A multimodal tensor-based late fusion approach for satellite image search in Sentinel 2 images

Ilias Gialampoukidis, Anastasia Moumtzidou, Marios Bakratsas, Stefanos Vrochidis, and Ioannis Kompatsiaris

Centre for Research and Technology Hellas
Information Technologies Institute
Thessaloniki, Greece
{heliasgj, moumtzid, mbakratsas, stefanos, ikom}@iti.gr

Abstract. Earth Observation (EO) Big Data Collections are acquired at large volumes and variety, due to their high heterogeneous nature. The multimodal character of EO Big Data requires effective combination of multiple modalities for similarity search. We propose a late fusion mechanism of multiple rankings to combine the results from several uni-modal searches in Sentinel 2 image collections. We fist create a K-order tensor from the results of separate searches by visual features, concepts, spatial and temporal information. Visual concepts and features are based on a vector representation from Deep Convolutional Neural Networks. 2D-surfaces of the K-order tensor initially provide candidate retrieved results per ranking position and are merged to obtain the final list of retrieved results. Satellite image patches are used as queries in order to retrieve the most relevant image patches in Sentinel 2 images. Quantitative and qualitative results show that the proposed method outperforms search by a single modality and other late fusion methods.

Keywords: Late fusion · Multimodal search · Sentinel 2 images

1 Introduction

The amount of Earth observation (EO) data that is obtained increases day by day due to the multitude of sources orbiting around the globe. Each satellite image has a collection of channels/bands that provide a variety of measurements for each place on Earth. This advance of the satellite remote sensing technology has led to quick and precise generation of land cover maps with concepts (snow, rock, urban area, coast, lake, river, road, etc.) that distinguish the characteristics of the underlying areas, and provide beneficial information to global monitoring, resource management, and future planning.

Searching in large amounts of EO data with respect to a multimodal query is a challenging problem, due to the diversity and size of multimodal EO data, combined with the difficulty of expressing desired queries. The multimodal character of satellite images results from the various number of channels (e.g. Red, Green, Blue, NIR, SWIR, etc.) and associated metadata (date, time, geographical location, mission, etc.). Each satellite image can be considered as a collection of satellite image patches with semantic information about each one of them, as concepts correspond to each patch (e.g. urban area, rock, water, snow, etc.). The main challenge in searching for similar satellite image patches seems to be the combination of multiple heterogeneous features (modalities) that can be extracted from collections of satellite images (e.g. low-level visual descriptors, high-level textual or visual features, etc.). The aforementioned combination process is known as multimodal fusion. The effective combination of all available information (visual patterns and concepts, spatial and temporal) results to more effective similarity search, in the case of multimodal items such as Sentinel 2 images, or patches of them. The representation of each modality is also challenging, due to the availability of several Neural Network architectures that provide feature vectors and are trained to extract concepts.

Our contribution is summarized as follows. First, we propose a novel late fusion mechanism that combines K modalities through a K-order tensor. This tensor is generated by the results of multiple single-modality searches, which then provides the final merged and unified list of retrieved results through its 2D tensor surfaces. In addition, we propose a custom neural network for concept extraction in satellite image patches which outperforms similar and standard Neural Network architectures.

The paper is organised as follows. Section 2 presents relevant works in multimodal fusion for similarity search in information retrieval. Section 3 presents our proposed methodology, where each single modality provides a list of retrieved results and the K lists of rankings are fused. In Section 4 we describe the dataset we have used, the settings, as well as quantitative and qualitative results. Finally, Section 5 concludes our work.

2 Related Work

There are two main strategies for multimodal fusion with respect to the level, at which fusion is accomplished. The first strategy is called early fusion and performs fusion at the feature level [5, 12], where features from the considered modalities are combined into a common feature vector. Deep learning [7] makes use of deep auto-encoders to learn features from different modalities in the task of cross-modal retrieval. Similarly, [18] proposed a mapping mechanism for multimodal retrieval based on stacked auto-encoders. This mechanism learns one stacked auto-encoder for each modality in order to map the high-dimensional features into a common low-dimensional latent space. Modality-specific feature learning has also been introduced in [17], based on a Convolutional Neural Network architecture for early fusion. The second strategy is the late fusion that fuses information at the decision level. This means that each modality is first learned separately and the individual results are aggregated into a final common decision [19]. An advantage of early fusion inspired approaches [4] is the fact that it utilises the correlation between multiple features from different modalities at an early stage. However when the number of modalities increases, there is a decrease in their performance due to the fact that this makes it difficult to learn the cross-correlation among the heterogeneous features. On the other hand, late fusion is much more scalable and flexible (as it enables the use of the most suitable methods for analysing each single modality) than early fusion. With respect to graph-based methods and random-walk approaches [1] present a unifying multimedia retrieval framework that incorporates two graph-based methods, namely cross-modal similarities and random-walk based scores. However, the fusion is performed at the similarity level, before the retrieval of multiple ranked lists.

In remote sensing image retrieval task both traditionally extracted features and Convolutional Neural Networks (CNN) have been investigated with the latter ones presenting performance advantage. CNN models that aim for both classification prediction and similarity estimation, called classification-similarity networks (CSNs), outputs class probability predictions and similarity scores at the same time [10]. In order to further enhance performance, the authors combined information from two CSNs. "Double fusion" is used to indicate feature fusion and score fusion. Moreover, [11] proposed a feature-level fusion method for adaptively combining the information from lower layers and Fully Connected (FC) layers, in which the fusion coefficients are automatically learned from data, and not designed beforehand. The fusion is performed via a linear combination of feature vectors instead of feature concatenation. Another work is that of [16], who performed multiple SAR-oriented visual features extraction and estimated the initial relevance scores. For the feature extraction, they constructed two bag-of-visual-words (BOVWs) features for the SAR images and another SAR-oriented feature, the local gradient ratio pattern histogram. The authors calculated a set of initial relevance scores and constructed the modal-image matrix, then they estimated the fusion similarity and eventually re-ranked the results returned based on this similarity. The work of [9] uses multiple type of features to represent high-resolution remote sensing images. One fully connected graph and one corresponding locally connected graph were constructed for each type of feature. Furthermore, a fused graph was produced by implementing a cross-diffusion operation on all of the constructed graphs. Then, from the fused graph, the authors obtained an affinity value between two nodes that directly reflects the affinity between two corresponding images. Eventually, in order to retrieve the similar images retrieval, the affinity values between the query image and the other images in the image dataset are calculated. K-order tensors appear also in graph-based fusion mechanisms, as in [2], mainly for early fusion of multiple modalities for the creation and learning of a joint representation learning.

Contrary to these approaches, we perform an unsupervised late fusion of multiple rankings, without the construction of a joint representation learning at an early stage. Our late fusion approach first aims to optimize each single-modality search, either with existing features or with customized Deep Neural Network architectures. Our late fusion approach is agnostic to the representation of each modality as a vector an is easily adaptable to any meta-search engine.

3 Methodology

For the retrieval of similar-to-a-query q content in satellite image collections $\mathcal{S}$, different modalities are combined, each one representing a different aspect of the satellite images. The considered modalities are: i) visual features, ii) visual concepts, and iii) spatiotemporal information (geographical location and time).

The overall framework (Figure 1) involves initially a multimodal indexing scheme of each satellite image as an item with multiple modalities, such as visual features from several channels, visual concepts, spatial and temporal information. Each modality provides a similarity score and a ranked list of retrieved items that need to be combined so as to obtain a unique ranked list of satellite image patches. The query in the image collection per modality then provides a ranked list of items which are relevant to the query q, and a tensor is created (Figure 2). Thirdly, a bi-modal fusion of the retrieved results follows for each 2D surface of the created tensor and the rankings are merged in a late fusion approach, as shown in Figure 3.

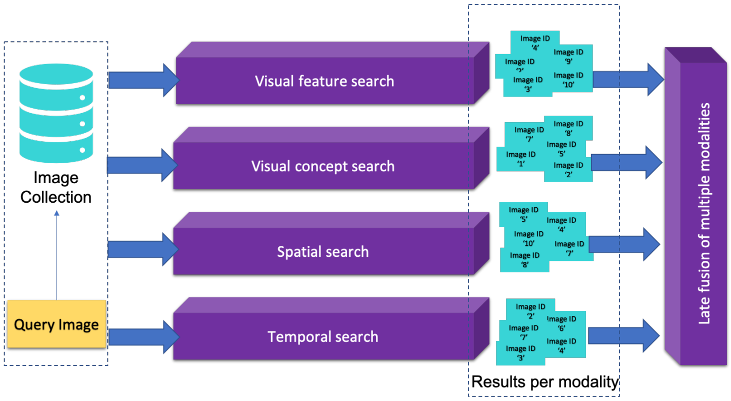


Fig. 1. Overall framework of our proposed retrieval of multiple modalities.

3.1 Late fusion of multiple modalities

The proposed approach fuses the output of K modalities, where $K > 2$. For each modality, we have N retrieved results and thus we have K such lists. We set as $\mathbf{L}$ the K-order tensor of the retrieved lists, $l_1, l_2, \ldots, l_K$. A single element $\mathbf{L}_{r_1,r_2,\ldots,r_K}$ of $\mathbf{L}$ is obtained by providing its exact position through a series of indices $r_1, r_2, \ldots, r_K$, defined as follows:

$$\mathbf{L}_{r_1,r_2,\ldots,r_K} = \begin{cases} 1, & \text{if the same element is ordered as } r_1 \text{ in list } l_1, \\ & \text{ as } r_2 \text{ in list } l_2, \ldots, \text{ and as } r_K \text{ in list } l_K \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Given the created tensor $\mathbf{L}$ (2), one tensor 2D surface is defined for each pair of modalities $(m_1, m_2), m_1 \leq m_2, 1 \leq m_1, m_2, \leq K$.

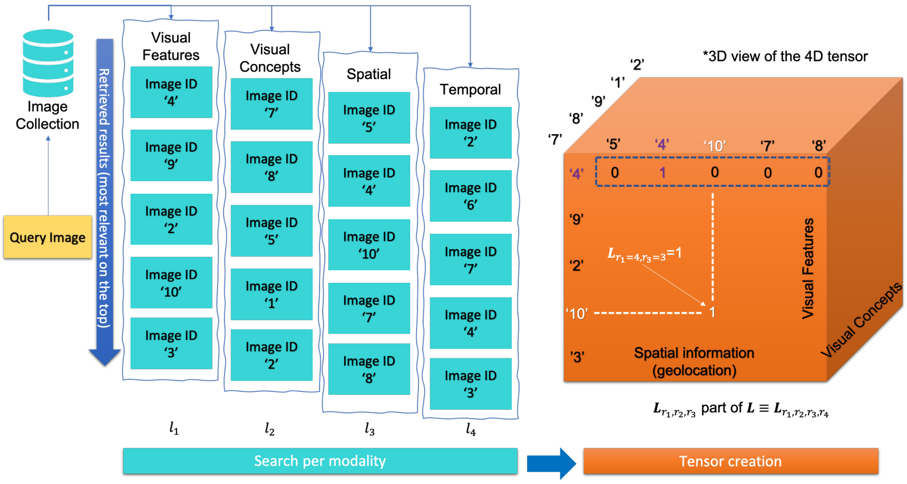


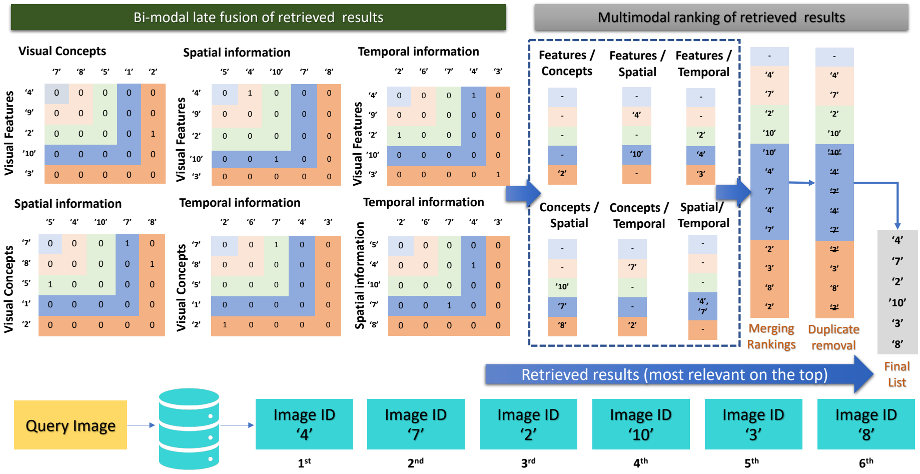
Fig. 2. Tensor creation from multiple lists of rankings.

For each tensor surface $\mathbf{L}(m_1, m_2)$, we denote by $\mathbf{L}_{r_{m_1} \leq j, r_{m_2} \leq j}(m_1, m_2)$ the tensor surface that is created only by the top-j retrieved results for the modalities m_1 and m_2. Similarly, we denote by $\mathbf{L}_{r_{m_1} \leq j-1, r_{m_2} \leq j-1}(m_1, m_2)$ the tensor surface that is created only by the top-$(j-1)$ retrieved results for the modalities m_1 and m_2. We create the list P_j, by keeping only the elements of the matrix $\mathbf{L}_{r_{m_1} \leq j, r_{m_2} \leq j}(m_1, m_2)$ which are not elements of $\mathbf{L}_{r_{m_1} \leq j-1, r_{m_2} \leq j-1}(m_1, m_2)$:

$$P_j = \mathbf{L}_{r_{m_1} \leq j, r_{m_2} \leq j}(m_1, m_2) \ominus \mathbf{L}_{r_{m_1} \leq j-1, r_{m_2} \leq j-1}(m_1, m_2) \tag{2}$$

If $\max\{P_j\} = 1$ then there is an element that appears for the first time in more than one modalities, with rank j. In Figure 3 we illustrate the passage from bi-modal fusion through tensor 2D surfaces to the final multimodal ranking of the retrieved results. For the tensor 2D surface that is extracted from visual concepts and visual features on the top-left we get the sequence of lists $P_1 = \{0\}$, $P_2 = \{0, 0, 0\}$, $P_3 = \{0, 0, 0, 0, 0\}$, $P_4 = \{0, 0, 0, 0, 0, 0, 0\}$, and $P_5 = \{0, 0, 1, 0, 0, 0, 0, 0, 0\}$. For the multimodal ranking of the visual concepts and visual features we get $\max\{P_j\} = 0$ for $j = 1, 2, 3, 4$, and $\max\{P_5\} = 1$, so the image ID '2' is temporarily ranked as 5^{th} in the "Features/Concepts" list. The same procedure is followed for all pairs of modalities. Afterwards, the image IDs in each position provide altogether a merged rankings ordered list as shown in Figure 3, where duplicates are removed and the final list is obtained.

In the following we present the uni-modal search per modality, before the creation of the unifying tensor $\mathbf{L}_{r_1, r_2, \ldots, r_K}$.


Bi-modal late fusion of retrieved results
Multimodal ranking of retrieved results

Visual Concepts

Visual Features	'7'	'8'	'5'	'1'	'2'
'4'	0	0	0	0	0
'9'	0	0	0	0	0
'2'	0	0	0	0	1
'10'	0	0	0	0	0
'3'	0	0	0	0	0

Spatial information

Visual Features	'5'	'4'	'10'	'7'	'8'
'4'	0	1	0	0	0
'9'	0	0	0	0	0
'2'	0	0	0	0	0
'10'	0	0	1	0	0
'3'	0	0	0	0	0

Temporal information

Visual Features	'2'	'6'	'7'	'4'	'3'
'4'	0	0	0	1	0
'9'	0	0	0	0	0
'2'	1	0	0	0	0
'10'	0	0	0	0	0
'3'	0	0	0	0	1

Spatial information

Visual Concepts	'5'	'4'	'10'	'7'	'8'
'7'	0	0	0	1	0
'8'	0	0	0	0	1
'5'	1	0	0	0	0
'1'	0	0	0	0	0
'2'	0	0	0	0	0

Temporal information

Visual Concepts	'2'	'6'	'7'	'4'	'3'
'7'	0	0	1	0	0
'8'	0	0	0	0	0
'5'	0	0	0	0	0
'1'	0	0	0	0	0
'2'	1	0	0	0	0

Temporal information

Spatial information	'2'	'6'	'7'	'4'	'3'
'5'	0	0	0	0	0
'4'	0	0	0	1	0
'10'	0	0	0	0	0
'7'	0	0	1	0	0
'8'	0	0	0	0	0

Features / Concepts	Features / Spatial	Features / Temporal	Concepts / Spatial	Concepts / Temporal	Spatial/ Temporal
-	-	-	-	-	-
-	'4'	-	-	'7'	-
-	-	'2'	'10'	-	-
-	'10'	'4'	'7'	-	'4', '7'
'2'	-	'3'	'8'	'2'	-

Merging Rankings	Duplicate removal	Final List
-	-	'4'
'4'	'4'	'7'
'7'	'7'	'2'
'2'	'2'	'10'
'10'	'10'	'3'
'10'	~~'10'~~	'8'
'4'	~~'4'~~	
'7'	~~'7'~~	
'4'	~~'4'~~	
'7'	~~'7'~~	
'2'	~~'2'~~	
'3'	'3'	
'8'	'8'	
'2'	~~'2'~~	

Fig. 3. Merging multiple lists from tensor surfaces.

3.2 Visual similarity search

For an effective search with respect to visual information, a suitable vector representation of the multi-channel Sentinel 2 image is required. After the transformation of each satellite image patch into embedding vectors, Euclidean distance calculations between the query q and the collection $\mathcal{S}$ represent visual similarity with respect to the content. Several visual feature representations are explored and the results are presented in Section 4. The visual similarity search is based on feature vectors that are extracted from pretrained networks and then Euclidean distance calculation follows on the top-N results. Feature vectors from Sentinel 2 images are extracted from specific intermediate layers of pretrained VGG19, ResNet-50 and Inception-ResNet-v2 networks [14]. Since ImageNet is a dataset of RGB images we created an input dataset of same type of images, i.e. Red (band 4), Green (band 3) and Blue (band 2) Sentinel-2 bands so as to form 3-channel patches. In VGG-19 convolutional neural network features are extracted from fc1 (dense) and fc2 (dense) layers, with feature size of 1 x 4096 float numbers per patch. In ResNet-50 features are extracted from avg_pool (GlobalAveragingPooling2) layer, with feature size of 1 x 2048 float numbers per patch. Finally, in Inception-ResNet-v2, which is a convolutional neural network with 164 layers the network has an image input size of 299-by-299.

3.3 Visual Concept search

A Custom Deep Neural Network is created for the extraction of visual concept vectors and therefore to support the visual concept search by Euclidean distance. The network has a structure that resembles VGG architecture (Figure 4). It contains blocks of convolutional layers with 3x3 filters followed by a max pooling layer. This pattern is repeating with a doubling in the number of filters with each block added. The model will produce a 7-element vector with a prediction between 0 and 1 for each output class. Since it is a multi-label problem, the sigmoid activation function was used in the output layer with the binary cross entropy loss function. For input we tested with both 3 channel images (as done with the pretrained networks) and also with images that consisted of 5 bands of Sentinel 2 images, i.e. the Red (band 4), Green (band 3), Blue (band 2), with the addition of NIR (band 8) and SWIR (band 11) for the 5-channel input.

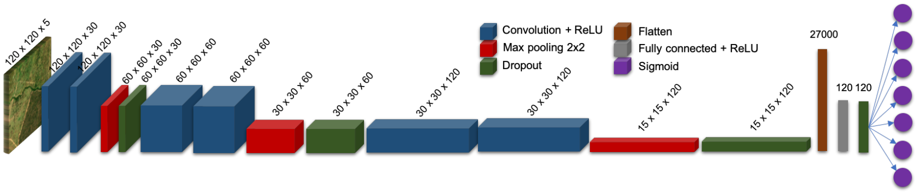


Fig. 4. The custom DCNN for visual concept search in 5-channel Sentinel 2 images.

3.4 Spatial and temporal search

Given a query image q, we exploit the "datetime" and "geolocation" metadata for spatiotemporal search. Our purpose is to maximise the proximity with respect to time and location between the query satellite image patch and the retrieved items. Our images are indexed in a MongoDB[1], which allows spatial search through the *geoNear* function. This function returns items ordered from the nearest to farthest from a specified point, i.e. the centroid of the query satellite image patch. Regarding the temporal information which is indexed in *IsoDate* form, sorting by timestamp provides a list of items which are temporaly close to the query image. Each modality allows unimodal search to retrieve a single list o returns a ranked similarity list. Performing late fusion on the formed lists returns the final sorted list with the closest images to the given query.

4 Experiments

4.1 Dataset Description

The BigEarthNet [15] dataset was selected for our experiments. The dataset contains ground-truth annotation about Sentinel 2 level-2A satellite images and consisted of 590,326 patches. Each image patch was annotated by the multiple land-cover classes (i.e., multi-labels) that were extracted from the CORINE

[1] https://www.mongodb.com/

Land Cover (CLC) inventory of the year 2018. Based on the available Corine land cover classes we group the closely related sub-classes of the CLC, forming seven major classes. We selected around 130,000 patches, of resolution 120 x 120 pixels in order to preserve a balance among the number of items of the different classes/concepts: 1) class “rice” 2) class “urban” merges “Continuous urban fabric” and “Discontinuous urban fabric”; 3) class “bare rock”; 4) class “vineyards”; 5) class “forest” merges “Broad-leaved forest”, “Mixed forest”, and “Coniferous forest”, 6) class “water” merges “Water courses”, “Water bodies” and “Sea and ocean” classes; 7) class “snow”.

4.2 Settings

For training of the custom DCNN and all pretrained deep neural networks we used Keras. Ssatellite image metadata and the extracted feature and concept vector are stored in MongoDB, that also allows spatial and temporal search. We select 10 test images for each of the seven classes parsed from the Corine Land Cover inventory, and thus we ended up with 70 test image patches. The procedure followed for obtaining the similarity according to the visual content involves the following steps: a) we extract feature vectors for each patch of the dataset, including the test images, b) we calculate the distance between the query image and the rest images of the dataset, c) we retrieve the images with the lowest distance from the query-test patch, and d) we calculate the mAP for the top 30 results. The learning rate for the 5-channel custom DCNN is 0.0005, the batch size is 256, and we used Adam optimizer with 200 epochs. To obtain the best possible results we enabled dropout regulation. We used the models with the best validation scores at a 5-Fold Cross-Validation.

4.3 Results

For the fusion of the results we tested our algorithm against three seminal rank fusion algorithms, namely Borda count [3], Reciprocal [13] and Condorcet [6] fusion, since they are suitable for Big Multimedia Data search [8].

Table 1. Comparison of fusion methods with mean average precision (mAP)

Classes / Method	Ours	Borda	Reciprocal	Condorcet
forest	89.56%	88.38%	60.11%	52.85%
rice	97.05%	98.92%	39.51%	66.61%
rock	62.90%	64.69%	26.53%	20.88%
snow	91.46%	89.44%	67.04%	15.22%
urban	79.96%	74.90%	53.72%	29.03%
vine	88.40%	88.25%	28.22%	19.54%
water	97.35%	97.97%	78.42%	76.05%
mAP:	**86.67%**	86.08%	50.51%	40.03%

For the evaluation of the various fusion methods we used the mean Average Precision (mAP) metric on the top-30 results that were retrieved. The comparison among our fusion method and the seminal methods of Borda, Condorcet and Reciprocal rank fusion, are shown in Table 1.

For obtaining these late fusion results, we performed experiments to identify the best performing unimodal search in visual features and concepts. The results for the pretrained and the custom neural networks are shown at Table 2 using Mean Average Precision as metric and are computed against the Corine Land Cover (CLC) annotation. The VGG19 architecture provides the optimal features for the multimodal retrieval problem. ResNet50 comes second and the Inception_v2 underperforms. Regarding the Visual concept similarity, the mAP results for the pretrained and the custom neural networks are shown at Table 3. The concepts are extracted in this case directly by the last prediction layer. The 5-channel custom DCNN obtains the best mAP results. Although it was expected that adding more channels in the DCNN architecture would lead to better performance, we observe in Table 2 that the pretrained network performs better than the custom.

Table 2. Mean average precision comparison on Feature extraction of seven classes among Pretrained networks, custom DNN and Colour Histogram methodologies.

	Pretrained Deep Neural Networks				Custom Deep Neural Network			
classes	VGG19 fc2	VGG19 flatten	ResNet50 avg_pool	Inception ResNetV2 avg_pool	5 bands flatten	5 bands dense	3 bands flatten	3 bands dense
				top #10				
forest	83.02%	81.17%	81.66%	63.70%	76.52%	80.38%	49.22%	50.89%
rice	86.79%	75.28%	57.68%	29.21%	25.89%	17.41%	30.40%	11.57%
rock	62.21%	76.38%	59.04%	58.09%	58.37%	52.96%	86.56%	60.44%
snow	86.37%	43.96%	91.85%	88.46%	74.93%	87.79%	48.03%	79.57%
urban	68.22%	45.46%	68.25%	73.43%	73.71%	66.53%	34.60%	42.77%
vine	74.74%	76.07%	67.85%	42.75%	45.44%	47.78%	59.67%	39.51%
water	98.78%	100.00%	100.00%	96.20%	100.00%	97.11%	95.22%	92.68%
mAP	**80.02%**	71.19%	75.19%	64.55%	64.98%	64.28%	57.67%	53.92%
				top #20				
forest	78.72%	80.07%	77.63%	62.08%	76.12%	70.72%	45.98%	51.55%
rice	82.09%	72.58%	49.74%	31.58%	21.01%	15.58%	30.40%	12.80%
rock	50.41%	62.01%	51.59%	50.85%	46.30%	44.57%	83.94%	54.99%
snow	81.07%	44.04%	90.92%	88.09%	74.52%	81.62%	49.20%	66.76%
urban	61.27%	40.80%	64.92%	70.20%	69.26%	60.82%	30.85%	38.54%
vine	65.77%	70.44%	61.53%	41.98%	41.55%	43.53%	44.75%	34.45%
water	98.83%	100.00%	99.66%	97.00%	99.89%	96.58%	96.01%	92.27%
mAP	**74.02%**	67.13%	70.86%	63.11%	61.24%	59.06%	54.45%	50.19%

We demonstrate the top-10 retrieved results for our proposed approach in Figure 5. The satellite image patch on the left is the query and the retrieved re-

Table 3. Mean average precision comparison on Concept extraction of seven classes among Pretrained networks, custom DNN and Colour Histogram methodologies.

Pretrained Deep Neural Networks				Custom Deep Neural Network		
classes	VGG19 predictions	ResNet50 fc1000	Inception ResNetV2 fc1000	classes	5 bands dense (last)	3 bands dense (last)
			top #10			
forest	63.59%	71.28%	66.67%	forest	80.38%	49.22%
rice	19.60%	3.25%	34.16%	rice	17.41%	30.40%
rock	22.70%	14.13%	30.68%	rock	52.96%	86.56%
snow	63.48%	84.47%	91.14%	snow	87.79%	48.03%
urban	55.66%	69.17%	58.85%	urban	66.53%	34.60%
vine	44.78%	29.43%	48.55%	vine	47.78%	59.67%
water	93.44%	97.47%	99.77%	water	97.11%	95.22%
mAP	51.89%	52.74%	61.40%	mAP	**64.28%**	57.67%
			top #20			
forest	61.38%	66.58%	60.16%	forest	70.72%	45.98%
rice	18.49%	3.54%	27.07%	rice	15.58%	30.40%
rock	22.98%	15.85%	28.73%	rock	44.57%	83.94%
snow	62.33%	83.91%	87.47%	snow	81.62%	49.20%
urban	54.03%	65.62%	54.58%	urban	60.82%	30.85%
vine	39.65%	28.15%	43.32%	vine	43.53%	44.75%
water	93.49%	96.73%	98.52%	water	96.58%	96.01%
mAP	50.33%	51.48%	57.12%	mAP	**59.06%**	54.45%

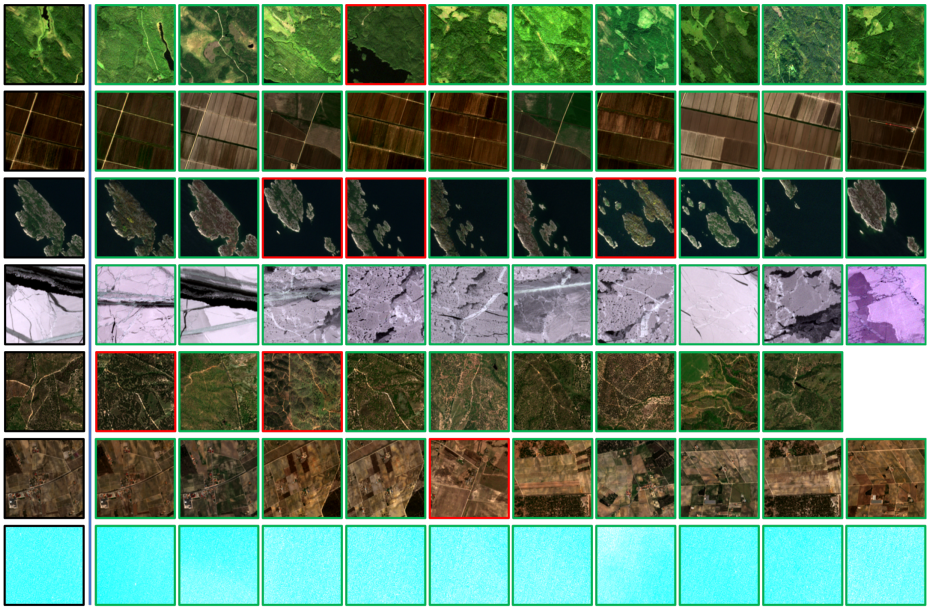


Fig. 5. Query and top-10 retrieved results with the proposed late fusion approach.

sults follow. The results show that, for the urban query, most of the misclassified results are rice and these two classes resemble to each other making difficult for the DCNNs to discriminate among them. Moreover, some of the retrieved images are of the forest class, because in some cases they depict sparse country-side areas mixed with snow. Furthermore, in the "vineyards" query, almost all the misclassified images were actually urban patches, and there is great similarity between these two classes, and even for a human it is difficult to classify. Finally, rock queries are mostly rocky areas near water, resulting to fetching many water patches.

5 Conclusions

In this work we proposed a novel late fusion method that combines the outputs of K ranking lists using a K-order tensor approach. The method is agnostic to the representation of each modality in each unimodal search. However, we examined the performance of several DCNN architectures and we proposed one for concept search in 5-channel Sentinel 2 image patches. Satellite images contain more than the three optical RGB channels that can also be exploited in unimodal similarity search. Our overall framework uses 5 channels from Sentinel 2 image patches to extract concepts, and combines them with visual features and spatiotemporal information to allow multimodal similarity search scenarios. Finally, the results show the importance of combining multiple modalities of an image in similarity search and we illustrate the top-10 results per late fusion method and per query.

Acknowledgements

This work was supported by the EC-funded projects H2020-832876-aqua3S and H2020-776019-EOPEN.

References

1. Ah-Pine, J., Csurka, G., Clinchant, S.: Unsupervised visual and textual information fusion in cbmir using graph-based methods. ACM Transactions on Information Systems (TOIS) **33**(2), 1–31 (2015)
2. Arya, D., Rudinac, S., Worring, M.: Hyperlearn: a distributed approach for representation learning in datasets with many modalities. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2245–2253 (2019)
3. Aslam, J.A., Montague, M.: Models for metasearch. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 276–284 (2001)
4. 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)
5. Caicedo, J.C., BenAbdallah, J., González, F.A., Nasraoui, O.: Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. Neurocomputing **76**(1), 50–60 (2012)

6. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 758–759 (2009)
7. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 7–16 (2014)
8. Gialampoukidis, I., Chatzilari, E., Nikolopoulos, S., Vrochidis, S., Kompatsiaris, I.: Multimodal fusion of big multimedia data. Big Data Analytics for Large-Scale Multimedia Search pp. 121–156 (2019)
9. Li, Y., Zhang, Y., Tao, C., Zhu, H.: Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion. Remote Sensing **8**(9), 709 (2016)
10. Liu, Y., Chen, C., Han, Z., Ding, L., Liu, Y.: High-resolution remote sensing image retrieval based on classification-similarity networks and double fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 1119–1133 (2020)
11. Liu, Y., Liu, Y., Ding, L.: Scene classification based on two-stage deep feature fusion. IEEE Geoscience and Remote Sensing Letters **15**(2), 183–186 (2017)
12. Magalhães, J., Rüger, S.: An information-theoretic framework for semantic-multimedia retrieval. ACM Transactions on Information Systems (TOIS) **28**(4), 1–32 (2010)
13. Montague, M., Aslam, J.A.: Condorcet fusion for improved retrieval. In: Proceedings of the eleventh international conference on Information and knowledge management. pp. 538–548 (2002)
14. Moumtzidou, A., Bakratsas, M., Andreadis, S., Karakostas, A., Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I.: Flood detection with sentinel-2 satellite images in crisis management systems. ISCRAM 2020 Conference Proceedings – 17th International Conference on Information Systems for Crisis Response and Management pp. 1049–1059 (2020)
15. Sumbul, G., Charfuelan, M., Demir, B., Markl, V.: Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 5901–5904. IEEE (2019)
16. Tang, X., Jiao, L.: Fusion similarity-based reranking for sar image retrieval. IEEE Geoscience and Remote Sensing Letters **14**(2), 242–246 (2016)
17. Wang, J., He, Y., Kang, C., Xiang, S., Pan, C.: Image-text cross-modal retrieval via modality-specific feature learning. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. pp. 347–354 (2015)
18. Wang, W., Ooi, B.C., Yang, X., Zhang, D., Zhuang, Y.: Effective multi-modal retrieval based on stacked auto-encoders. Proceedings of the VLDB Endowment **7**(8), 649–660 (2014)
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)