Weakly supervised construction of a repository of iconic images

Lydia Weiland and **Wolfgang Effelsberg** and **Simone Paolo Ponzetto**
University of Mannheim
Mannheim, Germany
{lydia,effelsberg,simone}@informatik.uni-mannheim.de

Abstract

We present a first attempt at semi-automatically harvesting a dataset of iconic images, namely images that depict objects or scenes, which arouse associations to abstract topics. Our method starts with representative topic-evoking images from Wikipedia, which are labeled with relevant concepts and entities found in their associated captions. These are used to query an online image repository (i.e., Flickr), in order to further acquire additional examples of topic-specific iconic relations. To this end, we leverage a combination of visual similarity measures, image clustering and matching algorithms to acquire clusters of iconic images that are topically connected to the original seed images, while also allowing for various degrees of diversity. Our first results are promising in that they indicate the feasibility of the task and that we are able to build a first version of our resource with minimal supervision.

1 Introduction

Figurative language and images are a pervasive phenomenon associated with human communication. For instance, images used in news articles (especially on hot and sensitive topics) often make use of non-literal visual representations like iconic images, which are aimed at capturing the reader's attention. For environmental topics, for instance, a windmill in an untouched and bright landscape surrounded by a clear sky is typically associated by humans with environmental friendliness, and accordingly causes positive emotions. In a similar way, images of a polar bear on a drifting ice floe are typically associated with the topic of global warming (O'Neill and Smith, 2014).

But while icons represent a pervasive device for visual communication, to date, there exists to the best of our knowledge no approach aimed at their computational modeling. In order to enable the overarching goal of producing such kind of models from real-world data, we focus, in this work, on the preliminary task of semi-automatically compiling an electronic database of iconic images. These consist, in our definition, of images produced to create privileged associations between a particular visual representation and a referent. Iconic images are highly recognizable for media users and typically induce negative or positive emotions that have an impact on viewers' attitudes and actions. In order to model them from a computational perspective, we initially formulate iconic image acquisition as a clustering task in which, given a set of initial, manually-selected 'seed' images – e.g., a photo of a polar bear on a drifting ice floe for the topic of global warming, a smokestack for the topic of pollution, etc. – we use their associated textual descriptions in order to collect related images from the Web. We then process these images using state-of-the-art image understanding techniques to produce clusters of semantically similar, yet different images depicting the same topic in an iconic way.

The acquisition of a database of iconic images represents the first step towards a full-fledged model to computationally capture the phenomenon of iconic images in context. Our long-term vision is to cover all three aspects of *content* (what makes an image iconic?), *usage* (in which context are iconic images used?), and *effects* (which negative/positive emotions do iconic images evoke on viewers?) of iconic images. To make this challenging problem feasible, we opt in this preliminary step for an approach that views the task of understanding iconic images as the ability to build a dataset for further research.

Proceedings of the 25th International Conference on Computational Linguistics, pages 68–73,
Dublin, Ireland, August 23-29 2014.

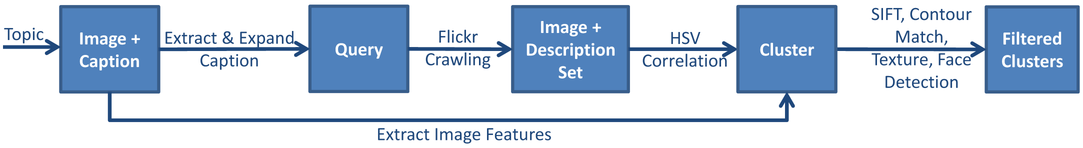

Figure 1: Our framework for the semi-automatic acquisition of iconic images from the Web.

2 Method

Our method for the semi-automatic acquisition of iconic images consists of five phases (Figure 1):

Seed selection. In the first phase of our approach we provide our pipeline with human-selected examples of iconic images that help us bootstrap the image harvesting process. To this end, we initially focus on a wide range of twelve different abstract topics that can be typically represented using iconic images (Table 1). Selecting initial examples describing a visual iconic representation of a topic can be a daunting task, due to their volatile and topic-dependent nature. In this work, we explore the use of Web encyclopedic resources in order to collect our initial examples. We start with the encyclopedic entries from National Geographic Education[1], an on-line resource in which human expert editors make explicit use of prototypical images to visually represent encyclopedic entries like "agriculture", "climate change", etc.. For instance, the encyclopedic entry for "air pollution" contains images of smokestacks, cooling towers release steam, and so on (cf. Table 1). We use these (proprietary) images to provide us with human-validated examples of iconic images, and use these to identify (freely available) similar images within Wikipedia pages based on a Google image search restricted to Wikipedia – e.g., by searching for `smokestack site:wikipedia.org`. We then use Wikipedia to create an initial dataset of iconic visuals associated with the textual descriptions found in their captions.

Text-based image search. In the next step, we make use of a query-by-text approach in order to collect additional data and enlarge our dataset with additional images depicting iconic relations. To this end, we start by collecting the entities annotated within the image captions (e.g., "Cumberland Power Plant at Cumberland City"), and manually determine their relevance to the associated topic (e.g., smokestacks and air pollution). This is because, to build a good query, we need to provide the search systems with a good lexicalization (i.e., keywords) of the underlying information need (i.e., the topic). Consequently, we extract entities from each caption of our initial set of images and use these to query additional data. For each seed, we generate a query by concatenating the entity labels in the captions and send it to Flickr[2]. We then filter the data by retaining only photos with title, description, and tags where both, tags and description (caption and title) contain the query words. This method provides us with around 4000 additional images and text pairs.

Image clustering. Text-based image search results can introduce noise in the dataset, e.g., cases of 'semantic mismatch' where the caption and tags do not appropriately describe the scene found in the image. In this work, we explore the application of image clustering techniques to cope with this issue. For each topic we start with a set of instances made up of the seed images and the crawled one, and group them into clusters based on image similarity measures. Clusters are built by calculating the linear correlation – i.e., which we take as a proxy for a similarity measure – from the HSV-histograms of each image, and applying the K-Means algorithm. Clustering on the basis of HSV-histograms does not take into account the semantic content of images, since images with different meanings can still have the same HSV-histogram. Nevertheless, this approach makes it possible to spot those outliers in the image sets that do not relate well to the other images retrieved with the same query.

Image filtering. The next processing step focuses instead on rule-driven filtering to improve the initial clustering-based filtering. We first apply a face detection and HoG (histogram of gradients) descriptor for

[1] `http://education.nationalgeographic.com`
[2] `http://flickr.com`

Topic	Themes of seed images
Adaption	hummingbird, king snake, koala
Agriculture	cattle, ploughing, rice terraces, tropical fruits
Air	balloon, sky view
Air Pollution	smokestack, Three Mile Island, wildfire
Biodiversity	Amazonas, blue starfish, cornflowers, fungi, Hopetoun Falls
Capital	Capitol Hill, Praça Dos Três, Washington Monument
Climate	Mykonos (mild climate), Sonoran Desert, tea plantation (cool climate)
Climate Change	polar bear, volcano, dry lake
Climate Refugee	climate refugees from Indonesia, Haiti, Pakistan, etc.
Ecosystem	bison, flooded forest, Flynn Reef, harp seal, rainforest, thorn tree
Global Warming	deforestation, flooding, smokestack
Greenhouse Effect	smokestack, steam engine train (smoke emissions)

Table 1: Overview of our covered topics and the themes associated with their seed images.

detecting people (Viola and Jones, 2001; Dalal and Triggs, 2005)[3]. Next, we filter our data as follows. If faces or people are recognized in the picture, and the caption is judged to be related to entities of type person (e.g., farmers iconically depicting the topic of agriculture), the instance is retained in the dataset. On the other hand, if faces and/or people are recognized, but the caption is not related to entities of type person (e.g., a blue linckia, which is a physical object), we filter out the image from the dataset.

Image matching. The filtered clusters we built so far are still problematic in that they do not account for diversity – i.e., we do not want as the outcome of our method to end up with clusters made up only of various pictures of the very same object (e.g., the cooling towers of the Three Mile Island power complex for the topic of air pollution, possibly seen from different perspectives, times of the day, etc.). That is, in our scenario we would like to promote heterogeneous clusters which still retain a high-level semantic match with the initial seeds (e.g., smokestacks or cooling towers belonging to different plants). To this end, we explore in this work an approach that leverages different image matching methods together at the same time to automatically capture these visual semantic matches.

Initially, for each cluster we select the image that minimizes the sum over all squared distances from the other images in the cluster. That is, given a cluster $C = \{c_1 \dots c_n\}$, we collect the image $\hat{c} = \arg\min_{c_i \in C} \sum_{c_j \in C - \{c_i\}} (c_i - c_j)^2$. We call this the *prototype* of the cluster. Several image processing methods are then used to compare the prototype of each cluster with the original seed images, with the aim to detect high-level content similarity (i.e., distinct, yet similar objects such as the smokestacks of different plants, etc.) and account for diversity with respect to our initial seeds. The first method is a template matching approach, based on minimum and maximum values of gray levels, which, together with their location are used to detect similar textures. The matching method is based on a correlation coefficient matching (Brunelli, 2009). In parallel, we explore an alternative approach where images and prototypes are compared using SIFT-based features (Lowe, 2004). Finally, we apply a contour matching method: we use a manually chosen threshold of the largest 10% of contours of an image to reduce the noise from non-characteristic contours like dots, points or other smaller structures. The matching of contours is based on rotation invariant moments (Hu, 1962). When a good match is found, bounding boxes are drawn around the contours.

The three methods provide evidence for potential matches between regions of each input prototype and seed pair. Information from each single method is then combined by intersecting their respective outputs: i) the patch, where the template matching is found is compared against the coordinates where relevant SIFT features are detected (SIFT-Template); ii) the template matching patch is tested for intersection with the bounding boxes of the matched contours (Template-Contour); iii) the bounding boxes of the contours

[3] We focus on face and people detection since these are both well studied areas in computer vision for which a wide range of state-of-the-art methods exist.

Topic	2-matches			all matches		
	P	R	F_1	P	R	F_1
Adaption	100.0	57.2	72.8	66.7	10.9	18.7
Agriculture	50.0	35.4	41.4	0.0	0.0	0.0
Air	84.2	75.6	79.7	66.7	15.8	25.5
Air Pollution	65.9	83.7	73.7	65.1	32.4	43.3
Biodiversity	54.0	40.6	46.3	34.4	8.1	13.1
Capital	61.7	54.6	57.9	50.6	12.6	20.2
Climate	93.7	81.6	87.2	89.1	20.0	32.7
Climate Change	88.5	78.1	83.0	50.0	21.4	30.0
Climate Refugee	40.0	50.0	44.4	0.0	0.0	0.0
Ecosystem	73.7	61.7	67.2	43.3	11.4	18.0
Global Warming	65.9	71.0	68.3	43.0	21.7	28.8
Greenhouse Effect	100.0	81.6	89.9	100.0	34.2	51.0

Table 2: Performance results per topic on iconic image detection (percentages).

are checked for relevant SIFT features (SIFT-Contour). Finally, we group together the prototype with the seed icon of the corresponding topic in case at least two or three of the single matching strategies in i–iii) identify the same regions in the images. This process is repeated until all prototypes have been examined: prototypes for which the no match can be found are filtered out as being not iconic.

3 Evaluation

Dataset statistics. We first provide statistics on the size of the datasets created with our approach. Using HSV correlation we initially generate 1232 clusters with an average size of 27.37 elements per cluster. Additional filtering based on at least two of our image matching methods produces 870 clusters (19.33 elements on average), whereas the more restrictive clustering based on all three methods gives 261 small-sized clusters of only 5.8 instances on average. This is because, naturally, applying matching-based filtering tends to produce a smaller number of clusters with fewer elements.

Gold standard and filtering evaluation. To produce a gold standard for our task, we annotated all of the 4,000 images we retrieved from Flickr. Each image is associated with a keyword query (Section 2): accordingly, we annotated each instance as being iconic or not with respect to the topic expressed by the keywords – e.g., given a picture of Hopetoun Falls, whether it captures the concept of waterfall or not. This is because, in our work, we take keywords as proxies of the underlying topics (e.g., biodiversity is depicted using waterfalls): in this setting, negative instances consist of mismatches between the query text and the picture – e.g., a photography taken near Hopetoun Falls, showing beech trees and thus capturing a query search for “forest” rather than “waterfalls”.

We next evaluate our system on the binary classification task of detecting whether an image is iconic or not. In our case, we can quantify performance by taking all images not filtered out in the last step of image matching (and thus deemed as iconic in the final system output), and comparing them against our gold-standard annotations. This way we can compute standard metrics of precision, recall and balanced F-measure. Our results indicate that combining the output of two image matching techniques allows us to reach 59.5% recall and 68.5% precision, whereas requiring all three methods to match reduces precision (46.9%) while drastically decreasing recall (14.3%). The results show that our system is precision-oriented, and that filtering based on the combination of all methods leads to an overall performance degradation. This is because requiring all methods to match gives an over-constrained filtering: our methods, in fact, tend to match all together only with those images which are highly similar to the seeds, thus not being able to produce heterogeneous clusters.

We finally compute performance metrics for each single topic in turn, in order to experimentally investigate the different degrees of performance of our system, and determine whether some topics are more difficult than others (Table 2). Our results indicate that some topics are indeed more difficult than others – e.g., our system exhibits perfect precision on “adaptation” and “greenhouse effect” vs. much poorer one on “biodiversity” or “climate refugee”. This is because some topics are bootstrapped from less heterogeneous, and hence ‘easier’, sets of seed images (e.g., all smokestacks, as in “greenhouse effect”, are very similar to each other). In general, this seems to point out that one of the key challenges in our scenario is to produce highly precise clusters, while allowing for image diversity as a trade-off.

Error analysis. We finally looked at the output of our system, in order to better understand its performance, as well as problems and future challenges. Examples of a few sample clusters are shown in Figure 2. These clusters show that, thanks to our method, we are able to collect quite diverse, yet iconic images retaining a topical affinity with the original seeds – e.g., the poster on fighting deforestation or the drawing used to depict air pollution. Due to the noise of our base image processing components, however, we also suffer from wrong matches such as the picture of a mobile phone for the topic of wildfire, where the meaning of a rapidly spreading conflagration is related to air pollution, whereas the mobile phone is not. Based on a random sample of 10% of the output clusters, we manually identified the main sources of errors as related to: i) false image matching due to problems with contour detection; ii) SIFT performing best for detecting different images of the same objects, but exhibiting lower performance on the more complex task of detecting similar objects; iii) we applied our image matching methods using default parameters and thresholds: further improvements could be obtained by in-domain tuning.

4 Conclusions

In this work, we presented some initial steps in developing a methodology to computationally model the challenging phenomenon of iconic images. More specifically, we focused on the task of building a repository of iconic images in a minimally supervised way by bootstrapping based on images found on Web encyclopedic resources.

As future work, we plan to better combine heterogeneous information from text and images, as well as use deeper representations for both information sources – cf. joint semantic representations such as specific LDA-based topic models and bags of visual (SIFT) features (Rasiwasia et al., 2010; Feng and Lapata, 2010). This, in turn, can be applied to a variety of different tasks such as the automatic semantification of captions, query generation and expansion. On the computer vision side, we are instead particularly interested in exploring region-growing algorithms to detect textured or homogeneous regions, and to allow for a segmentation of textured regions without contours, e.g., a cloudy sky or a view of a forest landscape.

Downloads The dataset presented in this paper is freely available for research purposes at `https://madata.bib.uni-mannheim.de/87/`.

References

Roberto Brunelli. 2009. *Template Matching Techniques in Computer Vision: Theory and Practice*. Wiley Publishing.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, pages 886–893.

Yansong Feng and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Proc. of HLT ’10*, pages 831–839.

Ming-Kuei Hu. 1962. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

S. O’Neill and Nicholas Smith. 2014. Climate change and visual imagery. *Wiley Interdisciplinary Reviews: Climate Change*, 5(1):73–87.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proc. of MM ’10*, pages 251–260.

Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR*, pages 511–518.

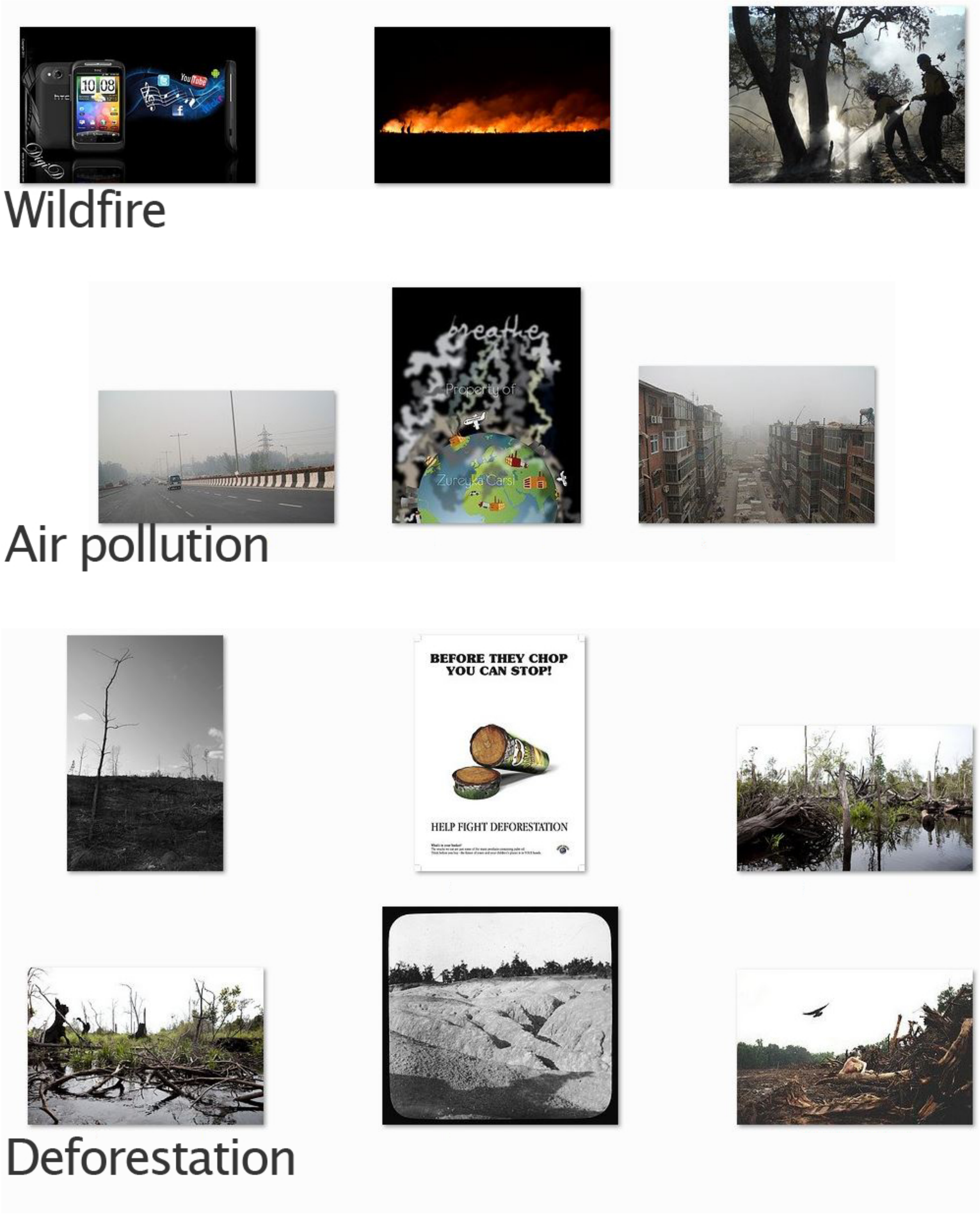

Figure 2: Sample iconic image clusters. Above a poor cluster on wildfire, below two good clusters on pollution and deforestation.