1 Million Captioned Dutch Newspaper Images

Desmond Elliott[*†] **and Martijn Kleppe**[‡]
[*]ILLC, University of Amsterdam; [†]CWI; [‡]Erasmus University Rotterdam
d.elliott@uva.nl, kleppe@eshcc.eur.nl

Abstract

Images naturally appear alongside text in a wide variety of media, such as books, magazines, newspapers, and in online articles. This type of multi-modal data offers an interesting basis for vision and language research but most existing datasets use crowdsourced text, which removes the images from their original context. In this paper, we introduce the KBK-1M dataset of 1.6 million images in their original context, with co-occurring texts found in Dutch newspapers from 1922 - 1994. The images are digitally scanned photographs, cartoons, sketches, and weather forecasts; the text is generated from OCR scanned blocks. The dataset is suitable for experiments in automatic image captioning, image–article matching, object recognition, and data-to-text generation for weather forecasting. It can also be used by humanities scholars to analyse photographic style changes, the representation of people and societal issues, and new tools for exploring photograph reuse via image-similarity-based search.

Keywords: Digital Humanities, Digitised Newspapers, Language and Vision

1. Introduction

Vision and language datasets typically consist of high-quality born-digital photographs paired with crowd-sourced descriptions (Rashtchian et al., 2010; Hodosh et al., 2013a; Young et al., 2014; Chen et al., 2015b). Descriptions obtained through crowdsourcing have several advantages: they usually describe what can be seen in the image, reducing the difficulty of visually grounding the words in the image, and it is easy to quickly collect multiple reference texts. However, crowdsourcing takes an image outside its original context, rendering the intentions of the photographer inaccessible. For example, does Figure 1 depict the world-famous Notre Dame Cathedral in Paris or automobiles waiting at a traffic light? It is difficult to be sure, but crowdsourcing affects our ability to situate the image in a wider context. An alternative to crowdsourced descriptions is sourcing images that naturally co-occur with text, such as on photo sharing sites, social media, or in newspapers (Feng and Lapata, 2008; Ordonez et al., 2011; Chen et al., 2015a). The captions in these datasets often refer to specific people (politicians, movie stars, etc.), emotions, or places (Paris, my back garden, etc.) that are difficult to ground against the images.

In this paper we introduce the KBK-1M dataset of 1.6 million captioned newspaper images. Contrary to other visual datasets such as WikiMedia Commons, the KBK-1M dataset contains images and captions that have been created by professional photographers and journalists and they are situated in their original context in the newspaper. The dataset covers 72 years of printed newspapers from a variety of media outlets, giving a broad topical scope to the collection. The images are not digital-born, which causes interesting and unusual noise artifacts due to the age and quality of scanning printed paper. The images cover black-/white and colour photographs, sketches, line-drawn cartoons, and line-drawn weather forecasts; see Figure 3 for examples from each category. The image captions were extracted via optical character recognition over known boundary boxes.

We anticipate this new resource will offer a wide-range of opportunities for interdisciplinary research. Natural Language Processing (NLP) and Computer Vision (CV) researchers will find a challenging dataset for automatic image captioning and multimodal image–sentence matching. Humanities scholars will find a longitudinal large-scale collection of images in context, that may enable the development of big-data approaches to photographic and media studies.

2. The KBK-1M Dataset

KBK-1M is a collection of 1,603,396 captioned images extracted from digitised newspapers stored in the Dutch National Library (KB) Newspaper Archive. We extracted images and texts from issues printed between 1922 - 1994 because previous research shows an increase in the number of published photographs in Dutch newspapers in the second-half of the twentieth century (Kester and Kleppe, 2015). To the best of our knowledge, the BBC News Database is the only directly comparable dataset (Feng and Lapata, 2008). The size, diversity of source material, and timespan of our dataset make it unique among vision and language datasets; see Table 1 for a comparison of KBK-1M against other similar datasets of images paired with sentences.

In the Newspaper Archive of the KB, a newspaper issue is stored as a set of scanned pages, with one high-resolution JPEG per newspaper page. Each page is associated with a set of metadata files which describe the locations of any known images, captions, and article blocks. The image and caption locations were manually annotated by trained workers as part of a previous project at the library, we use these annotations without further processing. The article bodies and caption text is available as automatic OCR-processed output. The images are high-resolution scanned images, however, they lack the clarity of modern digital photographs due to the aged printed paper source material. Figure 2 shows an overview of how we transformed the raw source material into the dataset.

Listing 1 shows an example of the JSON annotation format used in the dataset. We serialised the caption, the title of

1. A blue smart car parked in a parking lot.
2. Some vehicles on a very wet wide city street.
3. Several cars and a motorcycle are on a snow covered street.
4. Many vehicles drive down an icy street.
5. A small smart car driving in the city.

Figure 1: The "Notre Dame Effect": is this a photograph of the Notre Dame Cathedral or automobiles waiting at a traffic light? None of the crowdsourced descriptions for image 56305_z in the MS COCO dataset mention the iconic landmark.

Dataset	Images	Texts	Source
KBK-1M	1,603,396	Title and caption	Dutch newspapers
BBC News (Feng and Lapata, 2008)	3,361	Title, caption, body	BBC News Online
FiNGS (Kleppe, 2012)	5,344	Caption, descriptions	Dutch Textbooks
SBU Captioned Photos (Ordonez et al., 2011)	1,000,000	Caption	Photo uploader
Deja-Images (Chen et al., 2015a)	440,000	Caption	Photo uploader
Flickr30K (Hodosh et al., 2013b)	30,000	Five descriptions	Crowdsourced
MS COCO (Chen et al., 2015b)	164,062	Five descriptions	Crowdsourced

Table 1: Comparison of several news image datasets, user-captioned images, and crowd-sourced described images. For a more comprehensive overview, see Ferraro et al. (2015) and Bernardi et al. (2016).

the newspaper issue, the date of publication, and the identifiers of the content and text blocks in the original repository metadata document. Each newspaper issue is stored with a unique identifier linking an image – caption pair directly back to the newspaper issue ID from the Newspaper Archive. The images are broken down by year of publication and are available as year-by-year ZIP files at the National Library of the Netherlands. A request for access can be submitted to `dataservices@kb.nl` in order to comply with Dutch copyright agreements.

3. Use Cases

We present three use cases for the KBK-1M dataset, covering computer vision (CV), Natural Language Processing (NLP) and Humanities research. We welcome new use-cases from these and other communities.

3.1. CV & NLP: Automatic Image Captioning and Multimodal Ranking

Image-to-text generation is the task of automatically generating the description of an image using only the visual input. This task has received a great deal of attention in recent years (Farhadi et al., 2010; Yang et al., 2011; Li et al., 2011; Mitchell et al., 2012; Yatskar et al., 2014; Elliott and de Vries, 2015; Vinyals et al., 2015; Karpathy and Fei-Fei, 2015; Mao et al., 2015) and represents a strong challenge at the intersection of CV and NLP research. Most datasets for this task are based on crowdsourced descriptions, from which visual features and words can more easily co-reference. In the KBK-1M dataset there is no guarantee the captions will refer to anything depicted in the image. Therefore, it is not possible to directly learn literal description models. Furthermore, off-the-shelf object recognition methods are unlikely to work on scanned black-and-white images.

An alternative use for this dataset is Multimodal Ranking. This is the task of finding the sentence that best describes an image, and vice-versa (Hodosh et al., 2013b). The KBK-1M dataset contains images paired with captions, making it suitable for large-scale multimodal matching experiments.

3.2. NLP: Data-to-Text Generation

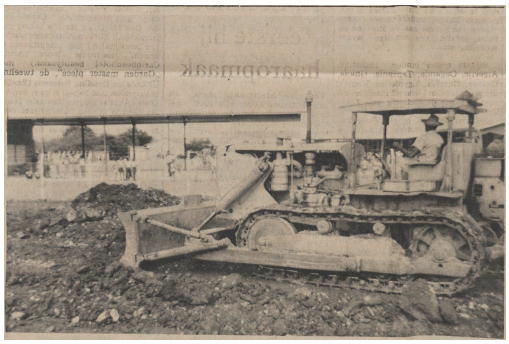
Data-to-Text Generation is a well-studied problem in Natural Language Generation (Konstas and Lapata, 2012, inter-alia). The task is framed as transforming database records into natural language, such as WeatherGov records with structured weather observation data. One possible use for the KBK-1M dataset is the development of a new method of Data-to-Text Generation, conditioned on weather-forecast sketches. The newspaper weather forecasts contain rich visual data, including isobar charts, temperatures, weather fronts, and major city names. Gold-standard weather data can be found in the formal records from the Royal Netherlands Meteorological Institute (KNMI) Data Centre.

3.3. Humanities: Image Reuse Through Time

Within the Humanities domain, photographs are increasingly being considered as historical source (Burke, 2006), as means to study photographic style changes (Manovich, 2009), and as source for studies on the representation of people and processes (Kleppe, 2013). One particular application of the KBK-1M dataset is the study of the reuse of

Figure 2: An example of an image and caption extracted from the front page of the January 27th 1951 issue of the De nieuwsgier. The image (red) and caption (black) are gold-standard annotated in the KB Newspaper Archive. We automatically extract the image and text from inside these annotations and save the resulting content as JPEG and JSON data.

(a) Flintstones. (Source: De Telegraaf, 06-01-1969)



(b) De graafwerkzaamheden voor -de bouw van vier school lokalen, toegezegd door de heer H. Rochcreau van het EEG-ontwikkelingsfoads tijdens diens bezoek aan Bonaire, zijn reeds begonnen. (Source: Amigoe di Curacao : weekblad voor de Curacaosche eilanden, 19-02-1968)

(c) Weer in Europa (Source: De Tijd, 02-12-1967)

Figure 3: The KBK-1M dataset contains comic strips, portrait and scene photographs, political cartoons, and weather maps.

photographs. Recent research has shown that the visualisation of news through photographs has exploded since the second half of the 20th century (Kester and Kleppe, 2015). This has also given rise to the recontextualisation of photographs in different contexts. However, the analysis of the reuse of imagery is currently cumbersome and labour-intensive because researchers need to manually study the source material. Nevertheless, workflows emerging in the field of Digital Humanities offer the possibility to refine this process. In particular, it has become feasible to use tools based on deep learning techniques from the field of computer vision. It is now possible to reliably recognise thousands of objects in images in ways that are similar to the way in which in natural language words, sentences, or paragraphs can be analysed without the prior need to manually define input features (LeCun et al., 2015). The latter tech-

Listing 1: JSON Metadata Format


```
{
    "caption": "\n\t\n\tDe Burmese minister van
        buitenlandse zaken vertoeft te Djakarta
        met enige leden van zijn staf; hij werd
        op Kemajoran verwelkomd door
        Indonesische autoriteiten en de Burmese
        ambassadeur. (Ipphos) j\n",
    "paper_title": "De nieuwsgier",
    "page": 1,
    "content_block_url": "http://imageviewer.kb.
        nl/ImagingService/imagingService?id=ddd
        :010474896:mpeg21:p001:image&colour=
        fefe56&s=0&x=2463&y=4545&w=1589&h=790",
    "text_block": "ddd:010474896:mpeg21:a0018",
    "content_block": "P1_CB00003",
    "date": "1951/01/27 00:00:00",
    "image_name": "1951/DDD:ddd:010474896:mpeg21/
        p001-P1_CB00003.jpg",
    "jp2_url": "http://imageviewer.kb.nl/
        ImagingService/imagingService?id=ddd
        :010474896:mpeg21:p001:image",
    "alto_url": "http://resolver.kb.nl/resolve?
        urn=ddd:010474896:mpeg21:a0018:ocr"
}
```


nique is relevant for the exploration of photo archives based on image caption text. However, in order to develop and deploy these techniques, a large dataset is needed. Given the size and multimodality of the KBK-1M dataset, researching the reuse of photographs by using automated recognition of photographs has now become feasible. Arguably the research agenda of the Digital Humanities is dominated by text-based research (Zundert and Dalen-Oskam, 2014; Ordelman et al., 2014). KBK-1M has the potential of contributing to the enrichment of this agenda by facilitating research that requires the availability of contextualized image data.

4. Discussion and Conclusions

In this paper we introduced the KBK-1M dataset of newspaper images paired with captions. The collection spans Dutch printed newspapers from 1922 - 1994, encompassing over 1.6 million images. The size, variety, and longitudinal nature of KBK-1M offers a novel dataset for NLP and CV research, and also provides a rich image collection for humanities and media scholars.

KBK-1M provides a large-scale resource for future research on automatic image captioning, multimodal matching, and data-to-text generation. Historians and photographers will find a wealth of new data within the collection, that enables them to quickly identify relevant content without the painstaking manual analysis of individual newspapers. We encourage interested parties to develop new interfaces to the Newspaper Archive that leverage the image-text corpus we have presented here.

In future, we plan to release an even larger dataset of every gold-standard annotation in the Newspaper Archive from 1900 - 1994. We will also attempt to separate sketches (cartoons and weather reports) from the photographs, making it easier to work with specific subsets of the data.

Acknowledgements

Authors contributed equally while working as Researchers-in-Residence at the Koninklijke Bibliotheek. DE was supported by ERCIM ABCDE Fellowship 2014-23.

References

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *JAIR*, pages 409–442.

Burke, P. (2006). *Eyewitnessing: The uses of images as historical evidence*. Reaktion books.

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

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015b). Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Elliott, D. and de Vries, A. P. (2015). Describing images using inferred visual dependency representations. In *ACL*.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *ECCV*.

Feng, Y. and Lapata, M. (2008). Automatic image annotation using auxiliary text information. In *ACL*, pages 272–280.

Ferraro, F., Mostafazadeh, N., Huang, T.-H., Vanderwende, L., Devlin, J., Galley, M., and Mitchell, M. (2015). A survey of current datasets for vision and language research. In *EMNLP*, pages 207–213, Lisbon, Portugal.

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

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

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kester, B. and Kleppe, M. (2015). Acceptatie, professionalisering en innovatie. Persfotografie in Nederland 1837-2014. *Journalistieke Cultuur in Nederland*, pages 53–76.

Kleppe, M. (2012). Foto's in Nederlandse Geschiedenis Schoolboeken (FiNGS). http://dx.doi.org/10.17026/dans-zfn-u8k4.

Kleppe, M. (2013). *Canonieke Icoonfotos. De rol van (pers) fotos in de Nederlandse geschiedschrijving*. Eburon.

Konstas, I. and Lapata, M. (2012). Concept-to-text generation via discriminative reranking. In *ACL*, pages 369–378, Jeju Island, Korea.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. Y. (2011). Composing simple image descriptions using web-scale n-grams. In *CoNLL*.

Manovich, L. (2009). Cultural analytics: Visualising cultural patterns in the era of more media. *Domus,(Milan), March*.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. (2015). Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*.

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A. C., Yamaguchi, K., Berg, T. L., Stratos, K., Daume, III, H., and III (2012). Midge: generating image descriptions from computer vision detections. In *EACL*.

Ordelman, R., Kemman, M., Kleppe, M., and Jong, F. d. (2014). Sound and (moving images) in focus. How to integrate audiovisual material in Digital Humanities research. In *Digital Humanities 2014*.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *NIPS*.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *NAACLHLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *CVPR*.

Yang, Y., Teo, C. L., Daume, III, H., and Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *EMNLP*.

Yatskar, M., Vanderwende, L., and Zettlemoyer, L. (2014). See No Evil, Say No Evil: Description Generation from Densely Labeled Images. *SEM*.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

Zundert, J. v. and Dalen-Oskam, K. v. (2014). Digital Humanities in the Netherlands. *H-Soz-Kult*.