J Inf Process Syst, Vol.19, No.5, pp.602~613, October 2023
https://doi.org/10.3745/JIPS.04.0288

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Brief Survey into the Field of Automatic Image Dataset Generation through Web Scraping and Query Expansion

Bart Dikmans and Dongwann Kang*

Abstract

High-quality image datasets are in high demand for various applications. With many online sources providing manually collected datasets, a persisting challenge is to fully automate the dataset collection process. In this study, we surveyed an automatic image dataset generation field through analyzing a collection of existing studies. Moreover, we examined fields that are closely related to automated dataset generation, such as query expansion, web scraping, and dataset quality. We assess how both noise and regional search engine differences can be addressed using an automated search query expansion focused on hypernyms, allowing for user-specific manual query expansion. Combining these aspects provides an outline of how a modern web scraping application can produce large-scale image datasets.

Keywords

Image Dataset Generation, Query Expansion, Web Scraping

1. Introduction

In the information technology (IT) field, there is a consistent need for high-quality up-to-date image datasets. For machine vision or smaller student projects, an image dataset of sufficient quality is required. There are three primary methods for generating such datasets, namely, manual, semi-automatic, and automatic methods. Manual dataset generation is the most time-consuming but also the most effective method; every individual image is guaranteed to be of high quality. This is because the creator of the dataset can determine whether the image meets the selection criteria. An automatically generated dataset uses a set of algorithms to collect images. The most common approach is using a web scraper to obtain images of interest via a search engine. The semi-automatic process is characterized by two approaches. The most modern approach involves manually creating a small high-quality dataset, and then augmenting through the automatic method to gather a larger dataset. The other approach entails using an automatic algorithm to collect images and going through them to handpick the appropriate high-quality images. As increased optimization becomes more commonly required in practice, the need for new automatic methods will increase. This can be observed in the literature, such as, the commercially available study conducted by Rosebrok [1].

Simple web scraping methods proposed by Thomas and Mathur [2] and Glez-Pena et al. [3] provide

Manuscript received April 22, 2022; first revision December 19, 2022; accepted January 1, 2023.
* **Corresponding Author:** Dongwann Kang (dongwann@seoultech.ac.kr)
Dept. of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, Korea (bartjedi68@gmail.com, dongwann@seoultech.ac.kr)

an easy approach for automatically collecting an expansive image dataset based on any possible query. However, the precision of these methods is typically insufficient for large-scale use, as observed by by Schroff et al. [4]. This study proposes a fully automated image dataset collector that includes image ranking to avoid selecting inadequate images. However, the solution proposed by Schroff et al. [4] is constrained by certain search limits, yielding low accuracy with multiple tests. Over the next sections, we discuss web scraping search engines, query expansion, dataset quality, and possible issues that occur during web scraping. Finally, we will recommend creation guidelines for a fully automated image dataset generator.

2. Overview of Web Scraping-based Image Dataset Generation

2.1 Automated Dataset Generation

To achieve automatic dataset generation, two approaches have been proposed, namely, full-automatic and semi-automatic. Schroff et al. [4], suggested a completely automated system that combined an initially obtained result from a query through search engines and re-ranked them based on text/metadata surrounding these images. In their research, they compared the precision of Google images (32%), with that of their own fully automated system. Yao et al. [5] expanded on the study conducted by Schroff et al., adding query expansion by using Google Books Ngram Corpus (GBNC) [6]. Accordingly, not only the initially given query was searched but also its slight variations were considered. The query expansion used by Yao et al. [5] was primarily focused on using adjectives. For example, if the search query was “Zebra,” through GBNC this query would be expanded to “Young Zebra” or “Wild Zebra,” to receive more specific results. A prime example of semi-automatic dataset generation is mentioned in the research conducted by Zink [7]. In that research, initially, a smaller high-quality dataset was manually constructed, which was later used as a control set to verify the images obtained through a web scraper. Through their research, they demonstrated that the accuracy of semi-automatic dataset generation had the potential to reach that of professionally manually collected datasets. Another conclusion from their research is that when the number of images scraped increased, the accuracy also increased.

2.2 Query Expansion

Recently, there have been multiple developments in the field of semantic relations between words. The oldest approach, which is still frequently used, is WordNet [8]. This is a large English lexicon that contains many links between words, one of its most recent developments is Word2Vec. Word2Vec uses a Corpus, such as the earlier mentioned GBNC, to generate vectors that indicate relations between words [9]. As there are many similarities in their use, the comparative study conducted by Handler [10] underlines two differences between WordNet and Word2Vec. Firstly, Word2Vec can handle a much wider array of words when trained with the GBNC; however, WordNet contains a lower number of samples (100 billion compared to 116,000 examples). Secondly, Word2Vec is particularly strong at detecting, holonyms, meronyms, and hypernyms. An example of the data obtained by Word2Vec is presented in Fig. 1.

There is of course also the manual option of query expansion. With manual query expansion, the user

provides some extra terms that can be used for improving the primary query. Yu et al. [11] demonstrated that having humans assist in qualifying an image increases the quality of the later-trained dataset generated. This same principle could be applied to having a user modify an initial query before querying images obtained from the Internet.

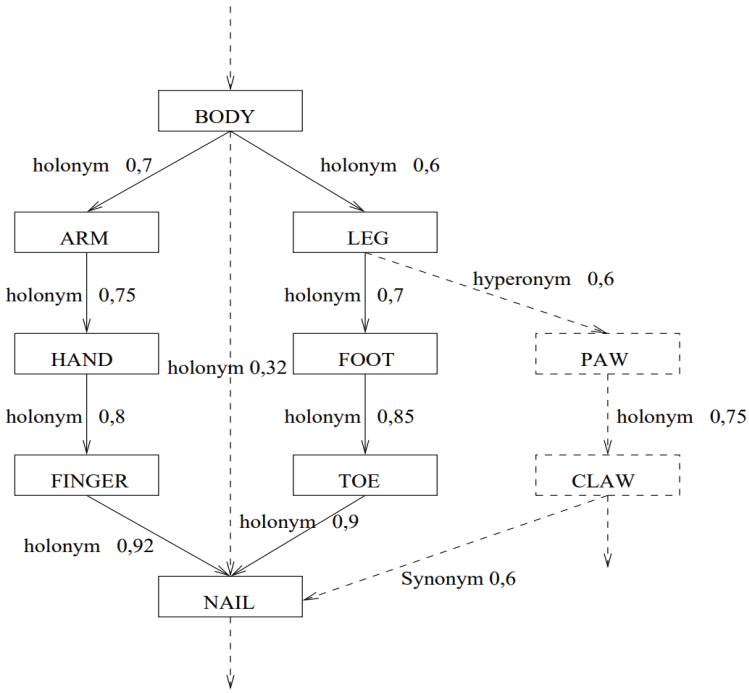

Fig. 1. An example Word2Vec datum demonstrating holonymy, synonymy, hypernymy, and meronomy constructed by Schwab and Lafourcade [12].

2.3 Web Scraping

Zhao [13] perfectly described the definition of web scraping as, "a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis." They also mentioned the related controversies, such as causing copyright infringement and potential distributed denial-of-service (DDOS) attacks via web scraping if not executed correctly. Based on this, a large collection of web scrapers has been created for varying purposes using different techniques. Sirisuriya [14] described nine different techniques that can be used for web scraping. The most popular method is using some form of web scraping program to obtain the desired dataset. Regarding these web scraping applications, there are many different options based on the desired dataset characteristics. As it stands, there seem to be at least over 20 different commonly used and readily available web scraping software versions. Comparative studies have been conducted regarding the performance and targets of these applications. Particularly Glez-Pena et al. [3] and Sirisuriya [14] conducted thorough research on the topic. Besides using existing frameworks, there exist numerous toolkits that allow a user to create their web scraping application. This allows the user to create a specific method for their project. Upadhyay et al. [15] described how a new and robust web scraping tool could be constructed for any task. Moreover, there exist numerous examples of researchers constructing web scrapers using various frameworks and

expanding them to their requirements. For example, Thomas and Mathur [2] used Scrapy as their primary tool for text-based web crawling, whereas Zink [7] expanded on the toolkit of iCrawler owing to its ease of modification and support for multiple search engines. Even though there exist various web scrapers, in the end, all of them follow the basic principle depicted in Fig. 2.

2.4 Dataset Quality

Dataset quality is significantly challenging to correctly assess automatically because it is characterized by many aspects of a dataset. Pipino et al. [16] defined 16 possible dimensions that can be used to determine the data size quality. In their research, they concluded that within these possible 16 dimensions there was no "one size fits all" solution that can be used to determine the dataset quality. However, within commercial sources [17-20] there seems to be a more conclusive answer.

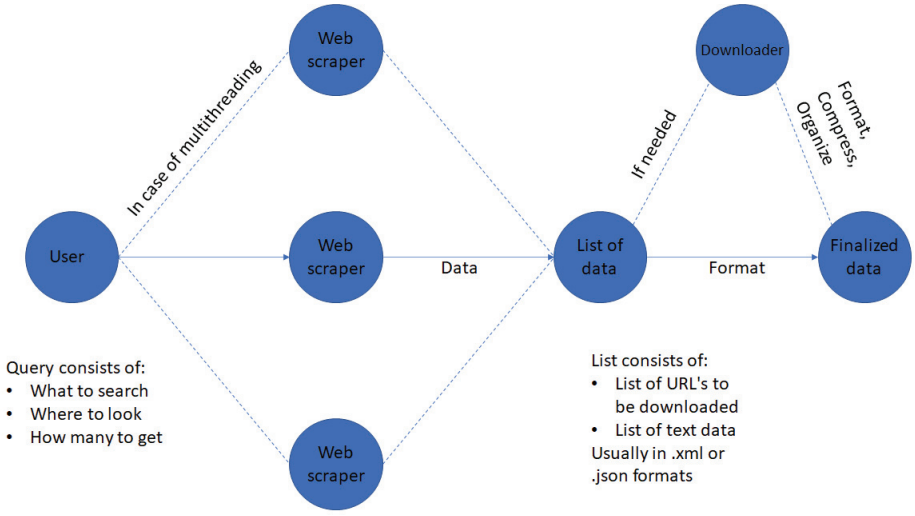

Fig. 2. Basic web scraping system.

The five main factors that have been determined from these commercial sources have been mentioned as important, but not decisive, by Pipino et al. [16] as well. The factors are listed as follows:

- Characteristic, how are the data measured?
- Accuracy, is every detail within the information correct?
- Completeness, how comprehensive is the information?
- Relevance, is this information needed?
- Timeliness, are the data up to date?

Image datasets follow these five characteristics as well; however, they add two more. Firstly, the image quality should be as high as possible. This means that the images should possess high resolution, and there should be as little use of blur or other manipulation as possible [21-23]. Secondly, the image dataset should consist of high-quality unique images. Both commercial [24,25] and academic research [20,22,26] indicate that the ideal threshold is a minimum of 1,000 images for any dataset to possess reasonable quality. The research conducted by Zink [7] also demonstrated a correlation between the number of images within a dataset and the accuracy of the image recognition dataset. Images must be unique to avoid a bias whenever the dataset gets used, which is described by both Rosebrock [27] and Hofesmann [28].

3. Discussion

Based on the concepts introduced in the previous section, we define the requirements that a fully automated image web scraper should contain.

3.1 Collecting an Image Dataset

Collecting an image dataset for any use must consider the seven characteristics that were mentioned before. When using an automated web scraper to collect an image dataset via search engines, the relevance, timeliness, quality, and reliability are taken care of by the used search engines as demonstrated by Zhang and Rui [29], which are hard to manually adjust. Completeness is something that a fully automated system cannot assess. However, as mentioned by Zink [7], if the quantity of images increases, the completeness will increase as well. When it comes to quantity, this can simply be a user-defined parameter that the web scraper tries to obtain. The main problem with achieving a high quantity is that every search engine only displays a relatively low amount. For example, Google images display a maximum of 400 results [30]. Therefore, for achieving a higher quantity of unique images, multiple search engines must be used. Considering even higher quantities of images needed for the dataset, query expansions can be used for different images. When evaluating the accuracy of the dataset, any of the current state-of-the-art neural networks can use the dataset to train and assess the scraped dataset.

3.2 Dealing with Noise

Within any dataset, there is a probability for “noise” to appear. Noise dramatically decreases the classification accuracy of the created model. Zhu and Wu [31] identifies two types of noise within datasets, class, and attribute noise, both significantly impacting the accuracy of the final model. Class noise occurs when an item in the dataset has been misclassified, whereas attribute noise occurs when a specific value corresponding to an item is incorrect. An image classification example is presented in Fig. 3, which depicts the image search results regarding the color red. In this example, the class noise is a completely wrong image whereas the attribute noise image yields a blue square.

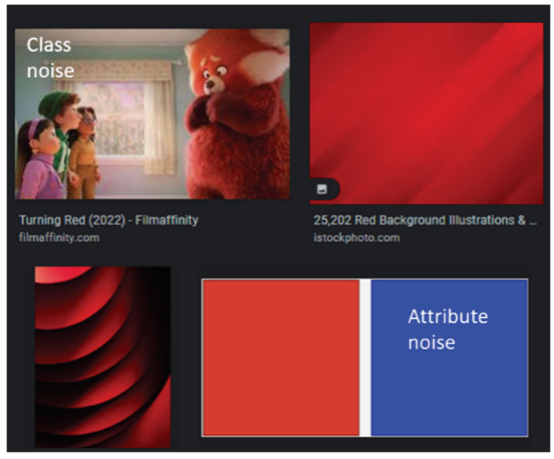

Fig. 3. An example of class and attribute noise that occurs when searching for images of red.

The authors of [32] compared 79 studies and techniques regarding noise identification and solving noise. Almost all considered techniques involve data preprocessing. Tableau [33] defined data preprocessing, also known as data cleaning, as the process where the data contained in a collected dataset is processed to remove any items that contain noise.

Within a fully automated environment, checking for noise is extremely difficult. This is because checking for noise in a large dataset using an algorithm requires human intervention. Within an automatically gathered image dataset, the preferred method of data preprocessing is demonstrated by Zink [7]. They used a method where a smaller high-quality dataset was used to check all images in the dataset to verify if they fit the input query. Within a fully automated system, there is no proven way to guarantee that an image is correctly classified; therefore, this method seems impossible. A potential method to create a smaller high-quality dataset is by obtaining the first few items from a search query, assuming they are without noise, as they are the best results provided by the search engine, and then generating the high-quality model to preprocess the remaining gathered images.

3.3 Regional Differences

All search engines suffer from regional differences. Country-based bias on search results, where certain search results are not shown, is proven by both Vaughan and Thelwall [34] and Mowshowitz and Kawaguchi [35]. When this is combined with language-based differences, which are more prevalent with image searches, a single query can result in a completely different set of results, as shown in Fig. 4. Based on the combination of bias and country-specific terms, search engines return different results on each query for each country.

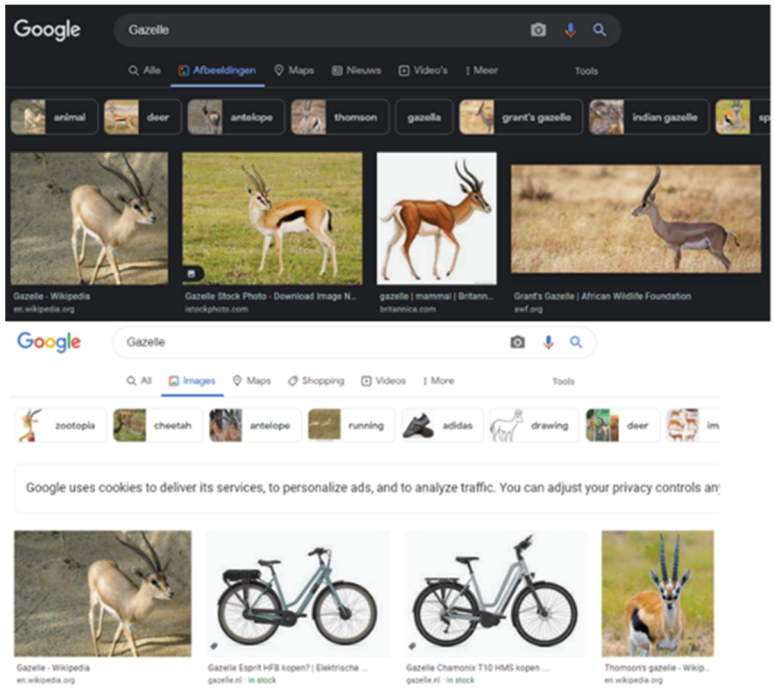

Fig. 4. A comparison between the search results of Gazelle initiated from Korea and the Netherlands. There is a difference as Gazelle is a famous bike brand in the Netherlands.

The main solution to this problem would be to implement a system that enforces the search query to always be constructed from a specific country using a VPN or allow the user to clarify its initial search query with supporting queries. These supporting queries would “guide” the search engine to the specific user-desired query. This can be done in both a positive and negative sense, as most modern search engines support a hyphened (-) query, which would omit certain results [36]. Based on this a user can prevent any problems caused by their location. An example, presented in Fig. 4, would be the main query of Gazelle supported by the queries of Animal and Africa so that the search engine is guaranteed to return the correct data.

3.4 Query Expansion

As discussed in Section 2.2, the most optimal way to automatically establish links between words is using a trained Word2Vec model. When a query expansion algorithm, such as Word2Vec, is used to enhance a search query, there are potentially thousands of links between each word. As mentioned earlier, the easiest links that can be established using Word2Vec are synonyms, holonyms, meronyms, and hypernyms. For image classification, a synonym should be avoided as it leads to potentially completely unrelated queries. For example, YourDictionary [37] lists the top-rated synonyms for a Labrador dog as Golden Retriever, Spaniel, and Dachshund, which are completely different dog breeds. Searching for these synonyms would experience a higher probability of noise. The search query behavior of a holonym does correctly help specify the primary query; however, it leads to a loss in image compilation. For example, Word2Vec demonstrates that one of the direct holonyms for a Labrador is the snout. Searching for a Labrador snout correctly returns images of the Labrador and its snout but it primarily returns images focused on its snout. Regarding meronyms and hypernyms, using both once- and twice-removed hypernyms yields search results leading to more detailed images of the initial query. All these examples can be seen in Fig. 5.

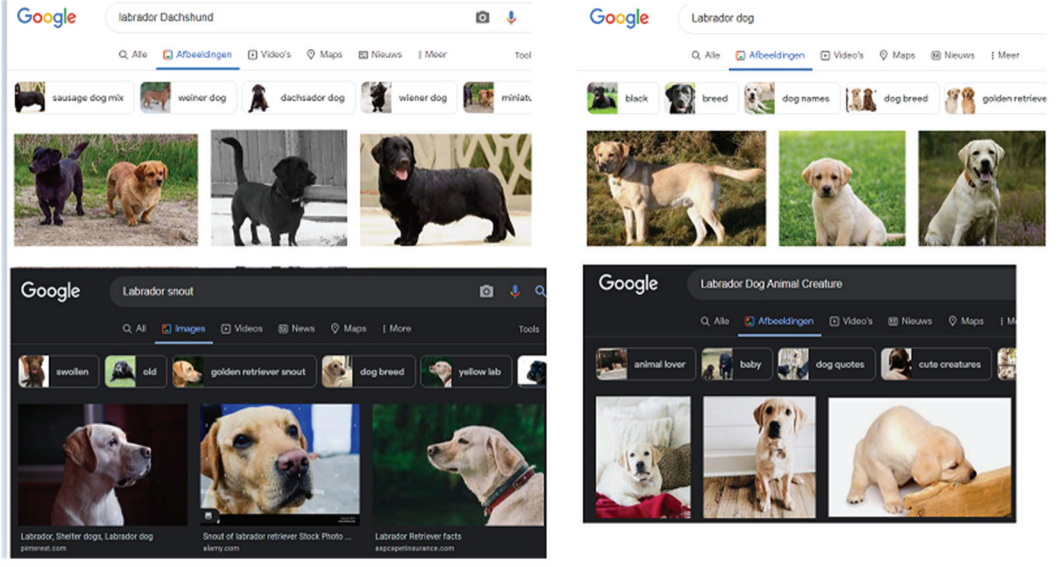

Fig. 5. Examples of synonym, holonym, and hypernym expanded searches.

However, there is still a chance that hypernyms can be incorrect. This can occur when a word can belong to multiple collections. For example, Labrador is a dog but there also exists a Labrador Island, which results in an existing hypernym link to Island. If this hypernym is used for expanding the search query, it would result in incorrect data. A solution to this problem would be to use a manual secondary query provided by the user and check the link between secondary queries and the hypernym of the primary query. In this case, a user-provided primary query of Labrador could be supported by Dog and Pet as secondary queries. By testing if a link exists between the primary query's holonyms and its secondary queries, it is possible to determine if the holonym can be used to expand the search query.

3.5 Copyright Infringement

Rappaport et al. [38] stated that even though web scraping is currently legal, laws regarding the potential copyright infringement caused by web scraping are rapidly changing. Paul [39] proposed multiple tips for limiting potential copyright infringement. For an automated web scraper for images, the main tip that can be applied is to focus on using public data and APIs to collect the data. Accordingly, a web scraper should primarily be limited to searching through search engines in a way that limits the requests sent. This is because search engines are also actively trying to remove copyrighted materials from their web page, as can be seen in Google Support [40]. Preferably the dataset collected using web scraping would be discarded after its immediate use. When generating a model, this could be implemented in the web scraping application so that the images can be immediately discarded and only the trained model persists. This would result in the lowest possible copyright infringement. If the user wishes for the image data to remain in their local storage, they must actively select an option to do so.

4. Conclusion and Future Work

In conclusion, there is a plethora of possible web scraping applications, as demonstrated in the studies conducted by Glez-Pena et al. [3] and Sirisuriya [13]. However, options for image web scraping are limited. Each research conducted on image web scraping [2,5,7] had used a different framework to construct a web scraper in a certain way to download images. None of the solutions offered in the exiting research provide a high-accuracy solution for fully automated web scraping, including query expansion. Accordingly, there is an opportunity to create a modern web scraping application that considers these requirements while implementing both query expansion and noise filtration measures.

Based on the research conducted on the web scraping field, we determined that a fully automated web scraper, such as that proposed by Yao et al. [5], can be created using both manual and automated query expansion to generate image datasets. Owing to advancements in search engine behavior, implementing a noise-checking code may not be necessary. However, if the resulting accuracy is too low, a noise management method can be applied to further improve the web scraper. The preferred web scraper, using some of the concepts discussed earlier, would attain a simple user flow, as depicted in Fig. 6.

After creating a web scraping tool that combines all these technologies, most modern machine learning models can be used to test the accuracy of the created datasets, which can in turn be compared to the results obtained by Zink [7] or image datasets already available online through ImageNet. If the

corresponding results are adequate, the proposed web scraper could be made available for research purposes, while taking measures to minimize any potential copyright infringement during automated web scraping.

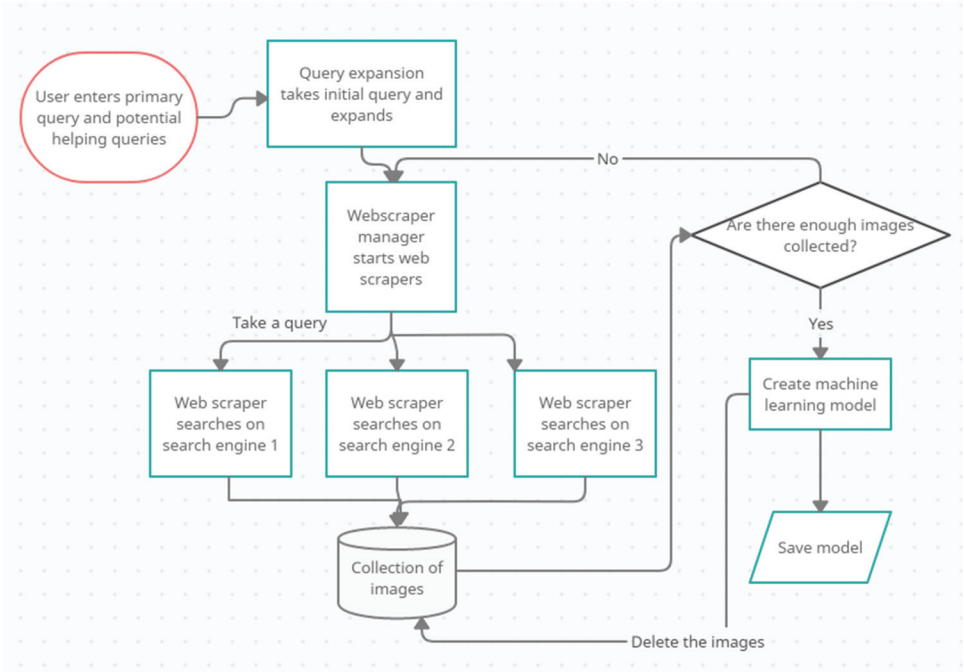

Fig. 6. Potential fully automated image dataset model.

Acknowledgement

This study was supported by the Research Program funded by the Seoul National University of Science and Technology (SeoulTech).

References

[1] A. Rosebrock, "How to create a deep learning dataset using google images," 2017 [Online]. Available: https://pyimagesearch.com/2017/12/04/how-to-create-a-deep-learning-dataset-using-google-images/.

[2] D. M. Thomas and S. Mathur, "Data analysis by web scraping using Python," in *Proceedings of 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2019, pp. 450-454. https://doi.org/10.1109/ICECA.2019.8822022

[3] D. Glez-Pena, A. Lourenco, H. Lopez-Fernandez, M. Reboiro-Jato, and F. Fdez-Riverola, "Web scraping technologies in an API world," *Briefings in Bioinformatics*, vol. 15, no. 5, pp. 788-797, 2014. https://doi.org/10.1093/bib/bbt026

[4] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 754-766, 2011. https://doi.org/10.1109/TPAMI.2010.133

[5] Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, and Z. Tang, "Automatic image dataset construction with multiple textual metadata," in *Proceedings of 2016 IEEE International Conference on Multimedia and Expo (ICME)*, Seattle, WA, 2016, pp. 1-6. https://doi.org/10.1109/ICME.2016.7552988

[6] Y. Lin, J. B. Michel, E. A. Lieberman, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the Google Books Ngram Corpus," in *Proceedings of the ACL 2012 System Demonstrations*, Jeju, South Korea, 2012, pp. 169-174.

[7] J. M. Zink, "Automated dataset generation for image recognition using the example of taxonomy," 2018 [Online]. Available: https://arxiv.org/abs/1802.02207.

[8] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995. https://doi.org/10.1145/219717.219748

[9] Google, "Word2Vec documentation," 2013 [Online]. Available: https://code.google.com/archive/p/word2vec/.

[10] A. Handler, "An empirical study of semantic similarity in WordNet and Word2Vec," Master's thesis, University of New Orleans, New Orleans, LA, USA, 2014 [Online]. Available: https://scholarworks.uno.edu/td/1922/.

[11] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: construction of a large-scale image dataset using deep learning with humans in the loop," 2015 [Online]. Available: https://arxiv.org/abs/1506.03365.

[12] D. Schwab and M. Lafourcade, "Hardening of acception links through vectorized lexical functions," 2002 [Online]. Available: https://www.researchgate.net/publication/2543982_Hardening_of_Acception_Links_Through_Vectorized_Lexical_Functions.

[13] B. Zhao, "Web scraping," in *Encyclopedia of Big Data*. Cham, Switzerland: Springer, 2017, pp. 1-3. https://doi.org/10.1007/978-3-319-32001-4_483-1

[14] D. S. Sirisuriya, "A comparative study on web scraping," 2015 [Online]. Available: http://ir.kdu.ac.lk/handle/345/1051.

[15] S. Upadhyay, V. Pant, S. Bhasin, and M. K. Pattanshetti, "Articulating the construction of a web scraper for massive data extraction," in *Proceedings of 2017 2nd International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, 2017, pp. 1-4. https://doi.org/10.1109/ICECCT.2017.8117827

[16] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211-218, 2002. https://doi.org/10.1145/505248.506010

[17] S. Shen, "7 steps to ensure and sustain quality data," 2019 [Online]. Available: https://towardsdatascience.com/ 7-steps- to-ensure-and-sustain-data-quality-3c0040591366.

[18] HeavyAI, "Data Quality FAQ," 2022 [Online]. Available: https://www.heavy.ai/technical-glossary/data-quality#:~:text=Data%20that%20is%20deemed%20fit,data%2C%20and%20poor%20data%20security.

[19] R. L. Sarfin, "5 Characteristics of data quality," 2022 [Online]. Available: https://www.precisely.com/blog/data-quality/5-characteristics-of-data-quality.

[20] C. Stedman and J. Vaughan, "Data Quality," 2022 [Online]. Available: https://www.techtarget.com/searchdatamanagement/definition/data-quality.

[21] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proceedings of 2016 8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1-6. https://doi.org/10.1109/QoMEX.2016.7498955

[22] Z. Chen, W. Lin, S. Wang, L. Xu, and L. Li, "Image quality assessment guided deep neural networks training," 2017 [Online]. Available: https://arxiv.org/abs/1708.03880.

[23] Z. Chen, W. Lin, S. Wang, L. Xu, and L. Li, "Image quality assessment based label smoothing in deep neural network learning," in *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 6742-6746. https://doi.org/10.1109/ICASSP.2018.8461630

[24] Picsella, "How to ensure image dataset quality for image classification," 2023 [Online]. Available: https://www.picsellia.com/post/image-data-quality-for-image-classification.

[25] A. Mikhailiuk, "Deep image quality assessment," 2021 [Online]. Available: https://towardsdatascience.com/deep- image -quality-assessment-30ad71641fac.

[26] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36-47, 2020. https://doi.org/10.1109/TCSVT.2018.2886771

[27] A. Rosebrock, "Detect and remove duplicate images from a dataset for deep learning," 2020 [Online]. Available: https://pyimagesearch.com/2020/04/20/detect-and-remove-duplicate-images-from-a-dataset-for-deep-learning/.

[28] E. Hofesmann, "Find and remove duplicate images in your dataset," 2021 [Online]. Available: https://towardsdatascience.com/find-and-remove-duplicate-images-in-your-dataset-3e3ec818b978#:~:text=Images%20with%20a%20low%20uniqueness,train%2Ftest%20your%20model%20on.

[29] L. Zhang and Y. Rui, "Image search: from thousands to billions in 20 years," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 9, no. 1s, article no. 36, 2013. https://doi. org/10.1145/2490823

[30] Google Search Help, "Q&A about image search (reply: Bluequoll)," 2020 [Online]. Available: https://support.google.com/websearch/thread/32492691/image-search-returns-almost-nothing-compared-to-what-it-used-to-on-about-any-subject-what-gives?hl=en.

[31] X. Zhu and X. Wu, "Class noise vs. attribute noise: a quantitative study," *Artificial Intelligence Review*, vol. 22, pp. 177-210, 2004. https://doi.org/10.1007/s10462-004-0751-8

[32] S. Gupta and A. Gupta, "Dealing with noise problem in machine learning data-sets: a systematic review," *Procedia Computer Science*, vol. 161, pp. 466-474, 2019. https://doi.org/10.1016/j.procs.2019.11.146

[33] Tableau, "Guide to data cleaning: definition, benefits, components, and how to clean your data," 2023 [Online]. Available: https://www.tableau.com/learn/articles/what-is-data-cleaning.

[34] L. Vaughan and M. Thelwall, "Search engine coverage bias: evidence and possible causes," *Information Processing & Management*, vol. 40, no. 4, pp. 693-707, 2004. https://doi.org/10.1016/S0306-4573(03)00063-3

[35] A. Mowshowitz and A. Kawaguchi, "Measuring search engine bias," *Information Processing & Management*, vol. 41, no. 5, pp. 1193-1205, 2005. https://doi.org/10.1016/j.ipm.2004.05.005

[36] University of Florida Libraries, "Google Guide: basic search tips," 2023 [Online]. Available: https://libguides.uwf.edu/c.php?g=215353&p=1420921.

[37] YourDictionary, "Labrador synonyms," 2023 [Online]. Available: https://thesaurus.yourdictionary.com/labrador.

[38] D. A. Rappaport, P. I. Altman, K. Handshumacher, "To scrape or not to scrape: the potential legal implications of using web scraping for market research," *Hedge Fund Law Report*, 2021 [Online]. Available: https://www.akingump.com/a/web/soxXRQ6Nw48FehNvwpdjJ1/2jiuhx/hflr-reprint-to-scrape-or-not-to-scrape-rappaport-altman-handschumacher-4819-0662-7801-v1.pdf.

[39] T. Paul, "is web scraping legal? A guide to understanding legality on web scraping," 2020 [Online]. Available: https://www.blog.datahut.co/post/is-web-scraping-legal.

[40] Google, "Removing content on Google," 2023 [Online]. Available: https://support.google.com/legal/troubleshooter/1114905?hl=en.

Bart Dokmans **https://orcid.org/0000-0001-8518-7406**

He received his B.S. of Game Development in the Amsterdam University of Applied Sciences in 2019 and is currently finishing up his M.S. in the Seoul University of Science and Technology. He is a full stack developer for T-MC in the Netherlands.

Dongwann Kang **https://orcid.org/0000-0001-7210-4595**

He is currently an assistant professor in the Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, Korea. He received the Ph.D. degree from Chung-Ang University, South Korea, in 2013, where he has been a research fellow, until 2015. He was a lecturer of Undergraduate Interdisciplinary Program in Computational Sciences, Seoul National University, South Korea, from 2014 to 2015; a lecturer with the Department of Multimedia, Sookmyung Women's University, South Korea, in 2014; a visiting researcher, from 2015 to 2018, and a Marie Sklodowska-Curie Fellow of the Faculty of Science and Technology, Bournemouth University, UK, in 2018. His research interests include non-photorealistic rendering and animation, computer vision, affective computing, and computational aesthetics.