Visual and textual analysis of social media and satellite images for flood detection @ multimedia satellite task MediaEval 2017

Konstantinos Avgerinakis[1], Anastasia Moumtzidou[1], Stelios Andreadis[1],
Emmanouil Michail[1], Ilias Gialampoukidis[1], Stefanos Vrochidis[1], Ioannis Kompatsiaris[1]
[1]Centre for Research & Technology Hellas - Information Technologies Institute, Greece
koafgeri@iti.gr,moumtzid@iti.gr,andreadisst@iti.gr
michem@iti.gr,heliasgj@iti.gr,stefanos@iti.gr,ikom@iti.gr

ABSTRACT

This paper presents the algorithms that CERTH team deployed in order to tackle disaster recognition tasks and more specifically Disaster Image Retrieval from Social Media (DIRSM) and Flood-Detection in Satellite images (FDSI). Visual and textual analysis, as well as late fusion of their similarity scores, were deployed in social media images, while color analysis in the RGB and near-infrared channel of satellite images was performed in order to discriminate flooded from non-flooded images. Deep Convolutional Neural Network (DCNN), DBpedia Spotlight and combMAX was implemented to tackle DIRSM, while Mahalanobis Distance-based classification and morphological post-processing were applied to deal with FDSI.

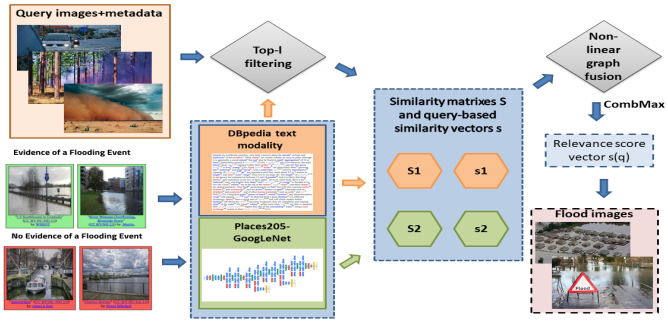

Figure 1: Block diagram of our multimodal retrieval system

1 INTRODUCTION

Security, surveillance and more specifically disaster prediction and classification from social media and satellite sources have raised a lot of interest in the computer science the last decade. The unobtrusive and abundant nature of these data rendered them as one of the most valuable sources to extract and deduct early warning or identification of an ongoing or eminent disaster.

Multimedia satellite task is a challenge of MediaEval that comprises of two tasks: (a) Disaster Image Retrieval from Social Media (DIRSM) and (b) Flood-Detection in Satellite Images (FDSI). DIRSM provides a great amount of social media images (YFCC100M-Dataset) and their metadata (Flickr), while FDSI is comprised of a large amount of 4 colour-channel, 3 for the RGB spectrum and 1 for the near-infrared, satellite images from PlanetLabs [5]. Both tasks ask from the participants to leverage any available technology so as to determine whether a flood event occurs in the provided test data. As far as visual data are concerned, a flood event is considered when an image shows an "unexpected high water level in industrial, residential, commercial and agricultural areas". The reader is suggested to read [1] for further information about the contest and the provided data.

In this work, CERTH presents its algorithms for DIRSM and FDSI subtasks. For flood recognition in images, CERTH uses the output of the last pooling layer of a trained GoogleNet [4] for global keyframe representation and trains an SVM classifier to recognize images that are related to a flooding event. Textual information is also retrieved by leveraging the metadata of the social media images by using DBpedia Spotlight annotation tool [2]. Both of these modalities are fused with a novel multimodal approach which combines non-linear graph-based fusion [3] with combMax scoring. For FDSI subtask CERTH performs a Mahalanobis distance classification and several morphological and adaptive filters, so as to separate flood from non-flood areas inside satellite image scene.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

2 APPROACH

2.1 Flood detection from social media (DIRSM)

Social media were crawled in this task so as to acquire images and text about flood scenarios. For that purposes, two modalities were deployed and fused with a non-linear graph-based fusion approach.

The first modality concerned visual analysis and more specifically flood detection inside image samples by adopting a Deep Convolutional Neural Network (DCNN) framework. GoogleNet [4] was trained on 5055 ImageNet concepts, and the output of the last pooling layer with dimension 1024 was used as a global keyframe representation. The provided development set was then splitted into two subsets and used to train an SVM classifier and define its optimal parameters: t (defines the kernel type) and g (gamma in kernel function). The best results were achieved for $t = 1$ (polynomial function) and $g = 0.5$. The test environment that CERTH built, included the evaluation of the precomputed features provided from the Multimedia-Satellite challenge (i.e. acc, gabor, fcth, jcd, cedd, eh, sc, cl, and tamura) and DCNN features that were produced from the $Places205 - GoogLeNet$ network by fusing the features from the convolutional layers $3a$ and $3b$. SVM classifiers were trained for all of these features and results showed that the proposed DCNN feature outperformed most of them significantly.

The second modality concerns the detection of flood-related text in social media metadata. For that purposes, DBpedia Spotlight [2]

was adapted so as to detect flood, water and related keyphrases that were acquired from the training set metadata (i.e. title, description, user tags). A disambiguation algorithm followed up to compare the aforementioned phrases with the collection, using Jaccard similarities. The similarity scores of the two modalities were also combined with the use of a late fusion approach that uses non-linear graph based techniques (random walk, diffusion-based) in a weighted non-linear way [3]. The top-l multimodal objects are filtered with respect to textual concepts, leading to $l \times l$ similarity matrixes S_1, S_2 and query-based $l \times 1$ similarity vectors s_1 and s_2. More specifically, 10 positive examples were selected from the training set as queries so as to acquire 10 ranked lists and by using combMAX late fusion to get the final list of relevant-to-the-flood multimodal objects. The overall block diagram of this approach is depicted in Fig. 1.

2.2 Flood detection from satellite images (FDSI)

Satellite images were collected from PlanetLabs [5] so that we can evaluate our localization algorithm in real case scenarios. Localization is based on a Mahalanobis classification framework and post-processing morphological operations.

Mahalanobis distances with stratified covariance estimates were computed to train our classifier by randomly selecting 10000 samples (RGB and infrared pixels) from each 7 sets of satellite images, leading into a final population of 70000 samples. Linear, diagonal linear, quadratic and diagonal quadratic discriminant functions were also computed, but Mahalanobis distances achieved the highest classification results. For every image of the testing set all pixels of the image were extracted, creating a four dimensional (R,G,B,NI) testing set consisting of 102400 samples (320 pixels $\times$ 320 pixels) per image. The final outcome was a binary mask that denoted 1 for flooded pixels and 0 for non-flooded ones.

Post-processing was then deployed on the acquired binary masks, in order to eliminate erroneous areas that resulted from the noisy nature of the dataset. A global filter was initially deployed on the binary mask so as to eliminate population of flood-denoted pixels that as a whole did not surpass the 5% of the image size. Similarly, a local filter followed up so as to eliminate the connected components of flood-denoted areas that did not surpass the size of 10 pixels. Image dilation and erosion was finally applied around each pixel and its surrounding area (circular cell with radius of 4 pixels) to eliminate small areas that were falsely denoted as flood, but simultaneously preserve the larger ones.

3 RESULTS AND ANALYSIS

Social media results for flood situations (DIRSM) are gathered in Table 1. Two retrieval approaches were used; (a) single cutoff scheme that returns the top-480 most similar samples and (b) multiple cutoff scheme that combines results from 4 different thresholds equal to 50, 100, 250, 480 by averaging their scores so as to conclude into a final list.

It is obvious that multiple cutoffs worked better than a single. Furthermore, we can observe that visual modality surpassed the textual by far and this is mainly attributed to the fact that some keywords related to flood and water might be found under several irrelevant contexts, leading text retrieval to very low accuracy rates. Fusion is also affected by the low performance of the textual modality and cannot leverage or complement the visual information in the final deduction, leading to lower accuracy rates than visual does.

Table 1: CERTH results in DIRSM task

Modalities	single cutoff	several cutoffs
Visual	78.82%	92.27%
Textual	36.15%	39.90%
Fusion	68.57%	83.37%

Table 2: CERTH results in FDSI task

loc01	loc02	loc03	loc04	loc05	loc06	loc07
81.71%	68.33%	82.08%	47.01%	45.84%	64.92%	56.27%

Results from Satellite images (FDSI) are presented in Table 2. The accuracy rates are quite diverse amongst them as we acquired very high rates in some locations such as loc01 and loc03, while other ones such as loc04 and loc05 were too low. From our point of view, this is attributed to the colour nature of the data in these areas, as in the former the separation of water was clear, while in the latter non-flood areas had similar colour with the flood ones. Furthermore, groundtruth masks included some non-flood pixels as flood and as the nature of our algorithm is pixel-wise they were misclassified as positive samples lead to poor performance models. Overall, our classifier lead to 74.67% localization accuracy rate.

4 DISCUSSION AND OUTLOOK

Multimedia satellite challenge gave as the opportunity to test our algorithm in real case disaster scenarios. Social media and satellite sources proved extremely valuable and helped us separate flood scenarios from others. The high average precision rate that visual features achieved proves that computer vision community can become ever more helpful in disaster detection and it is clear now that can surpass the ambiguity that text can introduce in the decision feature. On the other hand, satellite images proved quite noisy and require deeper investigation in the future.

As a future work, we plan to adopt deeper techniques that exist in the literature to recognize and discriminate places from each other, while we also plan to investigate hybrid representations that combine shallow with deep features so as to achieve even higher precision rates in the visual part of the system. Text approaches should undoubtedly revised and get tailored to disaster related scenarios, while fusion approaches that consider "semantic filtering" stages based on textual concepts will be revised. Regarding FDSI, we plan to build a shallow/deep representation scheme that will leverage both texture (i.e. LBP) and deep features so as to learn to separate flood from non-flood areas even more effectively.

ACKNOWLEDGMENTS

This work is supported by beAWARE project, partially funded by the European Commission (H2020-700475).

REFERENCES

[1] Benjamin Bischke, Patrick Helber, Christian Schulze, Srinivasan Venkat, Andreas Dengel, and Damian Borth. The Multimedia Satellite Task at MediaEval 2017: Emergence Response for Flooding Events. In *Proc. of the MediaEval 2017 Workshop* (Sept. 13-15, 2017). Dublin, Ireland.

[2] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

[3] Ilias Gialampoukidis, Anastasia Moumtzidou, Dimitris Liparas, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2017. Multimedia retrieval based on non-linear graph-based fusion and partial least squares regression. *Multimedia Tools and Applications* (2017), 1–21.

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions.. In *CVPR*. IEEE Computer Society, 1–9. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#SzegedyLJSRAEVR15

[5] Planet team. 2017. Planet Application Program Interface: In Space for Life on Earth. (2017).