Estimating Gender and Age of Web Page Visitors from the Way They Use Their Mouse

Krátky Peter, Chudá Daniela
Faculty of Informatics and Information Technologies
Slovak University of Technology
Ilkovičova 5, 842 16 Bratislava, Slovakia
{peter.kratky, daniela.chuda}@stuba.sk

ABSTRACT

Biometric data are affected by physiological properties of people, including gender or age. The paper describes an experiment of discovering gender and age in computer mouse movement data that might be notably beneficial for profiling anonymous visitors browsing the Web. The proposed method extracts features, such as velocity, path straightness or pauses duration, that are used by a multiclassifier system to make an estimate. Age category estimation shows encouraging results of the early method, especially for statistical analysis of a website audience.

Keywords

soft biometrics; user modeling; gender estimation; age estimation; mouse movement features

1. INTRODUCTION

In the field of biometrics, the simplest and most natural way of describing individuals is using *soft biometric traits* - evident features, such as gender, height, age, etc. Such physiological properties are also hidden in automatically and more sophistically extracted features from, for example, photographs, voice records or fingerprints. Research in soft biometrics is quite rare for biometric data acquired from standard input devices (keyboard, mouse), but could have a great impact as those are commonly used by website visitors to navigate through pages. Estimating gender or age from mouse data could be utilized for content recommendation to a new user whose profile has not been created. Another application could be statistical analysis of visitors and audience interested in particular pages for the owner of the website.

In our work, we propose a novel approach to profile gender and age group of website visitors based on mouse movement data. It shows that the performance of the method as a standalone solution for instant personalization is rather low, having precision 60% and 67% of determining gender and age group respectively (both binary classification problems) after browsing 10 web pages. On the other hand, we stress the possibility of estimating the dominant age group of website audience with accuracy 90%.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889384.

2. RELATED WORK

Extraction of soft biometrics has been studied mostly with purpose of filtering large databases when searching for identity or improving biometric systems when additional inferred features could boost precision of authentication. Soft biometric traits were discovered in various types of data often used in identity verification systems, such as face photography, fingerprint, iris or gait data [2].

In addition to these biometric data sources, there are standard input devices which become popular in research as no additional hardware is needed. Fairhust and Da Costa-Abreu [3] used keystroke dynamics data (timings of typing on a keyboard) to predict gender. Various classifiers (k-Nearest Neighbors, Naive Bayes, Decision Trees) reached precision above 68% and even better results (over 90%) were achieved using their combination into a multiclassifier system. Similar research on keystroke data was conducted by Idrus et al. [4]. Not only estimation of gender, but also age category, handedness and typing with one/two hands were studied, all with encouraging results. As the papers are aimed at improving security, the datasets contain repeated typing of passphrases. Our work is aimed at navigating on the Web, thus we focus on mouse data of free movements.

3. METHOD

The proposed method for estimating category of a user (gender, age) works in three stages. First, mouse data are recorded including time, event type (movement/click), x and y coordinates of the cursor and page where it occurred.

Then, a mouse movement instance represented by several features is created for each visited URL by the user. The features are calculated based on distance, angle and time duration between consecutive points. Basic features are mean and standard deviation of *velocity*, *curvature*, *angular velocity* characterizing movement within a single visited page. In addition to these features, we calculated total *duration* and *length*, number of *inflection points*, *straightness* property, *perpendicular* distance (deviation from the shortest path), *jitter* (ratio of smoothed and real path), number of *curves*, *clicks* count, *duration of button pressed*, *pauses* counts, *duration of pauses* when interacting with the page.

Finally, a trained classifier determines the category based on movement instances from multiple browsed pages by a user with majority voting scheme applied. For each classifi-

cation problem, multiple classifiers are trained and combined using sum fusion scheme. The classifiers operate only on a subset of features relevant for distinguishing the categories.

Apart from well-known techniques such as k-Nearest Neighbors, Logistic Regression, Logical Model Tree and Support Vector Machines, we also adopted a method based on distributions comparison we previously used for recognizing individuals [1]. The training data are divided into several model groups each containing data exclusively for one of the categories. When estimating category of a user, measured data are compared to the model groups using Kolmogorov-Smirnov test (suitable for asymmetric and skewed data). The k groups with the lowest K-S statistics determine the resulting class.

4. EXPERIMENTS AND RESULTS

To evaluate the method we conducted an uncontrolled experiment in a large e-shop. We obtained anonymized mouse data from a sample of browsing desktop/laptop users with provided gender and age during a period of 6 days. The dataset contains 42411 browsed pages in total from 1494 users within the e-shop domain, and a median visitor viewed 18 pages. To collect data, we implemented a logging (JavaScript) module that records mouse events (with uneven sampling rate) and sends them to server backend.

We divided the dataset into the training set containing logs from 4 days and the testing set with logs from the other 2 days. Only the users with at least 15 browsed pages (62% of all visitors) were chosen for training. To make the set balanced, an equal count of males and females was ensured (318 users together) and exactly 15 viewed pages for each user were kept. Similarly, the testing set contained other 152 users, males and females equally present. The two sets constituted of disjoint sets of users.

As for estimating gender, various classifiers were trained and the best one reached F-score 0.54 obtained by exploiting movement data from a single page (one movement instance). Figure 1a shows performance of a multiclassifier system combining 4 classifiers (each with F-score over 0.5). We compared precision for different number of visited pages, running the evaluation 10 times while randomly selecting user data to approximate results. At least nine visited pages were necessary so that the gender category estimation was as good as 0.6.

Similarly, the evaluation tests were run for age category estimation (Figure 1b). Two age categories were created, specifically, below and above median - 33 years old. A single classifier was able to estimate age from a single movement instance with F-score 0.6, while multiclassifier system reached 0.67 based on 10 visited pages.

In the second part of the study, we experimented with estimation of what is the major category viewing the website. We used the same classifiers as in the previous part, but this time the test set contained all users having at least 10 pages viewed (more than 86% of all visitors), specifically 260 visitors. We randomly chose 61 users, performed the classification and consequently determined the major category. Over a large number of runs we counted the number of cases where the majority category was successfully identified. The majority of males/females visitors was determined with 70% success rate and younger/older much better - 90% success rate.

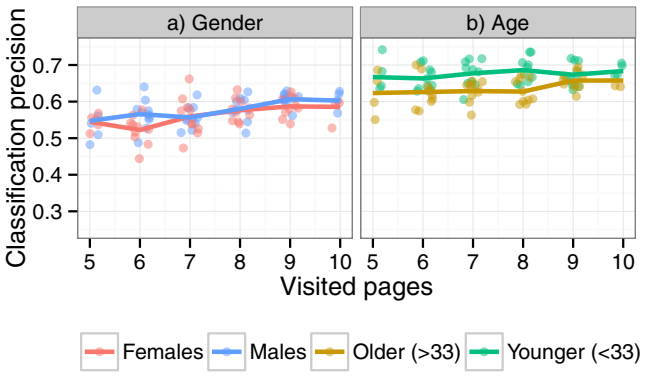

Figure 1: Performance of gender and age classification of individual visitors.

The obtained results might be biased as the correctness of user categories (gender, age) in the anonymized dataset could not have been checked as well diversity of hardware was present. Since the dataset was obtained from the one specific domain, the robustness of the method needs to be further examined in heterogeneous environment. Due to different types (news, e-shop, etc.) and layouts of websites the classifiers might be retrained for a specific target.

5. CONCLUSIONS

In the paper we presented a new approach to estimate gender and age category of the web visitors based on mouse movement characteristics. The proposed method for profiling individuals does not provide sufficient precision, but it could be used as an additional component for improving user modeling in personalized systems. Also, utilizing the method for statistical analysis of website audience shows promising results.

6. ACKNOWLEDGMENTS

This work was partially supported by the grant No. VG 1/0646/15.

Special thanks to Ján Sukeník and Tomáš Ulej, Martinus.sk, Slovakia, for building the dataset.

7. REFERENCES

[1] D. Chudá and P. Krátky. Grouping Instances in kNN for Classification Based on Computer Mouse Features. In *Proceedings of the 16th International Conference on Computer Systems and Technologies*, pages 214–220, 2015.

[2] A. Dantcheva, C. Velardo, A. D'Angelo, and J.-L. Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011.

[3] M. Fairhurst and M. Da Costa-Abreu. Using keystroke dynamics for gender identification in social network environment. In *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, pages 1–6, London, 2011. IET.

[4] S. Z. S. Idrus, E. Cherrier, C. Rosenberger, and P. Bours. Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords. *Computers & Security*, 45:147–155, 2014.