

Early and Late Fusion of Multiple Modalities in Sentinel Imagery and Social Media Retrieval

Wei Yao¹, Anastasia Moutzidou², Corneliu Octavian Dumitru¹, Stelios Andreadis², Ilias Gialampoukidis², Stefanos Vrochidis², Mihai Datcu¹, and Ioannis Kompatsiaris²

¹ Remote Sensing Technology Institute, German Aerospace Center (DLR), Germany
{Wei.Yao, Corneliu.Dumitru, Mihai.Datcu}@dlr.de

² Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Greece {moutzid, andreadisst, heliasgj, stefanos, ikom}@iti.gr

Abstract. Discovering potential concepts and events by analyzing Earth Observation (EO) data may be supported by fusing other distributed data sources such as non-EO data, for instance, in-situ citizen observations from social media. The retrieval of relevant information based on a target query or event is critical for operational purposes, for example, to monitor flood events in urban areas, and crop monitoring for food security scenarios. To that end, we propose an early-fusion (low-level features) and late-fusion (high-level concepts) mechanism that combines the results of two EU-funded projects for information retrieval in Sentinel imagery and social media data sources. In the early fusion part, the model is based on active learning that effectively merges Sentinel-1 and Sentinel-2 bands, and assists users to extract patterns. On the other hand, the late fusion mechanism exploits the context of other geo-referenced data such as social media retrieval, to further enrich the list of retrieved Sentinel image patches. Quantitative and qualitative results show the effectiveness of our proposed approach.

Keywords: Multimodal data fusion · Sentinel imagery retrieval · Social media retrieval · Earth Observation · Big Data

1 Introduction

The number of Earth Observation (EO) data is increasing rapidly due to the large number of space missions that were launched during the past years. Moreover, the fact that there are EO data that are freely available to the scientific community (e.g., data from the Copernicus missions), opens up the horizons for using them in several applications. Furthermore, the advancements in the domain of satellite remote sensing helped in producing quick and precise land cover maps that allowed us to identify target categories such as snow, rocks, urban areas, forests, and lakes. We use that to capture the characteristics of the underlying areas, and eventually exploit this information to assist in global monitoring and future planning. One major challenge is the lack of training datasets

for building well-performing models using shallow and deep learning models. To that end, an active learning method is proposed. Active learning is a form of supervised machine learning. The learning algorithm is able to interactively interrogate a user (as an information and background knowledge source) to label image patches with the desired outputs. The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels, if it is allowed to choose the data from which it learns. This operation with the active learning procedure is presented in Fig. 1.

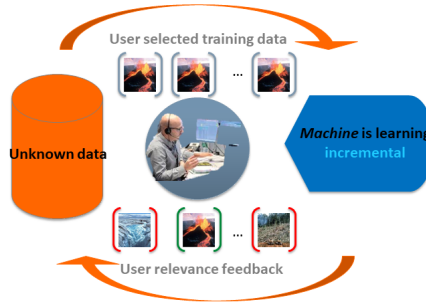


Fig. 1. The active learning concept.

The inclusion of crowdsourced geo-referenced data sources, through the retrieval of social media data, semantically enriches the retrieved results from satellite image content. Twitter is a popular platform, where a set of keywords, locations, and user accounts can be defined to formulate a query in order to obtain relevant information to a concept or event. Such information is integrated with the retrieval of satellite image patches, combining the results from remotely-sensed images with images and text descriptions from citizen observations and user-generated online content.

Our contribution can be summarized as follows:

- Retrieve satellite images using an active learning technique
- Extend satellite image retrieval with social media posts

The paper is organised as follows. Section 2 presents relevant works in multimodal fusion for the two main fusion strategies. Section 3 presents our proposed methodologies, one based on early fusion of data and the other on late fusion. In Section 4, we describe the datasets that we have used, the settings, and also the quantitative and qualitative results. Finally, Section 5 concludes our work.

2 Related Work

Over the years, two main strategies for fusing multimodal information have been identified [2]. The first strategy is known as early fusion; it is realized at feature

level, where features from multiple modalities are combined into a common feature vector, while the second strategy, known as late fusion, fuses information at the decision level.

In our previous investigation in data fusion [15], the data representation as Bag-of-Words has been discussed, using a clustering of various modalities and an application of Bayesian inference for fusing clusters into image classes. In addition, the work in [4] presents the extraction of different information modalities from the same observation and fusion for enhanced classification. Recently, within the framework of the CANDELA project³, we implemented the merging of different Sentinel-1 and Sentinel-2 bands [18]. Furthermore, during the Living Planet 2019 Conference⁴, a semantic level fusion for Synthetic Aperture Radar (SAR) images has been discussed. By exploiting the specific imaging details and the retrievable semantic categories of TerraSAR-X and Sentinel-1 images, we obtained semantically-fused image classification maps that allow us to differentiate several coastal surface categories [7].

Active learning has important advantages when compared with Shallow Machine Learning or Deep Learning methods, as presented in Table 1.

Table 1. Comparison of different learning schemes

Key Performance Indicator	Shallow ML	Deep Learning	Active Learning
Training data volume	Medium (GB)	Very high (PB)	Very small (0.1KB)
Trained data volume	Large (GB-TB)	Very high (PB)	Large (GB-TB)
No. of classes	Up to 100	Up to 100	Any, user-defined
Classification accuracy	Avg. 85%	Avg. 90%	Avg. 85%
Training speed	Medium (hour)	Slow (days)	Fast (minutes)

Active learning methods include Relevance Feedback and Cascaded Learning, see Algorithm 1. It supports users to search for images of interest in a large repository. A Graphical User Interface (GUI) allows the automatic ranking of the suggested images, which are expected to be grouped in a class of relevance. Visually supported ranking allows enhancing the quality of search results after giving positive and negative examples. During the active learning process, two goals are achieved: 1) learning the targeted image category as accurately and

³ <https://www.candela-h2020.eu/>

⁴ <https://lps19.esa.int/>

as exhaustively as possible, and 2) minimising the number of iterations in the relevance feedback loop.

Algorithm 1: Active Learning Algorithm

Data: Sentinel-1, Sentinel-2 image pair with fused feature vectors

Result: semantic annotation stored in DMDB

initialization;

while *user is not satisfied with the annotated results* **do**

 user selects new positive and negative images;

 calculate and show classification result;

 get **relevance feedback** (display ranked boundary images);

 start **cascaded learning** process as follows;

if *user is satisfied and there is a finer image grid* **then**

 go to the next image grid;

 set constraint on available image patches (only patches within
 previous annotated grid will be taken into account);

 current section becomes this one;

else

 go back to the beginning of current section;

However, the involvement of social media queries requires multimodal fusion mechanisms that are able to combine textual, visual, and spatiotemporal information. As it is already mentioned, our late-fusion techniques involve fusing information at decision level. This means that initially, each modality is learned separately and then the individual results are combined in order to reach a final common decision. Most of the late-fusion methods for retrieval are, in general, unsupervised techniques that use the document rank and score to calculate the decision. For example, in [19], the authors propose a multimodal knowledge-based technique in order to retrieve videos of a particular event in a large-scale dataset. The authors consider several modalities including speech recognition transcripts, acoustic concept indexing, and visual semantic indexing, which are fused using an event-specific fusion scheme. In [11], the authors describe a system for retrieving medical images. The system considers textual and visual content, separately as well as combined, using advanced encoding and quantisation by combining the results of the modalities in a sequential order. Specifically, the textual modality returns a list of results that is re-ranked based on the visual modality. The work of [9] retrieves text-image pairs, where queries of the same multimodal character are evaluated. Moreover, in [13], the authors combine data from Twitter along with Sentinel-1 images in order to increase the accuracy of the state-of-the-art remote sensing methods related to snow depth estimation. In [1] the authors present a method using social media content to address the problem of flood severity by checking both the text and the image in order to classify articles as flood event-related. Also, the visual features extracted from the images were used to identify people were standing in flooded area. Recently,

the EOPEN project⁵ has demonstrated the fusion of multiple modalities in the context of EO and non-EO data [8]. Contrary to these approaches, we use a tensor-based late fusion mechanism that aims to complement satellite image search with social media data for related concepts, such as food, flood, city, etc.

3 Methodology

3.1 Early Fusion in satellite image retrieval

Our Early Data Fusion aims at a better understanding of a scene from observations with multiple different sensors. In the particular case of the data fusion in CANDELA, the objective is to obtain a better classification of the Earth’s surface structures or objects using Sentinel-1 (S-1) and Sentinel-2 (S-2) observations. The design of the data fusion methods shall exploit the characteristics of the different sensing modalities. Table 2 is summarizing the main aspects of the complementarity of Sentinel-1 and Sentinel-2 observations.

Table 2. The complementarity between Sentinel-1 and Sentinel-2 images.

Criteria	Sentinel-1 (SAR)	Sentinel-2 (multispectral)
Sensor type	Active	Passive
EM spectrum	C-band	Blue to IR (13 bands)
Operation	Day/Night	Day
Dependence on cloud cover	No	Yes
Vegetation signatures	Low sensitivity	Good diversity of classes
Ocean/sea	Waves and currents	Water colour
In-land waters	Low backscatter	Diversity of spectral signatures, water colour
Urban constructions	Strong signatures	variable depending on may parameters
Soil	Moisture and roughness	Spectral signatures (colour)
Relief	Strong dependence	Moderate dependence
Snow/ice	Classification based on EM properties	Reduced separability, confusion with clouds

Based on these assets, the data fusion was designed using three important paradigms. Firstly, the fusion is performed at the level of image patch features; secondly, the classification is performed by an active machine learning paradigm, and finally, the classifier results are semantic annotations stored in a database.

The early fusion is performed at image feature level, so as to combine the very particular signatures of the scene for the two observation modalities, namely multispectral and SAR imaging. Our feature extractor for Sentinel-1 data is the Adapted Weber Descriptor [5]. Comparing to the original WLD feature, the adapted WLD includes not only local statistics but also local structure information, resulting in an improved performance to characterize SAR signatures by minimizing the noise effect of speckle. The feature extracted from Sentinel-2 is the multi-spectral histogram, since it contains the statistical physical information of the Sentinel-2 multispectral signatures. The two features are concatenated and become a fused descriptor of the Earth’s land observed by the two sensors.

⁵ <https://eopen-project.eu/>

The classifier is chosen to be an Active Machine Learning tool [3] based on a Support Vector Machine (SVM), allowing the user to select the training samples in an appropriate manner to avoid any contradiction which may occur from the different sensor signatures. The result of the classification is stored into a database as semantic annotation, thus enabling further analyses and the export of the information for integration or a next level of fusion with non-EO data. Fig. 2 depicts the software architecture of the Data Fusion module in the back end and in the front end. There are three layers which define a complete process: the platform layer as back-end, the user machine layer as front-end, and the transfer layer via an Internet connection.

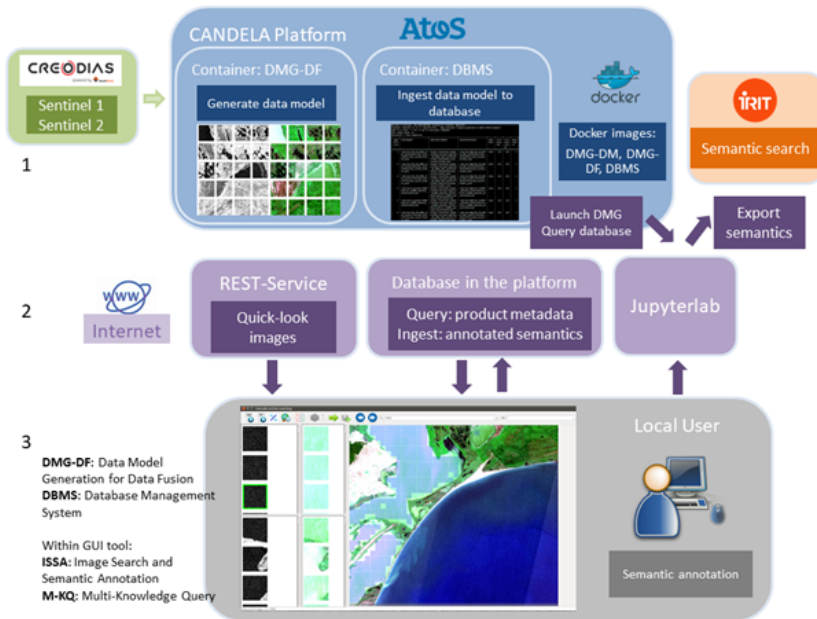


Fig. 2. Architecture of the data fusion module on the platform and front-end.

In the platform layer, Sentinel-1 and Sentinel-2 products are accessed by using the symbolic CreoDIAS⁶ links which are provided for the platform. Users start the Data Model Generation for the Data Fusion Docker container and it runs for one Sentinel-1 product and one Sentinel-2 product simultaneously. As a pre-processing step, the two products should be geometrically co-registered. The results (extracted metadata, cropped image patches, and extracted features for the patches) are ingested into the MonetDB⁷ database “candela” on the platform. The generated quick-look images are published on the platform to be

⁶ <https://creodias.eu/>

⁷ <https://www.monetdb.org/>

downloaded by local users via a Representational state transfer (RESTful) service. The Database Management System (DBMS) provides high-speed storage for real-time interactive operation during active learning and data fusion. This is the actionable information of the Data Fusion component. The framework (Fig. ??) provides the following front-end functionalities to the user: **Image Search and Semantic Annotation** (ISSA): image mining, query-by-example, retrieval and adding of semantic annotation to EO image products; **Multi-Knowledge and Query** (M-KQ): multimodal queries based on selected product metadata, image features, and semantic labels; and **System Validation**: supports the evaluation of the retrieval and classification performance. The Data Fusion module evolves from EOLib⁸.

In support of the semantic annotations, a hierarchical two-level semantic catalogue has been ingested into the “candela” database, which allows users to select the appropriate label during semantic annotation by using the active learning tool. In the case of Copernicus (e.g., S-1 and S-2), level-1 labels are the most general categories: *Agriculture*, *Bare Ground*, *Forest*, *Transportation*, *Urban Area*, and *Water Bodies*; while level 2 consists of more detailed labels, concerning each general level, respectively. In addition, because of the diversity of structures in an image, after choosing a specific general-level label, an extra user-defined label annotation function is allowed, so that new land cover or land use cases can be described according to the user’s own definition. This is different from a fixed classification system, and particularly useful in the case of evolving land cover patterns, e.g., floods.

3.2 Late-Fusion Approach to retrieve relevant social media content

The late-fusion approach retrieves social media posts that are similar to a given tweet by considering its different modalities, i.e., textual information, visual features and concepts, and spatiotemporal information. The late-fusion mechanism consists of the following phases: 1) description of the multimodal query q using a set of modalities; 2) querying the indexing schemes that are used to allow a fast and efficient retrieval for each modality in order to get ranked lists of retrieved items along with the similarity scores of the query tweet q to the available pool of tweets — these lists are used for creating a 4D tensor; and 3) two-step fusion procedure that initially involves a bi-modal fusion of the retrieved results for each 2D surface of the created tensor, followed by a merging of the produced rankings to get the final list of retrieved tweets (Fig. 3).

The proposed late-fusion approach fuses the output of four modalities. The algorithm comprises the following steps:

1. Retrieval of N results per modality, which eventually leads to four such lists from unimodal searches.
2. Creation of a fourth-order \mathbf{L} tensor by considering the aforementioned lists. The dimension of the binary tensor \mathbf{L} is (l_1, l_2, l_3, l_4) . The value of the single

⁸ <http://wiki.services.eoportal.org/tiki-index.php?page=EOLib>

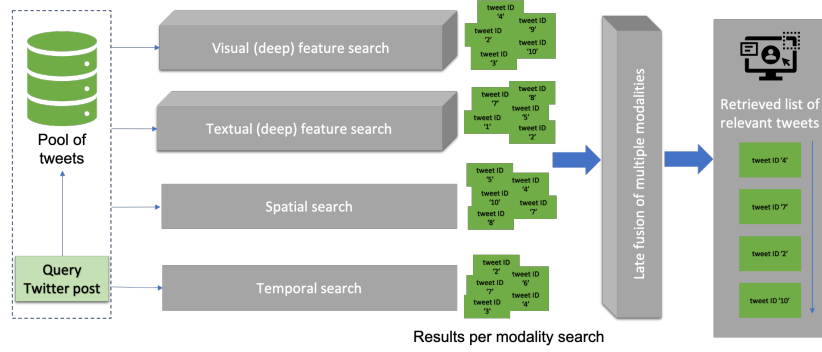


Fig. 3. Late-fusion framework for multimodal retrieval of geo-referenced tweets.

elements results from the following rule:

$$\mathbf{L}_{(\dots, r_i, \dots, r_k, \dots)} = \begin{cases} 1, & \text{if the same element is ordered as } r_i \text{ in list } l_i, \\ & \text{and ordered as } r_k \text{ in list } l_k \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

3. Creation of one 2D tensor surface for each pair of modalities $(i, k), i \leq k$.
4. For each 2D tensor surface, get the list of tweets that are jointly ranked higher by minimising the position in which $\mathbf{L}(i, k) = 1$ (details in [8]).
5. Merging of the rankings to obtain the final list of tweet IDs.

Text similarity between two or more texts is the procedure of computing the similarity in meanings between them. Although there are several approaches that can be used for text similarity that involve text representation as a first step, the one considered in this work is an off-the-shelf text search engine, i.e., the Apache Lucene⁹. Apache Lucene is a full-text search engine that can be used for any application that requires full-text indexing, and searching capability. It is able to achieve fast search responses, as it uses a technique known as inverted index and avoids searching the text directly. The representation of the text modality also considers the state-of-art Bidirectional Encoder Representation from Transformers (BERT) algorithm[6], which includes an attention mechanism to learn contextual relations between words in a text. BERT is used to represent each tweet text into a deep representation that allows similarity search.

As far as visual information is concerned, both visual features and visual concepts are taken into consideration. The framework used in both cases, i.e., a deep convolutional neural network (DCNN), is the same, but the vectors used are taken from different layers of the network. Specifically, we used the fine-tuned 22-layer GoogleNet network [16] that was trained on the 345 SIN TRECVID concepts. Regarding the visual features, they are DCNN-based descriptors and are the output of the last pooling layer of the fine-tuned GoogleNet architecture

⁹ <https://lucene.apache.org/>

previously described. The dimension of the last pooling layer is 1024 and it is used as a global image representation. The selection of this layer was based on the results of an evaluation analyzing its runtime and quality within the VERGE system [14] that has participated in the Video Browser Showdown in 2018. The visual concept representation is a concatenated single vector with a length of 345 elements, as the output of the aforementioned GoogleNet network.

Fast retrieval of similar visual and textual content is achieved by constructing an inverted file and combining it with Asymmetric Distance Computation [10]. Then, the k -nearest neighbours between the query image and the collection are computed. Temporal metadata also accompany the query tweet q and exist as a *ISODate* datatype inside the MongoDB¹⁰ used for storing the tweets information. The inherent MongoDB sorting functions allow the retrieval of a list of items which are temporally close to the query tweet. Regarding the locations mentioned in a tweet, we extract the corresponding named entities using the BiLSTM-CNNs-CRF[12] model. The bidirectional Long Short-Term Memory part is responsible for encoding, a DCNN for extracting character level features, and a Conditional Random Field for decoding.

4 Experiments

4.1 Datasets description

For the demonstration and validation of the Data Fusion mechanism in satellite image search we use 33 Sentinel-1 and Sentinel-2 images. The average image size in pixels is $26,400 \times 16,600$ and $10,980 \times 10,980$ for Sentinel-1 and Sentinel-2, respectively. These satellite images cover an area of 350,000 km². One band has been considered from Sentinel-1 and four bands at 10 meter resolution from Sentinel-2 images. The patch size is 120×120 pixels and with one image grid level. For the total number of 340,587 patches, 7,478 samples have been annotated into several semantic labels (see Fig. 6 and Table 3).

Twitter is a suitable social media platform for testing fusion approaches since each tweet comprises several modalities. Specifically, a tweet contains a short text with not more than 140 characters that may contain non-standard terms, sometimes an image that is semantically related to the text, the date and time the tweet was posted, and any named entities of the type “location” that can be extracted from the text. Three datasets were used that include publicly available tweets retrieved via the Twitter Streaming API¹¹. The datasets were created by collecting tweets that included the words “alluvione” (i.e., flood in Italian), “food”, and “lumi” (i.e., snow in Finnish). The total number of tweets selected in a period of three years for the three datasets are 1,000,383 for floods (IT), 120,666 for food (EN) and 66,175 for snow (FI), respectively. An example collected tweet, that can also be used as a query, is shown in Fig. 4.

¹⁰ <https://www.mongodb.com/>

¹¹ <https://developer.twitter.com/en/docs/twitter-api>



Fig. 4. Query tweet in English language that is related to “food”.

4.2 Results

Our final results and examples are presented in Table 3. We observe that the overall classification accuracy is up to 90%, even for a very small training data set and a maximum of three iterations during the active learning stage. Fig. 5 and 6 show a visual demonstration of the early fusion result. Five classes are discovered in the scenes of Munich: *Lakes*, *Mixed Forest*, *Mixed Urban Areas*, *Stubble*, and *Grassland*. The S-1 and S-2 images are the inputs for the Data Mining module to be fused, while CORINE land cover 2018 is provided as visual ground truth¹². Focusing on the urban area, data fusion achieves better results, because the multi-spectral signal together with the radar signal, which generates strong backscattering in the man-made construction areas, helps distinguish the urban signatures.



















Fig. 5. Munich, Germany as a data fusion example. Left: S-1 image, Middle: S-2 image, Right: CORINE land cover 2018.

As far as social media retrieval is concerned, we manually annotate the top-10 retrieved results for each method, and then calculate the average precision for each query and the mean average precision (mAP) for three queries for each method. Table 4 contains the average precision scores for the different similarity methods for each query and the mAP for each method.

We conclude that text modality doesn’t perform well when it isn’t fused with any other modality. However, in case of tweets, only text and temporal

¹² http://clc.gios.gov.pl/images/clc_ryciny/clc_classes.png

Table 3. Examples of the use of the Data Fusion component and overall performances.

Sentinel-1	Sentinel-2	Label	Accuracy
		Mixed Forest	90%
		Beach	80%
		Mixed Urban	90%
		Agriculture	70%
		Land	80%
		Hill	65%
		Low-Density Urban	80%
		Forest Spots	80%

information exist by default, so it is a very important modality to consider. Moreover, time modality has a better mAP compared with text, which can be explained easily, since we consider only the top-10 results. However, it is expected that if we retrieve the top- K results, this score (mAP) will fall for large values of K . Finally, visual features perform very well, since the modality searches for visually similar results using pre-trained models in larger image collections, but they cannot be used disregarding corresponding text. Fig. 4 is the example Twitter post query, while Fig. 7 and Fig. 8 provide the top-10 retrieved list of tweets. These lists can be compared to the results of the tensor-based multimodal approach in Fig. 9.

5 Conclusion

Active learning with a very small number of training samples allows a detailed verification of images. Thus, the results are trustable, avoiding the plague of training data based biases. Another important asset is the adaptability to user

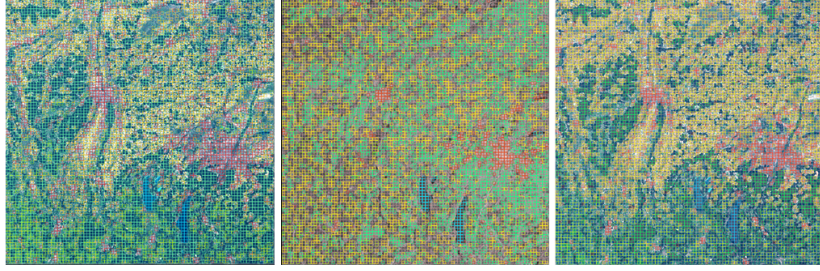


Fig. 6. Left: fusion results combining S-1 and S-2, Middle: classification results of S-1, Right: classification results of S-2.

Table 4. Average precision P@10 and mean Average Precision (mAP) of unimodal and multimodal searches.

Method	Flood, IT (P@10)	Food, EN (P@10)	Snow, FI (P@10)	mAP
Text	1.0	1.0	0.586	0.862
Spatiotemporal data	0.839	0.867	1.0	0.902
Visual Features	0.878	1.0	1.0	0.959
Visual Concepts	0.638	1.0	1.0	0.879
Multimodal fusion	0.906	1.0	1.0	0.969

conjectures. The EO image semantics are very different from other definitions in geoscience, as for example cartography. An EO image is capturing the actual reality on ground; a user can discover and understand it immediately, and extract its best meaning, thus enriching the EO semantic catalogue. With CANDELA platform as a back-end solution to support the query and ingestion of information into the remote database “candela” the early data fusion has been verified with various image pairs. The validation results show that the fused results generate more complete classification maps and perform very well even in challenging cases, such as *Beach*. The necessity to design and develop multimodal solutions is apparent also when combining EO with non-EO data, i.e. Twitter content in our case. Our presented method is able to effectively combine textual and visual information from tweets with other associated metadata, providing a search engine that can serve as an extension to satellite image search engines. In future, we plan on running more extensive experiments which involves evaluating the proposed late-fusion algorithm on large datasets that contain a variety of modalities and also testing it on significantly more queries. Finally, further integration and orchestration of EO and non-EO technologies is expected, with additional evaluation that also involves user satisfaction in the context of large-scale exercises in EU-funded projects.



Fig. 7. Top-10 the retrieved results with unimodal textual and temporal modalities.

Acknowledgements

This work has been supported by the EC-funded projects CANDELA (H2020-776193) and EOPEN (H2020-776019), and partly by the ASD HGF project. The content of this paper (DLR part) is mainly based on the results presented in the CANDELA Deliverable D2.8 [17].

References

1. Andreadis, S., Bakratsas, M., Giannakeris, P., et al.: Multimedia analysis techniques for flood detection using images, articles and satellite imagery. In: Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019. CEUR Workshop Proceedings, vol. 2670. CEUR-WS.org (2019)
2. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* **16**(6), 345–379 (2010)
3. Blanchart, P., Ferecatu, M., et al. Cui, S., Datcu, M.: Pattern retrieval in large image databases using multiscale coarse-to-fine cascaded active learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7**(4), 1127–1141 (2014)
4. Chaabouni-Chouayakh, H., Datcu, M.: Backscattering and statistical information fusion for urban area mapping using terrasar-x data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **3**(4), 718–730 (2010)
5. Cui, S., Dumitru, C.O., Datcu, M.: Ratio-detector-based feature extraction for very high resolution sar image patch indexing. *IEEE Geoscience and Remote Sensing Letters* **10**(5), 1175–1179 (2013)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)



Fig. 8. *Top-10* the retrieved results with unimodal visual feature & concept search.

7. Dumitru, C.O., Schwarz, G., Datcu, M.: Monitoring of coastal environments using data mining. In: Knowledge Extraction and Semantic Annotation (KESA 2018). pp. 34–39 (April 2018)
8. Gialampoukidis, I., Moutmtzidou, A., Bakratsas, M., et al.: A multimodal tensor-based late fusion approach for satellite image search in sentinel 2 images. In: 27th International Conference on Multimedia Modeling (MMM2021) (January 2021)
9. Gialampoukidis, I., Moutmtzidou, A., Liparas, D., et al.: Multimedia retrieval based on non-linear graph-based fusion and partial least squares regression. *Multimedia Tools and Applications* **76**(21), 22383–22403 (2017)
10. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* **33**(1), 117–128 (2010)
11. Kitanovski, I., Strezoski, G., Dimitrovski, I., Madjarov, G., Loskovska, S.: Multimodal medical image retrieval system. *Multimedia Tools and Applications* **76**(2), 2955–2978 (2017)
12. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* (2016)
13. Mantsis, D.F., Bakratsas, M., Andreadis, S., et al.: Multimodal fusion of sentinel 1 images and social media data for snow depth estimation. *IEEE Geoscience and Remote Sensing Letters* (2020)
14. Moutmtzidou, A., Andreadis, S., Markatopoulou, F., Galanopoulos, D., et al.: Verge in vbs 2018. In: In 24th International Conference on Multimedia Modeling. pp. 444–450. Springer (2018)
15. Palubinskis, G., Datcu, M.: Information fusion approach for the data classification: an example for ers-1/2 insar data. *International Journal of Remote Sensing* **29**(16), 4689–4703 (2008)

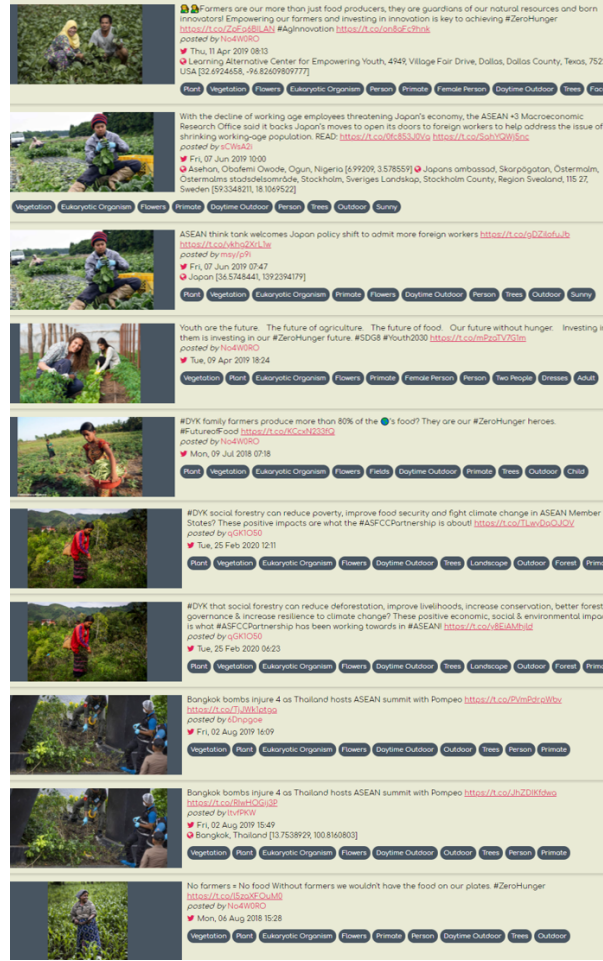


Fig. 9. Top-10 the retrieved results with multimodal fusion.

16. Pittaras, N., Markatopoulou, F., Mezaris, V., Patras, I.: Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In: International Conference on Multimedia Modeling. pp. 102–114. Springer (2017)
17. Yao, W., Dumitru, C.O., Datcu, M.: D2.8 data fusion v2, deliverable of the candela project, <https://www.candela-h2020.eu/content/data-fusion-v2>
18. Yao, W., Dumitru, C.O., Lorenzo, J., Datcu, M.: Data fusion on the candela cloud platform. In: European Geosciences Union (EGU) General Assembly - Big data and machine learning in geosciences (May 2020)
19. Younessian, E., Mitamura, T., Hauptmann, A.: Multimodal knowledge-based analysis in multimedia event detection. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. pp. 1–8 (2012)