Manual and Automated Labeling of Web User Interfaces for User Behavior Models

Anna Stepanova [a], Maxim Bakaev [a]

[a] *Novosibirsk State Technical University, Novosibirsk, 630073, Russia*

Abstract

The article contrasts manual and automated identification of elements in images of web user interfaces (UIs), which is essential for machine learning (ML) models that describe user behavior. We consider the principal advantages and disadvantages of the two methods and compare linear regression models. The constructed ML models describe users' subjective perception of web UIs in such dimensions as complexity, aesthetics and ordering. Somehow unexpectedly, the resulting R^2s of models built with certain factors obtained from automated labeling turned out to be slightly higher. Particularly, shares of text and images in the web UI, as well as the sizes of the elements, were rather influential. We believe that the main disadvantage of the manual labeling is the human factor, as mistakes made by the labelers and diversity of their outcome affect the quality of the models. In turn, the automated process has a number of drawbacks that must be taken into account and that we discuss in the paper. The results of our work might be of interest to both ML researchers and to usability engineers who seek to improve the subjective satisfaction of users with websites.

Keywords [1]

Image labeling, human-computer interfaces, machine learning, linear regression

1. Introduction

Any design object needs effective presentation, in which structuring of textual and visual information is highly important [1]. Many researchers and designers have been looking for the principles of harmonious organization of compositional elements in architecture and website design. For instance, visual appearance of web user interfaces (UIs) is known to affect behavior of users, and its analysis can help to improve usability and thus increase KPIs of the website, such as e.g. conversion rate [2]. The visual complexity assessment helps to identify and describe problems in the website UI. Visual complexity is affected by the number of elements in an object or image, their structural relations, the detail of the information that these elements provide, etc. [3]. It has been scientifically proven that aesthetic preferences for the visual complexity of web pages are influenced by users' age and previous experience [4]. However, our article focuses on the dependence of visual complexity in web UI screenshots: namely the common compositional elements in web pages (buttons, texts, lists, etc.), as well as multimedia elements (images, videos, etc.).

The identification of UI elements in a website page screenshot for further assessment of visual complexity can be obtained through either manual labeling or automated recognition process [5]. Automation of any process makes it possible to simplify it and helps to free a person from routine and tedious tasks, but often it involves additional costs and resources (time, labor, etc.), especially at the initial stage. Table 1 shows a comparison of the automatic and manual methods with respect to UI labeling.

Thus, the purpose of the current work is to determine the types of elements that affect the subjective perception of websites, as well as to compare the models built with the factors' values obtained via automated vs. manual labeling of web UI screenshots.

YRID-2020: International Workshop on Data Mining and Knowledge Engineering, October 15-16, 2020, Stavropol, Russia
EMAIL: s.nuta97@mail.ru (Anna Stepanova); bakaev@corp.nstu.ru (Maxim Bakaev)
ORCID: 0000-0003-0880-8760 (Anna Stepanova); 0000-0002-1889-0692 (Maxim Bakaev)

Table 1
Advantages and disadvantages of automated and manual methods

The labeling method	Advantages	Disadvantages
Manual	1) Higher accuracy in determining a UI element's type 2) More robust list of various UI elements' types can be used	1) It takes a large number of people to conduct the analysis 2) Human factor: grammatical errors, incorrect definition of the type of element, inconsistency in understanding, etc., which can distort the final result
Automated	1) The analysis process takes less time 2) No need to involve paid labelers	1) Image recognition is still computationally expensive and inaccurate in some aspects 2) The cost of implementing the code and its debugging 3) The need for training data, e.g. for detecting the UI elements' types

2. The Study Description

2.1. The Manual Labeling

In the experiment, the subjects were offered about 500 website interface screenshots and asked to label UI elements in them: highlight the element in a box and identify the elements' type. They were using a dedicated software tool, LabelImg (see in Fig. 1). In total, 11 human labelers took part in this activity, after providing informed consent.

Figure 1: Manual UI labeling with LabelImg software tool (the selection of UI type is at the bottom)

2.2. The Automated Labeling

In addition to the manual labeling of the screenshots, we also performed their automated analysis, using our dedicated Visual Analyzer (VA) software tool, available at http://va.wuikb.info/ and described in detail in [6]. It identifies UI elements in images based on previously trained ML models (see in Fig. 2), but our previous studies suggest that its accuracy is somehow deficient, particularly in determining the type of each UI element.

Figure 2: Automated UI labeling with our Visual Analyzer tool (an example)

In Table 2 we show the types of UI elements that had been identified in manual and automated labeling that we performed. As one can seen from it, the main difference between the two methods lies in determining the types of elements. In the manual labeling, the participants were able to rather successfully identify 26 different UI types, while for the automated labeling, there were 8 types, resulting from the pre-trained ML models. In addition, due to high visual diversity of today's web designs, the accuracy of the automated type detection was the most problematic dimension in the semantic-spatial analysis that our VA tool performed (see [6] for more detail).

From the labeling datasets, the heights and widths of the elements were calculated (also by the elements' types: text paragraphs, buttons, images, panels, etc.), and there was also the data on the height and width of the screenshot. Based on this information, areas of each UI he element was calculated, as well as the share occupied by this element in the total screenshot space (by the elements' types: texts, images, background, etc.).

Thus, there were 2 groups of factors in the behavior models that we further constructed for complexity, aesthetics and orderliness that were the dependent variables in the study:

1) the number of elements: the number of elements of a certain type located in one screenshot;
2) the proportions: the shares of the total area of the elements' types to the total area of the screenshot.

Table 2
The UI elements' types used in the manual and the automated labeling of web UIs

Element type	Manual labeling	Automated labeling
Button	+	+
Text	+	+
Checkbox	+	+
Radio button	+	+
Dropdown	+	+
Label	+	+
Image	+	+
Link	+	+
Background image	+	-
Panel	+	-
Tabs	+	-
Paragraph	+	-
Imagelink	+	-
Textinput	+	-
Textblock	+	-
Buttonimage	+	-
Headline	+	-
Image_background	+	-
Symbol	+	-
Back	+	-
Selectbox	+	-
Scroll bar	+	-
Date	+	-
List	+	-
Pagination	+	-
Table	+	-

2.3. The Subjective Perception Evaluation

For each screenshot, we also had evaluations of complexity, orderliness and aesthetics, provided by another 137 participants (67 female, 70 male). The majority of them were Russians (89.1%), while the rest were from Bulgaria, Germany, South Africa, etc. More details on the participants and the procedure can be found in one of our previous works [7].

Each of the subjective perception dimensions was assessed on a scale from 1 to 10. Then the average values for each indicator were found using formulas (1)-(3), where i is the number of web pages, n_j is the number of participants who provided the evaluations for the i-th website.

$$y_{aest\ i} = \frac{\sum_1^n y_{aest}}{n_i}, \quad (1)$$

where $y_{aest\ i}$ – indicator of the aesthetics of the interface of the i-th web page, y_{aest} – estimation of the aesthetics of the interface of the i-th web page

$$y_{comp\ i} = \frac{\sum_1^n y_{comp}}{n_i}, \quad (2)$$

where $y_{comp\ i}$ – indicator of the complexity of the interface of the i-th web page, y_{comp} – estimation of the complexity of the interface of the i-th web page

$$y_{ord\ i} = \frac{\sum_1^n y_{ord}}{n_i}, \quad (3)$$

where $y_{ord\ i}$ – indicator of the ordering of the interface of the i-th web page, y_{ord} – estimation of the ordering of the interface of the i-th web page.

Complexity is the number of elements on the screen and their arrangement. Orderliness is an ordered data structure of the web interface that allows the user to easily find the information they need. Aesthetics is an assessment of the attractiveness of a product by the user. This indicator is important because the aesthetics of any web interface has a strong impact on the user, even when he tries to evaluate the functionality of the system [8].

3. Results

Using the SPSS statistical analysis software, we build linear regression user behavior models (as representing the most universal ML method) with the factors resulting from the manual and the automated labeling. The dependent variables in the models were the three subjective perception evaluation scores, while the independent variables were the factors resulting from the two labeling processes. The step-by-step method of the regression analysis allowed stepwise inclusion of the factors in the models, thereby discarding those that did not make a significant contribution to explaining the dependent variables.

The results of the regression analysis are presented in Tables 3-5, each corresponding to a different subjective perception dimension. The null hypothesis H_0 was as usual in the regression analysis, that the regression equation is not significant. For each of the dependent variable, the models turned out to be significant ($p < 0.05$), while the significances for the factors selected by the step-by-step method are shown in the respective columns of the tables.

Table 3

The regression analysis results for the *complexity* dependent variable

Method type	Factor	R^2	Non-standard coef.	Standard coef.	t	Significance
Automated	Count of labels	0.092	0.001	0.218	4.576	0.000
	Count of text	0.127	0.012	0.195	4.131	0.000
	Count of radio buttons	0.136	0.039	0.096	2.137	0.033
Manual	Number of elements on the page	0.062	0.001	0.172	3.427	0.001
	Share of images	0.082	0.005	0.153	3.156	0.002
	Share of panels	0.095	0.002	-0.118	-2.699	0.007
	Share of buttons	0.107	0.017	0.119	2.790	0.005
	Share of paragraphs	0.114	0.007	0.089	2.042	0.042
	Share of tabs	0.121	0.032	0.087	2.025	0.043

As one can note from Table 3, the perceived complexity of websites was most influenced by the amount of text, the number of radio buttons, the number of buttons and tabs. The subjective orderliness (Table 4) was most influenced by the proportion of radio buttons and buttons. The perception of aesthetics (Table 5) was most influenced by the proportion and amount of text on the page, the number of buttons, labels and images.

Table 4
The regression analysis results for the *orderliness* dependent variable

Method type	Factor	R^2	Non-standard coef.	Standard coef.	t	Significance
Automated	Share of text	0.027	1.221	-0.162	-3.604	0.000
	Screenshot width	0.048	0.000	.253	4.793	0.000
	Count of labels	0.080	0.001	-0.199	-3.763	0.000
	Count of buttons	0.096	0.005	-0.123	-2.722	0.007
	Share of radio buttons	0.107	14.489	-0.104	-2.305	0.022
Manual	Share of text	0.024	0.004	-0.179	-4.104	0.000
	Share of background images	0.051	0.001	0.138	3.115	0.002
	Share of links	0.067	0.011	-0.125	-2.828	0.005
	Share of "back" arrows	0.077	0.057	0.103	2.354	0.019

Table 5
The regression analysis results for the *aesthetics* dependent variable

Method type	Factor	R^2	Non-standard coef.	Standard coef.	t	Significance
Automated	Screenshot width	0.072	0.000	0.360	7.042	0.000
	Share of text	0.119	1.549	-0.207	-4.749	0.000
	Count of button	0.144	0.007	-0.147	-3.373	0.001
	Count of label	0.161	0.002	-0.155	-3.032	0.003
Manual	Share of background images	0.046	0.002	0.224	4.933	0.000
	Share of text	0.079	0.007	-0.281	-5.178	0.000
	Share of links	0.099	0.014	-0.123	-2.823	0.005
	Count of text	0.114	0.014	0.152	2.786	0.006
	Count of images	0.122	0.003	0.095	2.114	0.035

Despite the fact that the constructed models had rather low determination coefficients (R^2), it is still possible to draw a general conclusion about which elements affect the assessment of website perception among users, and whether the degree of this dependence is influenced by the labeling method (Table 6).

As one can see from Table 6, the share of text on the page was significant for all the three subjective perception dimensions. Moreover, the presence of the text had the greatest impact on the assessment of complexity and aesthetics. At the same time, the obtained values for the automated and the manual labeling did not differ much (4-10%), but with the automated labeling, the models' quality indexes, as represented by R^2s, were somehow superior.

Table 6
Comparison of the Regression Analysis Values

Dependent variable	Factor	R^2 (Automated method)	R^2 (Manual method)
Complexity	Count of text	0.127	0.114
Ordering	Share of text	0.027	0.024
Aesthetics	Share of text	0.119	0.114

To improve the quality of the constructed user behavior models, we tried reducing the number of factors and took only the common ones for the two methods, namely:

1. The total number of elements in the interface
2. The number of types of elements on the page
3. Percentage of space under the text
4. Percentage of space under images
5. Average height of elements
6. Average width of elements
7.

The average size was taken into account for elements such as labels, buttons, test blocks and images. For text and images, a size restriction has been imposed: no more than 300 px in height and no more than 1000 px in width. The results of the regression analysis for the three dimensions of the subjective perception are shown in Table 7.

Table 7
Pivot table of regression analysis for a limited number of factors

Dependent variable	Factor	R^2 (Automatic method)	R^2 (Manual method)
Complexity	Number of elements on the page	0.071	0.062
	Average height of text	0.091	0.083
Orderliness	Average height of text	-	0.018
	Average width of buttons	-	0.033
	Average height of images	-	0.046
	Average width of labels	-	0.054
	Average height of labels	0.044	-
Aesthetics	Average height of button	-	0.054
	Average height of text	0.071	0.088
	Number of elements on page	0.057	-
	Average height of label	0.041	-

4. Conclusion

The study showed that the subjective assessment of the website perception is influenced by the amount of text on the page and the share of images: the more text on the page, the more complex and less aesthetic the website appears. However, this statement is not entirely correct, since in addition to

the share of the text area, the text style (font, size, etc.), and the presence of pictures that dilute the text, and many other factors are also important. The number of images, their size, quality and position on the web page affect the overall website subjective perception, including the assessment of the aesthetics and orderliness. As a general rule, these indicators are interchangeable: the lack of text is often compensated by a variety of graphic elements and images. Therefore, in order for the web UI to comply with usability standards and to be simpler and more understandable for its users, it is necessary not to clutter the interface with a large number of elements and break long texts into smaller parts. At the same time, the number of element types did not significantly affect any of the subjective perception dimensions.

The quality of the user behavior models built on the factors resulting from the automated labeling was slightly higher than for the manual one, which suggests feasibility of our UI visual analysis tool. In general, this may indicate that the automation of the labeling process makes sense, but it requires high costs and the presence of certain knowledge to implement it. After all, our VA software was initially trained with the data once provided by human labelers [6].

It should be noted that the constructed linear models have low R^2 coefficients, so they should not be used in production. The goal of our current study was merely to compare the two labeling methods, while for real user behavior modeling, more advanced ML methods and architectures should be used.

Our plans for further research include investigation of the effects of web page layouts (the size of the element, its type and occupied area) on the subjective assessment of the perception of the site, but also the colors, fonts, types of buttons, animations, etc.

5. Acknowledgment

The reported study was funded by RFBR according to the research project No. 19-29-01017.

6. References

[1] S.P. Rassadina, O.V. Ivanova. Assessment of visual perception of information design objects. Proc. XVI Int Conf on Cultural studies, philology, art history: urgent problems of modern science, 30-35 (2018). – In Russian.

[2] P.V. Pesterev, Influence of behavioral factors of ranking on the position of sites in search results. Bulletin of the Belgorod State Technological University, 2, 219-221 (2017). – In Russian.

[3] S. Kusumasondjaja, F. Tjiptono, Endorsement and visual complexity in food advertising on Instagram. Internet Research, 29 (4), 659-687 (2019).

[4] H.F. Wang, C.H. Lin, An investigation into visual complexity and aesthetic preference to facilitate the creation of more appropriate learning analytics systems for children. Computers in Human Behavior, 92, 706-715 (2019).

[5] M.V. Tsyplyaev, N.A. Vinokurov, System and method for selecting significant page elements with implicit indication of coordinates for identifying and viewing relevant information. Patent, RU 2708790 C2 (2019).

[6] M. Bakaev, S. Heil, V. Khvorostov, M. Gaedke, Auto-extraction and integration of metrics for web user interfaces. Journal of Web Engineering, 17(6&7), 561-590 (2018).

[7] M. Bakaev, M. Speicher, S. Heil, M. Gaedke, I Don't Have That Much Data! Reusing User Behavior Models for Websites from Different Domains. Proc. 20th International Conference on Web Engineering (ICWE 2020), 146-162 (2020). Springer, Cham.

[8] S. Triberti, A. Chirico, G. La Rocca, G. Riva, Developing emotional design: Emotions as cognitive processes and their role in the design of interactive technologies. Frontiers in psychology, 8, 1773 (2017).