GRAPH-BASED EARLY-FUSION FOR FLOOD DETECTION

Rafael de O. Werneck[1], *Icaro C. Dourado*[1], *Samuel G. Fadel*[1], *Salvatore Tabbone*[2], *Ricardo da S. Torres*[1]

[1] RECOD Lab, Institute of Computing (IC), University of Campinas (Unicamp), Campinas, SP, Brazil
[2] Université de Lorraine-LORIA UMR 7503, Vandoeuvre-lès-Nancy, France

ABSTRACT

Flooding is one of the most harmful natural disasters, as it poses danger to both buildings and human lives. Therefore, it is fundamental to monitor these disasters to define prevention strategies and help authorities in damage control. With the wide use of portable devices (e.g., smartphones), there is an increase of the documentation and communication of flood events in social media. However, the use of these data in monitoring systems is not straightforward and depends on the creation of effective recognition strategies. In this paper, we propose a fusion-based recognition system for detecting flooding events in images extracted from social media. We propose two new graph-based early-fusion methods, which consider multiple descriptions and modalities to generate an effective image representation. Our results demonstrate that the proposed methods yield better results than a traditional early-fusion method and a specialized deep neural network fusion solution.

Index Terms— graph-based fusion, early fusion, flood detection, MediaEval, image representation

1. INTRODUCTION

Natural disasters caused 306 billion dollars in damage in the United States of America in 2017,[1] and it may rise with global warming, increasing the intensity of heavy rainstorms [1]. In this scenario, it is fundamental to create monitoring systems that help authorities define appropriate strategies for damage control prevention and for victims' assistance. Among the different natural disasters, flooding is one of the most harmful and costly, as it destroys buildings, devastates agricultures, and threatens human lives [2].

However, traditional hydrological monitoring systems during floods have limited use in emergency response, due to, among other factors, ground inaccessibility or lack of aerial information [3]. Meanwhile, smartphones can provide an increase of documentation, dissemination, and communication of flooding events in social media streams. This new source of information may provide a much denser coverage of the natural disaster, and also document the impact of the disaster on human lives [4]. Also, handling multiple and complementary data modalities (e.g., text, images, videos) can help in the interpretation of flooding events. However, the accuracy and validity of these data may be questionable [5].

The literature considers the use of social media in the detection of natural events from different perspectives. Basnyat et al. [6] investigated a multi-modal approach using Twitter text and images to assess flood impacts. Twitter text was clustered using Latent Semantic Analysis into three clusters (help, damage, and casualties), and images were processed using Discrete Cosine Transformations to be classified into *water*, *nowater*, and *others*. Wang et al. [7] explored computer vision to classify natural events. They combined text content, based on a codebook containing the 1000 most frequent tags, for which each image has a vector indicating the presence or absence of the tag, with image content features learned using a Convolutional Neural Network (CNN).

The MediaEval initiative in 2017 also payed attention to this challenging detection problem. It proposed a task related to the retrieval of multimedia content from social media streams that are associated with flooding events (Disaster Image Retrieval from Social Media) [8]. One of the studies developed in the context of MediaEval 2017 refers to the work of Bischke et al. [9]. They proposed to extract visual features using CNNs and metadata features trained in a Word2Vec, with weights defined in terms of tf-idf. They also concatenated the above representations for multi-modality-based experiments. Ahmad et al. [10] also proposed the use of CNNs. They extracted eight feature vectors, that were fed into ensembles of Support Vector Machines. For textual metadata, they used a Random Tree classifier. In the multimodal approach, the classification scores were combined for both methods using Induced Ordered fusion scheme and Particle Swarm Optimization. Avgerinakis et al [11] proposed a CNN framework, using the GoogLeNet architecture to classify visual features only. To detect flooding using textual data, they adapted the DBpedia Spotlight, followed by a disambiguation algorithm using Jaccard similarities. They also performed a late fusion method to combine both modalities with a non-linear graph-based technique. Nogueira et al. [12]

Thanks to CNPq, CAPES (grant #88881.145912/2017-01), FAPESP (grants #2017/24005-2, #2017/16453-5, #2016/18429-1, #2014/12236-1, #2014/50715-9, #2013/50155-0, and #2013/50169-1) agencies for funding.

[1] https://www.nytimes.com/2018/01/08/climate/2017-weather-disasters.html (As of Jan. 2018).

also employed CNNs (ResNet [13] and GoogLeNet architectures) to classify visual data. They used a Relation Network (RN) to learn the co-occurrence of words in the metadata, and also a ranked solution, in which they used a rank aggregation technique on the best three pairs of text representation models and distance functions. Finally, they concatenated the RN and the CNN to devise a multimodal approach.

In this paper, we present two graph-based early-fusion approaches that combine different features and/or modalities, and apply them in a scenario of flooding detection in social media streams. We provide a better joint representation of the image considering different modality descriptions that complete each other's view. A graph representation is used to encode existing relations among representations in multiple feature/modality spaces. This graph is projected into a graph codebook, generating a final joint vector representation. These approaches can be applied for any feature extraction framework that provides multiple representations associated with the same or different modalities. Experiments show that the graph-based representation is more effective than traditional baselines from the literature. Moreover, in some cases, our approaches perform better than a recent deep neural network approach where feature description pairs are learned.

2. GRAPH-BASED EARLY-FUSION METHODS

Our motivation is based on previous works [14] where we propose a discriminant and efficient representation based on local structures of an image combining graphs with the BoW model. We introduced two Bag-of-Graphs (BoG)-based models that generate a meaningful vocabulary describing the main local patterns of a set of objects. We presented formal definitions, introducing concepts and rules that make these models flexible and adaptable for classification problems.

In this perspective, we propose in this paper two graph-based early-fusion methods which extend the BoG approach to create a joint representation of multiple descriptions and/or modalities. The fusion scheme aims to encode existing relationships between different features of objects.

2.1. Bag of KNN Graphs

Bag of KNN Graphs (BoKG) considers multiple features or modalities originated from a same object. This approach first builds a graph, where a vertex represents the object and edges connect their multiple representations associated with different feature spaces. In the following, this graph is enriched by adding edges that connect each object with its k-nearest neighbors according to each representation. The weights of edges connecting vertices within the same feature space are defined as the similarity score among object features. The weights of edges among vertices of different feature space, in turn, are based on the identification of the k nearest neighbors

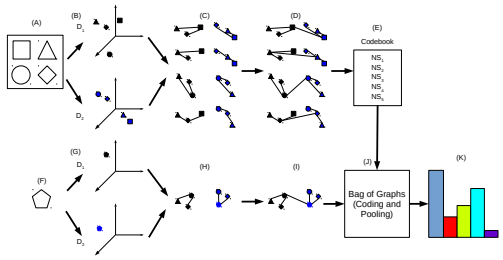

Fig. 1. Bag of KNN Graphs.

of each vertex, and on the use of a ranked-list-based similarity function.

Figure 1 illustrates this process. Given the graphs defined for each object, we apply the bag approach to describe the object represented by this graph. First, a collection of objects (A) is described by different description schemes (B). For each object, its k-nearest neighbors are determined for each description and a graph is created connecting vertices associated with objects and their neighbors (C). Then, we connect the different features (i.e., points in different feature spaces) of an object with an edge. The weight of the edge is defined by the similarity between the ranked lists of the objects connected by the edge (D). Next, we extract node signatures from all object graphs. We use the same definition of node signature as [14], which is composed of the feature of the vertex, its degree, and the features of its adjacent edges. These node signatures are used to create the codebook of the bag approach (E). For a new object (F), the same process is repeated. It is characterized by the description approaches (G), and has its graph created, considering as the nearest neighbors the objects in the collection (H). Edge weights are again computed by the similarity of ranked lists (I). Finally, we extract all the node signatures from this object graph to perform the coding and pooling steps of the bag approach (J) and thus generate its final vector representation (K).

2.2. Bag of Cluster Graphs

We also proposed another extension for the Bag-of-Graphs approach, a Bag of Cluster Graphs (BoCG). In this extension, given multiple representations, a unique graph is created. In this graph, objects represented within the same feature space are first clustered into n clusters. Cluster centroids represent the vertices of the final graph. Next, for each object in the collection, we find the clusters in the different feature spaces to which this object representation is assigned. Later, edges are created, connecting centroids of clusters to which the object belongs. The edge weight is defined as the ratio of the number of objects belonging to the two vertices of the edge, by the total number of objects in the collection.

Figure 2 illustrates this approach. First, a collection of objects (A) is described using two or more description meth-

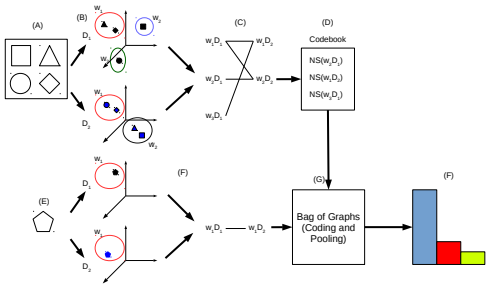

Fig. 2. **Bag of Cluster Graphs.**

ods (e.g., D_1 and D_2). Then, we create clusters of these features (B) and use their centroids to represent the vertices of the final graph (C). Next, for each object in the collection, we find the clusters to which each object is assigned. Then, we connect the clusters that are associated with the same object (C). Later, we extract node signatures from the graph created. These node signatures are also clustered to construct the codebook of the Bag-of-Graphs approach (D), in which each object is represented by the node signatures of each description. Given a new object (E), we apply the same steps to generate the node signatures. We predict in which clusters the new object features are, following the edges between these cluster vertices, and extract the node signatures from this object graph (F). Finally, coding and pooling methods (G) are applied to generate the final feature.

3. EXPERIMENTS & RESULTS

3.1. Dataset

The dataset used in this work was a dataset developed for the Multimedia Satellite Task at MediaEval 2017 in its first subtask: Disaster Image Retrieval from Social Media. This dataset is composed of 6,600 Flickr images extracted from the YFCC100M-Dataset [15], in which the images with *flooding* tags were selected and refined by human annotators. Both images and metadata were available in the challenge. The images were divided into two separated sets, development and test, with 5,280 and 1,320 images, respectively.

To evaluate our proposals before testing it, we also split the development set into training and validation sets, with a ratio of 80/20 (4,224 in the training set and 1,056 in the validation set). Thus, we can train our proposed methods, and select their best configuration to test, following the evaluation protocol adopted in the MediaEval competition.

3.2. Features & Baselines

We used the visual features provided by the organization of the challenge. These visual features were extracted with the LIRE library[2] with default parameters. The provided visual features are: AutoColorCorrelogram (ACC) [16]; EdgeHistogram (EH)[3]; Color and Edge Directivity Descriptor (CEDD) [17]; ColorLayout (CL)[3]; Fuzzy Color and Texture Histogram (FCTH) [18]; Joint Composite Descriptor (JCD) [19]; Gabor [20]; ScalableColor (SC)[3]; and Tamura [21]. For the textual data provided, we use 2GRAMS with Term Frequency, followed by a PCA for reducing the dimensionality.

We also included the concatenation of the provided features as a baseline early-fusion method to compare with the graph-based early-fusion methods proposed in this paper.

Feature	AP@50 (%)
ACC	50.55
CEDD	58.17
CL	47.26
EH	**69.03**
FCTH	59.53
Gabor	24.84
JCD	60.58
SC	5.63
Tamura	15.11
2GRAMS_TF (PCA)	**81.47**

Table 1. Average precision for the baselines.

3.3. Evaluation

The MediaEval 2017 contest proposed the use of Average Precision@X, with several cutoffs (50, 100, 250, and 480), for the correctness of retrieved images in the experiments. This metric scores the proportion of relevant images among the top-X retrieved images, also taking their order into account. Here we present our results considering the top-50 retrieved images. For the baselines and the proposed approaches, we performed experiments with a two-class SVM classifier (with linear kernel and $C = 1$).

3.4. Results

3.4.1. Baselines

First, we computed the results considering all features provided, using the validation set. Table 1 shows the Average Precision @ 50 (AP@50). EdgeHistogram obtained the best AP for the provided image features, with a precision of 69.03%, and 2GRAMS_TF (PCA), a text descriptor, obtained an AP of 81.47%.

As baselines, we also considered the concatenation of these features. Table 2 shows the results considering the concatenation of the provided features normalized between 0.0

[2] `http://www.lire-project.net/` (As of Nov. 2017).

[3] `https://mpeg.chiariglione.org/standards/mpeg-7/visual` (As of Nov. 2017).

Concatenation	AP@50 (%)
ACC & CEDD & CL & EH & FCTH & Gabor & JCD & SC & Tamura	82.25
EH & 2GRAMS_TF (PCA)	68.27

Table 2. Results of the concatenation of all features and our best modalities features as baseline.

Features	AP@50 (%)
ACC & CEDD & CL & EH & FCTH & Gabor & JCD & SC & Tamura	81.11
EH & 2GRAMS_TF (PCA)	86.90

Table 3. Bag of KNN Graphs (BoKG) results.

and 1.0, and the concatenation of the best provided image features with the textual one. The visual features introduced noise when concatenated with textual features, leading to worse precision scores.

The best baseline result, considering the concatenation of features, attained an AP@50 of 82.25%. Once the most promising features were identified, we evaluated the proposed approaches and compared with these baselines in the scenario of the challenge, i.e., using the validation and test sets.

3.4.2. Evaluation of Proposed Approaches

First, we used the validation set with the aim of identifying the best parameters for our approaches. The Bag of KNN Graphs uses two sets of nearest neighbors (with 10 and 20 neighbors), a Cosine similarity metric, a codebook of 500 node signatures randomly selected, the intersection of ranked lists as similarity function, hard assignment, and max pooling. For the Bag of Cluster Graphs, which has less parameters, we selected 1000 random features for each modal cluster and 2000 random node signatures. We also used hard assignment and max pooling. These parameters were selected as they provided the best results in experiments with the validation set.

Table 3 shows the results for the Bag of KNN Graphs compared with the baselines. As we can see, our BoKG approach performed similarly as the concatenation considering all provided features but, for the multiple modality combination, it performed better than baselines, showing that our proposed approach can provide an effective joint representation by combining different modalities (text and visual features) of the same object.

Table 4 presents the results obtained with the Bag of Cluster Graphs. This table shows that, although it did not perform better than the BoKN, our results for the multiple modalities joint representation also outperformed the baseline based on early-fusion concatenation. The results of the BoCG are below the one of BoKG because of its sparse final vector representation, as this approach uses less node signatures in the coding and pooling steps than the BoKG. The sparse features

Features	AP@50 (%)
ACC & CEDD & CL & EH & FCTH & Gabor & JCD & SC & Tamura	47.94
EH & 2GRAMS_TF (PCA)	73.85

Table 4. Bag of Cluster Graphs (BoCG) results.

Features	AP@50 (%)
ACC & CEDD & CL & EH & FCTH & Gabor & JCD & SC & Tamura	79.63
EH & 2GRAMS_TF (PCA)	75.55

Table 5. Results of the Relation Network deep approach.

provided less information for the classifier to train a separation model between considered classes.

3.4.3. Comparison with the Relation Network Approach

The Relation Network (RN) [22] is a recently proposed neural network, which learns to infer relationships between objects and produce decisions over them. The RN is composed of two neural networks (denominated f and g) whose parameters are learned together. The function g is used to encode the relationship between pairs of objects, while the function f takes the sum of all encodings as input and produces a decision over the entire collection. We used a Relation Network as a baseline, finding the relationship between the objects representations. Table 5 shows the Average Precision @ 50 for this deep network. The experiments with the RN used the parameters suggested by the authors: 128 epochs and a learning rate of $2.5 \cdot 10^{-4}$, and we also used the same training set of the proposed approaches. The provided results were not better than BoKG (see Table 3), as our approach enriched the final representation when considering the neighborhood of objects in different feature spaces.

4. CONCLUSIONS

In this paper, we present two new approaches based on Bag of Graphs to create a joint representation of multiples modalities and/or descriptions: Bag of KNN Graphs and Bag of Cluster Graphs. We validate these approaches in the flood detection scenario, proposed by the MediaEval 2017 contest. In this scenario, we show that our early-fusion approach outperforms the traditional concatenation fusion when dealing with multiple modalities. Our experiments also showed that our approach has better performance than a deep neural network (RN). For future work, we propose to compare our methods with other early-fusion methods, as well as other deep learning approaches that use early-fusion. We also plan to include deep features as input features/modalities of our approaches.

5. REFERENCES

[1] R. P. Allan and B. J. Soden, "Atmospheric warming and the amplification of precipitation extremes," *Science*, vol. 321, no. 5895, pp. 1481–1484, 2008.

[2] S. Martinis, A. Twele, and S. Voigt, "Towards operational near real-time flood detection using a split-based automatic thresholding procedure on high resolution terrasar-x data," *Natural Hazards and Earth System Sciences*, vol. 9, no. 2, pp. 303–314, 2009.

[3] T. De Groeve, "Flood monitoring and mapping using passive microwave remote sensing in namibia," *Geomatics, Natural Hazards and Risk*, vol. 1, no. 1, pp. 19–35, 2010.

[4] N. Tkachenko, S. Jarvis, and R. Procter, "Predicting floods with flickr tags," *PloS one*, vol. 12, no. 2, pp. 1–13, 2017.

[5] J. F. Rosser, D. G. Leibovici, and M. J. Jackson, "Rapid flood inundation mapping using social media, remote sensing and topographic data," *Natural Hazards*, vol. 87, no. 1, pp. 103–120, 2017.

[6] B. Basnyat, A. Anam, N. Singh, A. Gangopadhyay, and N. Roy, "Analyzing social media texts and images to assess the impact of flash floods in cities," in *International Conference on Smart Computing*. IEEE, 2017, pp. 1–6.

[7] J. Wang, M. Korayem, S. Blanco, and D. J. Crandall, "Tracking natural events through social media and computer vision," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1097–1101.

[8] B. Bischke, P. Helber, C. Schulze, S. Venkat, A. Dengel, and D. Borth, "The multimedia satellite task at mediaeval 2017: Emergence response for flooding events," in *Proceedings of the MediaEval 2017 Workshop*, Ireland.

[9] B. Bischke, P. Bhardwaj, A. Gautam, P. Helber, D. Borth, and A. Dengel, "Detection of flooding events in social multimedia and satellite imagery using deep neural networks," in *Working Notes Proc. MediaEval Workshop*, 2017, p. 2.

[10] K. Ahmad, P. Konstantin, M. Riegler, N. Conci, and P. Holversen, "Cnn and gan based satellite and social media data fusion for disaster detection," in *Working Notes Proc. MediaEval Workshop*, 2017, p. 2.

[11] K. Avgerinakis, A. Moumtzidou, S. Andreadis, E. Michail, I. Gialampoukidis, S. Vrochidis, and I. Kompatsiaris, "Visual and textual analysis of social media and satellite images for flood detection@ multimedia satellite task mediaeval 2017," in *Working Notes Proc. MediaEval Workshop*, 2017, p. 2.

[12] K. Nogueira, S. G. Fadel, Í. C. Dourado, R. de O. Werneck, J. A.V. Muñoz, O. A.B. Penatti, R. T. Calumby, L. T. Li, J. A. dos Santos, and R. da S. Torres, "Data-driven flood detection using neural networks," in *Working Notes Proc. MediaEval Workshop*, 2017, p. 2.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[14] F. B. Silva, R. de O. Werneck, S. Goldenstein, S. Tabbone, and R. da S. Torres, "Graph-based bag-of-words for classification," *Pattern Recognition*, vol. 74, pp. 266–285, 2018.

[15] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[16] J. Huang, S.R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun 1997, pp. 762–768.

[17] S. Chatzichristofis and Y. Boutalis, "Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *Computer Vision Systems*. 2008, pp. 312–322, Springer.

[18] S. Chatzichristofis and Y. Boutalis, "Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval," in *Workshop on Image Analysis for Multimedia Interactive Services*, May 2008, pp. 191–196.

[19] S. Chatzichristofis, Y. Boutalis, and M. Lux, "Selection of the proper compact composite descriptor for improving content based image retrieval," in *International Association of Science and Technology for Development*, vol. 134643, p. 064.

[20] J. G. Daugman, "Complete discrete 2-d gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[21] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 6, pp. 460–473, June 1978.

[22] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in Neural Information Processing Systems 30*, pp. 4967–4976. 2017.