Enhancing Flood Impact Analysis using Interactive Retrieval of Social Media Images

Björn Barz, Kai Schröter, Moritz Münch, Bin Yang, Andrea Unger, Doris Dransch and Joachim Denzler

Abstract The analysis of natural disasters in a timely manner often suffers from limited sensor data. This limitation could be alleviated by leveraging information contained in images of the event posted on social media platforms, so-called "Volunteered Geographic Information (VGI)". To save the analyst from manual inspection of all images posted online, we propose to use content-based image retrieval with the possibility of relevance feedback for retrieving only relevant images of the event. To evaluate this approach, we introduce a new dataset of 3,710 flood images, annotated by domain experts regarding their relevance with respect to three tasks (determining the flooded area, inundation depth, water pollution). We compare several image features and relevance feedback methods on that dataset, mixed with 97,085 distractor images, and are able to improve the precision among the top 100 results from 55% to 87% after 5 rounds of feedback.

Björn Barz, Joachim Denzler
Friedrich Schiller University Jena, Computer Vision Group, Ernst Abbe Platz 2, 07743 Jena, Germany
✉ bjoern.barz,joachim.denzler@uni-jena.de

Kai Schröter, Moritz Münch
Deutsches GeoForschungsZentrum, Sec. Hydrology, Telegrafenberg, 14473 Potsdam, Germany
✉ kai.schroeter,moritz.muench@gfz-potsdam.de

Bin Yang, Andrea Unger, Doris Dransch
Deutsches GeoForschungsZentrum, Sec. Geoinformatics, Telegrafenberg, 14473 Potsdam, Germany
✉ bin.yang,andrea.unger,doris.dransch@gfz-potsdam.de

1 Introduction

The rapid analysis of recent or current natural disasters such as floods is crucial to provide information about impacts as a basis for efficient disaster response and recovery. With an increasing availability of data and information channels, the identification and exchange of core information and best possible up-to-date information is essential for disaster response (Turoff, 2002; Comfort et al, 2004). For recovery and improved disaster risk reduction, comprehensive image-based documentations of disaster dynamics help to gain insights and to improve our understanding of system behavior during extreme events. This knowledge is important to review and adapt flood prevention and protection concepts which are the basis to mitigate adverse consequences from flooding. However, event analyses often suffer from limited or insufficient data (Poser and Dransch, 2010; Thieken et al, 2016). For the case of flood mapping, traditional measuring devices as, for instance, water level gauges are expensive and hence only coarsely distributed. Malfunction and uncertainties of recordings during extreme events are known issues.

On the other hand, there is usually a large amount of complementary information that could be derived from images posted by volunteers on social media platforms (Assumpção et al, 2018). Since most modern consumer devices are GPS-enabled and store the geographical location where an image has been taken in its metadata, these images could be used to derive information about the flood at locations where the sensor coverage is insufficient, to substitute failures of measurements, or to complement other pieces of information (Schnebele and Cervone, 2013; Fohringer et al, 2015). The users of social media platforms posting images about natural disasters can thus be considered as *human sensors* providing so-called *volunteered geographic information (VGI)* (Goodchild, 2007), from which different types of information can be extracted for being combined with the data obtained from traditional sensors.

These data are potentially useful during all stages of disaster management (Poser and Dransch, 2010): To prepare for the case of a natural disaster, sufficient data about past events are necessary. During an event, the rapid availability of social media images would leverage monitoring the extent and intensity of the disaster and the current status of response activities. For post-disaster recovery, on the other hand, up-to-date damage estimates are required for financial compensation, insurance payouts, and reconstruction planning.

However, the sheer amount of almost 6,000 tweets posted every second on Twitter alone (Krikorian, 2013) renders inspecting all of them intractable, even when the set of images is restricted to those within a certain region and time-frame. Therefore, an automated filter retrieving only those images that are relevant for the analysis is highly desirable.

The notion of *relevance* is usually not fixed but depends on the objective currently pursued by the analysts. In the case of flood impact analysis, hydrologist might sometimes be interested in determining whether a certain area is flooded or not, which might be difficult to detect based on just a few water level measurements or due to mobile flood-protection walls that may alter the flooding process and expected inundation areas. However, while VGI images can be of great benefit for this task, just retrieving all images of the flooding is not sufficient in general. Because at another point of time, the information objective of the analysts might be to determine inundation depth as a key indicator of flooding intensity. In this regard, a different set of images would be relevant, showing visual clues for inundation depth such as partially flooded traffic-signs or people walking through the water. Another example is the task of determining the type and degree of water pollution from images, which changes the set of relevant image features drastically.

These image characteristics are sometimes difficult to verbalize in natural language. An example image, however, can often capture the search objective much more easily. Moreover, text-based search always runs the risk of missing relevant images with insufficient textual descriptions. Thus, we propose an approach based purely on the image content.

Since all the information objectives an analyst might have in mind constitute an open set, it is not possible to train a fixed set of classifiers for distinguishing between relevant and irrelevant samples. Instead, we propose an *interactive image retrieval* approach to assist the analyst in finding those images that are relevant with respect to the current task.
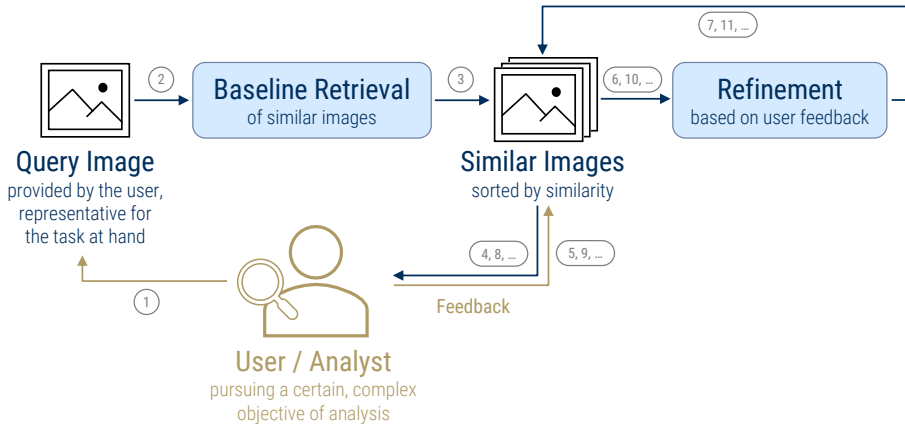
Figure 1: Schematic illustration of our interactive image retrieval process.

This procedure is illustrated in Fig. 1: The user first provides a so-called *query image* that should capture the search objective reasonably well. The system then extracts image features for this query and compares it with all other images in the database or social media image stream using the Euclidean distance between images in the feature space. The result is a list of retrieved images, ranked by their proximity to the query. This procedure is known as *content-based image retrieval* (Smeulders et al, 2000) and has been an active topic of research since 1993 (Niblack et al).

However, the results of this baseline retrieval will be suboptimal in most cases, since it is based on just a single query image. Thus, the system enables the user to flag some of the retrieved images as relevant or irrelevant. This information will then be incorporated to obtain a refined list of retrieval results, which should match the search interest pursued by the user more precisely. This step can be repeated until the user is satisfied with the result.

In this work, we investigate how to construct an image retrieval pipeline that is suitable for retrieving flood images by comparing several types of features extracted from deep neural networks for the baseline retrieval and various approaches to incorporate relevance feedback. To enable a quantitative evaluation, we introduce a novel dataset comprising 3,435 images of the European Flood 2013 from Wikimedia Commons plus 275 images showing water pollution from various sources. All images have been annotated by domain experts with respect to their relevance regarding three pre-defined information objectives.

The remainder of this paper is organized as follows: We will first briefly review related work on using VGI images for disaster management in Sect. 2. Our novel flood dataset is introduced in Sect. 3 and various baseline retrieval and relevance feedback methods are described and evaluated in Sect. 4. Section 5 concludes this work and discusses directions for future research.

2 Related Work

Many approaches for leveraging VGI from social media focus on linguistic patterns (e.g., Ireson, 2009), text-based classification (e.g., Sakaki et al, 2010; Yin et al, 2012) and keyword-based filtering (e.g., Vieweg et al, 2010; Fohringer et al, 2015).

Similar to our motivation, Schnebele and Cervone (2013) used volunteered data which have been retrieved using the photo, video and news Google search engines for a flood in Memphis (US) in May 2011. These information have been combined with remote sensing, digital elevation and other data to produce flood extent and flood hazard maps.

Twitter messages have been used by Brouwer et al (2017) to estimate flooding extents. While this approach uses only Twitter text message contents, it applies a set of keywords to filter relevant tweets. Geolocation information is derived from location references contained in the tweet.

Fohringer et al (2015) proposed to derive information about flood events from images posted on Twitter or Flickr and found them to contain "additional and potentially even exclusive information that is useful for inundation depth mapping". Likewise, Rosser et al (2017) retrieve geo-tagged imagery from Flickr using a defined study area and time window in combination with the keyword "flood". In this approach, only the image location is used to delineate flooded areas in combination with other data sources. However, both works do not employ any automatic image-based filtering but collect all tweets containing some predefined keywords and then analyze the relevance of all images included in these tweets manually. On the one hand, this tedious process is prohibitive for rapid flood impact estimation due to the time needed for inspecting all images. Further, the initial keyword-based filtering involves the risk of missing a large portion of relevant images due to the lack of matching keywords in the text.

We show how these issues can be overcome using computer vision techniques for filtering based on the image content only.

3 A Dataset for Flood Image Retrieval

A quantitative evaluation of our interactive image retrieval approach demands a sufficient number of both flood and non-flood images. In addition, we need to know for each image whether it is relevant for a certain task or not.

While obtaining non-flood-related images is rather easy, since any existing dataset such as, for example, the Flickr100k dataset (Philbin et al, 2007) comprising 100,031 images from Flickr could be used for that, finding a sufficient number of images relating to a certain flood event is more difficult. We used Wikimedia Commons as a source for flood images, since it already provides dedicated categories for major flood events, the images are released under a permissive Creative Commons license, and many of them contain geotags.

In the following subsections we describe how we collected flood images and annotated them with respect to their relevance regarding three exemplary information objectives. The dataset, including metadata and annotations, can be obtained at `https://github.com/cvjena/eu-flood-dataset`.

3.1 Collecting Flood Images

The Wikidata project strives towards creating machine-readable representations of all structured information present in Wikipedia. This information can be queried fully automatically using the SPARQL query language, which allows retrieving a list of all flood events recorded on Wikipedia with an associated Wikimedia Commons category. We can then use the Wikimedia Commons API to fetch all images and their metadata from those categories automatically. With a large margin, the highest number of images is available for the Central Europe floods of 2013[1], which comprises 3,855 images in total (as of July 2017). We hence decided to use this event as a basis for our flood dataset. After excluding sub-categories that relate exclusively to public transportation during the flood but do not show actual flooding, a total of 3,435 images remain.

However, these images do not show any water pollution, in which we are also interested. Thus, we added another set of 275 images to the dataset, which we

---

[1] `https://commons.wikimedia.org/wiki/?curid=26466898`
(accessed: July 21st, 2017)

have collected manually from the web by querying image search engines for the names of recent major oil spill events. To this end, we have again used a list of oil spills provided on Wikipedia[2].

3.2 Relevance Annotations

For a quantitative evaluation and comparison of several image retrieval methods, we need to simulate the behavior of a user of our proposed interactive image retrieval system. To enable such a simulation, we have defined a set of three common tasks, which could be pursued by a hydrologist using the system:

Flooded vs. dry Does the image help to determine whether a certain area is flooded or not? Usually, one would assume flooding of a certain area based on the intersection of the water level height and the elevation of the terrain. However, the area might actually be dry due to a flood-protection wall, for example. An image considered as relevant would show the boundary between flooded and dry areas. Images that do not show any inundation at all are considered not relevant. While these could be used to track the spread of the flood at a certain location over time, we only consider the individual relevance of images in this work, ignoring aspects that might become relevant when compared with other images in the dataset.

Inundation depth Is it possible to derive an estimate of the inundation depth from the image due to visual cues such as, for example, traffic signs or other structures with known height? If there is no flooding at all, the image is considered as not relevant for inundation depth.

Water pollution Does the image show any pollution substances? The focus is on heavy contamination by chemical substances such as oil, for example.

---

[2] https://en.wikipedia.org/w/index.php?title=List_of_oil_spills (accessed: May 30th, 2018)
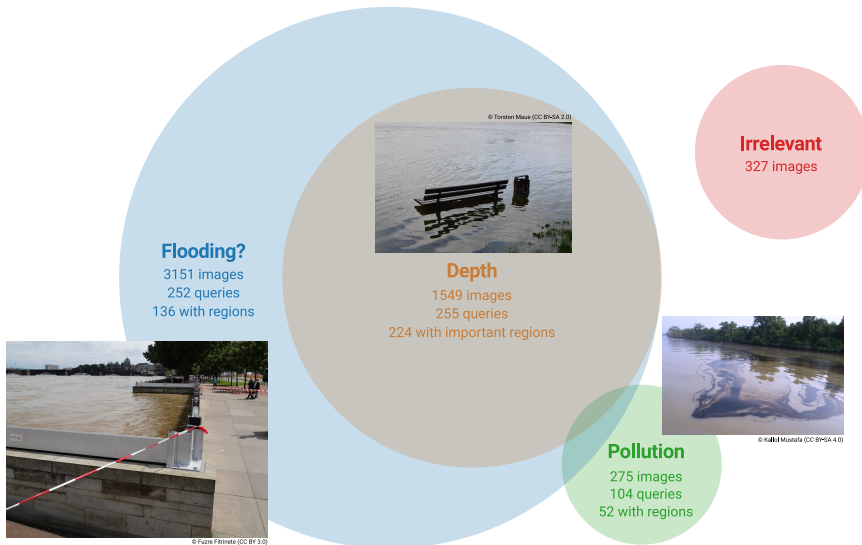
Figure 2: Venn diagram of the sets of images per task in our novel dataset.

Each image in the dataset has been assigned to one of several domain experts for annotation. Any image could be relevant for one, multiple, or none of the tasks described above.

Figure 2 shows the number of images marked as relevant for each task, the overlap of the categories, and an example image for each task. 9.5% of the images were found not to show any flooding situation despite being associated with the flood in Wikimedia Commons. We assigned the special label "irrelevant" to these images and treat them in the same way as the distractor images from the Flickr100k dataset (see Sect. 4).

Due to limited resources, we only obtained a single annotation for each image. However, an additional domain expert was asked to assure the quality of random samples from the set of annotations and to select between 100 and 250 ideal query images for each task that reflect the search objective well and could be used as initial query images for our image retrieval approach.

3.3 Important Image Regions

(a) Flooded vs. dry.   (b) Inundation depth.   (c) Water pollution.

Figure 3: Examples for annotations of important image regions.

Besides relevance annotations for images as a whole, we have also asked one domain expert to highlight important regions on some of the images selected as queries for each task. This aims to account for the fact that the relevance of a certain image is often due to a particular small part of the image without which it would not be relevant at all, e.g., partially flooded traffic signs in the case of the inundation depth task. We also allowed the expert to mark multiple relevant regions per image and to create groups of regions that have to be present together in a single image for being relevant. Example annotations are shown in Fig. 3b. We do not make use of these region-level information in our image retrieval system at the moment, but plan to do so in the future.

4 Interactive Image Retrieval

In the following, we describe and compare several methods for the two components of our interactive image retrieval pipeline depicted in Fig. 1: Constructing a feature space for the baseline retrieval of similar images and incorporating relevance feedback provided by the user.

All methods are evaluated on a combination of our novel flood dataset introduced in Sect. 3 and images from Flickr100k (Philbin et al, 2007) as distractors. While Flickr100k comprises a total of 100,031 images, we excluded those tagged with "river" or "water", since some of them show flooding situations.

After this, 97,085 distractor images remain, which we do not expect to show flooding given their tags. The set of flood-related images from our novel dataset hence accounts for as few as 4% of the combined dataset.

We employ the *normalized discounted cumulative gain* (Järvelin and Kekäläinen, 2002) among the top 100 results (NDCG@100) as performance metric, which does not only measure the fraction of relevant images among the top 100 results but considers their order as well, assigning higher weights to earlier positions in the ranking. For a query q and a ranked list of $n \geq k$ retrieved images with relevance labels $y_i \in \{0, 1\}, i = 1, \ldots, n$, the NDCG@$k$ is defined as:

$$\text{NDCG@}k(y_1, \ldots, y_n \mid q) = \left(\sum_{i=1}^{k} \frac{y_i}{\log_2(i + 1)} \right) \Big/ \left(\sum_{i=1}^{\min\{k, |R(q)|\}} \frac{1}{\log_2(i + 1)} \right), \tag{1}$$

where $R(q)$ denotes the set of all images relevant for the query q. The best NDCG hence is 1.0 and the worst is 0.0. We cap the ranking at $k = 100$ since the advantage of our image retrieval system for finding relevant flood images vanishes if the user has to inspect more than 100 results.

In the following, we always report the average NDCG@100 over all 611 query images from our dataset identified as suitable by the domain experts, which are issued as individual queries to the system. Images from the dataset are considered as relevant with respect to a certain query if they are assigned to the label for which the query image has been selected as "ideal example".

4.1 Baseline Retrieval

The main challenge of content-based image retrieval (CBIR) is constructing a feature space where similar images lie close together, so that retrieval can be performed by searching for the nearest neighbors of the query in that space. The notion of similarity is often fuzzy and depends on the application. This relation is most often defined as two images either showing the same object, objects of the same class, or being "visually similar", which is difficult to formalize.

Traditional CBIR approaches usually consist in detecting invariant keypoints in an image, extracting handcrafted local descriptors from the neighborhood of these keypoints, embedding them in a high-dimensional space, and finally

aggregating them into a single global image descriptor. A good summary of these approaches has been given by Babenko and Lempitsky (2015).

In the past few years, however, such approaches have been outperformed by deep-learning-based image features extracted from a convolutional neural network (CNN) (LeCun et al, 1989) pre-trained on a classification task. A CNN typically consists of a sequence of convolution operations with learned filters and non-linear activation functions in-between. After certain layers, the feature map is sub-sampled using local pooling operations. The result of the last convolutional layer is hence a low-resolution map of features for different regions of the image. These local feature vectors are then aggregated by averaging with either uniform or learned weights and fed through a sequence of so-called fully-connected layers, which essentially realize a multiplication of the features with a learned matrix followed by a non-linear activation function. In classification scenarios, the output of the final layer is interpreted as the logits of a probability distribution over the classes. The entire network is trained end-to-end using backpropagation (LeCun, 1985), so that all the intermediate feature representations are learned from data and optimized for the task at hand, where the degree of abstraction from visual to semantic features increases with the depth in the network (Zeiler and Fergus, 2014).

Surprisingly, Babenko et al (2014) found these features, which they extracted from the fully-connected layers of a pre-trained network, to also perform competitively for the task of content-based image retrieval. The traditional CBIR approaches with handcrafted features were finally outperformed using local CNN features extracted from the last convolutional layer (Babenko and Lempitsky, 2015). These need to be aggregated into a global image descriptor first, which provides additional leeway for adapting the pre-trained features to the retrieval scenario. Besides simple average and maximum pooling, a variety of sophisticated pooling functions has been proposed in the past few years. In this work, we evaluate the following ones:

Average Pooling (avg) Uniform average (Babenko and Lempitsky, 2015).

Partial Mean Pooling (PMP) Averaging over the top 10% highest activations per channel (Zhi et al, 2016). This combines average and maximum pooling.

Generalized-Mean Pooling (GeM) Using the L^p-norm of each spatial feature map (Radenović et al, 2018), which generalizes between average (for $p = 1$)

and maximum (for $p \rightarrow \infty$) pooling. We have empirically found $p = 2$ to work well on our dataset.

Adaptive Co-Weighting (adacow) Combination of spatial and channel-wise weighting, where the spatial weights are based on the sum of activations at each position and channel weights are determined in a way so that frequently occurring bursty features get a low weight (Wang et al, 2018).

In all cases, we extract the local features to be aggregated from the last convolutional layer of the so-called *VGG16* convolution neural net architecture (Simonyan and Zisserman, 2014), pre-trained[3] for classification on millions of images from the ImageNet dataset (Russakovsky et al, 2015). This is the network architecture that has initially been used by Babenko et al (2014) and Babenko and Lempitsky (2015) for the first CBIR approaches using neural features and has remained popular until today. We also evaluate global image features extracted from the first fully-connected (FC) layer of the same CNN, as done by Babenko et al (2014). This corresponds to a complex aggregation function with learned weights for both feature dimensions and spatial positions. Regardless of the aggregation function being used, we always L^2-normalize the final global image descriptors, which has proven to be beneficial for image retrieval, because the direction of high-dimensional feature vectors often carries more information than their magnitude (Jégou and Zisserman, 2014; Horiguchi et al, 2019)

Besides the use of pre-trained CNNs for feature extraction, neural networks trained end-to-end specifically for image retrieval have shown superior performance recently. In this regard, we evaluate the approach of Gordo et al (2017), who extended a ResNet-101 architecture (He et al, 2016) with R-MAC pooling (sum-pooling over maximum-pooled features from several regions of interest; Tolias et al, 2016), followed by PCA and L^2-normalization. This network has been trained for image retrieval on a landmarks dataset using a triplet loss, which enforces similar images to be closer together in the feature space than dissimilar ones. We denote this approach as "Deep R-MAC"[4].

---

[3]   The   pre-trained   VGG16   model   can   be   obtained   at http://www.robots.ox.ac.uk/vgg/research/very_deep/ (accessed July 10th, 2019).

[4] The pre-trained Deep R-MAC model can be obtained at https://github.com/figitaki/deep-retrieval (accessed July 10th, 2019).
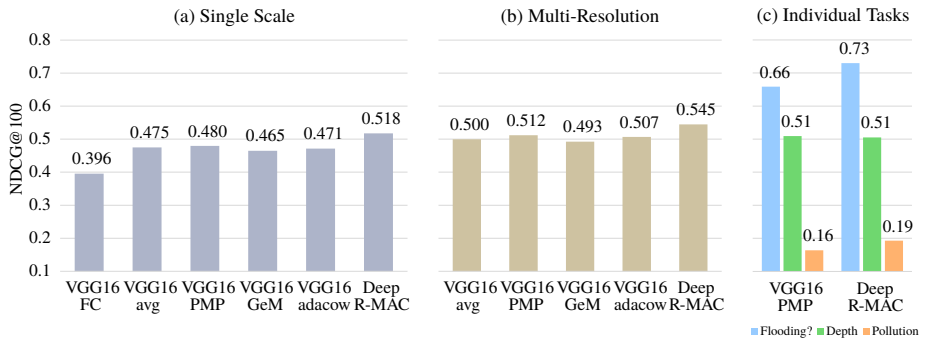
Figure 4: Baseline retrieval performance.

The performance comparison in Fig. 4a shows that aggregated convolutional features perform significantly better than features extracted from fully-connected layers which are presumably already too class-specific. The choice of the pooling method makes only a slight difference, while PMP performed best. Features from the Deep R-MAC network fine-tuned for object retrieval provided even better performance than VGG16, resulting in an NDCG@100 of 51.8% for the simple baseline retrieval.

For these experiments, all images have been resized so that their larger side is 512 pixels wide, except for fully-connected (FC) pooling, which only works with rather small images of size 224×224 due to the fixed number of learned weights. Following Gordo et al (2017) we have also evaluated averaging image descriptors extracted from 3 differently scaled versions of the same image, where we resized the larger side to 550, 800, and 1050 pixels.

The results in Fig. 4b show that the use of multiple resolutions leads to an absolute improvement of NDCG@100 by about 3%, regardless of the features.

Since the number of images and queries per task in our flood dataset is not balanced (cf. Fig. 2), we also report the per-task performance of the two best-performing types of features in Fig. 4c. Obviously, finding images relevant for pollution is much more difficult than the other two tasks, which we do not solely attribute to the small number of relevant images, but also to the fact that images of oil films are easily confused with photos of abstract art from Flickr.

4.2 Relevance Feedback

Approaches for incorporating relevance feedback into image retrieval can usually be divided into four categories:

Query Point Movement The query vector is modified based on the feedback, e.g., by averaging over the features of all images marked as relevant (Rocchio, 1971). These approaches belong to the oldest ones in information retrieval, but since their use is very limited, we will not address them in this work.

Probabilistic The distribution of the probability that a particular image is relevant given the feedback is estimated. Here, we investigate the simple kernel density estimation (KDE) method proposed by Deselaers et al (2008).

Classification A classifier is trained for distinguishing between relevant and irrelevant images. In this work, we investigate two approaches for this: an Exemplar-LDA classifier (Hariharan et al, 2012) and a support vector machine (SVM, Cortes and Vapnik, 1995), falling back to a One-Class SVM (Schölkopf et al, 2001) if only positive feedback is given.

Metric Learning A new metric $d : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ is applied to the D-dimensional feature space $\mathbb{R}^D$, minimizing the distance between relevant images and maximizing the distance between relevant and irrelevant ones. Many approaches use a Mahalanobis metric of the form $d_M(x_1, x_2) = (x_1 - x_2)^\top M(x_1 - x_2)$ and learn a positive semi-definite matrix $M \in \mathbb{R}^{D \times D}$. This is equivalent to a linear transformation of the data into a new space where the Euclidean distance corresponds to d_M in the original space.

In this work, we investigate the feature weighting approach of Deselaers et al (2008), the diagonal variant of MMC (Xing et al, 2003), and information-theoretic metric learning (ITML, Davis et al, 2007). The first two approaches learn a diagonal matrix M, which corresponds to a weighting of individual features, but use different objectives and optimization algorithms: Deselaers et al (2008) minimize the ratio of distances between similar and dissimilar samples using gradient descent, while Xing et al (2003) employ a convex optimization objective minimizing the distance of similar samples while keeping dissimilar ones away by at least a fixed radius. ITML (Davis et al, 2007), in contrast, learns a full matrix M, so that similar pairs are

closer than a certain threshold and dissimilar ones are apart by at least another threshold. This is possible despite the high dimensionality of the feature space and limited annotations thanks to regularization towards the Euclidean distance as a prior metric. We choose the two thresholds needed for ITML on a per-query basis as follows: All pairs of relevant images should be closer to each other than half the distance between the query and the first irrelevant retrieval result. The distance of all images tagged as irrelevant to any relevant image should be greater than the 95th percentile of the distances between the query and all images in the dataset.

Additionally, we investigate combinations of the three metric learning methods with the KDE-based approach of Deselaers et al (2008).

The annotations of our flood dataset allow us to completely simulate the feedback process for a quantitative evaluation: For all the 611 images denoted as ideal queries, we first perform the baseline retrieval and then mark 10 random images out of the top 100 results either as relevant or irrelevant according to their labels. This is repeated for a total number of 10 feedback rounds and the retrieval quality is evaluated after each round in terms of the NDCG@100. To get an impression of the variance of the results, we repeat the entire experiment 10 times with different random sub-samples of 75% of the dataset.

Based on the findings from the previous section, we use the two best-performing types of features (Deep R-MAC and PMP on the last convolutional layer of VGG16), averaged over multiple image scales.
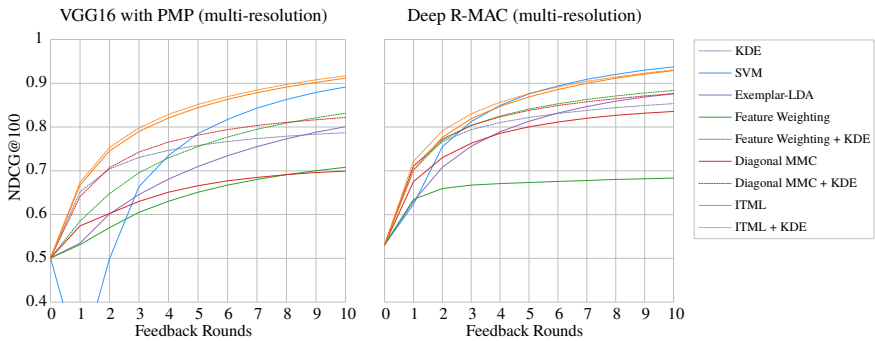
Figure 5: Comparison of methods for incorporating relevance feedback.

The results averaged over the 10 repetitions are shown in Fig. 5, where round 0 denotes the performance of the baseline retrieval. It can be seen that the simple KDE method already performs quite well, especially better than the two feature weighting techniques. The two classification-based approaches (SVM and Exemplar-LDA) apparently suffer from the limited amount of annotations and behave extremely unstable during the first rounds. ITML, on the other hand, provides superior performance from the beginning and leads to an NDCG@100 of 86.9% after 5 rounds of feedback and 92.9% after 10 rounds. Though KDE can be added on top of any other method, the benefit when combined with ITML is too marginal to justify the computational overhead.

After 5 rounds of feedback, the SVM-based approach starts to outperform ITML slightly when using Deep R-MAC features. However, the performance during the first few rounds is of greater importance, since we do not expect most users to regularly spend more than five rounds of feedback for refining the results. During the early iterations, however, SVM performs worst among all methods. It might hence be an interesting direction for future work to investigate how ITML and SVM can be combined to improve performance at all stages of the process.

The maximum standard deviation of all methods and all feedback rounds over the 10 repetitions was 1.2%. We conducted a paired Student's t-test to assess the significance of the differences between the methods in Fig. 5 at a significance level of 5%. At the final feedback round, all differences are significant except that between Exemplar-LDA and Diagonal MMC + KDE when using Deep R-MAC features. At the first round, all differences are significant for VGG16 features and all besides that between ITML and Feature Weighting + KDE for Deep R-MAC features. For VGG16 features, ITML and ITML + KDE performed significantly better than all other methods across all rounds. With the Deep R-MAC features, ITML coincided with Feature Weighting + KDE at round 1 and with SVM at round 4, but was otherwise significantly different from the rest. SVM started to perform significantly better than the rest from round 7 on. Besides that, Feature Weighting, Feature Weighting + KDE, and Diagonal MMC, performed significantly different from the rest in at least 9 of 10 rounds.

5 Conclusions and Future Work

We have proposed an interactive image retrieval approach with relevance feedback for finding flood images on online image platforms that are relevant for a particular information interest. To evaluate our approach, we have presented a novel dataset comprising 3,710 flood images annotated with relevance labels regarding three exemplary search objectives and important image regions.

For the baseline retrieval, Deep R-MAC features (Gordo et al, 2017) averaged over multiple image scales perform best. Convolutional features extracted from other networks not fine-tuned for object retrieval can also perform well when aggregated using partial mean pooling (Zhi et al, 2016).

Regarding the incorporation of relevance feedback, an SVM-based approach provides the best performance in the long run, but needs a substantial amount of feedback for being useful. Information-theoretic metric learning (Davis et al, 2007), on the other hand, provides superior performance during the early feedback rounds and remains competitive with SVM later on. Finally, the simple KDE method of Deselaers et al (2008) has turned out to be a quick and decent baseline as well, which is particularly easy to implement and combine with existing frameworks. Using relevance feedback, the average NDCG@100 can be improved from 55% yield by the baseline retrieval to 87% after five rounds and 93% after ten rounds of feedback, which we expect to be useful for hydrologists to find relevant images quickly.

In the future, we would like to investigate how the selection of important image regions can be integrated as an additional component of the system to improve the relevance of the retrieved images even further, as proposed by Freytag et al (2015), for example. It seems also appealing to combine ITML with the SVM-based approach to improve the performance at all stages of the feedback process. Moreover, it seems promising to apply active learning methods for asking the user for feedback regarding certain actively selected images from which the system expects the most benefit. Finally, the interactive image retrieval system should be integrated into a visual analytics interface providing data from other sensors as well, enabling a case-study on a more recent flood event with real users.

References

Assumpção TH, Popescu I, Jonoski A, Solomatine DP (2018) Citizen observations contributing to flood modelling: opportunities and challenges. Hydrology and Earth System Sciences 22(2):1473–1489. ISSN: 1027-5606, DOI: 10.5194/hess-22-1473-2018.

Babenko A, Lempitsky V (2015) Aggregating local deep features for image retrieval. In: IEEE International Conference on Computer Vision (ICCV), IEEE, pp. 1269–1277. ISBN: 978-1-467383-91-2, ISSN: 2380-7504, DOI: 10.1109/ICCV.2015.150.

Babenko A, Slesarev A, Chigorin A, Lempitsky V (2014) Neural codes for image retrieval. In: European Conf. Computer Vision (ECCV), Springer, Cham, pp. 584–599. ISBN: 978-3-319105-89-5, DOI: 10.1007/978-3-319-10590-1_38.

Brouwer T, Eilander D, Van Loenen A, Booij MJ, Wijnberg KM, Verkade JS, Wagemaker J (2017) Probabilistic flood extent estimates from social media flood observations. Natural Hazards and Earth System Sciences 17(5):735–747, Copernicus Publications. DOI: 10.5194/nhess-17-735-2017.

Comfort LK, Ko K, Zagorecki A (2004) Coordination in Rapidly Evolving Disaster Response Systems: The Role of Information. American Behavioral Scientist 48(3):295–313. ISSN: 0002-7642, DOI: 10.1177/0002764204268987.

Cortes C, Vapnik V (1995) Support-vector networks. Machine learning 20(3):273–297, Kluwer Academic Publishers. ISSN: 0885-6125, DOI: 10.1007/BF00994018.

Davis JV, Kulis B, Jain P, Sra S, Dhillon IS (2007) Information-theoretic metric learning. In: Proceedings of the 24th international conference on Machine learning, ACM, New York, ICML '07, pp. 209–216. ISBN: 978-1-595937-93-3, DOI: 10.1145/1273496.1273523.

Deselaers T, Paredes R, Vidal E, Ney H (2008) Learning weighted distances for relevance feedback in image retrieval. In: International Conference on Pattern Recognition (ICPR), IEEE, pp. 1–4. ISBN: 978-1-424421-74-9, ISSN: 1051-4651, DOI: 10.1109/ICPR.2008.4761730.

Fohringer J, Dransch D, Kreibich H, Schröter K (2015) Social media as an information source for rapid flood inundation mapping. Natural Hazards and Earth System Sciences 15(12):2725–2738. DOI: 10.5194/nhess-15-2725-2015.

Freytag A, Schadt A, Denzler J (2015) Interactive Image Retrieval for Biodiversity Research. In: Pattern Recognition, Gall J, Gehler P, Leibe B (eds), Springer International Publishing, Cham, pp. 129–141. ISBN: 978-3-319249-47-6, DOI: 10.1007/978-3-319-24947-6_11.

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221. ISSN: 1572-9893, DOI: 10.1007/s10708-007-9111-y.

Gordo A, Almazán J, Revaud J, Larlus D (2017) End-to-End Learning of Deep Visual Representations for Image Retrieval. International Journal of Computer Vision 124(2):237–254. ISSN: 1573-1405, DOI: 10.1007/s11263-017-1016-8.

Hariharan B, Malik J, Ramanan D (2012) Discriminative decorrelation for clustering and classification. In: European Conference on Computer Vision, Springer, Berlin, pp. 459–472. ISBN: 978-3-642337-64-2, DOI: 10.1007/978-3-642-33765-9_33.

He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. ISBN: 978-1-467388-51-1, ISSN: 1063-6919, DOI: 10.1109/CVPR. 2016.90.

Horiguchi S, Ikami D, Aizawa K (2019) Significance of Softmax-based Features in Comparison to Distance Metric Learning-based Features. Transactions on Pattern Analysis and Machine Intelligence, IEEE. ISSN: 0162-8828, DOI: 10.1109/TPAMI. 2019.2911075.

Ireson N (2009) Local community situational awareness during an emergency. In: 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, pp. 49–54. DOI: 10.1109/DEST.2009.5276763.

Järvelin K, Kekäläinen J (2002) Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) 20(4):422–446, ACM, New York, NY, USA. ISSN: 1046-8188, DOI: 10.1145/582415.582418.

Jégou H, Zisserman A (2014) Triangulation embedding and democratic aggregation for image search. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 3310–3317. ISBN: 978-1-479951-18-5, ISSN: 1063-6919, DOI: 10.1109/CVPR.2014.417.

Krikorian R (2013) New Tweets per second record, and how! URL: `https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html` [accessed 2018-08-30].

LeCun Y (1985) Une procedure d'apprentissage pour reseau a seuil asymmetrique (A learning scheme for asymmetric threshold networks). In: Proceedings of Cognitiva 85.

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1(4):541–551. DOI: 10.1162/neco.1989.1.4.541.

Niblack CW, Barber R, Equitz W, Flickner MD, Glasman EH, Petkovic D, Yanker P, Faloutsos C, Taubin G (1993) QBIC project: Querying images by content, using color, texture, and shape. In: Storage and retrieval for image and video databases, International Society for Optics and Photonics, Vol. 1908, pp. 173–188. DOI: 10. 1117/12.143648.

Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object Retrieval with Large Vocabularies and Fast Spatial Matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 1–8. ISSN: 1063-6919, DOI: 10.1109/CVPR.2007.383172. Dataset online available: `http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/flickr100k.html`.

Poser K, Dransch D (2010) Volunteered geographic information for disaster management with application to rapid flood damage estimation. Geomatica 64(1):89–98. URL: `http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:240838`.

Radenović F, Tolias G, Chum O (2018) Fine-tuning CNN Image Retrieval with No Human Annotation. Transactions on Pattern Analysis and Machine Intelligence 41(7):1–1, IEEE. ISSN: 0162-8828, DOI: 10.1109/TPAMI.2018.2846566.

Rocchio JJ (1971) Relevance feedback in information retrieval. The SMART retrieval system: Experiments in automatic document processing, pp. 313–323, Prentice-Hall Inc.

Rosser JF, Leibovici D, Jackson M (2017) Rapid flood inundation mapping using social media, remote sensing and topographic data. Natural Hazards 87(1):103–120, Springer, Netherlands. ISSN: 0921-030X, DOI: 10.1007/s11069-017-2755-0.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3):211–252. ISSN: 1573-1405, DOI: 10.1007/s11263-015-0816-y.

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: Proceedings of the 19th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '10, pp. 851–860. ISBN: 978-1-605587-99-8, DOI: 10.1145/1772690.1772777.

Schnebele E, Cervone G (2013) Improving remote sensing flood assessment using volunteered geographical data. Natural Hazards and Earth System Sciences 13(3):669–677, Copernicus Publications. DOI: 10.5194/nhess-13-669-2013.

Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. Neural computation 13(7):1443–1471, MIT Press. DOI: 10.1162/089976601750264965.

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Smeulders AW, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22(12):1349–1380, IEEE. ISSN: 0162-8828, DOI: 10.1109/34.895972.

Thieken AH, Bessel T, Kienzler S, Kreibich H, Müller M, Pisi S, Schröter K (2016) The flood of June 2013 in Germany: how much do we know about its impacts? Natural Hazards and Earth System Sciences 16(6):1519–1540. ISSN: 1684-9981, DOI: 10.5194/nhess-16-1519-2016.

Tolias G, Sicre R, Jégou H (2016) Particular object retrieval with integral max-pooling of CNN activations. In: International Conference on Learning Representations (ICLR), San Juan.

Turoff M (2002) Past and Future Emergency Response Information Systems. Communications of the ACM 45(4):29–32. DOI: 10.1145/505248.505265.

Vieweg S, Hughes AL, Starbird K, Palen L (2010) Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, CHI '10, pp. 1079–1088. ISBN: 978-1-605589-29-9, DOI: 10.1145/1753326.1753486.

Wang J, Zhu J, Pang S, Li Z, Li Y, Qian X (2018) Adaptive Co-weighting Deep Convolutional Features For Object Retrieval. CoRR abs/1803.07360.

Xing EP, Jordan MI, Russell SJ, Ng AY (2003) Distance metric learning with application to clustering with side-information. In: Advances in neural information processing systems, MIT Press, Cambirdge, NIPS'02, pp. 521–528.

Yin J, Lampert A, Cameron M, Robinson B, Power R (2012) Using Social Media to Enhance Emergency Situation Awareness. IEEE Intelligent Systems 27(6):52–59. ISSN: 1541-1672, DOI: 10.1109/MIS.2012.6

Zeiler MD, Fergus R (2014) Visualizing and Understanding Convolutional Networks. In: European Conference on Computer Vision (ECCV), Springer, Cham, pp. 818–833. ISBN: 978-3-319105-89-5.

Zhi T, Duan LY, Wang Y, Huang T (2016) Two-stage pooling of deep convolutional features for image retrieval. In: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, pp. 2465–2469. ISBN: 978-1-467399-61-6, ISSN: 2381-8549, DOI: 10.1109/ICIP.2016.7532802.