Doing Visual Big Data – Creating the KBK-1M Dataset Containing 1,6 Million Newspaper Images Available for Researchers

Martijn Kleppe
National Library of the Netherlands (KB)
martijn.kleppe@kb.nl

Desmond Elliott
University of Amsterdam (UvA)
d.elliott@uva.nl

The visualisation of news through photographs has exploded since the second half of the 20th century (Kester & Kleppe 2015). However, methods that are employed to analyse the (re)use of visual materials are labour-intensive because Humanities researchers tend to analyse their sources manually (Burke 2001). To estimate the increase in the use of pressphotographs in Dutch newspapers, Kester & Kleppe (2015) e.g manually analysed a sample of 385 newspapers and 5.877 press photographs over the period 1870-2013. To find the recurring use of photographs in Dutch history textbooks, Kleppe (2012) followed a same approach by manually analysing over 5.000 photographs in 400 history textbooks, creating the 'Foto's in Nederlandse Geschiedenisschoolboeken (FiNGS) (Photos in Dutch History textbooks) dataset (Kleppe 2013b).

Even though manually created and annotated datasets such as FiNGS contain rich & well-annotated data, their scope remains limited given its labour-intensive creation and analyses process. To find the recurring imagery in the FiNGS dataset, Kleppe (2012) e.g. manually created and assessed all images and metadata, leading to inevitabel human errors. However, digitised historical imagery is increasingly becoming available, allowing researchers to undertake the first steps in the field of 'Visual Big Data', following the footsteps of Barry Salt's study on the characteristics of opening shots of 20th century films (Salt 1974) and Scott McCloud work on the visual language of Japanese manga and comics from the West (McCloud 1993). More recent, the work of Lev Manovich on exploring large scale visual datasets such as Manga Comics (2012), Time covers, and selfies (Manovich & Tifentale 2015) is seen as a new way of what he calls doing 'cultural analytics' (Manovich 2012).

While the focus of these studies is on characteristics of the images, other studies using large scale image dataset focus on the recurrance of imagery in different types of contexts, aiming 1) to assess the impact of scholarly images online (Kousha 2010), 2) to analyse the reuse of digital images of cultural and heritage material on the internet (Terras 2013) or within a closed dataset (Resig 2014; Reside 2014) and 3) to detect poetic content in historical newspapers (Lorang et al 2015).

To cater their research questions, these scholars all created visual datasets on their own. However, large datasets containing photographs that are free available for researchers are scarce. Only within the Computer Vision and Natural Language Processing community we found some datasets

(Ordonez et al, 2011; Chen et al, 2015a; Chen et al, 2015; Hodosh et al., 2013), but these are mainly created for training purposes of algorithms, not for Humanities research questions.

Therefor this poster presents the KBK-1M dataset, that was created specifically for (Digital) Humanities researchers. This dataset contains a collection of 1.603.395 captioned images extracted from Dutch digitised newspapers stored in the Dutch National Library (KB) Newspaper archive of the period 1922-1994. On our poster, we will describe how we obtained the images, what types of research questions it could tailor and how researchers can obtain the dataset for their research purposes.

References

Burke, P. (2001) *Eyewitnessing. The uses of images as historical evidence*. London: Cornell University Press.

Chen, J., Kuznetsova, P., Warren, D. S., and Choi, Y. (2015a) 'Deja image-captions: A corpus of ´ expressive descriptions in repetition', *NAACL*, pp. 504–514.

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollar, P., and Zitnick, C. L. (2015b) 'Microsoft COCO captions: Data collection and evaluation server', *CoRR*, abs/1504.00325.

Hodosh, M., Young, P., and Hockenmaier, J. (2013) 'Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics', *Journal of Artificial Intelligence Research*, 47, pp. 853–899.

Kester, B., & Kleppe, M. (2015) 'Acceptatie, professionalisering en innovatie. Persfotografie in Nederland, 1837-2014', In Bardoel, J. & Wijfjes, H., *Journalistieke Cultuur in Nederland*, pp. 53-76. Amsterdam: Amsterdam University Press.

Kleppe, M. (2013a) *Canonieke Icoonfoto's. De rol van (pers)foto's in de Nederlandse geschiedschrijving*. Delft: Eburon.

Kleppe, M. (2013b) *Foto's in Nederlandse Geschiedenisschoolboeken (FiNGS)* http://www.persistent-identifier.nl/?identifier=urn:nbn:nl:ui:13-l37n-bi

Kleppe, M. (2012) 'Wat is het onderwerp op een foto? De kansen en problemen bij het opzetten van een eigen fotodatabase',T*ijdschrift voor Mediageschiedenis*, 2 (14), pp. 93 - 107.

Kousha, K., Thelwall, M., Rezaie, S. (2010) 'Can the impact of scholarly images be assessed online? An exploratory study using image identificationtechnology', *Journal of the American Society for Information Science and Technology* 61 (9), pp. 1734-1744.

Manovich, L. (2009) 'Cultural analytics: Visualing cultural patterns in the era of more media', *Domus* (923).

Manovich, L. Douglass, J., Zepel, T., (2012) *How to compare one million Images in Understanding Digital Humanities* http://softwarestudies.com/cultural_analytics/2011.How_To_Compare_One_Million_Images.pdf

Manovich. L. and Tifentale, A (2015) 'Selfiecity: Exploring Photography and Self-Fashioning in Social Media', In Berry, D.M., Dieter, M., *Postdigital Aesthetics: Art, Computation and Design*, pp. 109-122. New York: Palgrave Macmillan.

McCLoud, S. (1994) *Understanding Comics: The Invisible Art.* New York: Harper Perennial.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011) 'Im2text: Describing images using 1 million captioned photographs', *NIPS*.

Reser, G., & Bauman, J. (2012) 'The Past, Present, and Future of Embedded Metadata for the Long-Term Maintenance of and Access to Digital Image Files', *International Journal of Digital Library Systems (IJDLS),* 3(1), 53-64.

Reside, D. (2014), 'Using Computer Vision to Improve Image Metadata'. *Digital Humanities 2014*. http://dharchive.org/paper/DH2014/Paper-294.xml

Resig, J. (2013) *Using Computer Vision to Increase the Research Potential of Photo Archives* http://ejohn.org/research/computer-vision-photo-archives/#analysis-implementations

Salt, B. (1974) 'The Statistical Style Analysis of Motion Pictures', *Film Quarterly* 28 (1), pp. 13-22.

Smith, J. R. (2013) Riding the multimedia big data wave', *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval – SIGIR '13*, New York: ACM Press doi:10.1145/2484028.2494492

Terras, M. M., Kirton, I. (2013) 'Where do images of art go once they go online? A Reverse Image Lookup study to assess the dissemination of digitized cultural heritage' *Selected papers from Museums and the Web North America*, pp. 237 – 248.