ACM

2 Penn Plaza, Suite 701
New York, New York 10121-0701

TRANSMITTAL FORM

Journal TWEB Art. 1

Author Wu et al.

march

Month/Issue Vol. 7 no. 1 ~~Feb~~'13

No. of Manuscript Pages 33

No. of Online-Only Pages 0

Total No. of Pages 33

Tables (No. of Tables) 10

Figures (No. of Figures) 14

Special Instructions:

OU WU,
WEIMING HU, Chinese Academy of Sciences
and LEI SHI, Yahoo! Beijing Research Center

Left Running Head:
O. Wu et al.

Right Running Head:
Measuring the Visual Complexities of Web Pages

Notes for page 1: This work is partly supported by NSFC (grant no. 60825204, 60672040, 61003115, 60935002), and the National 863 High-Tech Program of China (grant no. 2012AA012504).

Authors' addresses: O. Wu (corresponding author), W. Hu, NLPR, Institute of Automation, Chinese Academy of Sciences, China; email: wuou@nlpr.ia.ac.cn; L. Shi, Yahoo! Beijing Research Center, China.

Code:

© 2013 ACM 1559-1131/2013/~~02~~03-ART01 $15.00
DOI: http://dx.doi.org/10.1145/2435217.2435216
DOI: http://dx.doi.org/10.1145/2435217.2435216

Running Foot (all pages)
ACM Transactions on the Web, Vol. 7, No. 1, Article 1, Publication date: ~~February~~ March 2013

Insert A

ACM Reference Format:

Wu, O., Hu, W., and Shi, L. 2013. Measuring the visual complexities of web pages. ACM Trans. Web 7, 1. Article 01 (~~February~~ March 2013) 000 pages.

DOI: http://dx.doi.org/10.1145/2435217.2435216
DOI: http://dx.doi.org/10.1145/2435217.2435216.

Measuring the Visual Complexities of Web Pages

see hardcopy for new author affiliations

Ou Wu, NLPR, Institute of Automation, Chinese Academy of Sciences, wuou@nlpr.ia.ac.cn
Weiming Hu, NLPR, Institute of Automation, Chinese Academy of Sciences, wmhu@nlpr.ia.ac.cn
Lei Shi, Yahoo! Beijing Research Center, lshi@yahoo-inc.com

Visual complexities (VisComs) of Web pages significantly affect user experience, and automatic evaluation can facilitate a large number of Web-based applications. The construction of a model for measuring the VisComs of Web pages requires the extraction of typical features and learning based on labeled Web pages. However, as far as the authors are aware, little headway has been made on measuring VisCom in Web mining and machine learning. The present ~~paper~~ provides a new approach combining Web mining techniques and machine learning algorithms for measuring the VisComs of Web pages. The structure of a Web page is first analyzed, and the layout is then extracted. Using a Web page as a semi-structured image, three classes of features are extracted to construct a feature vector. The feature vector is fed into a learned measuring function to calculate the VisCom of the page.

In the proposed approach of the present study, the type of the measuring function and its learning depend on the quantification strategy for VisCom. Aside from using a category and a score to represent VisCom as existing work, this study presents a new strategy utilizing a distribution to quantify the VisCom of a Web page. Empirical evaluation suggests the effectiveness of the proposed approach in terms of both features and learning algorithms.

article

Categories and Subject Descriptors: H.4.m [**Information Systems**]: Miscellaneous; H.2.8 [**Data-based Applications**]: Data Mining

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Visual complexity (VisCom), Quantification, Web mining, Feature, Learning.

lc x3

see production sheet

~~**ACM Reference Format:**~~
~~Wu, O., Hu, W., and Shi, L. 201x. Measuring the Visual Complexities of Web pages. ACM Trans. the Web 1, 1, Article 1 (January 20xx), 33 pages.~~
~~DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000~~

1. INTRODUCTION

Web pages have become indispensable for acquiring information in everyday life, and they also serve as the user interfaces of the Internet. Naturally, there is an increasing need to design visually appealing and easily interactive Web pages. Visual complexity (VisCom) plays an important role in the perception of visual stimuli. Existing studies [Berlyne 1974][Geissler et al. 2006][Pandir and Knight 2006] have revealed that a relationship exists between VisCom and user experiences (e.g., visually pleasant) of Web pages. Tuch et al. [2009] conducted various experiments to evaluate the impacts of VisCom on users when visiting Web pages. Their findings demonstrated that VisComs of Web pages have multiple effects on human cognition and emotion, including the experience of pleasure and arousal, task performance, and so on. Specifically, a negative linear relationship exists between VisCom and affective valence, whereas a positive correlation exists between VisCom and arousal ratings. Users performed better on search and recognition tasks on low visually complex pages. Further, another study reported that the simplicity of interaction with Web pages plays an important role in improving user experience [Harper et al. 2009], and is a desirable advantage for Web-based applications [Song 2007].

Understanding and measuring VisComs of Web pages is important for both Web design and Web-based applications. Web designers have become increasingly concerned about creating pages that are visually attractive and simple. As designers themselves do not always have similar impressions with the users [Park et al. 2004], surveys of the

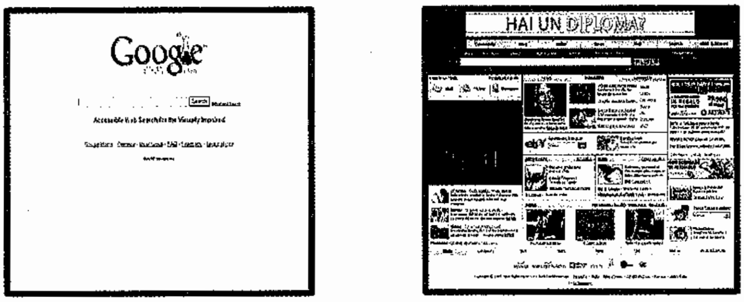


Fig. 1. A low-VisCom page (left) and a high-VisCom page (right).

perceptions of population samples aid in evaluating Web design in terms of VisCom. Nevertheless, the entire evaluation procedure can be costly and slow. Hence, an objective third-party Web VisCom evaluation tool, which is capable of producing reliable estimates of the VisComs of the designing Web pages, can help designers obtain feedback at reduced design costs and time. In addition, an effective VisCom measurement can help researchers interested in interaction of cognition and Web use [Harper et al. 2009; Tuch et al. 2009; ~~Tuch et al.~~ 2011]. Tuch et al. [2009] considered the JPEG size of a page's transformed image as the VisCom of that page. However, this simple strategy may lead to unreliable conclusions. As our experiments suggest, using the JPEG size to measure VisCom is inaccurate for some Web pages. More accurate VisCom measures can help cognitive researchers who analyze Web processing obtain more reliable results.

Figure 1

Figure 1 shows two Web pages with distinct VisComs. Typically, VisComs are related to human empirical observations that are unavoidably subjective [Annett 2002]; therefore, there is some doubt as to whether or not VisComs can be measured. Song [2007] analyzed VisComs of Web pages using the Gestalt principle and concluded that although the VisComs of Web pages are perceived subjectively, they can be measured reliably. Actually, VisCom is a quantifiable property [Donderi 2006] that has been measured in such diverse areas as images [Rosenholtz et al. 2007; Forsythe 2009], icons [Forsythe et al. 2003], visual advertisement [Pieters et al. 2010], and 3D graphics [Gero and Kazakov 2004]. These previous measurement methods motivate us with the idea that measuring Web page VisComs is feasible and is a worthwhile topic of study.

Automatic measurement of VisComs of Web pages can be reduced to a Web mining problem. However, current Web mining mainly involves in the exploration of content, usage, and structure [Cheng and Cant-Paz 2010; Jiang et al. 2010; Liu 2007]. VisCom measurement for the Web has only been investigated in human-computer interaction (HCI) literature. The Visual Complexity Rankings and Accessibility Metrics (ViCRAM) project[1] launched by the University of Manchester is the pilot work in Web VisCom measurement. The ViCRAM project aims to improve accessible Web design by describing the VisComs of Web pages. Although the VisCRAM project is a pioneer work that has obtained a number of valuable conclusions, several problems still deserve further study. One primary problem is that the number of features is limited. For instance, some important factors such as colors and textures are not considered. In addition, the measurement model construction relies heavily on conventional polynomial regression methods, resulting in a weak generalization capability of the constructed models.

Footnote 1

[1] http://hcw.cs.manchester.ac.uk/research/vicram/

To advance Web VisCom research and construct a generalized and more accurate VisCom measurement model, this paper proposes a new VisCom measuring approach by integrating studies in HCI, and methodologies in Web mining and machine learning. Specifically, a new layout extraction algorithm is presented to analyze the structure of a Web page, diverse features motivated by existing studies in HCI [Michailidou et al. 2008; Donderi 2006; Song 2007; Harper et al. 2009], design [Ahmad et al. 2008; Park et al. 2004], Web mining [Cai et al. 2003b; Song et al. 2004; Kim and Wilhelm 2008; Wu et al. 2011], and computer vision [Hasler and Susstrunk 2003; Geusebroek and Smeulders 2005; Rosenholtz et al. 2007] fields are extracted. Moreover, theoretically well-founded machine learning algorithms are utilized to cope with the learning of the measuring function. The conducted experiments show that the proposed approach achieves better results than the HCI method as well as three other methods used for comparison.

article

Studies on HCI mainly investigate on primary factors (features) from carefully designed psychological experiments, whereas studies on machine learning more focus on feature extraction and model learning. Hence, compared with conventional HCI papers, the present paper introduces technical details of the employed features and learning algorithms before describing the actual experiments. The remainder of the current paper is organized as follows: Section 2 briefly reviews related literature. Section 3 introduces the materials and problems. Sections 4 describes the framework and details of the proposed approach, including VisCom labeling, feature extraction, and learning algorithms. Section 5 reports the experimental evaluation results. Finally, Section 6 gives the conclusions as well as the limitations of this work and several topics for possible future studies.

work

article

?

2. RELATED WORK

VisCom measurement is an emerging important topic in multiple disciplines including the HCI and design fields. This section briefly reviews previous studies on the VisComs of Web pages. The current study relies on Web mining and machine learning techniques, so several related topics will be reviewed. Additionally, several studies, which also investigate the subjective issues, will be briefly introduced.

2.1. Visual complexity measurement

cap x2

As previously mentioned, current studies on VisCom are primarily from the ViCRAM project. In a recent study of ViCRAM [Michailidou et al. 2008], Michailidou et al. utilized three features, namely, top left corner counts (TLC), word counts (W), and image counts (I) to construct a linear model expressed as: $VisCom = 1.743 + 0.097TLC + 0.053W + 0.003I$, where TLC represents the number of distinct sections a Web page is organized into. The following steps are undertaken in counting TLC of a Web page [Michailidou 2009]: (1) chunk rendering of the page is created based on cues, such as background colors, headings, stand-alone images, and visual lines or borders; (2) the page is divided into boxes, each of which contains a section or subsection; and (3) the top left corner of a box is counted provided that its left and top sides are not adjacent or have a common side with another box. Word counts comprise all texts used to present any type of information on the page, including texts from menu lists and within images. Image counts contain any image on the page such as advertisements, logos, and decorative images. These three features describe several key factors related to the visual presentation of a Web page, such as color, layout, texts, and images.

Measuring VisCom is also investigated in the area of images [Rosenholtz et al. 2007; Forsythe 2009], electronic displays [Donderi 2006], and 3D graphics [Gero and Kazakov 2004]. Rosenholtz et al. [2007] proposed two classical methods for image VisCom

Footnote 2

measurement: subband entropy (SE) and feature congestion (FC)[2]. SE is based on the perception that VisCom is related to the number of bits required for subband (wavelet) image coding. A larger number of required bits result in greater VisCom. FC is seen as the difficulty measurement of adding a new salient item to an image. A higher difficulty value indicates a higher VisCom. An easier method to measure VisCom is to utilize the file sizes of digital images after compression (e.g., jpeg and zip). A larger file size indicates higher complexity [Stickel et al. 2010]. Donderi [2006] compared several popular compression formats for VisCom assessment, and experimental results indicated that JPEG is closer to user experiences among others, although it cannot sufficiently reflect the experienced complexity of users. Gero ~~et al.~~ [2004] investigated the VisCom of 3D graphics. The geometry of 3D graphics can be represented by a graph, and the complexity of the graph is then calculated based on the probability distribution of different node types in the graph.

cap ×4

s

and Kazakov

2.2. Web mining

cap

Web mining aims to discover useful information from the Web data. According to analysis targets, Web mining can be divided into content, usage, and structure mining [Liu 2007]. Web content mining is the process for discovering useful information from the text, image, audio, or video data on the Web. Web usage mining is a process to extract useful information about what users are looking for on the Internet based on Web logs, ~~i.e.~~ user access history. Web structure mining is the process to use graph theory to analyze the node and connection structure of a Web site. Web mining has been successfully applied in many areas such as Web search and electronic commerce.

that is;

A large number of recent studies [Cai et al. 2003b; 2003a; Wu et al. 2011] have given attention to the visual presentation of Web pages. Cai et al. [2003b] introduced a visual-based page segment (VIPS) algorithm to determine the structure of a Web page. Although this algorithm is part of Web structure mining, the VIPS algorithm is primarily based on many Web appearance cues, such as split lines, decorative images, colors, and fonts. In [Cai et al. [2003a], visual appearance cues are utilized to distinguish different parts of a Web page and find useful content blocks of Web pages. Wu et al. [2011] proposed an automatic approach for determining whether a page is aesthetic. These studies are summarized into a new Web mining division, ~~i.e.~~ Web appearance mining, which considers the appearance of Web pages as the (partial) analysis target and focuses on discovering useful information (e.g., useful content blocks in [Cai et al. [2003a]) based on Web appearance. In comparison, the current study also focuses on the appearance of a Web page. Therefore, the present work belongs to the area of Web appearance mining. Web appearance mining is very likely to become increasingly important because the appearances of Web pages significantly affect the interaction between users and such pages.

; / [

; that is;

; / [

2.3. Machine learning

cap

Machine learning deals with the design and development of *prediction* models (or functions) that allow computers to evolve behaviors based on training data such as sensor data or databases (An analysis of the primary difference between model construction in HCI and machine learning is given in Appendix A). For example, given a facial database, one can construct a face recognition model based on machine learning theories. Machine learning can be roughly divided into supervised, semi-supervised, and

lc

; / [

; / [

[2] In [Rosenholtz et al. [2007], SE and FC are primarily proposed to measure the visual clutter of images. However, Rosenholtz et al. discussed the definition of VisCom and the features of their proposed algorithms. They concluded that the proposed methods (i.e., SE and FC) can also be used in measuring VisCom. Interested readers can refer the fourth paragraph of the sixth section in [Rosenholtz et al. [2007].

unsupervised learning [Mitchell 1997]. In supervised learning, a prediction model is learned (constructed) to map inputs (called features) to desired outputs (called labels) based on training data of input-output pairs. Once the prediction model is learned, a new input can be fed into the model and the output is generated. Supervised learning can also be divided into sub-areas, namely, classification and regression. In classification, labels are categorical; in regression, labels are continuous. In this study, for example, if we aim to predict whether or not a Web page is visually complex, a model is constructed to classify a Web page into either a "Visually complex" or a "Visually simple" category. In other words, the labels are categorical. If we aim to predict the VisCom score of a Web page, a model is constructed to score the Web page. In this case, the labels to be used are continuous real numbers. lc x 2

Machine learning pays much attention to the generalization capability of a prediction model, which refers to the prediction performance on new, unseen data prediction. Two supervised machine learning algorithms, namely, random forest (RF) [Breiman 2001] and support vector machine (SVM) [Vapnik 1998] are demonstrated to have good generalization capability as reported in previous literature. These algorithms are used in this study and are briefly introduced in Appendix B. cap x 2 cap x 3

2.4. Automatic measurement of human feelings

cap x 3

With user experience receiving an increasing amount of attention, many studies have been conducted on the automatic evaluation of subjective human feelings on images, videos, texts, and Web pages using computational approaches. For instance, Datta et al. [2006] proposed a computational aesthetic approach to evaluate the visual quality of images. Ninassi et al. [2009] investigated visual quality assessment for videos. Pitler and Nenkova [2008] extracted 32 features for a text document and constructed a classifier to assess the readability of the text document. Zheng et al. [2008; 2009] designed elaborate experiments to collect user ratings and utilized low-level features to estimate user feelings and perceptions on interfaces and Web pages based on machine learning approaches. Their experimental results suggest that there are interesting patterns in the relationship between low-level features of Web pages and design-relevant dimensions. Wu et al. [2011] constructed a linear regression model to score the visual aesthetics of Web pages. The aforementioned studies profoundly illuminated the current study, particularly on the primary procedures used in the former, which are similar to the latter. One example is feature extraction, which is one of the areas considered in the current study that has been directly motivated by previous studies on images and Web pages.

3. MATERIALS AND PROBLEMS

3.1. Web pages and participants

cap x 2

Homepages are usually effectively designed to attract users because they give users the first impression of a Web site. In this ~~paper~~, 1300 homepages were collected, mainly from company, university, and personal sites among others. Given that all the participants were Chinese, no Chinese Web pages were selected. This is a common practice that designers use to prevent human evaluation from being interfered by the content [Thomas and Tullis 1998]. The URLs of all the experimental Web pages are available in an online Appendix (Appendix C). After completing the download, each page was pre-processed and their features were extracted using the feature algorithm introduced in Section 4.3. articles

The participants were recruited using mail advertising from our experimental laboratory. To ensure that all selected participants had normal perception on VisCom, each candidate was asked whether the page of www.baidu.com was visually simpler than www.sohu.com. If the answer was Yes, the candidate was selected. Finally, seven members from the laboratory and five friends of the laboratory members were selected as participants. The seven members from the laboratory were all Ph.D. students in Computer Science. The other five members were students specializing in Computer Science, Agriculture, and Electrical Engineering. Among the participants, seven were males and five were females, aged between ages 20 and 30. All the participants accessed Web pages at least every week.

" / "

lc x 2

lc x 5

3.2. Rating procedure

cap

In Web perception experiments [Papachristos et al. 2006], human labeling interfaces are generally well-designed to keep the perception of participants as close as possible to their perception when they access the pages freely. Therefore, we designed a labeling platform which shows a page similar to a Web browser with the scoring box located at the bottom of the interface as shown in Fig. 2. For each Web page to show, the platform tries to ensure that the page's presentation is as close to that in a standard Web browser as possible. Therefore, for long Web pages, users scroll the pages and find the rating boxes at the bottom of the interface. The platform ran on a Dell OPTIPLEX 320 PC with a resolution of 1280 x 1024 pixels.

Figure 2

ure

TR

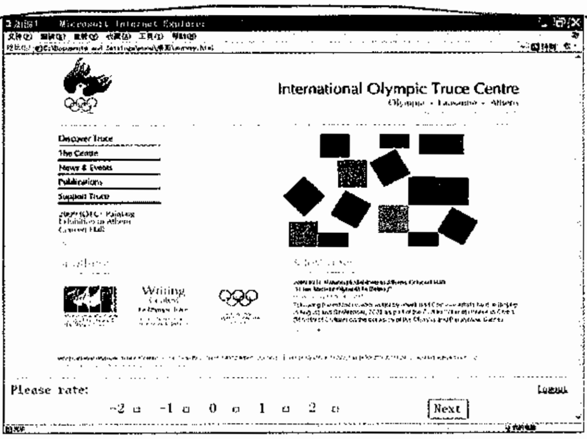


Fig. 2. The human labeling platform.

Considering that the concept of VisCom is not difficult to understand and we wanted the participants to access the pages freely, we did not give special instructions to the participants but only a simple training on how to use the labeling platform. Each participant was allowed to view one page within 5s and rate the page within an integral score from the set of basic ratings $\{-2, -1, 0, 1, 2\}$[3], where a score of 2 indicates that the VisCom is very high, whereas 2 indicates that the VisCom is low.

Footnote 3

minus sign

During the user-rating session, a Web page is randomly loaded. The participant then views the page and subsequently rates it. After rating, the participant clicks "next" to load the succeeding random page. If the participant fails to rate a page within a fixed amount of time, a page is randomly selected among the un-rated pages and is loaded automatically. After all the 1300 pages were rated, the participant's rating task

[3] There is no general standard about the setting of the number of basic ratings. In existing studies including [Michailidou et al. 2008; Annett 2002; Pitler and Nenkova 2008], five, seven, ten, and eleven are usually chosen. Since it is difficult to say which number is the best without any further information, five is selected in this study.

3 / [/]; / [/
]; and / [

is concluded. Each participant was required to rate all the 1300 pages within four hours. When all the participants finished their rating tasks, each Web page received 12 VisCom ratings.

3.3. Problems

The aim of this study is to construct a VisCom measurement model, which can predict the VisCom of a new Web page, based on the collected 1300 Web pages and their associated VisCom ratings. There are several key problems. (1) How to represent a Web page using a set of quantitative features? (2) How to learn a measurement model with good generalization capability? (3) Once a measurement model is constructed, how to validate its effectiveness?

The ~~above~~ problems are addressed by using a typical supervised machine learning approach together with Web mining and computer vision techniques. Generally, a typical supervised machine learning approach consists of three major stages: data preparation, training, and testing. The data preparation stage extracts features and collects user ratings to compile training and testing data. The training stage constructs a prediction model based on the training data. The testing stage evaluates the performance of the learned prediction model based on the testing data. The following section introduces our methodologies. preceding / # #

TR

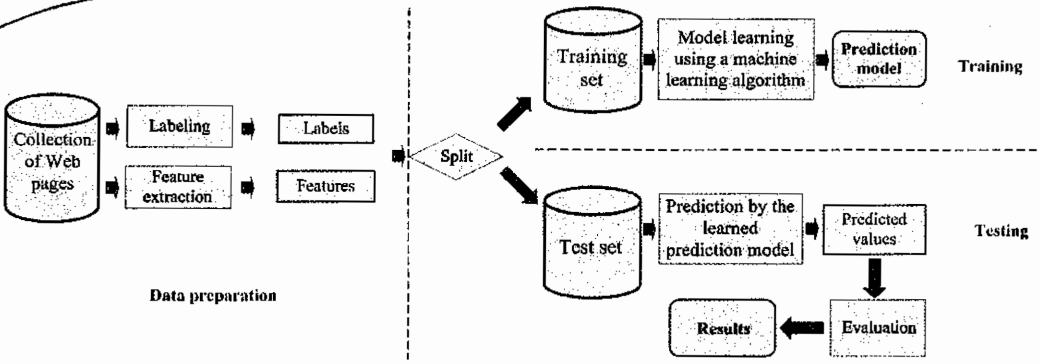


Fig. 3. Overview of the proposed approach.

4. METHODOLOGIES

4.1. Overview of the Proposed Approach

Figure 3

Our proposed approach falls under a conventional supervised machine learning approach. Therefore, the construction of a VisCom measurement function (a prediction model) by the proposed approach conforms to the main flow of a machine learning approach. The outline is shown in Fig. 3.

sure

As can be seen, the input of the data preparation stage is a collection of Web pages enclosed within the leftmost cylinder (Fig. 3). Initially, each Web page is labeled, and a number of measurable properties (features) are extracted from each Web page. All the extracted features of the Web pages and their labels comprise the labeled data set. The data set is then split into two sets, namely, training and test sets. In the training stage, the input is the training set. A (supervised) machine learning algorithm is utilized to obtain a prediction model (i.e., a VisCom measurement function) based on the training set. The output is a prediction model whose input and output comprise features of a Web page and a VisCom label, respectively. In the testing stage, the input / sure is the test set, and the learned prediction model is initially run on the features of each Web page in the test set. A predicted VisCom label for each test Web page is then obtained. Subsequently, the predicted labels are evaluated using an evaluation criterion (e.g., classification accuracy) based on the original labels. The final output suggests the performance of the learned prediction model.

Two points distinguish the proposed approach from the other HCI studies. (1) Our approach conforms to a standard machine learning framework, so theories in machine learning can help construct an effective VisCom measurement function with good generalization capability. (2) Unlike previous studies that involve limited number of factors, the proposed approach leverages Web mining and computer vision techniques to extract an extensive range of features. These features can represent a Web page in many cues such as color, layout, texture, and other visual factors.

a

4.2. VisCom Labeling

As introduced in Section 3, considering that the VisCom of a Web page is subjective and different people may perceive different or even opposite VisComs of the same page, each collected page is repeatedly labeled by 12 participants using several basic ratings. When all the participants finished their rating tasks, each Web page received 12 user ratings. The histogram of user ratings reflects how users perceive the VisCom of a Web page. For example, Figure 4 shows the histograms of user ratings of the two pages in Fig. 1. The two histograms suggest that all users agree that the left page has a lower VisCom. Other exemplar histograms of user ratings of four Web pages in our collection by twelve participants are shown in Fig. 5. The following part discusses the quantification of raw user ratings into a VisCom label.

Figure 4

Figure 5

ure

ure

TR

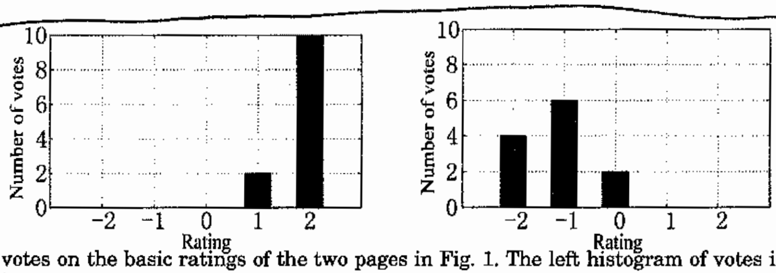


Fig. 4. The votes on the basic ratings of the two pages in Fig. 1. The left histogram of votes is for the left page in Fig. 1

⊙

Existing studies [Datta et al. 2006; Pedro and Siersdorfer 2009] usually use a categorical label or a score to quantify the user ratings. Despite the reasonableness of both the category and score quantifications, in some cases, a single quantity is insufficient to capture the true nature of the subjectivity of VisCom. It is observed from Fig. 5 that the disagreements between participants are relatively high. Some participants labeled a page as very visually complex (2), while some others labeled the same page as very visually simple (2). Figure 6 shows the histogram of variances of the ratings by participants for our collected 1300 Web pages. The variances of user ratings reflect the inter-rater disagreement. Most pages' rating variances are larger than 0.5. Therefore, the inconsistenc between VisCom scores by human cannot be ignored. However, both the category and score strategies do not consider this issue.

Figure 6

ure

“ / ” / minus sign

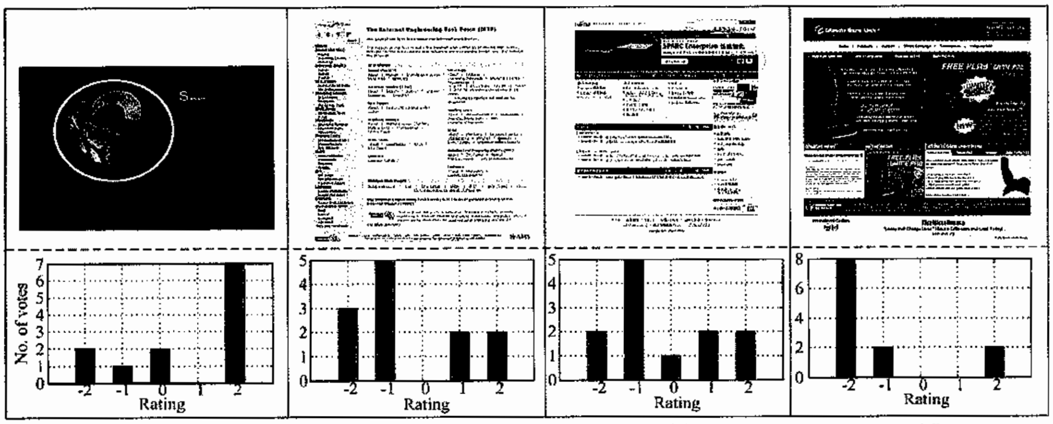
“ / ”

y / s

To this end, we proposed a new quantification strategy that directly applies the normalized histogram of user ratings as the VisCom label of a Web page. The quantification is as follows

7

$$y_k = (p_{k1}, \cdots, p_{ki}, \cdots, p_{kZ}), \tag{1}$$



(a) (b) (c) (d)

Fig. 5. The vote histograms on basic VisCom ratings of four Web pages.

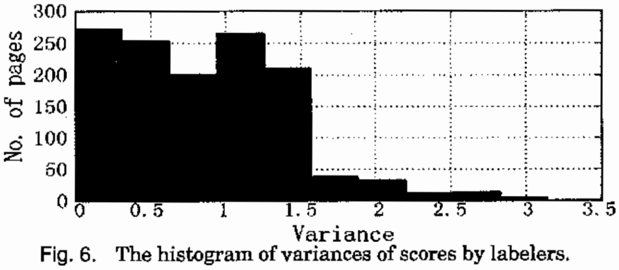


Fig. 6. The histogram of variances of scores by labelers.

where p_{ki} is the proportion that the i-th basic rating is chosen by users.

$$p_{ki} = \frac{\text{The number of users who chose the } i\text{-th basic rating to rate the } k\text{-th page}}{\text{The number of users}} \qquad (2)$$

When the number of users approaches infinity, y_k becomes the *distribution* of the perceived VisCom for the k-th Web page. Hence, this new strategy is called distribution quantification strategy. Using this strategy, the labels of the four pages in Fig. 5 are (0.1667, 0.0833, 0.1667, 0, 0.5833), (0.2500, 0.4167, 0, 0.1667, 0.1667), (0.1667, 0.4167, 0.0833, 0.1667, 0.1667), and (0.6667, 0.1667, 0, 0, 0.1667), respectively.

ure

Distribution quantification can capture the subjective nature of VisCom better in three aspects. (1) It is consistent with the subjectivity nature of VisCom that a page's VisCom can be perceived differently. (2) A distribution seems like a soft label of a page, which is similar in spirit with the fuzz set theory that arises from human subjectivity [Zadeh 1965]. Each entry of y_k reflects the approximate possibility that the VisCom belongs to the corresponding rating. (3) It provides a clear picture of how people perceive the VisCom of a page, and carries more information than a category and a score. As each entry of y_k is approximately continuous and locates in $[0, 1]$, the learning for this new quantification is called *VisCom distribution regression*.

Since a distribution is more suitable, then why are classification and score regression still explored in the study? Although distribution contains more information, learning a distribution measuring function is also more difficult and requires more human labelers, which is a practical challenge in real use. Section 4.4 will describe the learning algorithms for the measuring functions under the three quantifications.

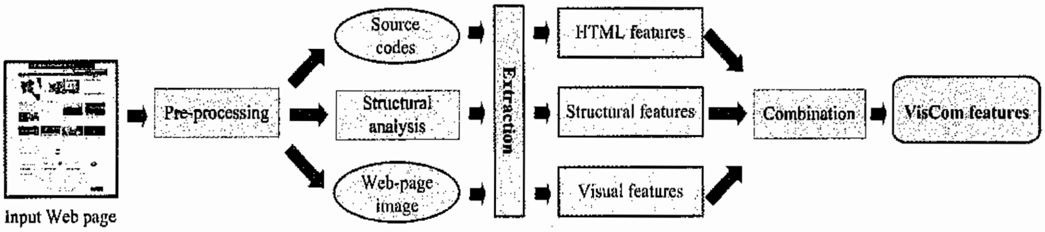


Fig. 7. The feature extraction for a Web page in the proposed approach.

Figure 7

4.3. Feature extraction

cap

Web page structural elements (e.g., text, links and images) and their visual characteristics (e.g., overall color scheme) determine the VisCom level of a Web page [Michailidou et al. 2008]. To extract both structural and visual features of a Web page, a new feature extraction procedure is proposed (Fig. 7). First, the source codes of an input Web page are obtained, and the page is transformed into an image (called Web page image). The page is then segmented, and the layout is constructed for the page. Three classes of features are extracted, namely, source-code (also called HTML features in this study), structural, and visual features. HTML features refer to the quantities of Web page elements (e.g., texts, links), structural features reflect the layout of a Web page, and visual features represent the color and the texture of the transformed Web page images. These three classes of features characterize the five main parts of a Web page: texts, images, links, background, and layout.

cure / =

#

#

In visual feature extraction, each transformed Web page image is represented both in Red-Green-Blue (RGB) and Hue-Saturation-Value (HSV) color spaces in the calculation requirement. In total, 44 features are extracted and denoted as $\{f_i | 1 \leq i \leq 44\}$. The succeeding section explains the details of each of the feature extraction steps.

#

4.3.1. Structural analysis. Michailidou et al. [2008] concluded that there is a strong correlation between the layouts of Web pages and their perceived VisComs. Hence, the primary goal of structural analysis is to extract the layout of a Web page. The obtained layout will be used to extract layout features and to aid in the extraction of visual features. Based on the definition of layout in design research [Ahmad et al. 2008], the layout of a page in this study is defined as a set of unoverlapped large rectangular blocks that (approximately) cover the whole page. These rectangular blocks are also called layout blocks. Figure 8 gives three layout examples for three pages, respectively.

cap

Figure 8

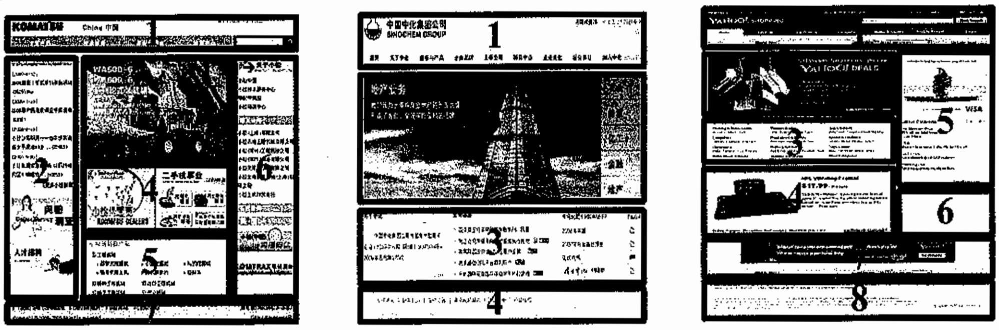


Fig. 8. Three Web pages and their extracted layouts (rectangles with heavy black lines) using V-LBE. There are seven, four, and eight layout blocks in the left, middle, and right pages, respectively.

TR

Numerous well-known Web page segmentation algorithms are proposed in previous literature [Cai et al. 2003b; Kohlschtter and Nejdl 2008]. These algorithms represent a segmented page using a tree. Song et al. [2004] directly adopted the leaf nodes (blocks) as the layout blocks of a Web page. The sizes of leaf nodes vary substantially, and some leaf nodes' sizes are very small, so a new heuristic layout extraction algorithm based on the Web page segmentation is introduced instead. Given that the present study explores the visual aspects of Web pages, the visual-based page segment (VIPS) [Cai et al. 2003b] algorithm is selected as the basic segmentation algorithm. A brief introduction of VIPS can be found in Appendix D.

Our algorithm is called **V**IPS based **L**ayout **B**lock **E**xtraction algorithm (V-LBE). The VIPS algorithm is first performed to segment a Web page to derive the VIPS tree. In this step, the parameter Permitted Degree of Coherence(*PDoC*) of VIPS is set to be large enough to ensure the smallest possible leaf node block granularity of the VIPS tree. Based on the VIPS tree, V-LBE selects all the layout block candidates whose sizes are above a threshold (τ_1) and then deletes or inserts blocks to construct a set of unoverlapped large blocks which (approximately) cover the whole page. The steps of V-LBE are shown in Algorithm 1. In our experiments, τ_1 is heuristically set as 1/9 of the whole page size, while τ_2 is heuristically set as 1/36 of the whole page size. The reason why 1/9 and 1/36 are selected is as follows. A Web page can be divided into 3*3 blocks or 6*6 blocks. At first, only the blocks whose sizes are above 1/(3*3) = 1/9 of the whole page are taken as major blocks or layout blocks. As many small blocks are deleted in this step, the left blocks cannot cover the whole page. Hence, in the second step, new blocks are generated in order to cover the uncovered page. Finally, if the new generated blocks are smaller than 1/(6*6)=1/36 of the whole page, then these new blocks are not taken as major blocks, namely, layout blocks.

no bold

allcaps

Algorithm 1: V-LBE

Input: a Web page, two thresholds τ_1 and τ_2.
Output: a set of layout (rectangular) blocks.
Steps:

1. Segment the Web page into a block tree using the VIPS method described in [Cai et al. 2003b]. The parameter *PDOC* is set to be large enough to ensure the smallest possible leaf node block granularity.
2. Access each node of the tree and select the nodes whose areas are equal to, or bigger than, the threshold τ_1. These selected nodes also consist of a new tree T_{new}.
3. Access each non-leaf node of T_{new}. If the node's children do not cover it, new nodes are generated as the node's children such that the node can be covered by its children.
4. Delete T_{new}'s leaf nodes whose areas are below τ_2, and output the rest of the leaf nodes' rectangular blocks.

Once the layout blocks are obtained, relative positions of the blocks can be easily inferred. An adjacent matrix ($\boldsymbol{A}$) is used to describe the relationships between blocks: $\boldsymbol{A}_{ij} = 1$ if the i-th block and the j-th block are adjacent, while $\boldsymbol{A}_{ij} = 0$ otherwise. Take the middle page in Fig 8 as an example, its adjacent maxtrix ($\boldsymbol{A}$) is $[0, 1, 0, 0; 1, 0, 1, 0; 0, 1, 0, 1; 0, 0, 1, 0]$.

sure

Let us suppose a Web page image's layout blocks are $\{B_1, \ldots, B_i, \ldots, B_M\}$. The following subsection describes the details of candidate features for VisCom measurement.

Table I. Definition of HTML features

cap

f_1	number of texts
f_2	number of linked texts
f_3	f_2/f_1
f_4	number of fonts
f_5	number of font sizes
f_6	average font size
f_7	number of font weights
f_8	average font weight
f_9	number of tables
f_{10}	number of background colors

Table 1

4.3.2. HTML Feature Extraction. The HTML features are all extracted directly from the source codes of Web pages. Details of these features are described in Table I. The features f_6 and f_8 are calculated as

$$f_6 = \sum_i fz_i \cdot p_i, \qquad f_8 = \sum_j fw_j \cdot q_j \tag{3}$$

where fz_i represents the i-th font size used in a page, p_i is the proportion of the texts with the i-th font size, fw_i represents the i-th font weight used in a page, and q_j is the proportion of the texts with the j-th font weight. Methods in HCI field focus on this class of features.

Figure 9

Table 2

4.3.3. Structural Feature Extraction. The structure of a Web page significantly affects its visual presentation and thus the perceived VisCom. The VIPS tree describes the detailed structure of a page, while the layout describes the overall structure. Intuitively, the more complex the VIPS tree, the more visually complex the page is Figure 9 shows two pages with distinct VIPS trees. The left page has lower VisCom and also a lower complex VIPS tree. However, studies on tree complexity are rare. We note that there are studies on graph complexity [Kim and Wilhelm 2008]. To measure the complexity of a graph, some basic structural information such as numbers of nodes and edges are usually used. Motivated by the graph complexity measurement, the features (f_{11}-f_{18}) shown in Table II are extracted to describe the complexity of a VIPS tree. Two features (f_{19}, f_{20}) are used to represent the layout.

Some features that describe texts and images in a page are also taken as structural features for they are obtained based on the VIPS tree. Four features (f_{21}-f_{24}) characterize the text distribution. For images, the number of images that provide information, instead of the number of all images used in HCI studies, is considered because there are a large number of quite small images only for decoration. The number of informative images (f_{25}) can be approximately obtained based on the VIPS tree. Each leaf node of the VIPS tree is described by a set of metadata. In the metadata, the value of the attribute of ContainImg denotes the number of images contained in the leaf node. Then, the sum of the values of the ContainImg attributes of all the leaf nodes of the VIPS tree is taken as the number of informative images (f_{25}). The rest features (f_{26}, f_{27}) are inspired by Schaik and Ling 1991].

4.3.4. Visual Feature Extraction. Colors and their organization are also key issues that affect the VisCom of an image [Rosenholtz et al. 2007]. Several attributes are utilized to characterize the color present in a Web page image.

Brightness (Bri) Three bright features (f_{28}, f_{29}, f_{30}) are used to describe the average brightness, brightness difference (among adjacent blocks), and brightness variance, respectively. The brightness of a pixel is its V component in the HSV color space. Let $Bri(B_i)$ be the average brightness of pixels in a Web page image layout block B_i.

ital. no bold

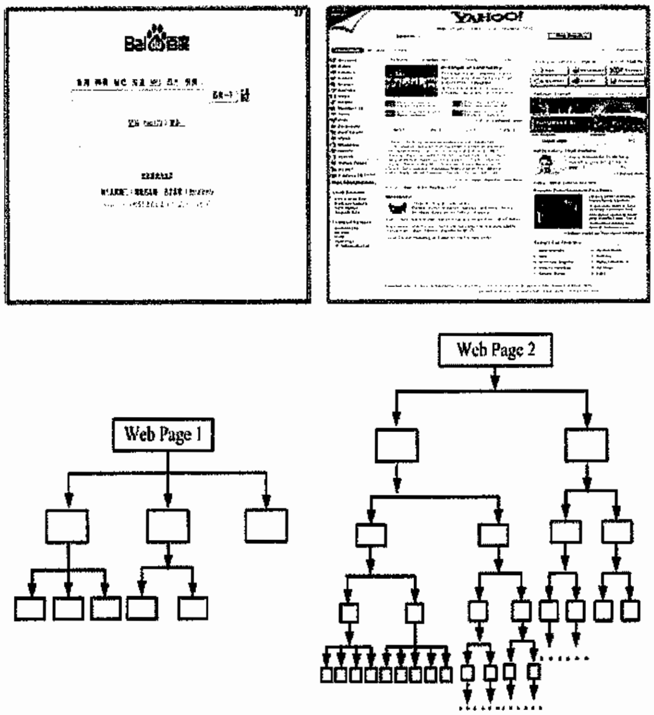


Fig. 9. The VIPS trees of two Web pages.

Table II. Definition of structural features

cap x 2

f_{11}	number of leaf nodes
f_{12}	number of layers
f_{13}	number of non-leaf nodes
f_{14}	number of nodes that have two children
f_{15}	number of nodes that have three children
f_{16}	number of nodes that have four children
f_{17}	number of nodes that have five children
f_{18}	number of nodes that have more than five children
f_{19}	number of layout blocks
f_{20}	number of pairs of adjacent layout blocks
f_{21}	number of text leaf nodes
f_{22}	number of texts in proportion to the whole page
f_{23}	total text area in proportion to the whole page
f_{24}	number of texts in proportion to the total text area
f_{25}	number of informative images
f_{26}	page's width + page's height
f_{27}	aspect ratio (page's height / page's width)

The three features are calculated as

$$f_{28} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i Bri(B_i), \tag{4}$$

where λ_i is the area proportion of B_i in the Web page image. #

$$f_{29} = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{M} A_{ij} |Bri(B_i) - Bri(B_j)| \tag{5}$$

$$f_{30} = \mathrm{Var}(Bri(B_i)). \tag{6}$$

Hue This factor is one of the main properties of a color. Three hue features ($\boldsymbol{f}_{31}, \boldsymbol{f}_{32}, \boldsymbol{f}_{33}$) are used to represent the average hue, hue difference, and hue variance, respectively. They are calculated similarly to Eqs. (4-6) by replacing $Bri(B_i)$ with the average hue value of the i-th block.

ital. no bold

} /N {

Colorfulness (Col) An efficient colorfulness evaluation algorithm is proposed by Hasler [Hasler and Susstrunk 2003]. The algorithm first calculates the opponent color space where for a pixel (R, G, B), its new coordinates are

ital. no bold) /⊙

and / [

$$rg = R - G, \qquad yb = 0.5(R + G) - B. \tag{7}$$

For block B_i, once the new coordinates (rg and yb) are obtained, the variance and mean of the rg and yb components can be obtained and represented by $\sigma_{rg}(B_i), \sigma_{yb}(B_i), \mu_{rg}(B_i), \sigma_{yb}(B_i)$, respectively. Then the colorfulness of an image block is calculated using

?

$$Col(B_i) = \sigma_{rgyb}(B_i) + 0.3\mu_{rgyb}(B_i), \tag{8}$$

$$\sigma_{rgyb}(B_i) = \sqrt{[\sigma_{rg}(B_i)]^2 + [\sigma_{yb}(B_i)]^2}, \tag{9}$$

$$\mu_{rgyb}(B_i) = \sqrt{[\mu_{rg}(B_i)]^2 + [\mu_{yb}(B_i)]^2} \tag{10}$$

⊙

Then three features ($\boldsymbol{f}_{34}, \boldsymbol{f}_{35}, \boldsymbol{f}_{36}$) are calculated similarly to brightness to characterize the average colorfulness, colorfulness difference, and blocks' colorfulness variance, respectively, by replacing $Bri(B_i)$ with $Col(B_i)$ in Eqs. (4-6).

} /N / {

Texture Texture is another important factor related with visual emotion. In the visual arts, texture refers to the surface quality perception of an artwork. Geusebroek et al. [2005] reported a six-stimulus basis for stochastic texture perception by considering the contrast and grain size of an image. The contrast of the image is indicated by the width of the distribution β, and the grain size is given by γ, which is the peakedness of the distribution. Hence, a higher value for γ indicates a smaller grain size (more fine textures), while a higher value for β indicates more contrast. For both parameters, three features are extracted similarly to brightness. Hence there are six features ($\boldsymbol{f}_{37}$-$\boldsymbol{f}_{42}$).

ital. no bold) /⊙

and Smeulders

Compressed File Size The compressed file size (JPEG size) ($\boldsymbol{f}_{43}$) of a Webpage image is taken into account. Normalized JPEG size ($\boldsymbol{f}_{44}$) is also considered, which is the ratio of the JPEG size to the whole Webpage image's area.

ital. no bold) /⊙

/

4.3.5. The Algorithmic Steps of the Feature Extraction. The sketch of the feature extraction is summarized in Algorithm 2. Step 4 occupies most of the processing time. To accelerate the extraction approach, large Webpage images are scaled down at first.

#

Once the features of a Web page are extracted using Algorithm 2, a feature vector is obtained for the Web page. The feature vector is the form of $\{f_1, f_2, ..., f_{44}\}$. To construct a map from the feature vector to the VisCom, machine learning algorithms will be used. The following subsection describes the learning algorithms used in this work.

4.4. Machine Learning Algorithms in Training

This subsection discusses the learning algorithms used in the training stage of the proposed approach. All three VisCom quantification strategies (i.e., category, score, and distribution) are investigated. The first two strategies have already been investigated in previous studies, whereas the investigation of the third strategy is pioneered by the current research. For the first two strategies, two well-known learning algorithms allcaps

```
Algorithm 2: Feature Extraction
Input: a Web page with its source codes.
Output: features ($\{f_i|1 \leq i \leq 44\}$).
Steps:
1. Extract the HTML features ($\{f_i|1 \leq i \leq 10\}$).
2. Apply VIPS and V-LBE (Algorithm 1) to segment the page and extract the layout blocks.
3. Extract the layout features ($\{f_i|11 \leq i \leq 27\}$).
4. Transform the Web page into an image; extract the brightness features ($\{f_i|28 \leq i \leq 30\}$),
   hue features ($\{f_i|31 \leq i \leq 33\}$), colorfulness features ($\{f_i|34 \leq i \leq 36\}$), texture features
   ($\{f_i|37 \leq i \leq 42\}$) and the compressed file size features ($f_{43}$ and $f_{44}$) based on the extracted
   layout blocks.
```

Table III. Two learning algorithms for each quantification strategy.

cap ×5

Quantification	Measuring function	Algorithm
Category	Classifier	Random forest, Support vector machine
Score	Regression function	Random forest, Support vector regression
Distribution	Distribution regression function	Neural network, SVDR

Table 3

were introduced and compared. For the third strategy, a new regression algorithm was introduced to address the learning of the measuring function.

Table III lists the learning algorithms for the three different VisCom quantification strategies and measuring functions. When VisComs are quantified using categorical labels, the measuring function becomes a classifier that can predict whether or not a Web page is visually complex. When VisComs are quantified using scores, the measuring function becomes a regression function that can score the VisCom of a page, whereas if they are quantified using distributions, the measuring function becomes a distribution regression function that can predict the VisCom distribution on the basic ratings for a page.

4.4.1. Classification. Theoretically, most existing learning algorithms can be leveraged to train the classifier. As previously mentioned, the random forest (RF) [Breiman 2001] and support vector machine (SVM) [Vapnik 1998] will be used and compared in the experiments.

cap ×2

cap ×3

4.4.2. Score Regression. Similarly, most regression algorithms can be used. RF and SVM can also address regression problems, so these methods are still used in score regression. RF is also called random forest regression (RFR), and SVM is also called support vector regression (SVR).

cap ×6

4.4.3. Distribution Regression. If Eq. (1) is used, the target value y_k is a vector instead of a single quantity. The learning is a multi-input multi-output regression problem. Conventional regression algorithms cannot be directly used. Two separate algorithms based on neural network (NN) [Rumelhart et al. 1986] and structural SVM [Tsochantaridis et al. 2004] are applied.

cap ×2

Distribution Regression based on NN. NN can be easily used in distribution regression. At first a back-propagation (BP) NN [Rumelhart et al. 1986] is learned with the following loss function

ital. no bold

cap ×2

$$l(y, \widehat{y}) = \left\|y - \widehat{y}\right\|_2^2, \tag{11}$$

where y is the desired output and $\widehat{y}$ is the prediction. As $\widehat{y}$ should satisfy each entry is non-negative and $\left\|\widehat{y}\right\|_1 = 1$, the output of the NN should be normalized. First, each entry of output $\widetilde{y}$ is maximum-minimum normalized transformed into [0, 1]; then, the output is normalized for the sum of all entries to equal 1.

Ital- no bold lo-x2

Distribution Regression based on Structural SVM (SVDR). Structural SVM is a general framework that addresses structural output learning problems. We adapt it to a new algorithm called SVDR (support vector distribution regression) to learn the distribution regression function. SVDR aims to learn a discriminate function $F : X \times Y \rightarrow R$ over input/output pairs, where X and Y represent the input and output spaces, respectively. With F, the measuring function φ (or distribution regression function) is

$$\varphi(x) = \arg\max_{y \in Y} F(x, y). \tag{12}$$

$F(x, \cdot)$ can be seen as a matching function. Ideally, the maximum of $F(x, \cdot)$ is at the desired output y for an input x. F is usually assumed linear in some combined feature representation of inputs and outputs $\Psi(x, y)$,

$$F(x, y : \mathbf{w}) = < \mathbf{w}, \Psi(x, y) >, \tag{13}$$

where w denotes the parameter vector to learn. The specific form of Ψ depends on the nature of the problem. Once Ψ is defined, w can be obtained by using mathematical optimization algorithms. Please refer to Appendix E for details.

5. EXPERIMENTS

Compared with existing methods, the proposed approach extracts extensive features and utilizes machine learning to construct the measuring function. Hence, the experiments aim to investigate: (1) whether or not the extracted features are effective, (2) which machine learning algorithm is more effective for VisCom measurement, and (3) whether or not the proposed approach outperforms existing methods. Section 5.1 proposes several hypotheses related to these problems, Section 5.2 introduces the experimental setup, Section 5.3 presents the experimental results, and Section 5.4 discusses the experimental results and analyzes the features.

As VisCom measurement for Web pages is a relatively new topic, studies on it are rare. Therefore, there is no much choice for competing methods. As introduced in Section 2.1, the HCI method is proposed by the ViCRAM project which initiates the study of VisCom measurement for Web pages. The FC and SE methods are two state-of-the-art image VisCom measurement algorithms. The JPEG method[4], though it seems very simple, has been used by researchers in previous related studies. Therefore, the above four methods were compared in our experiments. Appendix F provides a brief introduction of the codes and the software program of these methods.

ot

Footnote 4

preceding

5.1. Hypotheses

The evaluation of the feature extraction and learning algorithms is based on the following hypotheses.

"H" is itaic no bok

— **H1.** RF is more appropriate than SVM in terms of VisCom classification.
— **H2.** The proposed approach outperforms existing state-of-the-art studies in terms of VisCom classification.
— **H3.** RF is more appropriate than SVM in terms of the VisCom scoring.
— **H4.** The proposed approach outperforms existing state-of-the-art studies in terms of VisCom scoring.
— **H5.** The proposed algorithm SVDR outperforms BP NN in terms of VisCom distribution regression.

[4] For a Web page, it can be transformed into an image in the JPEG format using a Web browser. Then the image's file size is obtained and used as the indication of the VisCom of that page.

ital no bold

— **H6:** The features are distributed in a linear space.

Section 5.3 describes the experiments conducted to test the above hypotheses. Specifically, Section 5.3.1 discusses the tests for **H**1 and **H**2, where the proposed approach with two classification algorithms (RF and SVM) is compared against the four existing methods. Section 5.3.2 discusses the tests for **H**3 and **H**4, where the proposed approach with two score regression algorithms (RFR and SVR) is compared with the four existing methods. Section 5.3.3 discusses the tests for **H**5, where the proposed approach with two distribution regression algorithms, namely, neural networks (NN) and SVDR, is compared. In all experiments, two well-known dimension reduction techniques are applied to test **H**6: principal component analysis (PCA) and kernel PCA (KPCA) [Cao et al. 2003]. In addition, important features among all the 44 mentioned features are analyzed and discussed in Section 5.4.

preceding

cap × 4

5.2. Experimental Setup

5.2.1. Experimental Data. As previously introduced in Section 3, there are 1300 Web pages and each page receives 12 user ratings. Each Web page was transformed to a feature vector using the feature extraction method in Algorithm 2. Each page's user ratings are transformed into a categorical label, a score, and a distribution, respectively. The transformed labels, scores, and distributions are the desired outputs in the experiments of VisCom classification, scoring, and distribution regression, respectively. The features and the desired outputs of the 1300 pages construct the experimental data.

Footnote 5

According to the introduction in Section 4.2 for VisCom quantification, in classification, pages with average ratings $\leq 0 - \delta/2$ were placed into the high-VisCom category[5], whereas pages with average ratings $\geq 0 + \delta/2$ were taken as low VisCom. High-VisCom pages are viewed as positive samples. In score regression, the average of user ratings for each page was taken as the overall rating of the page. In distribution regression, the range of rating grades [-2, 2] is five, whereas the number of labelers is only 12. Therefore, the score range had to be transformed into a new basic rating set {-1, 0, 1}, wherein the labelers' rating scores within {-2, -1} become -1, and rating scores within {1, 2} become 1. The VisCom distributions for each page are then calculated using Eqs. (1) and (2). Six Web pages with low/medium/high average VisCom scores are shown in Fig. 10.

minus sign

Figure 10

ure

5.2.2. Evaluation Criteria. Evaluation criteria should be introduced to measure the predicted results (output by a VisCom measurement model learned on training data) on test data. Classical measures for binary classification [Fawcett 2006], such as true positive rate (**TPR**, or sensitivity), true negative rate (**TNR**, or specificity), micro-accuracy (**Acc**), were borrowed to evaluate the VisCom classification results. Let NH be the number of high-VisCom pages and NL be the number of low-VisCom pages in the test set. Let NPH be the number of high-VisCom pages that predicted as high-VisCom and NPL be the number of low-VisCom pages that predicted as low-VisCom in the test set. Then

/ cap × 3

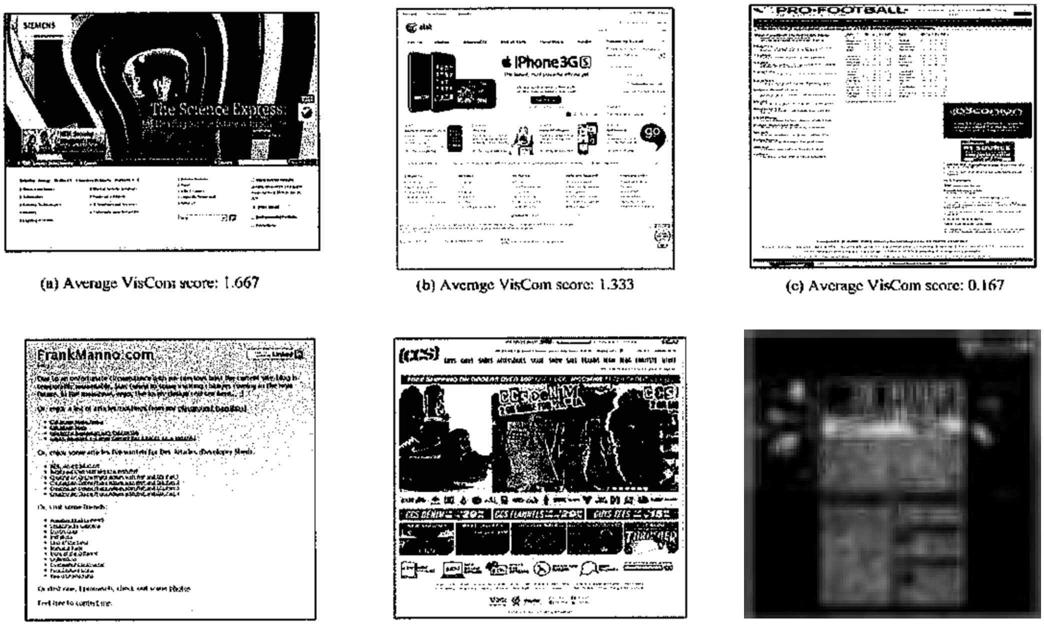
no bold / cap × 3

no bold

$$TPR = \frac{NPH}{NH} \tag{14}$$

$$TNR = \frac{NPL}{NL} \tag{15}$$

[5] Rating 0 represents the page is neither visually complex nor visually simple. δ is a manually determined parameter and indicates the gap in the average ratings between the high-VisCom and low-VisCom categories. For instance, if $\delta/2$ is set as 0.2, then pages whose average ratings are lower than -0.2 are placed into the high-VisCom category, and pages whose average ratings are not lower than 0.2 (0+0.2) are placed into the low-VisCom category. The gap in the average ratings between the two classes is 0.4 ($\delta/2$ + $\delta/2$).



(a) Average VisCom score: 1.667

(b) Average VisCom score: 1.333

(c) Average VisCom score: 0.167

(d) Average VisCom score: 0

(e) Average VisCom score: -1.333

(f) Average VisCom score: -1.667

Fig. 10. Six Web pages (ATT, SIEMENS, FrankManno, pro-football-reference, CCS Direct LLC, and Vision Festival) with low to high VisComs according to human ratings. A larger average VisCom score indicates a low visual complexity.

$$Acc = \frac{NPH + NPL}{NH + NL} \qquad (16)$$

⊙

For the scoring, the residual sum-of-squares error **(RSSE)** was applied to evaluate the performances

Cap x 4 / no bol

$$RSSE_r = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - f_r(x_i))^2, \qquad (17)$$

where N is the number of test pages, x_i is the feature vector of the i-th page in the test set, y_i is the desired output (i.e., the average of user ratings) of x_i, and $f_r(x_i)$ is the predicted VisCom score of x_i. In the VisCom distribution regression, the residual sum-of-squares error **(RSSE)** was also applied to evaluate the performances with a slight variation

no bold

$$RSSE_{dr} = \frac{1}{N-1}\sum_{i=1}^{N}\|y_i - f_{dr}(x_i)\|_2^2, \qquad (18)$$

where y_i is desired output (i.e., the normalized histogram of user ratings) of x_i and $f_{dr}(x_i)$ is the predicted VisCom distribution of x_i.

5.2.3. Comparison of the Competing Methods in an Experimental Session. To fairly compare the five methods (HCI, FS, SE, JPEG, and our proposed approach), they were run on the same set of Web pages that had VisCom labels. As illustrated in Fig. 3 in Section 4.1, the learned VisCom measurement model was evaluated on the testing stage. Therefore, the other four competing methods were also evaluated on the testing stage by replacing the step involving the "Prediction by the learned prediction model" with

ure

the step of "Prediction by a competing method (e.g., HCI)". The results of each of the competing methods were recorded and used in the performance comparison.

VisCom scoring was considered as an example to describe the comparison in detail. The 1300 rated pages were **randomly divided** into two subsets with equal sizes. One subset was taken as the training set, while the other was taken as the test set. The training set was used to train a VisCom scoring function based on the proposed approach. Subsequently, all the methods, including the VisCom scoring function by the proposed approach, took turns to predict the VisCom score of each Web page in the test set. The regression error for each competing method was then calculated using Eq. (17) based on the VisCom labels and the predicted scores. The ~~above~~ **random division** and the successive training and testing were repeated five times, and the regression error for each method was recorded at each time. Hence, for each method, five regression errors were recorded. Finally, the average regression error for each method was calculated, and the results were used to compare the five competing methods.

no bold

aforesaid

no bold

Furthermore, we performed the Wilcoxon rank sum test [Lam and Longnecker 1983] to judge the significance of the performance difference between two competing methods. If the difference was significant, we then concluded that the method with better results (e.g., a lower average regression error) was superior to the other one. Hence, the comparison between two methods, denoted as A and B, throughout the experiments consisted of two steps. The first step directly compared their corresponding values on the evaluation criteria, with the assumption that B has larger classification accuracy or a lower scoring error. The second step leveraged the Wilcoxon rank sum test to check whether or not the difference was significant. If the Wilcoxon rank sum test suggests that the difference was significant, the conclusion that B outperforms A was obtained.

5.3. Experimental Results

5.3.1. VisCom Classification. The RF and SVM, as well as PCA and KPCA, are compared to obtain an initial picture of the approach. As a result, six combinations are produced, namely, RF, PCA+RF, KPCA+RF, SVM, PCA+SVM, and KPCA+SVM. For RF, only the number of trees in {10, 50, 100, 200, 300} is changed, and other parameters are default. For SVM, only results on the linear kernel are reported (results on the radius basis function kernel are similar). The parameter C of the linear kernel SVM is searched in {0.01, 0.1, 1, 10, 20, 50, 100}. This parameter controls the tradeoff between errors on training data and model complexity which is related to generalization capability. All parameters are obtained via then-fold cross validation.

Table IV. Classification results (%) of the proposed approach based on SVM, PCA+SVM, and KPCA+SVM

cap ×4 / 9

TR

$\delta/2$	SVM			PCA+SVM			KPCA+SVM		
	TPR	TNR	Acc	TPR	TNR	Acc	TPR	TNR	Acc
0	65.73	76.92	72.00	65.03	73.63	69.85	64.34	80.77	**73.54**
0.2	65.73	76.92	70.02	68.53	70.79	69.78	62.24	80.90	**72.59**
0.4	70.31	77.06	74.16	74.22	69.41	71.48	69.60	79.17	**75.09**
0.6	72.36	76.51	74.74	73.98	71.08	72.32	72.36	79.52	**76.47**
0.8	80.37	73.53	**76.54**	75.86	68.18	71.48	73.68	75.16	74.53
1	80.37	74.07	76.86	77.19	67.97	71.91	78.50	77.21	**77.78**

Table 4
Table 5

Performances of the approach under each combination are shown in Tables IV and V. The results on three evaluation criteria, including Acc, are presented. Acc refers to the average accuracy of the predicted results. Both true positive rate (TPR) and true negative rate (TNR) indicate the accuracy rate of each of the two VisCom class labels, i.e. high and low VisCom, respectively. For all three criteria, a higher value

cap ×6

; that is; / #

cap ×4)/9

Table V. Classification results (%) of the proposed approach based on RF, PCA+RF, and KPCA+RF.

$\delta/2$	RF			PCA+RF			KPCA+RF		
	TPR	TNR	Acc	TPR	TNR	Acc	TPR	TNR	Acc
0	69.23	83.52	**77.23**	68.53	77.47	73.54	60.84	79.67	71.38
0.2	72.03	81.46	**77.26**	67.13	76.4	72.27	65.73	81.46	74.45
0.4	76.56	7941	**78.19**	66.20	76.97	72.19	68.75	81.76	76.17
0.6	76.42	79.52	**78.20**	69.11	76.51	73.36	70.40	82.74	77.47
0.8	83.33	78.43	**80.52**	78.07	69.28	73.03	78.07	74.51	76.03
1	84.11	80.00	**81.82**	74.77	77.78	76.45	78.50	73.33	75.62

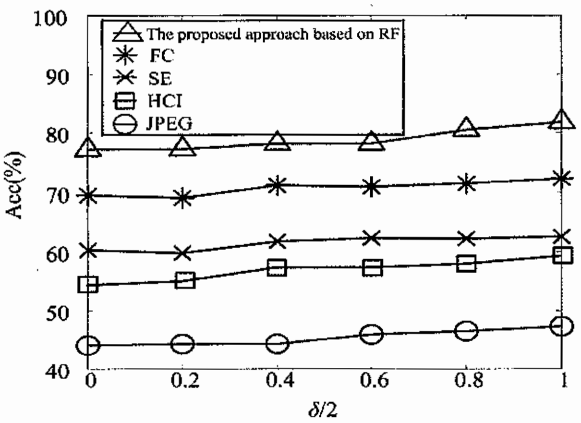


Fig. 11. The comparison of the proposed approach (with RF) and other methods on in terms of classification accuracy.

Figure 11

indicates better performance. Specifically, RF achieved the highest classification accuracies under all $\delta/2$ values. The proposed approach based on RF was then compared against the four existing methods, which produced an output score for an input Web page. First, the scores were linearly transformed into the range [-2, 2], after which the new scores were transformed into labels in comparison with $-\delta/2$ and $\delta/2$. The overall micro-accuracy (Acc) values are shown in Fig. 11. The item "The proposed approach based on RF" in Fig. 11 denotes the proposed approach when RF is used as the learning algorithm.

ure
ure

Based on the results in Tables IV and V, H1 can be checked. H1 states that in VisCom classification, RF is better than SVM. Under all the values of $\delta/2$, the Acc values of RF are higher than those of SVM. To check whether or not the differences are significant, the Wilcoxon rank sum test was performed. Results of the Wilcoxon rank sum test between RF and SVM show that the p-value remains at 0.002 (< 0.01). Consequently, H1 is valid and RF outperforms SVM in this domain. Using a similar comparison, we find that KPCA+RF outperforms PCA+RF, whereas KPCA+SVM outperforms PCA+SVM. The results further reveal that the original data are non-linearly distributed in the space. However, KPCA+RF is inferior to RF partially because performing integrated feature selection and feature reduction may discard a large amount of useful information.

Based on the results in Fig. 11, H2 can be checked. H2 states that in VisCom classification, the proposed approach is better than the existing state-of-the art methods. Under all the values of $\delta/2$, the Acc values of the proposed approach are consistently higher than those of the other four methods. After performing the Wilcoxon rank sum

ure

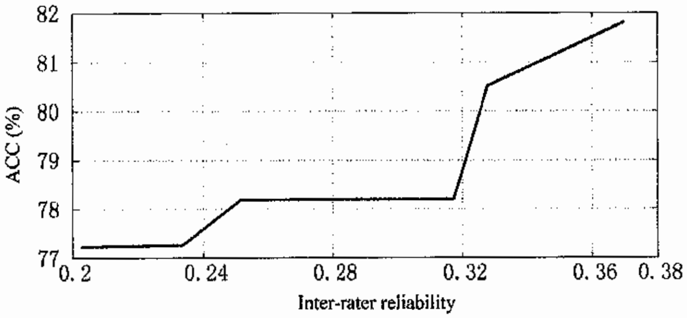


Fig. 12. The variations of classification accuracy (Acc) in terms of the inter-rater reliability.

Table VI. Regression errors of different score regression methods

cap ×5)

Methods	$RSSE_r$
RFR	**0.7484**
PCA+RFR	0.9365
KPCA+RFR	0.9449
SVR	0.9550
PCA+SVR	0.9339
KPCA+SVR	0.7967
HCI	1.0132
JPEG	1.1240
SE	1.0952
FC	1.0518

test, results between the proposed approach and each of the other competing methods show that the p-value remains at 0.002 (< 0.01). Further, the test results indicate that significant differences in accuracies exist among the various methods. Hence, H2 is supported.

Figure 12

Both Tables IV and V provide the results under different values of $\delta/2$. The overall performances (Acc) of most combinations show an upward trend with increasing $\delta/2$. We calculate the inter-rater reliability of the user ratings for pages with average scores outside ($-\delta/2$, $\delta/2$) using the Fleiss' kappa measure [Fleiss 1971]. Figure 12 shows the variations of classification accuracies in terms of inter-rater reliability. A clear increasing trend can be observed, and the correlation coefficient between accuracy and inter-rater reliability is 0.896. Therefore, we conclude that the pages whose VisComs have higher inter-rater reliability are easier to classify using a machine. In our point of view, the inter-rater reliability partially reflects the subjectivity. The pages that are more subjective are more difficult to classify using a machine which is based on objective criteria.

5.3.2. VisCom Scoring. The proposed approach is compared against the four existing methods in terms of scoring. RFR and SVR with linear kernel, KPCA, and PCA are applied, which still yields six combinations (i.e., RFR, PCA+RFR, KPAC+RFR, SVR, PCA+SVR, and KPCA+SVR). The parameters are searched in the same way as in VisCom classification. The scores achieved by HCI, JPEG, SE, and FC are linearly transformed into [-2, 2]. Table VI shows the achieved errors ($RSSE_r$) of different methods.

Table 6

Based on the ~~above~~ results, H3 and H4 can be checked. In Table VI, the proposed approach based on RFR achieves the lower regression error compared with all the other methods. The results of the Wilcoxon rank sum test between the proposed ap-

previous

proach and each of the other competing method show that the p-value remains at 0.008 (< 0.01). The test results indicate that RFR is superior to SVR in terms of VisCom scoring. Furthermore, the HCI algorithm is better than the other three existing algorithms (JPEG, SE, and FC). H4 is partially supported by the results in Table VI. Since the proposed approach with RFR obtains the lowest error, it is therefore superior to existing methods. However, when combined with PCA and KPCA, "PCA+RFR" is inferior to "PCA+SVR", while "KPCA+RFR" is inferior to "KPCA+SVR." The partial reason lies in the fact that performing integrated feature selection and feature reduction may discard a large amount of useful information. The results of the comparison are unsurprising because the proposed approach utilizes both HTML (primarily used in the HCI method) and visual (primarily used in FC and SE) features, as well as high-level structural information.

Footnote 6

Table 7

5.3.3. VisCom Distribution Regression. In this experiment[6], the performances of the proposed approach with two learning algorithms (NN and SVDR) are compared. For NN, a three-layer back-propagation neural network is employed. The node number in the hidden layer is selected from {50, 100, 150, 200, 300, 500}. The parameter C of linear kernel SVDR is determined from {0.01, 0.1, 1, 10, 20, 50, 100}. KPCA and PCA are still applied, which yields six combinations (i.e., NN, PCA+NN, KPCA+NN, SVDR, PCA+SVDR, KPCA+SVDR). Table VII shows the achieved regression errors of the proposed approach under the six combinations, as well as a random assignment strategy, in terms of $RSSE_{dr}$.

Table VII. Distribution regression errors of different distribute regression methods on the extracted features.

cap × 8

1R

Methods	$RSSE_{dr}$
NN	0.4688
PCA+NN	0.4631
KPCA+NN	0.4610
SVDR	**0.2036**
PCA+SVDR	0.2735
KPCA+SVDR	0.2495
Random	0.5857

H5 states that in the VisCom distribution regression, the proposed algorithm SVDR outperforms several heuristic algorithms that are based on the simple modification of classical regression methods. The results shown in Table VII support this hypothesis. A lower $RSSE_{dr}$ value indicates better performance, and SVDR achieves the lowest value. The results of the Wilcoxon rank sum test between the proposed approach based on SVDR and each of the other competing methods show that the p-value remains smaller than 0.016 (< 0.05). Thus, we concluded that SVDR has the best performance. The average $RSSE_{dr}$ values of the SVDR-series (SVDR, PCA+SVDR, and KPCA+SVDR) are nearly half of those of the NN-series (NN, PCA+NN, and KPCA+NN).

Table 8

For a detailed comparison of the six combinations, the residual sum-of-squares errors ($RSSEs$) of the three basic ratings (i.e., -1, 0, +1) were calculated. The results are listed in Table VIII. For the basic rating "0", the NN-series (NN, PCA+NN, and KPCA+NN) outperform the SVDR-series (SVDR, PCA+SVDR, and KPCA+SVDR). For the other two basic ratings, the SVDR-series obtain better results partially because for many training pages, VisCom distributions on the basic rating "0" approach zero. The

[6] Because the HCI, JPEG, SE, and FC methods output only a single score for each page, they are not compared in this experiment.

SVDR-series utilize the correlation between basic ratings and prefers predictions that have nonzero values on the basic rating “0”. Thus, the SVDR-series behave worse than the NN-series on the basic rating “0”. However, SVDR models the relationship between input and output and all constraints better. In addition, SVDR has better generalization capability than NN. Consequently, SVDR achieves better results. We further redefined the function Ψ in Eq. (19) in Appendix E) by removing $(y(1)y(2),\dots,y(Z-1)y(Z))$. The overall results of SVDR are inferior to the current form, which suggests the superiority of the current form of Ψ defined in Eq. (19). The SVDR-series achieves more flat errors on the three basic labels (“-1”, “0”, and “+1”) than the NN-series. As analyzed earlier, in distribution regression, SVDR considers the relationships among the quantities of the distribution over different basic ratings. As a result, the output distributions of SVDR will be flatter than those of NN.

TR

Table VIII. Regression erros of different distribute regression methods on each basic rating (-1, 0, +1) on the extracted features.

cap × 10

Methods	$RSSE_r$ on '-1'	$RSSE_r$ on '-0'	$RSSE_r$ on '+1'
NN	0.2214	0.0457	0.2017
PCA+NN	0.2236	0.0392	0.2003
KPCA+NN	0.2854	**0.0319**	0.1437
SVDR	0.0712	0.1176	0.0148
PCA+SVDR	**0.0145**	0.2501	**0.0089**
KPCA+SVDR	0.1419	0.0971	0.0105
Random	0.2315	0.1081	0.2461

In VisCom distribution regression, SVDR outperforms NN on extracted features because the former can model the learning problem better. Although distribution quantification appears to be more suitable in representing VisCom than category and score, the learning for a distribution regression function is more difficult than that for a classifier and a score regression function. In the distribution regression experiments, the performance of NN approaches the random method. Although SVDR achieves better results over NN and the random method, the former requires higher training time than the latter.

5.4. Discussion and feature analysis

cap × 2

The experiments provide several initial qualitative analyzes and results for different learning algorithms (i.e., RF and SVM), different VisCom quantification strategies, and comparisons with existing algorithms. The results support the hypotheses presented at the beginning of Section 5. The performance of the extracted features in this study is better than that of the features in the HCI method, indicating that extracting more advanced features improves VisCom measurement. For learning algorithms, RF (RFR) outperforms SVM (SVR) in terms of classification and score regression. In distribution regression, the proposed algorithm, SVDR, is better than NN. On both the VisCom classification and scoring, the proposed approach achieves better measurement performance than existing ~~existing~~ methods.

Now the features are analyzed. The features that are more discriminative in VisCom classification[7] were determined by applying Fisher's criterion [Duda et al. 2001] and the feature selection tool provided by LibSVM[8] to rank all extracted features according to their discriminative capabilities. Thus, two rank lists are obtained. RF also provides the feature ordering as output when training, so the rank list is, likewise, considered.

Footnote 7

Footnote 8

[7] As most feature evaluation algorithms are designed for classification, our features are also evaluated in terms of classification.

[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Table IX. Top 10 principal features in VisCom classification 9

cap x3 / 9

Rank	Feature
1	f_{43} (JPEG size)
2	f_{44} (normalized JPEG size)
3	f_{37} (average γ value in the texture features)
4	f_{28} (average brightness)
5	f_{25} (number of informative images)
6	f_{40} (average β value in the texture features)
7	f_{13} (number of non-leaf nodes of VIPS tree)
8	f_{30} (brightness variance)
9	f_{26} (page' width + page' height)
10	f_8 (average font weight)

As a consequence, three feature rank lists reflecting the features' discriminative capability are derived. To get a more appropriate and robust feature ordering, the three rank lists are fused to a new feature ordering via rank aggregation techniques [Dwork et al. 2001] which can combine different orderings into a consensus ordering.

Table 9

The top 10 features of the consensus ordering are the most important features which are listed in Table IX. The ordering is consistent with the studies and findings in the human-computer interaction (HCI) field. First, JPEG size (normalized JPEG size) is a factor of primary importance. JPEG is notably used in the measurement of VisCom solely in the area of human computer interaction. Hence, if a limitation is imposed to use only one factor to indicate the VisComs of Web pages, JPEG size appears to be the best choice. Second, the four features ($f_{37}, f_{28}, f_{40}, f_{30}$) are directly extracted from the Web page images. The high discriminative capabilities of these features suggest that transforming a Web page into an image and extracting features using computer vision techniques are helpful. In fact, previous studies reveal that textures with repetitive and uniformly oriented patterns were found to be less complex than disorganized patterns [Michailidou et al. 2008]. As such, an entirely black computer screen would be judged to be insignificantly less complex than a screen generated by a random collection of red, green, and blue pixels [Donderi 2006]. Texture (f_{37} and f_{40}) and brightness (f_{28} and f_{30}) do matter for VisCom.

#

=

#

In addition, f_{30} is among the top 10 crucial factors, which also indicates the usefulness of contrast features. Feature f_{25}, which describes the number of informative images, can also be observed as very discriminative. This finding is reasonable because informative images are usually more attractive to users and hence they play an important role on users' perception. Feature f_{13} represents the number of non-leaf nodes of the VIPS tree of a Web page. Intuitively, the number of non-leaf nodes partially reflects the hierarchical structure of a Web page. Hence, this feature is still reasonably highly discriminative. Thus, the conclusion of the current study is inconsistent with those of previous studies in [Michailidou et al. 2008] that support the proposition that the overall structural layout of a page is the most important factor in predicting user impressions.

} / [

Table 10

The correlation between features and VisCom scores is also investigated. The top 10 features that are most correlated to VisCom scores are ordered in Table X. The listed features are generally consistent with those listed in Table IX, especially since five features appear in both tables, ~~i.e.,~~ JPEG size (f_{43}), normalized JPEG size (f_{44}), average β value (f_{40}), number of non-leaf nodes of VIPS tree (f_{13}), and number of informative images (f_{25}). All features in Table X are negatively correlated to the VisCom scores. Alternatively, a larger value of the features in Table X denotes a higher VisCom, which is consistent with the definitions of these features. The feature that is most positively correlated to VisCom is the average brightness (f_{28}), which is also consistent with the observations on the six sample Web pages as shown in Fig. 10. The sample pages with

that is

Figure 10

ure

Table X. Top 10 most correlated features to VisCom scores

cap × 4

Rank	Feature
1	f_{43} (JPEG size
2	f_{40} (average β value)
3	f_{44} (normalized JPEG size)
4	f_{12} (number of layers of VIPS tree)
5	f_{13} (number of non-leaf nodes of VIPS tree)
6	f_{11} (number of leaf nodes of VIPS tree)
7	f_{10} (number of background colors)
8	f_{25} (number of informative images)
9	f_{42} (average β variance)
10	f_{21} (number of text leaf nodes of VIPS tree)

low VisComs are brighter than others. The average brightness feature (f_{28}) is selected as being among the top 10 important features in classification, because it is the most important feature to Web pages in the high-VisCom category. Six structural features ($f_{11}, f_{12}, f_{13}, f_{21}, f_{25}$) are among the top 10 most correlated features to VisCom score, indicating the importance of structural information in determining the VisCom scores of Web pages.

6. CONCLUSIONS

This paper initially investigated VisCom measurement for Web pages by integrating studies from the areas of HCI, Web mining, computer vision, and machine learning. Motivated by existing HCI studies (mainly from the ViCRAM [Michailidou et al. 2008] project) on VisCom, a number of features have been extracted to represent a Web page based on Web mining and computer vision techniques. These features describe a wide range of visual cues in a Web page, including text, images, structure, color, brightness, texture, and so on. The top 10 important features in VisCom classification are investigated, and the findings are consistent with existing HCI studies [Donderi 2006; Michailidou et al. 2008]. Moreover, this study also discovered the importance of several features (e.g., brightness variance and texture) that have not been used in previous VisCom measurement. Brightness variance and texture partially characterize diversity and density, respectively. According to the studies of VisCRAM, these are primary factors that affect VisCom. Half of the top 10 most correlated features to VisCom scores are the structural features from the VIPS trees. These findings indicate that the utilization of Web mining and computer vision techniques to extract the important features is helpful.

article

As another contribution, the present work discusses the subjective property of VisCom. The user rating results in our study show that there are often larger disagreements among the raters. Given that subjectivity should not be ignored in VisCom research, an attempt has been made to apply user rating distribution to quantify VisCom and capture subjectivity better than existing categorical or score representations. Two learning algorithms have been introduced for each of the three quantification strategies, i.e., VisCom classification, scoring, and distribution regression, respectively. The experimental results indicated that RF outperforms SVM in classification and scoring partially because of the implicit feature selection of RF.

; that is, / ?

Compared with state-of-the-art methods used in HCI and image processing fields, the proposed approach achieve the best performance. However, several limitations in the current study have been identified. (1) The background data of the participants in our experiments are not diverse. All of them are Chinese users aged in 20 30 and with good educational background. Therefore, their perceptions are not enough to represent the perceptions of all users worldwide. As a result, the learned VisCom measurement model may only be suitable for users with similar background as those of the partic-

s

–/N

ipants in our experiments. (2) The number of experimental Web pages is insufficient. From the perspective of machine learning, the training data should be substantially large in order to train an effective VisCom measuring function for Web pages. Although over 1000 Web pages have been used in the present work, these are still insufficient because of the diverse styles of Web pages. (3) The time consumed for the visual feature extraction is relatively high. Visual features are suggested to be very effective; unfortunately, the extraction procedure is highly time consuming. In related experiments performed with the present study, the average time required in transforming a Web page into a Web page image is 5 s in a 3.4 GHz PC with 1GB memory.

To address the above limitations, the following future work can be considered. (1) Collection of a greater quantity of Web pages. With the development of current Web crawling techniques, automatic collection of a large number of Web pages is relatively easy. To ensure that the collected pages thoroughly represent the Web pages all over the world, the styles of the collected pages should be as diverse as possible. Thereafter, the Web page genre classification techniques [Chen and Choi 2008] can be introduced to monitor the distribution of the different styles of the collected Web pages. If the proportion of pages belonging to a specific style is obviously small, it is expected that pages of that style are more crawled. (2) Recruit~~ment of~~ more users to rate the pages. The crowdsourcing Internet marketplace Amazon's Mechanical Turk [Amazon 2005], which receives increasing attention in machine learning in recent years, can be used to obtain labels from users from various background and age groups at a relatively low cost. With the help of Amazon's Mechanical Turk, we can collect VisCom rating data from users worldwide. With sufficient number of rating users, we can construct more effective VisCom measurement models as well as conduct more interesting VisCom studies. For example, we can investigate the relationships between VisComs and user language experiences, and explore distinct "VisCom perception groups" and construct personalized VisCom measurement models for different groups of users with similar VisCom perception. (3) Investigation of the subjective nature of VisCom. In our point of view, inter-rater reliability reflects the subjective nature of a Web page. Therefore, inter-rater reliability may help us analyze the subjectivity of VisCom. Furthermore, the VisCom of a Web page is subjectively viewed by different users. In other words, for a specific Web page, the VisCom depends not only on the Web page itself but also on the users. Therefore, we may consider both the feature of a Web page and the characteristics of the involved user. The combination of these features takes both Web pages and users into account. So far, previous studies on automatic measurement of human feelings have not considered this issue. Other possible future work is introduced in Appendix G.

just mentioned

#

/ ()

~~aforesaid~~

Concerning the lc /

We should lc /

e / has / d

Regarding the lc /

APPENDIX A: The primary difference between model construction in HCI and machine learning

cap × 6

no bold no ital

lc

In classical HCI studies, the primary goal of model (or function) construction is to ***reveal*** the mathematical relationships among the target value (e.g., VisCom) and the features (e.g., TLC, Word counts) of an object. As a byproduct, the constructed models can be used to ***predict*** the target value of a new object. In machine learning, the primary goal of model (or function) construction is to ***predict*** the target value of a new object. Nevertheless, if the model is an explicit function, the model also ***reveals*** the relationships among the target values and the features of an object.

Classical HCI focuses on description over prediction, whereas machine learning focuses on prediction over description. When a set of target values and features of an object (called training set in machine learning) is constructed, classical HCI attempts to construct a model that best fits for the target values and features, whereas machine learning attempts to learn (or construct) a model that has the best generalization capability. Therefore, in HCI, conventional regression algorithms are usually introduced to construct models, whereas in machine learning, numerous learning skills are usually taken into conventional regression algorithms to improve the model's generalization capability.

HCI studies have several advantages: (1) Domain knowledge is deeply analyzed, and (2) HCI models are usually explicit linear or polynomial function functions, whereas several machine learning models are implicit functions. Hence, HCI models make it easier to observe how the target values depend on the key features. The current work primarily aims to construct a model (function) to predict the VisComs of new Web pages, so the machine learning approach was employed. Nevertheless, considerable domain knowledge used in HCI is used to guide feature extraction. In our point of view, the construction of an effective VisCom measuring function relies on the combination of HCI, Web mining, computer vision, and machine learning studies.

APPENDIX B: Random forest and support vector machine

cap × 4

no bold

Random forest (RF): RF [Breiman 2001] is an ensemble classifier consisting of many decision trees [Schaik and Ling 1991] such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. This technique has several advantages: efficiency, embedding feature selection, and good generalization capability among others.

Support vector machine (SVM): This technique [Vapnik 1998] transforms data from the original space to a high or infinite dimensional space and constructs a hyperplane or set of hyperplanes in the transformed space, which can be used for classification, regression or other tasks. It has been widely used in many studies including text classification, image attractiveness prediction, and so on.

APPENDIX C: The experimental Web pages

cap × 2

For interested readers, the URLs of all the experimental Web pages are available at the online appendix.

APPENDIX D: A brief introduction of VIPS

cap × 2

cap × 3

The VIPS algorithm combines the document object model (DOM) tree and the visual cues (e.g., background color, font size, font weight, and so on) of a Web page to deduce the visual-based content structure of the page. First, the visual block extraction process is performed from the root node of the DOM tree to check each DOM node so as to judge whether it forms a single block or not. If so, a visual block, which can cover the node, is extracted; otherwise, the node's children will be checked in the same manner, and all the extracted blocks are placed into a pool. Secondly, visual separators among adjacent blocks are identified, and the tree structure of the blocks is constructed. Third, each visual block is checked to verify whether or not it meets the granularity requirement; if not, the block will be further partitioned iteratively. Finally, the vision-based block tree (called VIPS tree) for the Web page is output. The VIPS tree can be represented by VIPStree = {VB1, VB2, ..., VBn}, where VBi is the top-level visual block. VBi can also be represented by VBi={VBi_1, VBi_2, ..., VBi_m}, where VBi_k is the visual block of VBi. Figure 13 demonstrates a classical Web page segmentation example using VIPS from Cai et al. 2003b]. Figure 13(a) is a Web page. Figure 13(b) shows the visual blocks. Figure 13(c) shows the VIPS tree of the page. For each node, VIPS calculates the degree of coherence (DoC) to measure how coherent it is. Therefore, a parameter Permitted Degree of Coherence (*PDoC*) is used in VIPS to control the granularity (coherence) of the leaf node blocks. Different *PDoC*s correspond to different granularity levels of VIPS trees. The larger the *PDoC*, the finer the VIPS tree.

Figure 13

3 / [

cap × 2

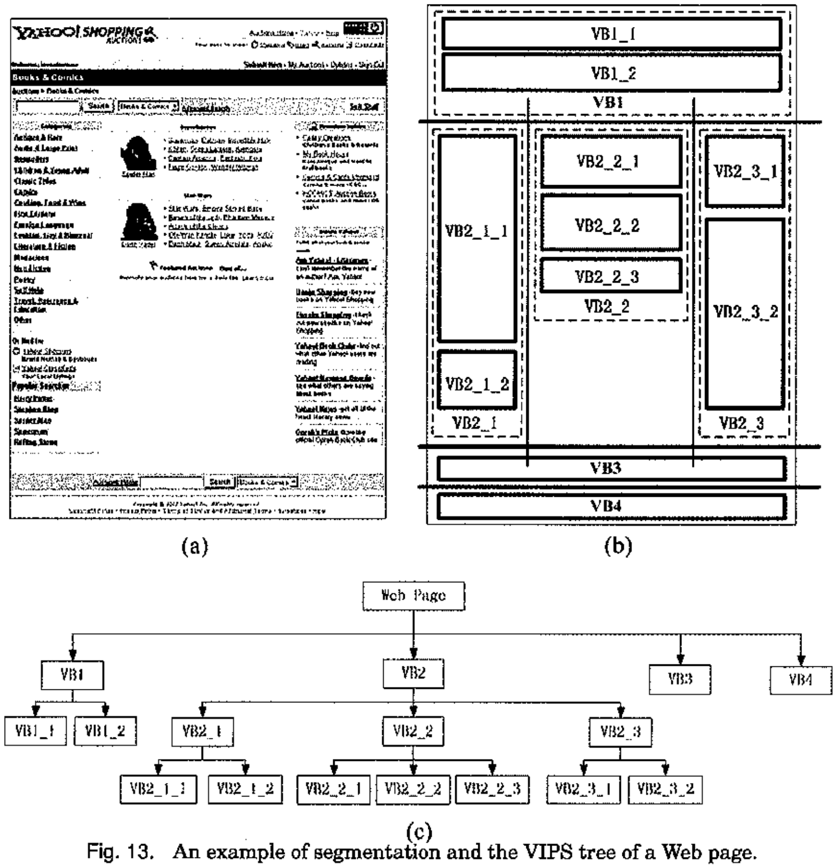


Fig. 13. An example of segmentation and the VIPS tree of a Web page.

APPENDIX E: The solution of *w* in Eq. (13)

Let V be the feature dimension. Motivated by the definition of Ψ in multiclass learning [Tsochantaridis et al. 2004], in our study, Ψ is defined as follows

$$\Psi(x, y) = (x \otimes y, y(1)y(2), \cdots, y(k)y(k+1), \cdots, y(Z-1)y(Z)) \tag{19}$$

where $x \otimes y((x(1)y(1), \cdots, x(i)y(j), \cdots, x(V)y(Z)))$ is the tensor product of x and y, which describes the relationships between input and output, and $(y(1)y(2), \cdots, y(k)y(k+1), \cdots, y(Z-1)y(Z))$ describes the correlation between adjacent basic ratings. The loss function in SVDR is the same as in Eq. (11). If $P(x, y)$ denotes the joint distribution, then the goal of SVDR is to find an optimal w such that the risk

$$R_p(\varphi) = \int_{X \times Y} l(y, \varphi(x)) dP(x, y) \tag{20}$$

is minimized. With the defined $\Psi(x, y)$ and $l(y, \widehat{y})$, the optimization for SVDR learning can be summarized as follows

$$\begin{array}{l} \min\limits_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum\limits_{i=1}^{N}\xi_i \\ s.t. \\ \quad \forall i \in [1,N], \xi_i \geq 0 \\ \quad \forall i \in [1,N], \forall y \in Y/y_i: \ <\mathbf{w}, \Delta\Psi_i(y)> \ \geq l(y_i,y) - \xi_i \end{array} , \tag{21}$$

Footnote 9

where $\Delta\Psi_i(y) = \Psi_i(x_i,y_i) - \Psi_i(x_i,y)$, C controls the model complexity[9], and ξ_i are slack variables.

allcaps

Algorithm 3: Update the working set (WS) for (x_i, y_i) in the t-th iteration

```
Input: $(x_i, y_i)$, $\epsilon$, $\mathbf{w}_{t-1}$, working set $WS_{t-1}(i)$.
Output: working set ($WS_t(i)$).
Steps:
1. Compute $\widehat{y} = \arg\max_{y\in Y} G(y)$, where $G(y) = l(y_i, y) - <\mathbf{w}_{t-1}, \Delta\Psi_i(y)>$.
2. Compute $\eta_i = \max\{0, \max_{y\in WSet_{t-1}(i)} G(y)\}$.
3. If $G(\widehat{y}) > \eta_i + \epsilon$ then $WS_t(i) = WS_{t-1}(i) \cup \{\widehat{y}\}$, else $WS_t(i) = WS_{t-1}(i)$.
```

However, as y is continuous, the constraints for each y_i in the optimization problem are infinite. To handle this problem, for each y_i, a small working set of most active constraints is constructed to replace the infinite constraints. Following the method proposed by Tsochantarids et al. [Tsochantaridis et al. 2004], the construction of the working set for the i-th training sample (x_i, y_i) is given by Algorithm 3. The maximum optimization problem in Algorithm 3 is as follows

$$\begin{array}{l} \max\limits_{y} \sum\limits_{v=1}^{V}\sum\limits_{\varsigma=1}^{Z} \mathbf{w}_{t-1}(Z(v-1)+\varsigma)x_i(v)y(\varsigma) + \sum\limits_{\varsigma=1}^{Z-1} \mathbf{w}_{t-1}(ZV+\varsigma)y(\varsigma)y(\varsigma+1) + \|y_i - y\|_2^2 \\ \quad s.t. \sum\limits_{\varsigma=1}^{Z} y(\varsigma) = 1, \quad y(\varsigma) \geq 0, \varsigma = 1, ..., Z \end{array} \tag{22}$$

This optimization problem can be solved via quadratic programming. Once the working sets for each training sample are obtained, the second class of constraints of Eq. (21) becomes

$$\forall i \in [1,N], \forall y \in WS_t(i): \ <\mathbf{w}, \Delta\Psi_i(y)> \ \geq l(y_i,y) - \xi_i. \tag{23}$$

Then w can be updated by solving the dual form of Eq. (21) using the cutting-plane algorithm [Franc and Sonnenburg 2008]. The entire iteration stops until the working sets for each sample remain unchanged.

The main steps of SVDR are shown in Algorithm 4.

APPENDIX F: The use of the codes or software in the experiments

cap ×4

For the VIPS algorithm, the Demo and Dynamic-link library (DLL) are available at `http://www.zjucadcg.cn/dengcai/VIPS/VIPS.html`. A print screen of the Demo is shown in Fig. 14. Users can input a value of the $PDoC$ in the range of $[1, 40]$ and then press the "VIPS" button to obtain the VIPS tree of the page. Based on the Dynamic-link library of VIPS, the VIPS tree (shown in Fig. 9) of a Web page can be obtained.

Figure 14

ure

ure

[9] C's value is required to manually set. In practice, its value can be searched via a classical machine learning experimental trick, namely, cross validation [CV].

=

2013

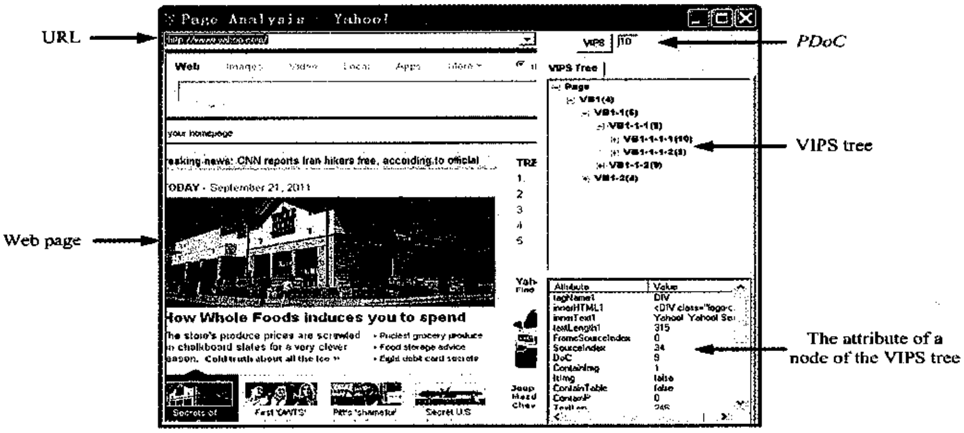
allcaps

```
Algorithm 4: SVDR
Input: $\{(x_1, y_1), \cdots, (x_N, y_N)\}$, $\epsilon$, $\mathbf{w}_0$, working set $WS_0(i) = null$, $t = 1$.
Output: The parameter vector $\mathbf{w}$.
Steps:

1. Compute the working set $WS_t(i)$ for each training sample $x_i$ using Algorithm 3. If for all
   $i \in [1, N]$, $WS_t(i) == WS_{t-1}(i)$, return the $\mathbf{w}_t$ and exit.

2. Replace the constrains in Eq. (21) with the working sets obtained in Step 1.

3. Optimize Eq. (21) using the cutting plane algorithm [Franc and Sonnenburg 2008]. Goto
   Step 1.
```



Attribute	Value
tagName1	DIV
innerHTML1	<DIV class="logo-c
innerText1	Yahoo! Yahoo! Se
textLength1	315
FrameSourceIndex	0
SourceIndex	34
DoC	9
ContainImg	1
IsImg	false
ContainTable	false
ContainP	0
TextLen	248

Fig. 14. A print scree of the Demo of VIPS.

For the RF algorithm, the introduction, codes, and guide are available at `http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`. For SVM, more than one type of software is available online. Among the available softwares, libSVM is one of the most widely used versions in the literature. The introduction, software and its guide are available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`. Users can use 3 the software to train and test classification or regression models. SVM is one of the most effective machine learning algorithms in recent years and is commonly used as a competing algorithm in pattern recognition and machine learning experiments. Users who plan to use LibSVM can refer to the online document at `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`. For SVDR, we implemented it based on the codes of structural SVM available at `svmlight.joachims.org/svm_struct.html`. For BP NN, we can run it directly using the functions provided by the software Matlab.

For the counting of the TLC of a Web page, interested readers can download the software of the ViCRAM tool according to [Michailidou 2009] (pp. 254]. For the methods 3 / [/3 page /] of SE and FC, the codes can be found at `http://Web.mit.edu/rruth/www/clutter.htm` (please refer to the bottom of that Webpage). #

APPENDIX G: Other possible future work

cap × 3

Some other future work which may be useful the VisCom measurement is as follows. (1) Selection of more effective features. A number of features used in this study are ital correlated. From a machine learning perspective, a smaller number of independent features could be more useful than a large number of correlated features. Combining more domain knowledge in HCI and feature selection techniques in machine learning will help us find a more effective feature subset. In addition, the costs of feature extraction should be considered during feature selection. Some features with high extraction cost can be omitted if they do not significantly affect the measurement performance. Indeed, studies about cost-awareness feature selection for Web page genre classification [Levering and Cutler 2009] have been conducted. These can help in the construction of a VisCom measurement model, which finds a balance between time cost and measurement performance. (2) Introduction of more sophisticated machine learning algorithms. As a machine learning approach, this study pays more attention to VisCom labeling and feature extraction than machine learning algorithms. In fact, many more sophisticated machine learning algorithms can be introduced to derive the VisCom measurement model. For example, if the number of collected Web pages is large (e.g., 10000), a participant cannot possibly rate all of the collected Web pages. In such cases, the participants can rate only a small portion of the collected Web pages, after which a semi-supervised machine learning algorithm can be used. In addition, an active learning strategy [Wang and Hua 2011] can also be employed to help users label more effectively. (3) Construction of application-oriented measuring functions. Measuring functions that can classify, score, or predict the score distribution of the VisCom of a Web page have been constructed in this study. In some concrete applications such as Web search, the orderings, instead of the categories or scores, are of great concern. Under such circumstances, constructing application-oriented VisCom ranking functions would be useful.

ital.

ital.

Acknowledgment

The authors would like to thank the associate editor and anonymous reviewers for their insight comments. The authors also would like to thank Dr Qiulei Dong, Qishen Wang, Jun Gao, Xinyu Zou, Qiang You, and Xue Mao in Institute of Automation Chinese Academy of Sciences for their helpful suggestions. This work is partly supported by NSFC (Grant No. 60825204, 60672040, 61003115, 60935002), and the National 863 High-Tech Program of China (Grant No. 2012AA012504)

ful

pg. 1 notes

REFERENCES

see hardcopy & Word .doc file for new References

~~AHMAD, A.-R., BASIR, O., HASSANEIN, K., AND AZAM, S. 2008. An intelligent expert systems approach to layout decision analysis and design under uncertainty. *Studies in Computational Intelligence (SCI) 97*, 321–364.~~

~~AMAZON. 2005. Amazon's mechanical turk. https://www.mturk.com/mturk/welcome.~~

~~ANNETT, J. 2002. Subjective rating scales: science or art? *Ergonomics 45,* 14, 966–987.~~

~~BERLYNE, D. 1974. *Studies in the New Experimental Aesthetics*. Hemi-sphere Publishing.~~

~~BREIMAN. 2001. Random forests. *Machine Learning 45*, 5–32.~~

~~CAI, D., YU, S., WEN, J.-R., AND MA, W.-Y. 2003a. Extracting content structure for web pages based on visual representation. In *Proc. the 5th Asia-Pacific web conference on Web technologies and applications.* 406–417.~~

~~CAI, D., YU, S., WEN, J.-R., AND MA, W.-Y. 2003b. Vips: a vision-based page segmentation algorithm. *Microsoft Technical Report* MSR-TR-2003-79.~~

~~CAO, L. J., CHUA, K. S., AND CHONG, W. K. 2003. A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomputing 55*, 1–2, 321–336.~~

~~CHEN, G. AND CHOI, B. 2008. Web page genre classification. In *Proceedings of the 2008 ACM symposium on Applied computing*. 2353–2357.~~

~~CHENG, H. AND CANT-PAZ, E. 2010. Personalized click prediction in sponsored search. *Proc. ACM International Conference on Web Search and Data Mining*, 351–360.~~

~~CV. http://en.wikipedia.org/wiki/cross-validation_*statistics*.~~

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. *Proc. European Conference on Computer Vision*, 288–301.

DONDERI, D. C. 2006. Visual complexity: A review. *Psychological Bulletin 132*, 73–97.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2001. *Pattern Classification* 2nd Ed. John Wiley & Sons, USA.

DWORK, C., KUMAR, R., NAOR, M., AND SIVAKUMARC, D. 2001. Rank aggregation methods for the web. In *Proc. the 10th International Conference on World Wide Web*. ACM, 613–622.

FAWCETT, T. 2006. An introduction to roc analysis. *Pattern Recognition Letters. 27*, 861–874.

FLEISS, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*, 5, 378–382.

FORSYTHE, A. 2009. Visual complexity: Is that all there is? In *Proceedings of HCII, LNAI 5639*. 158–166.

FORSYTHE, A., SHEEHY, N., AND SAWEY, M. 2003. Measuring icon complexity: An automated analysis. *Behavior Research Methods, Instruments, & Computers 32*, 2, 334–342.

FRANC, V. AND SONNENBURG, S. 2008. Optimized cutting plane algorithm for support vector machines. In *Proc. International Conference on Machine Learning*. 320–327.

GEISSLER, G. L., ZINKHAN, G. M., AND WATSON, R. T. 2006. The influence of home page complexity on consumer attention, attitudes, and purchase intent. *Journal of Advertising 35*, 2, 69–80.

GERO, J. S. AND KAZAKOV, V. 2004. On measuring the visual complexity of 3d objects. *Journal of Design Sciences and Technology 12*, 1, 35–44.

GEUSEBROEK, J. AND SMEULDERS, A. 2005. A six-stimulus theory for stochastic texture. *International Journal of Computer Vision 62*, 1-2, 7–16.

HARPER, S., MICHAILIDOU, E., AND STEVENS, R. 2009. Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Trans. on Applied Perception 6*, 2, Artical 10.

HASLER, S. AND SUSSTRUNK, S. 2003. Measuring colorfulness in real images. In *Proc. SPIE Electron. Imag: Hum. Vision Electron*. 87–95.

JIANG, D., PEI, J., AND LI, H. 2010. Web search/browse log mining: challenges, methods, and applications. In *Proc. International World Wide Web Conference*. 1351–1352.

KIM, J. AND WILHELM, T. 2008. What is a complex graph? *Physica A 387*, 2637–2652.

KOHLSCHTTER, C. AND NEJDL, W. 2008. A densitometric approach to web page segmentation. In *Proc. ACM International Conference on Information and Knowledge Management(CIKM)*. 1173–1182.

LAM, F. C. AND LONGNECKER, M. T. 1983. A modified wilconxon rank sum test for paired data. *Biometrika 70*, 510–513.

LEVERING, R. AND CUTLER, M. 2009. Cost-sensitive feature extraction and selection in genre classification. *Journal for Language Technology and Computational Linguistics 24*, 2, 57–72.

LIU, B. 2007. *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer.

MICHAILIDOU, E. 2009. *Visual Complexity Rankings and Accessibility Metrics*. PhD thesis, University of Manchester.

MICHAILIDOU, E., HARPER, S., AND BECHHOFER, S. 2008. Visual complexity and aesthetic perception of web pages. In *Proc. ACM International Conference on Design of Communication (SIGDOC)*. 215–223.

MITCHELL, T. M. 1997. *Machine learning*. McGraw Hill.

NINASSI, A., MEUR, O. L., OLIVIER, P. L., AND BARBA, D. 2009. Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Journal of Selected Topics in Signal Processing 3*, 253–265.

PANDIR, M. AND KNIGHT, J. 2006. Homepage aesthetics: the search for preference factors and the challenges of subjectivity. *Interacting with Computers 18*, 1351–1370.

PAPACHRISTOS, E., TSELIOS, N., AND AVOURIS, T. 2006. Bayesian modelling of impact of colour on web credibility. In *Proc. European Conference on Artifical Intelligence*. 41–45.

PARK, S., CHOI, D., AND KIM, J. 2004. Critical factors for the aesthetic fidelity of web pages: empirical studies with professional web designers and users. *Interacting with Computers 16*, 351–376.

PEDRO, J. S. AND SIERSDORFER, S. 2009. Ranking and classifying attractiveness of photos in folksonomies. In *Proc. International World Wide Web Conference*. 771–780.

PIETERS, R., WEDEL, M., AND BATRA, R. 2010. The stopping power of advertising: measures and efects of visual complexity. *Journal of Marketing 74*, 48–60.

PITLER, E. AND NENKOVA, A. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proc. the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 186–195.

ROSENHOLTZ, R., LI, Y., AND NAKANO, L. 2007. Measuring visual clutter. *Journal of Vision 7*, 2, 1–22.

RUMELHART, D. E., HINTON, G. E., AND WILLIAMS, R. J. 1986. Learning representations by back-propagating errors. *Nature 323,* 6088, 533–536.

SCHAIK, R. AND LING, J. 1991. The effects of screen ratio and order on information retrieval in web pages. *IEEE Trans. Systems, Man and Cybernetics 21,* 3, 660–674.

SONG, G. 2007. Analysis of web page complexity through visual segmentation. In *Proc. HCII, LNAI 4553.* 114–123.

SONG, R., LIU, H., WEN, J.-R., AND MA, W.-Y. 2004. Learning block importance models for web pages. In *Proc. International World Wide Web Conference.* 203–211.

STICKEL, C., EBNER, M., AND HOLZINGER, A. 2010. The xaos metric - understanding visual complexity as measure of usability. In *Proc. the 6th Symposium (USAB 2010) of the Workgroup HCI&UE of the Austrian Computer Society.* 278–290.

THOMAS, C. AND TULLIS, S. 1998. A method for evaluating web page design concepts. In *Proc. International conference on Human Factors in Computing Systems (CHI '98).* 323–324.

TSOCHANTARIDIS, I., HOFMANN, T., JOACHIMS, T., AND ALTUN, Y. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proc. International Conference on Machine Learning.* 104–112.

TUCH, A. N., BARGAS-AVILA, J., OPWIS, K., AND WILHEM, F. 2009. Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *Int. J. Human-Computer Studies 67,* 703–715.

TUCH, A. N., KREIBIG, S., ROTH, S., BARGAS-AVILA, J., OPWIS, K., AND WILHEM, F. 2011. The role of visual complexity in affective reactions to web pages: Subjective, eye movement, and cardiovascular responses. *IEEE Transactions on Affective Computing,* 1–9.

VAPNIK, V. 1998. *Statistical Learning Theory.* Wiley.

WANG, M. AND HUA, X.-S. 2011. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology 2,* 2, Article 10.

WU, O., CHEN, Y., LI, B., AND HU, W. 2011. Evaluating the visual quality of web pages using a computational aesthetics approach. In *ACM International Conference on Web Search and Data Mining (WSDM).* 337–346.

ZADEH, L. A. 1965. Fuzzy sets. *Information and Control 8,* 338–353.

ZHENG, X. S., CHAKRABORTY, I., LIN, J. J.-W., AND RAUSCHENBERGER, R. 2008. Developing quantitative metrics to predict users' perceptions of interface design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES).* 2023–2027.

ZHENG, X. S., CHAKRABORTY, I., LIN, J. J.-W., AND RAUSCHENBERGER, R. 2009. Correlating low-level image statistics with users - rapid aesthetic and affective judgments of web pages. In *Proceedings of the 27th international conference on Human factors in computing systems (CHI).* 1–10.

Received Mar. 2011; revised May 2012; accepted Oct. 2012

REFERENCES

AHMAD, A. -R., BASIR, O., HASSANEIN, K., AND AZAM, S. 2008. An intelligent expert systems approach to layout decision analysis and design under uncertainty. *Stud. Comput. Intell. 97*, 321–364.

AMAZON. 2005. Amazon's mechanical turk. https://www.mturk.com/mturk/welcome

ANNETT, J. 2002. Subjective rating scales: Science or art? *Ergonomics 45*, 14, 966–987.

BERLYNE, D. 1974. *Studies in the New Experimental Aesthetics*. Hemi-sphere Publishing.

BREIMAN, L. 2001. Random forests. *Mach. Learn. 45*, 1, 5–32.

CAI, D., YU, S., WEN, J. -R., AND MA, W. -Y. 2003a. Extracting content structure for web pages based on visual representation. In *Proceedings of the 5th Asia-Pacific Web Conference on Web Technologies and Applications*. 406–417.

CAI, D., YU, S., WEN, J. -R., AND MA, W.-Y. 2003b. Vips: A vision-based page segmentation algorithm. Tech. rep. MSR-TR-2003-79. Microsoft.

CAO, L. J., CHUA , K. S., AND CHONG, W. K. 2003. A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomput. 55*, 1–2, 321–336.

CHEN, G. AND CHOI, B. 2008. Web page genre classification. In *Proceedings of the ACM Symposium on Applied Computing.* 2353–2357.

CHENG, H. AND CNT-PAZ , E. 2010. Personalized click prediction in sponsored search. In *Proceedings of the ACM International Conference on Web Search and Data Mining.* 351–360.

CV. 2013. http://en.wikipedia.org/wiki/cross-validation statistics

DATTA, R., JOSHI, D., LI , J., AND WANG, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the European Conference on Computer Vision.* 288–301.

DONDERI, D. C. 2006. Visual complexity: A review. *Psychol. Bull. 132*, 1, 73–97.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2001. *Pattern Classification*, 2nd ed. John Wiley & Sons.

DWORK, C., KUMAR, R., NAOR, M., AND SIVAKUMARC, D. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*. ACM, 613–622.

FAWCETT, T. 2006. An introduction to roc analysis. *Pattern Recogn. Lett. 27*, 861–874.

FLEISS, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull. 76*, 5, 378–382.

FORSYTHE, A. 2009. Visual complexity: Is that all there is? *In Proceedings of the 13th International Conference on Human-Computer Interaction.* Lecture Notes in Artificial Intelligence, vol. 5639, Springer, 158–166.

FORSYTHE, A., SHEEHY, N., AND SAWEY, M. 2003. Measuring icon complexity: An automated analysis. *Behav. Res. Methods Instrum. Comput. 32*, 2, 334–342.

FRANC, V. AND SONNENBURG, S. 2008. Optimized cutting plane algorithm for support vector machines. In *Proceedings of the International Conference on Machine Learning.* 320–327.

GEISSLER, G. L., ZINKHAN, G. M., AND WATSON, R. T. 2006. The influence of home page complexity on consumer attention, attitudes, and purchase intent. *J. Advertising 35*, 2, 69–80.

GERO, J. S. AND KAZAKOV, V. 2004. On measuring the visual complexity of 3d objects. *J. Des. Sci. Technol. 12*, 1, 35–44.

GEUSEBROEK, J. AND SMEULDERS, A. 2005. A six-stimulus theory for stochastic texture. *Int. J. Comput. Vision 62*, 1-2, 7–16.

HARPER, S., MICHAILIDOU, E., AND STEVENS, R. 2009. Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Trans. Appl. Percept. 6*, 2, Artical 10.

HASLER, S. AND SUSSTRUNK, S. 2003. Measuring colorfulness in real images. *Proc. SPIE Electron. Imag:Hum. Vision Electron.* 87–95.

JIANG, D., PEI, J., AND LI, H. 2010. Web search/browse log mining: Challenges, methods, and applications. In *Proceedings of the International World Wide Web Conference*. 1351–1352.

KIM, J. AND WILHELM, T. 2008. What is a complex graph? *Phys. A 387*, 2637–2652.

KOHLSCHTTER, C. AND NEJDL, W. 2008. A densitometric approach to web page segmentation. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM).* 1173–1182.

LAM, F. C. AND LONGNECKER, M. T. 1983. A modified wilconxon rank sum test for paired data. *Biometrika 70*, 510–513.

LEVERING, R. AND CUTLER, M. 2009. Cost-Sensitive feature extraction and selection in genre classification. *J. Lang. Technol. Comput. Linguistics 24*, 2, 57–72.

LIU, B. 2007. *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer.

MICHAILIDOU, E. 2009. Visual complexity rankings and accessibility metrics. Ph.D. thesis, University of Manchester.

MICHAILIDOU, E., HARPER, S., AND BECHHOFER, S. 2008. Visual complexity and aesthetic perception of web pages. In *Proceedings of the ACM International Conference on Design of Communication (SIGDOC).* 215–223.

MITCHELL, T. M. 1997. *Machine Learning.* McGraw Hill.

NINASSI, A., MEUR, O. L., OLIVIER, P. L., AND BARBA, D. 2009. Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE J. Sel. Top. Sign. Proces. 3,* 2, 253–265.

PANDIR, M. AND KNIGHT, J. 2006. Homepage aesthetics: The search for preference factors and the challenges of subjectivity. *Interact. Comput. 18,* 6, 1351–1370.

PAPACHRISTOS, E., TSELIOS, N., AND AVOURIS, T. 2006. Bayesian modelling of impact of colour on web credibility. In *Proceedings of the European Conference on Artifical Intelligence.* 41–45.

PARK, S., CHOI, D., AND KIM, J. 2004. Critical factors for the aesthetic fidelity of web pages: Empirical studies with professional web designers and users. *Interact. Comput. 16,* 351–376.

PEDRO, J. S. AND SIERSDORFER, S. 2009. Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the International World Wide Web Conference.* 771–780.

PIETERS, R., WEDEL, M., AND BATRA, R. 2010. The stopping power of advertising: Measures and effects of visual complexity. *J. Market. 74,* 5, 48–60.

PITLER, E. AND NENKOVA, A. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* 186–195.

ROSENHOLTZ, R., LI, Y., AND NAKANO, L. 2007. Measuring visual clutter. *J.* ~~*Vis.*~~ Vision *7,* 2, 1–22.

RUMELHART, D. E., HINTON, G. E., AND WILLIAMS, R. J. 1986. Learning representations by back-propagating errors. *Nature 323,* 6088, 533–536.

SCHAIK, R. AND LING, J. 1991. The effects of screen ratio and order on information retrieval in web pages. *IEEE Trans. Syst. Man Cybern. 21,* 3, 660–674.

SONG, G. 2007. Analysis of web page complexity through visual segmentation. In *Proceedings of the 12th International Conference on Human-Computer Interaction.* Lecture Notes in Artificial Intelligence, vol. 4553, Springer, 114–123.

SONG, R., LIU, H., WEN, J. -R., AND MA, W. -Y. 2004. Learning block importance models for web pages. In *Proceedings of the International World Wide Web Conference.* 203–211.

SICKEL, C., EBNER, M., AND HOLZINGER, A. 2010. The xaos metric - Understanding visual complexity as measure of usability. In *Proceedings of the 6th Symposium of the Workgroup HCI & UE of the Austrian Computer Society (USAB '10).* 278–290.

THOMAS, C. AND TULLIS, S. 1998. A method for evaluating web page design concepts. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '98).* 323–324.

TSOCHANTARIDIS, I., HOFMANN, T., JOACHIMS, T., AND ALTUN, Y. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine Learning.* 104–112.

TUCH, A. N., BARGAS-AVILA, J., OPWIS, K., AND WILHEM, F. 2009. Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *Int. J. Hum. Comput. Stud. 67,* 703–715.

TUCH, A. N., KREIBIG, S., ROTH, S., BARGAS-AVILA, J., OPWIS, K., AND WILHEM, F. H. 2011. The role of visual complexity in affective reactions to web pages: Subjective, eye movement, and cardiovascular responses. *IEEE Trans. Affective Comput. 2,* 4, 230-236.

VAPNIK, V. 1998. *Statistical Learning Theory.* Wiley.

WANG, M. AND HUA, X.-S. 2011. Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell. Syst. Technol. 2,* 2, Article 10.

WU, O., CHEN, Y., LI, B., AND HU, W. 2011. Evaluating the visual quality of web pages using a computational aesthetics approach. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM).* 337–346.

ZADEH, L. A. 1965. Fuzzy sets. *Inf. Control 8,* 338–353.

ZHENG, X. S., CHAKRABORTY, I., LIN, J. J. -W., AND RAUSCHENBERGER, R. 2008. Developing quantitative metrics to predict users' perceptions of interface design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES).* 2023–2027.

ZHENG, X. S., CHAKRABORTY, I., LIN, J. J. -W., AND RAUSCHENBERGER, R. 2009. Correlating low-level image statistics with users- rapid aesthetic and affective judgments of web pages. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI).* 1–10.

Received February 2011; revised October 2012; accepted October 2012