Cross-year multi-modal image retrieval using siamese networks
Margarita Khokhlova, Valérie Gouet-Brunet, Nathalie Abadie, Liming Chen

To cite this version:

HAL Id: hal-02903434
https://hal.science/hal-02903434
Submitted on 21 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CROSS-YEAR MULTI-MODAL IMAGE RETRIEVAL USING SIAMESE NETWORKS

Margarita Khokhlova\textsuperscript{1,2}, Valérie Gouet-Brunet\textsuperscript{1}, Nathalie Abadie\textsuperscript{1}, Liming Chen\textsuperscript{2}

\textsuperscript{1}LASTIG, Univ Gustave Eiffel, ENSG, IGN, F-94160 Saint-Mande, France,
\textsuperscript{2}LIRIS/Lyon Centrale, 36 av Guy de Collongue, 69134 Écully, France

ABSTRACT

This paper introduces a multi-modal network that learns to retrieve by content vertical aerial images of French urban and rural territories taken about 15 years apart. This means it should be invariant against a big range of changes as the (natural) landscape evolves over time. It leverages the original images and semantically segmented and labeled regions. The core of the method is a Siamese network that learns to extract features from corresponding image pairs across time. These descriptors are discriminative enough, such that a simple kNN classifier on top, suffices as final geo-matching criteria. The method outperformed SOTA “off-the-shelf” image descriptors GEM and ResNet50 on the new aerial images dataset.

Index Terms— Siamese networks, multi-modal CBIR.

1. INTRODUCTION

Aerial images, such as images from satellites or other aerial imaging devices, are distinctly different from image datasets such as CIFAR \cite{1}, Imagenet \cite{2}, etc. These images are much more semantically similar in composition as they capture natural and urban landscapes, which are all made up of visually near-identical elements such as vegetation and man-made structures. Lately, a great volume of historical images was digitized, among them many aerial images through national surveys mainly from mapping agencies \cite{3}. They are a unique resource to study landscape evaluation, urbanization, land usage, historical events, and others. Alegoria project \cite{4} aims to create content-based image retrieval (CBIR) tools to help end users accessing such volumes of images. The difficulty is that many photographic materials are scarcely, or not at all annotated, which makes it hard to link them to modern photographic images of the same territory.

State-of-the-art (SOTA) approaches for image retrieval \cite{5,6} are designed to deal with cross-view and multi-modality challenges but are trained and tested on benchmarks composed of important and distinct man-made architectures \cite{7,8}, which are very different from aerial images, if only by the resolution and composition. It is, therefore, an open research question whether these methods can be used for cross-time aerial image retrieval. An additional question is whether the semantic information from cartographic maps can be beneficial for such a cross-time image retrieval task and if so, how it can be exploited and encoded by a descriptor? This paper aims to match urban and rural vertical aerial images, taken 15 years apart. Obviously, these images differ as the landscape evolves over time, which makes matching non-trivial. The core idea is to learn an embedding that retains all information required to recover the scene appearance through time and under varying acquisition conditions. The dataset contains images in pairs, one from each decade, next to manually labeled segmented semantic masks. An example of the data can be seen in Figure 1. The sought-after method must be able to distinguish between images that are semantically close, as all contain similar elements, yet also must be robust against change, the appearance and disappearance of objects over time and even seasonal effects. Lastly, as we are learning to match image pairs, each image pair is a class by itself. Unlike Imagenet object recognition tasks containing at least 500 images per class, we deal with only 2 training images per correspondence.

Our contributions are the following. Firstly, we evaluate the performance of existing methods for a new cross-time retrieval task using a dataset created for this purpose. We determine the most important data modality and evaluate several scenarios for multi-modal fusion. Secondly, we propose a novel descriptor for multi-modal data and fine-tune it on our dataset. The core of our method is formed by a Siamese network that takes image pairs as an input. The image pairs contain not only natural images but also the semantic labels corresponding to each image, which makes our approach multi-modal. The net outputs a single descriptor per image pair that captures the similarity whilst being robust against all changes occurred over time. This descriptor is low-dimensional but powerful enough such that the final classification whether an image pair is a temporal match may be done by a simple unsupervised k-nearest neighbors (kNN) method.

2. PROBLEM STATEMENT AND BACKGROUND

The dataset originates from French Mapping Agency (IGN) \cite{9} and contains vertical aerial images taken from three
French regions (Moselle, Bas-Rhin, and Meurthe-and-Moselle) in 2004 and 2019. Consequently, we call it FR-0419. The extra modality is formed by per pixel semantic annotations, similar to those found in traditional cartographic maps. The number of image pairs for the regions is 6000, 4430 and 5855.

We selected several categories of semantic objects (See Table 2). Note that the number of annotated objects may differ significantly, partly due to the different annotation strategies and partly due to landscape evolution. The aerial images are 50cm/pixel and used in patches of a square kilometer. We use perfectly matching image coordinates between the years - i.e. the image regions are aligned. This scenario is not realistic but allows us to test and demonstrate the robustness of descriptors against changes and evolutions in the landscape.

**Problem statement.** This paper ascertains to what extent existing descriptors may be used to match aerial images through time, and which information (visual, semantic, or both) is more relevant for this task. We test different fusion strategies to encapsulate all modalities into a single multi-dimensional descriptor. Images from 2019 are used queries against their 2004 counterparts, and different geographical regions are selected to form a training, validation, and testing sets. Lastly, we introduce our new Siamese net-based descriptor.

**Background.** Recently with the progress of segmentation Convolutional Neural Network (CNN) architectures the task of fully automated scene segmentation became possible [10]. Semantic maps are an incredible additional source of information that can potentially improve geolocalisation, cross-view and cross-time retrieval. Combining different information sources and modalities to improve (CNN) models was explored in [11, 12, 13, 14]. However, to the best of our knowledge, using multi-modal data to retrieve aerial images taken in different years, is novel. A related research problem is visual localization, where acquisitions differ in viewing conditions and suffer from a wide range of distortions [5, 6, 15] or extreme viewpoint changes [16, 17]. Traditionally the goal is to represent image features as robust feature vectors, contemporary work focuses on learning based methods [14, 5, 6]. However, current methods are designed to handle object-specific features and are not tailored to retrieve images over time, where the scenes might not contain a single outstanding key object being composed of repetitive man-made structures instead. Hence these methods cannot be applied straightforwardly in the case of large landscape changes over time along with non-characteristic (distinct) image features.

Aerial image descriptors can greatly benefit from the data of other modalities. This was successfully demonstrated by Audebert et al. [13], where better segmentation maps were obtained using an encoder-decoder architecture and images along with semantic data originating from OpenStreetMap [18]. Up to our knowledge, this is the only work that directly uses semantic labels along with images in an aerial image context. Li et al. [19] proposed a multi-modal late feature fusion-based framework to improve the geographic image annotation. However, they source from a single image, where the different modalities are simply features extracted by different algorithms. Chen et al. [20] propose a CBIR method benefiting from multi-modal (spatial and spectral) information content of Remote Sensing images, however, they use hand-crafted descriptors and focus on the retrieval task.

Siamese network architectures aim to construct an embedding, where two extracted features corresponding to the same identity are likely to be closer, than features from different identities [21]. They are a popular choice for problems dealing with so-called one-shot learning problems, when a single training sample is available for each class. The efficiency of Siamese networks was previously demonstrated for visual object tracking [22], person reidentification [23], cross-view image matching [24] and other tasks.

We propose to use a custom Siamese architecture to obtain an embedding, which encodes the visual features and learn to ignore any temporal induced changes. It allows us to train descriptors on our single-pair correspondence dataset, whilst the backbone architecture is simultaneously designed to benefit from multi-modal information.

**3. PERFORMANCE BASELINE**

Up to our knowledge, there are no publicly available pre-trained image descriptors, that are fine-tuned on aerial images, nor are there any quantitative studies comparing image presentations specifically for multi-modal aerial images. We, therefore, establish our own baseline performance benchmark using existing image descriptors.

The baseline is formed by Resnet [25] and GEM [5], both global in their nature. They are pre-trained on Imagenet [26] and Oxford5k [26] respectively. We use the output of the last convolutional layer followed by max-pooling to obtain a descriptor for Resnet and the pre-trained framework as provided by the authors for GEM. The first step is to comparatively evaluate methods, their combinations, and parameters for accurate cross-year image matching. The baseline study using the algorithm(s) is schematically depicted in Figure 2.

**Cross-time image retrieval setup.** As a benchmark, we evaluate the retrieval accuracy of the "off-the-shelf" descriptors for different data modalities. We resize the input images to the size of 512x512x3 for Resnet50 and 1024x1024x3 for GEM [5]. We evaluate three scenarios:

<table>
<thead>
<tr>
<th>Data type</th>
<th>Color RGB</th>
<th># 2004</th>
<th># 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>road</td>
<td>(255,165,0)</td>
<td>380731</td>
<td>326882</td>
</tr>
<tr>
<td>church &amp; chapel</td>
<td>(255,255,0)</td>
<td>1292</td>
<td>2195</td>
</tr>
<tr>
<td>fort &amp; blockhaus</td>
<td>(128,128,128)</td>
<td>481</td>
<td>734</td>
</tr>
<tr>
<td>other building</td>
<td>(255,0,0)</td>
<td>251294</td>
<td>3475104</td>
</tr>
<tr>
<td>water resource</td>
<td>(0,0,255)</td>
<td>28043</td>
<td>12040</td>
</tr>
<tr>
<td>sport ground</td>
<td>(138,43,226)</td>
<td>1409</td>
<td>2839</td>
</tr>
<tr>
<td>cemetery</td>
<td>(75,0,130)</td>
<td>928</td>
<td>1299</td>
</tr>
<tr>
<td>vegetation zone</td>
<td>(0,255,0)</td>
<td>224101</td>
<td>164435</td>
</tr>
<tr>
<td>railroad</td>
<td>(255,0,255)</td>
<td>3308</td>
<td>3972</td>
</tr>
</tbody>
</table>

**Table 1.** Main semantic categories stats. Moselle 2004-2019.
The weights of the convolutional fusion layer are pre-defined.

We use the Mean Average Precision (map@N) to evaluate the results: \( \text{map@N} = \frac{\sum_{m=1}^{M} \sum_{n=5}^{N} \delta}{r M N} \), where \( N = 5 \), \( r \) is equal to 1 if the retrieved image is correct and 0 otherwise and \( n \) is the order of retrieved images, \( M \) the total number of images. Tables 2 and 3 summarise the averaged map@N results. We experimented the retrieval with different distances (Euclidean, Cosine), RGB and grayscale semantic masks, different types of re-scaling techniques \((L_1, L_2, \text{standardization})\), and concatenation or sum of the obtained descriptors. We observed that different parameter combinations affect the final map@5 accuracy significantly yielding up to 5% variations. However, we did not establish a single common trend for the three regions tested apart of cosine distance constant outperforming the euclidean in the kNN matching stage. We, therefore, report the best map@5 score and a parameter set.

Baseline results. Firstly, descriptors based on semantic information give better results in terms of map@5 value than descriptors based on natural images. The best map@5 scores were obtained using both modalities which confirms that the additional information is beneficial for the cross-time image retrieval task. Moreover, the combination of visual and semantic data at the late stage allows improving the results even further for the descriptors tested. Overall, the obtained results are not good, showing the limitations of existing CNN-based approaches when dedicated training datasets are not available.

4. SIAMESE ARCHITECTURE

Our multi-modal Siamese network architecture is schematically illustrated in Figure 3. The architecture has two copies of the function \( G_W \), which share the same set of parameters \( W \), consists of two branches and and a distance module. A loss module is placed on top of the architecture.

The network architecture is designed to handle the multi-modal input and fine-tune the descriptors for the image retrieval task. The input to the CNN is a pair of multi-modal images \((X_1, L_1; X_2, L_2)\) and a label \( Y \). One branch processes corresponding image pairs originating from the same geographical zone through time (2019-2004), the other processes a non-corresponding pair from the same year (2019). The images are passed through, yielding two outputs \( G(X_1, L_1) \) and \( G(X_2, L_2) \). The cost module then generates the distance \( D_W(G_W(X_1, L_1), G_W(X_2, L_2)) \). The loss function combines the probability \( p \) predicted by the classification layer with the sigmoid activation based on \( D_W \) with label \( Y \) to produce the scalar loss value:

\[
\mathcal{L}(W) = -\frac{1}{M} \sum_{m=1}^{M} L(W, (Y, X_1, X_2, S_1, S_2)^m) \tag{1}
\]

\[
D_W(X_1, S_1, X_2, S_2) = |G_W(X_1, S_1) - G_W(X_2, S_2)| \tag{2}
\]

\[
\mathcal{L}(W, (Y, X_1, X_2, S_1, S_2)^m) = Y \log(p) + (1 - Y) \log(1 - p) \tag{3}
\]

where \( m \) is the number of pairs. The first layer of the network is a convolutional layer with pre-defined trainable weights which serves to warm-up the training. The backbone of the network is Resnet50 pre-trained on ImageNet. The output of the last convolutional layer of Resnet is passed through 3 convolutional layers \( C_1-C_3 \) followed by a fully connected layer. We found that using the \( \tanh \) activation and batch normalization in all the added convolutional layers gives the best result.

Mining so-called hard image pairs is essential to make proper training of Siamese nets possible [27]. We adopted the following learning strategy: every 5 epochs, the map@5 score is calculated for the training dataset. Hard image pairs are the ones that have wrong retrieved images (i.e retrieved images do not correspond to the same geographical zone).

The code and weights of the trained model are available\(^1\).

5. EXPERIMENTS

In this section, we describe the experimental setup used to compare our method to the baseline results obtained with “off-the-shelf” image descriptors on FR-0419 benchmark dataset. We fine-tune the proposed architecture in an end-to-end fashion using the images from the Moselle region for training, whereas Bas-Rhin is a validating set and Meurthe and Moselle form the testing set. Throughout all experiments, the architecture from Figure 3 is used and the input size of both aerial and semantic images is 256x256. The fusion by convolution is deployed to combine the multi-modal data and allow end-to-end training. The first convolutional layer is pre-initialized. The 3 convolutional layers \( C_1-C_3 \) on top of the Resnet have \( 3 \times 3 \) kernels and the number of filters equals 1024, 512, and 256. The final descriptor dimension is \( D^{128} \), next to testing the values of 256, and 512. See Table 5, \( R=128 \) dimensions generalizes the best.

The network learns to predict whether two multi-modal descriptors correspond to the same geographical zone through time, based on the \( L_1 \) distance between them. The idea is similar to contrastive loss [28] commonly employed in Siamese networks, but we found this approach to work better.

We use BCE loss, next to Adam optimizer with a fixed \( lr \) 8e-4 and decay 8e-7. We re-determine hard samples after every 5 epochs based on the training set map@5 scores. The kNN algorithm with cosine distance is used to retrieve

\(^1\)https://github.com/margokhokhlova/siamese_net
the most similar images given a query. During training, each batch is composed of 12 pairs of positive and negative images, half of which are randomly selected and half are the hard images. In each epoch we go through all training set images, each time selecting random negatives and adding hard-mining samples into a batch. Data augmentation consisted solely of vertical and horizontal image flips. The final model was trained 120 epochs. The map@5 score on the validation set determines the best descriptor parameters. We also cross-validated by swapping the regions for training/validation and test and re-training the net from zero. Once the net is tuned, we use one branch to calculate a multi-modal input descriptor.

Table 5 summarises the final map@5 scores obtained.

<table>
<thead>
<tr>
<th>Data</th>
<th>Norm</th>
<th>Fusion</th>
<th>Distance</th>
<th>R</th>
<th>Map@5 Training</th>
<th>Map@5 Validation</th>
<th>Map@5 Testing</th>
<th>Testing</th>
<th>Map@5 Training</th>
<th>Map@5 Validation</th>
<th>Map@5 Testing</th>
<th>Map@5 Training</th>
<th>Map@5 Validation</th>
<th>Map@5 Testing</th>
<th>Map@5 Training</th>
<th>Map@5 Validation</th>
<th>Map@5 Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>visual image</td>
<td>n/a</td>
<td>none</td>
<td>cosine</td>
<td>2048</td>
<td>0.54</td>
<td>0.63</td>
<td>0.59</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>semantic image</td>
<td>n/a</td>
<td>none</td>
<td>euclidean</td>
<td>2048</td>
<td>0.63</td>
<td>0.64</td>
<td>0.61</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vis + semantic</td>
<td>mean std</td>
<td>D concatenation</td>
<td>cosine</td>
<td>4096</td>
<td>0.66</td>
<td>0.69</td>
<td>0.71</td>
<td>0.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vis + RGB sem</td>
<td>none</td>
<td>image fusion</td>
<td>cosine</td>
<td>2048</td>
<td>0.51</td>
<td>0.52</td>
<td>0.56</td>
<td>0.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vis + RGB sem</td>
<td>none</td>
<td>late fusion</td>
<td>cosine</td>
<td>2048</td>
<td>0.75</td>
<td>0.73</td>
<td>0.84</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Off-the-shelf descriptor map@5 for pre-trained Resnet50 [25]. Best scores per per department are shown in bold.

Table 3. Off-the-shelf descriptor map@5 for pre-trained Gem [5], architecture resnet101-gem-reg-whiten.

6. CONCLUSION

In this study we tackle cross-time aerial image retrieval. We introduced a novel approach for learning from multi-modal data to fine-tune any CNN-based image descriptor. We performed a comprehensive comparison of different strategies to use multi-modal information and proposed a custom Siamese network architecture. The resulting descriptor is powerful enough to distinguish between images that are semantically close and is robust against evolutionary landscape changes through time. Experiments show that our method improves the baseline and outperforms SOTA image descriptors. We demonstrated how to use both image and semantic modalities in a single descriptor. In addition, the method is generalizable to any CNN-based feature extractor.

7. ACKNOWLEDGEMENTS

This work is supported by ANR, the French National Research Agency, within the ALEGORIA project, under Grant ANR-17-CE38-0014-01.
8. REFERENCES


