to draw users’ attention to the image. Since the tactic combination (Cim, Scr) is the most remarkable and efficient transition, once users compare annotated image with reference images, a following scroll bar could be recommended to guide users’ annotation process. For example, the slider could be set moving slowly in the scroll bar, or the scroll bar could be set in bold to invite users to choose it. The rating method should not be emphasized at the beginning of the annotation task, as its demand for cognitive effort could decrease users’ desire to refine their annotation results. Accordingly, it also could be recommended as an auxiliary method following Cim to present the annotation results.

Third, the quarter decreasing pattern indicates that refinement activity should be encouraged. After users generate initial relevance feedback, an explicit option of refinement needs to be offered to direct users to judge the relevance of an image in a fine-
grained way. With the evolution of the annotation phase, different types of support need to be provided according to users’ prior annotation behavior. Specifically, the rating method should be further minimized to reduce unnecessary transitions from NrF or NrS to Val because users become more incapable to judge image semantics from the number when the precision request for task grows higher, and because Val has not been proved to be an efficient tactic. Instead, reference images can be highlighted under a tiny scale environment for their perceptual feature. However, if users are uncertain about their annotation process, it is recommended to support them in applying additional methods to help them make accurate decisions. For example, the rating method could be introduced in an interface that only includes scroll bar or reference images to guide users who are not sure about their annotation results.

Apart from relevance feedback design implications, findings in this study are also helpful for researchers to establish a standardized digitized image corpus when conducting an image annotation experiment.

7. Conclusion

This study examined users’ annotation tactics, transitions of tactics, strategies, and time spent on tactics in three phrases in completing assigned image annotation tasks. The results of this study emphasize that the cognitive process of human interpretation of images is rather complex. Users do apply multiple interactive annotation methods in the process, which calls for researchers to take multi-interactive methods into consideration when conducting image annotation research. More important, four main contributions are highlighted. First, this study identified the most and least applied image annotation tactics and revealed the reasons behind the findings. Second, the frequency and probability of tactic transitions illustrated the dynamic changes of annotation process. Third, annotation strategies were identified based on the analysis of sequences of tactic transitions. These strategies represented the typical patterns of annotation-tactic transitions. Fourth, time data showed efficient tactic applications as well as two unique patterns: “quarter decreasing” and “transition cost”. Annotation patterns found in this study are further transformed into the discussion of annotation system design to facilitate users efficiently annotating images based on relevance feedback captured by the system.

Several limitations of the research must be acknowledged. First, only 90 participants within a university were recruited in this experiment study. Second, think aloud protocol recorded by Morae was used to capture participants’ cognitive states in the annotation process, but not every participant stated what they thought in detail. Third, the effects of user factors, for example, gender, major, educational level, or prior experience and task factors were not considered in this study. Future research can recruit participants with diverse backgrounds, investigate the effects of some user factors or task factors on the applications of tactics/strategies, and employ other cognitive data collection methods, such as eye tracking, electroencephalograph, and brain magnetic resonance imaging, to better understand human’s semantic interpretation process to images.

Acknowledgments

The authors gratefully acknowledge the financial support for this work provided by National Natural Science Foundation of China (No. 71273195, 71420107026), the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (No. 17JD870002), the National Basic Research Program of China (973 Program, No. 904171200), the Major Project of the Ministry of Education of China (No. 17JZD034), Higher Education Reform Research Project of Hubei Province (No. 20160678), and World Class Discipline Construction Project of School of Information Management of Wuhan University (No. 20180312).

References


