Icono Lausanne: Reorientation of the photographic archives of the Historical Museum

Boubacar Camara
Student ID: 262276

Msc Computer Science

EPFL

Semester Project

Professors: Frédéric Kaplan, Isabella di Leonardo
Supervisors: Rémi Petitpierre

Spring 2021

List of Figures

Table of Contents

Chapter 1

Introduction

The Lausanne historical museum has collected 38'680 historical images from the Lausanne region, mainly between the 18th and 21th century. This project consists in the development of semi-automatic methods to infer the position and orientation in which photographs have been taken. A successful geo-referencing of historical images would open up new possibilities to explore historical image datasets via projection onto 4D models or Augmented reality applications. This task is particularly challenging considering the low resolution of the images and the large size of the dataset. In this report, we will present the different steps the pipeline developed in order to geo-reference historical images. We will evaluate its precision on the site of Tribunal de Montbenon. Then we will use it to geo-reference images cluster from Place Saint-François and Place de la Riponne.

Chapter 2

Related Work

Image geo-referencing is a challenging task which has gained attention in the last decades considering the amount of images available on the internet as well as archive collections. A pillar work in the domain is the Photo-tourism project [5] where the authors built a system taking as input a large collections of unordered photographs from historical sites, and automatically compute each photo's viewpoint, build a sparse model of the scene and geo-reference the photos.

The core of their algorithm is a photogrametric pipeline where images are matched pairwise using SIFT features. These features are then fused into tracks, being sets of features visible from different images. Finally a structure from motion algorithm is applied in order to compute the relative position and orientation of the images. Images are then geo-referenced either via a manual alignment of the set of recovered cameras and pointcloud on an aerial 2D map of the target site. Automatized alignments techniques have been developed in a second phase.

However, the Photo-Tourism project and other initiatives aiming at enhancing a such pipeline use recent images with a reasonable quality, not photographs dating from more than 50 years. In their research, Maiwald et al. [4] have studied the usability of historical sites archive images in a photogrammetric pipeline, in order to generate a 3D model from the resulting point cloud and obtain the relative position of the cameras. They have shown this was possible, with a careful selection of images with a sufficient resolution, making sure that enough angles of the target site were covered and that the baseline between the images was not too large. The quality of the reconstruction could be improved by adding actual images to the pipeline when the observed architecture did not significantly change over time.

Chapter 3

Dataset

The MHL dataset contains 38'680 archive images with metadata fields. The images contain a checkerboard as shown in the image below and require a pre-processing phase to be extracted. This is possible by using dhSegment [3].

Each image in the dataset has 50 fields of metadata, such as the date of production, production technique, geolocation information (city, street name, street number), keywords, authors of the photograph etc. These metadata are however very heterogeneous as they have been entered manually by archivists. Some fields may be missing, or some values representing the same entity may be written in different ways. After a phase of data cleaning and standardization here are some statistics obtained:

- Total number of images: 38'680
- Number of street images: 28'686
- Number of images without geolocation information: 13'000

It can be observed that among these numbers about 10'000 images do not represent streets. These images were filtered out during the exploration phase because they represent posters, photographs of objects or furniture and are not useful for this project.

Moreover, the images are not distributed in a uniform way. As can be seen on Figure 3.1, streets near touristic places contain many more images than others. Given the

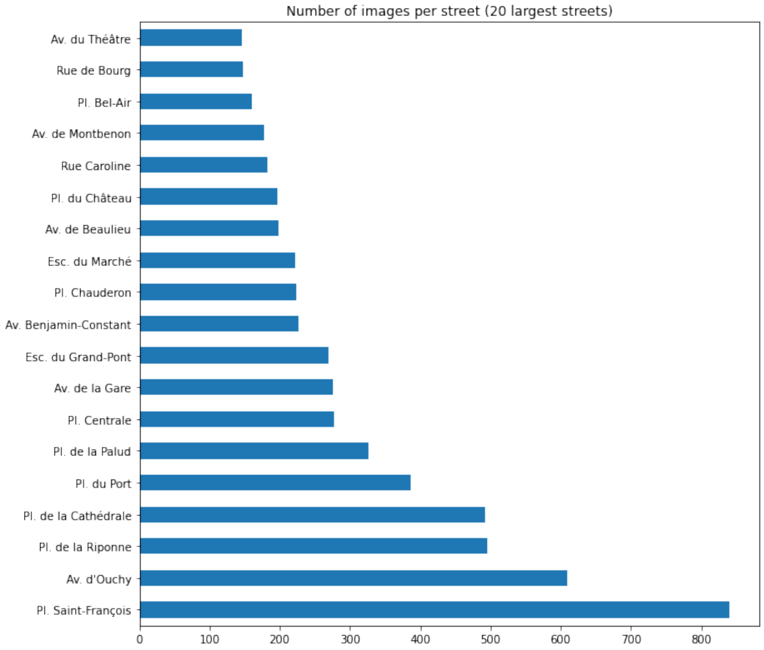


Figure 3.1: Images distribution in the 20 largest streets

large number of images available, it is necessary to be able to group them in clusters according to their degree of similarity.

Chapter 4

Methods

In this section, we will describe the different components of our semi-automatic pipeline to recover the position and orientation of the images. To this end, we identify images cluster, compute the relative poses of their images and align them using reference images annotated with their real word position and orientation.

4.1 Dataset exploration and clustering

To group similar images in clusters, we tried two approaches. The first is a clustering based on the keywords available in the metadata data using the Latent Dirichlet Allocation method, while the second is based on a clustering of features extracted with the neural network InceptionV3, using the Pixplot project.

4.1.1 Keyword based clustering

The latent Dirichlet Allocation (LDA) is a NLP topic modeling technique which given a set of documents, finds the distribution of each document across K topics discovered in unsupervised manner. As images are annotated with keywords, we can expect that images with similar keywords should have similar content in their images. Hence, applying LDA to the images keywords, LDA allows us to estimate the usability of the keywords in the discovery of clusters. We applied LDA to the 20 largest streets and generated an interactive web page where each cluster is rendered as a complete graph for visualization convenience. However, the results were not optimal as keywords were not discriminative enough to provide an efficient clustering.

4.1.2 Image features based clustering

Pixplot is a project aiming at facilitating the visualization of image databases containing thousands of documents. In order to so, image features are extracted using the InceptionV3 deep neural network. The dimensionality of these features is reduced to obtain two-dimensional points which are then clustered via KMeans. Images can be browsed in this two-dimensional space in a web interface. This process allows us to easily identify clusters of similar images. This approach provided satisfying results and we used it to identify image clusters for the rest of the project.

4.2 Image Geo-Referencing

We've seen how it is possible to identify image clusters. In order to geo-reference images, we will use the photogrammetry software Meshroom [2] to recover the relative position and orientation in wihch cameras have taken the pictures. By including geo-referenced images from google streetview in the input images, or manually referencing some of the images, we can obtain baseline points that can be used to estimate the geographical position of the reconstructed cameras during the photogrammetric pipeline and computing the alignment error.

4.2.1 Pose Recovery

Once a cluster of similar images has been obtained, we can begin the process of recovering the pose and orientation of the images. For this, we use the Meshroom framework. This allows us to apply a photogrammetry pipeline composed of SIFT feature extraction, image and feature matching, and finally a structure from motion step. Fundamentally, this is a classical photogrammetric pipeline where the features are organized into tracks and an incremental reconstruction is then launched to compute 3D coordinates of the images features, as well as the cameras intrinsics (focal and distorion factors) and extrinsic parameters (rotation matrix and camera position). Bundle adjustment is used in order to minimize the reprojection error of the observed 3D points onto cameras planes, hence refining the point cloud and cameras parameters. The reconstructions outputs the coordinates of the triangulated 3D points, the recovered cameras center and rotation matrix. Euler rotation angles can be extracted from the rotation matrix in order to obtain the cameras orientations.

Due to the low resolution of the images we use a Residual Dense Network for Single Image super-resolution [1] in order to double the resolution of the images and see if this can increase the pose recovery quality.

4.2.2 Pose alignment

We've seen how to obtain the relative pose and orientation of the cameras using a photogrammetric pipeline. Our goal is now to assign GPS coordinates to each recovered camera. In order to do so we need at least to know the latitude and longitude coordinates of at least two points which have been recovered during the photogrammetric reconstruction. We will use these coordinates as references to triangulate the GPS position and true orientation of the other recovered cameras. To include images with known location in the pipeline we have two options. The first is to download geo-referenced images from Google street view, and the second is to manually annotate some images of the MHL dataset with an estimation of their geographical position and orientation.

Google Street view image extraction

Google allows registered users to download Street view images at specified coordinates and orientation via its MAPS API. The camera orientation is defined as heading (angle with respect to the north), and pitch (vertical angle with respect to the horizon). Hence we use Google Maps in order to mark the places of interest where Google street view images are available and export them as KML file, a variant of the XML format used to represent geographical data. We then parse this file and download street view images at the specified locations, varying the camera heading. The pitch is kept fixed at 20°, 0 being the horizon. Finally, we store images metadata (id, GPS location, heading) on disk as well.

Google Earth image registration

An alternative is to manually annotate images the dataset images location and orientation. Google Earth offers the possibility to create image overlays by specifying the location, altitude, pitch and heading of the imported image. The resulting annotations can be exported as a KML file. This allows us to estimate the image geographical position of MHL images, and use this data as a baseline to measure the reliability of the alignment process.

Alignment

Our approach is to use two of the reconstructed cameras as baselines, for which we know their GPS location and heading, both obtained either by google street view or manual annotation. Given these two reference points, we need to find and affine transformation composed of translation, rotation and scaling, which maps the recovered cameras positions to their estimated latitude longitude coordinates. The following transformation is defined under the assumption that target zone on which replace the cameras is of a relatively small area.

Let r_1, $r_2 \in \mathbf{R^2}$, denote the two references cameras positions in the photogrammetric reconstruction space projected on the XY plane. We define their normalized vectors $n_1, n_2 \in \mathbf{R^2}$, and their corresponding annotated GPS location b_1, b_2. Let r be the position of another camera in the reconstruction space. Its estimated GPS location can expressed as a homogeneous vector b such that:

$$b = \begin{bmatrix} \cos\theta & -\sin\theta & b_{1_x} \\ \sin\theta & \cos\theta & b_{1_y} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_x - r_{1_x} \\ r_y - r_{1_y} \\ 1 \end{bmatrix} \tag{4.1}$$

where:

$$\theta = atan2(n_{1_y}, n_{1_x}) - atan2(n_{2_y}, n_{2_x}) \tag{4.2}$$

$$s = \frac{\|b_1 - b_2\|}{\|r_1 - r_2\|} \tag{4.3}$$

This is a simple change of coordinate system where r_1 and b_1 are considered as the systems origins.

Since we know the orientation of the reference cameras in both the real-word and the reconstruction space, we can trivially compute the true heading of the aligned cameras in the real-world.

Finally we generate a KML containing the position and orientation of all recovered cameras and import it in Google Earth for a 3D visualisation.

Chapter 5

Results

In a first phase, we have evaluated the performance of the image geo-referencing pipeline on the site of the Tribunal de Montbenon. Then we have applied it to a subset of images from two famous places in Lausanne: the Palais de Rumine located at Place de la Riponne, and Place Saint-François.

5.1 Tribunal de Montbenon evaluation

5.1.1 Setup

We have evaluated our pipeline on 39 images from the Tribunal de Montbenon from the MHL dataset before and after increasing their resolution with a super resolution neural network which reduces the images noise and doubles their dimensions. We used Google Earth to annotate the MHL images with their GPS position and heading orientation. Additionally, we have added 7 Google streetview images to the pipeline. Each Street view image has been captured from a different location and has been selected depending on its amount of content similarity with the MHL images.

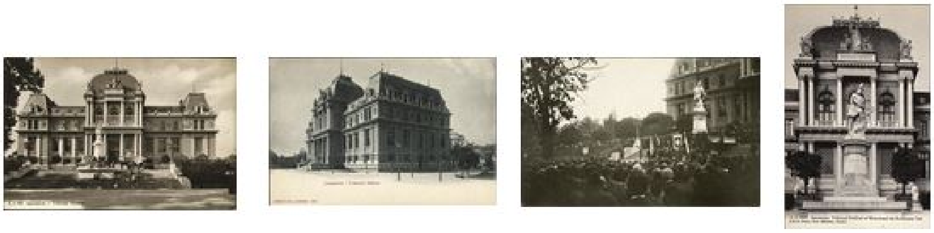


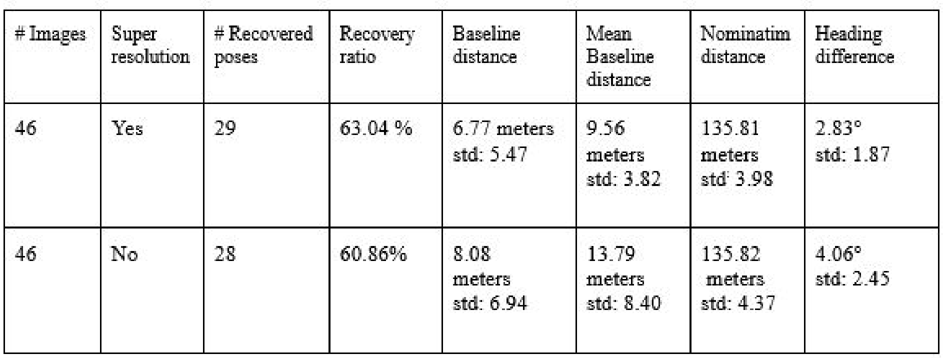
Figure 5.1: Sample MHL images from the Tribunal de Montbenon

5.1.2 Measures

During the pose recovery process only one Google street view image pose could be recovered. Hence, we evaluated the alignment process for all possible reference pairs r_1, r_2 where r_1 is the pose of the recovered Street view image, and r_2 is the pose of a MHL image. For each alignment, we have computed the following metrics:

- **Nominatim distance**: average distance between the aligned points and the location obtained by querying the Nominatim API with the images street address
- **Baseline distance**: average distance between the aligned points and their annotated position
- **Heading difference**: difference in degrees between the annotated heading and the computed one

5.1.3 Statistics



# Images	Super resolution	# Recovered poses	Recovery ratio	Baseline distance	Mean Baseline distance	Nominatim distance	Heading difference
46	Yes	29	63.04 %	6.77 meters std: 5.47	9.56 meters std: 3.82	135.81 meters std: 3.98	2.83° std: 1.87
46	No	28	60.86%	8.08 meters std: 6.94	13.79 meters std: 8.40	135.82 meters std: 4.37	4.06° std: 2.45

Figure 5.2: Precision measurements obtained for the alignment with the lowest baseline distance, as well as the mean baseline distance over all alignments

As we can see in Figure 5.2, using the super resolution increases the alignment precision. In particular, the lowest baseline distance reached by one of the alignments is 6.77 meters. There is a significant distance of 135.81 meters between the aligned points and the location obtained by querying Nominatim. We obtain a heading difference of 2.83 °which indicates that images have been well oriented. Averaging these results obtained on all possible reference pairs, we obtain an average mean baseline distance of 9.56 meters.

Without super resolution, the obtained distances are higher but still offer a decent overall precision.

5.1.4 Visualization

By visualizing the aligned points on a map (Figure 5.6) and in Google earth (Figure 5.8), we observed that our pipeline could estimate position the points located in front of the Tribunal de Montbenon. For the alignment of the upscaled images shown in Figure 5.8, it correctly estimated which images were taken perfectly in front of the stairs, or more on the right side. Images of the statue were replaced close to the stairs except one of them. Images taken from large angles similar to the second image in Figure 5.1, could not be recovered, due to their low number of feature matches with the other images. All images facing the building have been aligned. The alignment of the setup without super resolution was slightly less accurate when estimating the distance of some images with respect to the building. Furthermore it could not recover the tilted camera on the right image of Figure 5.8, near the stairs.

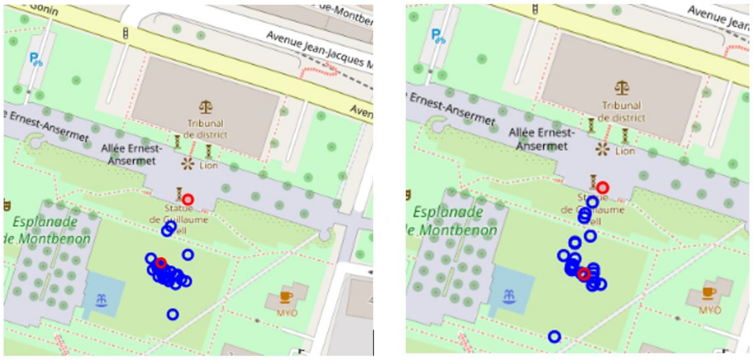


Figure 5.3: Distribution of the aligned images from Tribunal de Montbenon. In red are the reference points. From left to right: Aligned cameras with super resolution, Aligned cameras without super resolution

5.2 Application to other sites

We have applied a similar geo-referencing pipeline to two image clusters from Place Saint-Francois and Place de la Riponne. During our experiments, no Google Street

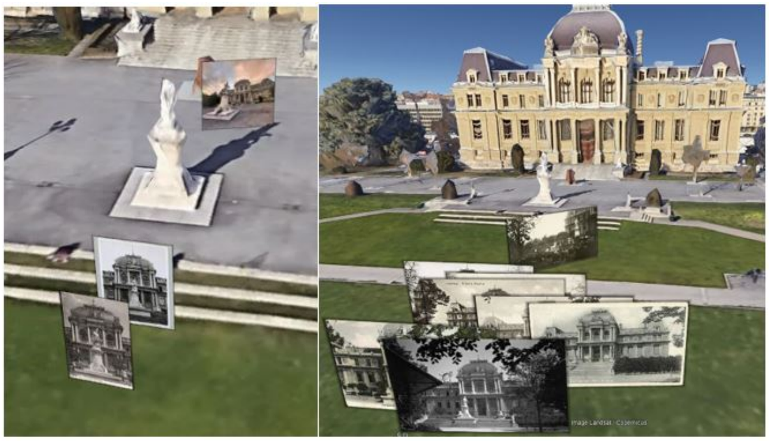
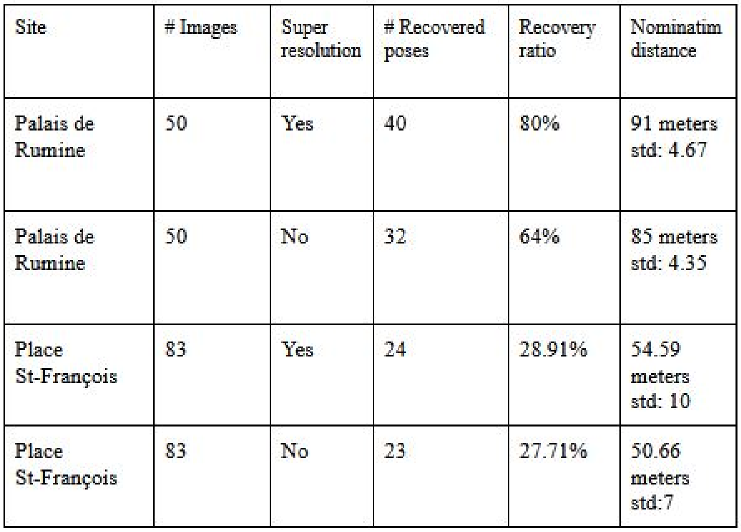


Figure 5.4: Alignment resulting from the super resolution setup imported in Google Earth

view image pose could be recovered via photogrammetry. Hence, for each cluster, we manually annotated two reference points with their GPS position and heading. These points were selected by visually inspecting the recovered poses and point cloud, and choosing the pair for which the relative poses were the most meaningful.

As shown in Figure 5.5 up-scaling the images increases the number of recovered poses for both sites.



Site	# Images	Super resolution	# Recovered poses	Recovery ratio	Nominatim distance
Palais de Rumine	50	Yes	40	80%	91 meters std: 4.67
Palais de Rumine	50	No	32	64%	85 meters std: 4.35
Place St-François	83	Yes	24	28.91%	54.59 meters std: 10
Place St-François	83	No	23	27.71%	50.66 meters std:7

Figure 5.5: Recovery ratio and Nominatim distance for Palais de Rumine and Place Saint-François. Google Streetview images are not considered

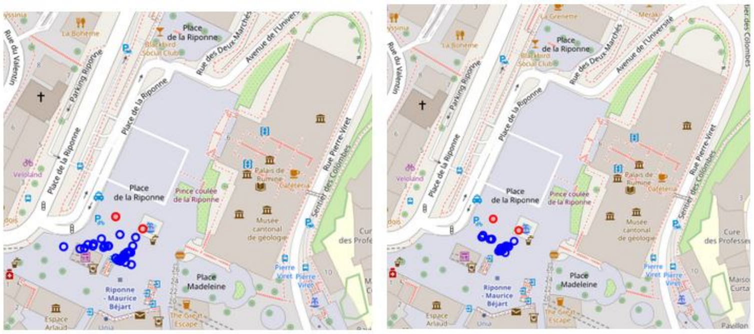


Figure 5.6: Distribution of the aligned images from Palais de Rumine. In red are the reference points. From left to right: Aligned cameras with super resolution, Aligned cameras without super resolution

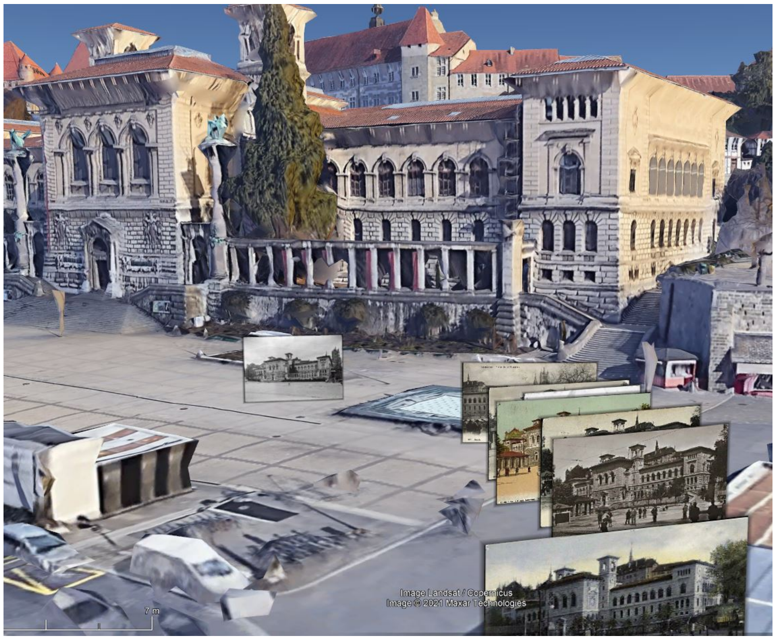


Figure 5.7: Place de la Riponne alignment resulting from the super resolution setup imported in Google Earth

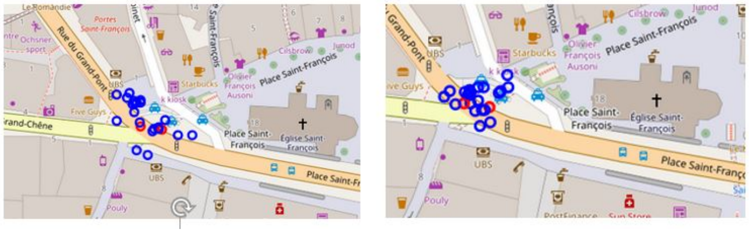


Figure 5.8: Distribution of the aligned images from Place Saint-François. In red are the reference points. From left to right: Aligned cameras with super resolution, Aligned cameras without super resolution

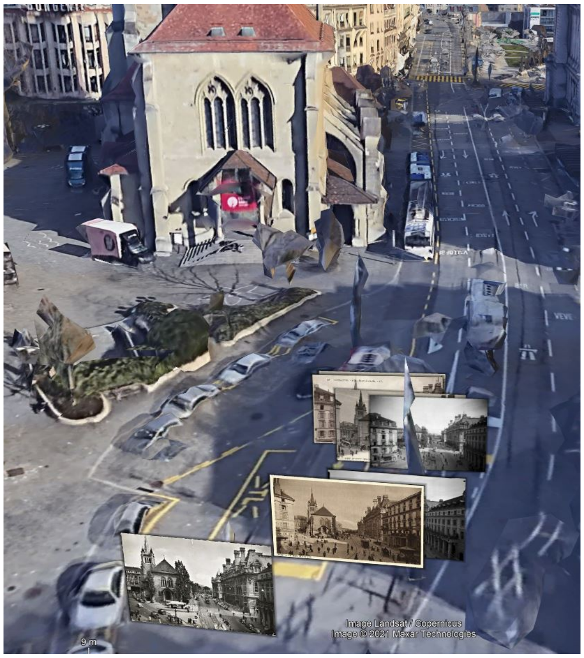


Figure 5.9: Aligned images from Place Saint-François

Chapter 6

Discussion

Our pipeline had a reasonable alignment precision on the site of Tribunal de Montbenon, which shows that it is possible to recover the viewpoint of historical buildings photographs. In particular, it is possible to accurately determine the relative position of the image with respect to the building. However, this works less well for images of Place Saint-François, as images from the studied cluster tend to have been taken from the very same axis. Their relative altitude however is accurately captured during the pose recovery, despite not being shown in the Google earth visualization. The super resolution neural network used to upscale the images, slightly sharpened the images edges but did not drastically improve the images quality. During experiments, the relative poses obtained were usually more meaningful when applying this pre-processing step and more cameras poses could be recovered. We suspect that the increase in images dimensions has an impact on the photogrammetric phase, as the number of tracks usually doubled as well.

Manual annotation of images was necessary to align images, as the pose of Google streetview images failed to be recovered, except one on the Tribunal de Montbenon.

Chapter 7

Conclusion

In this work, we developed a semi-automatic pipeline to infer the position and orientation in which historical photographs have been taken. We have shown that the pose of similar archive pictures can be computed using photogrammetry techniques, and aligned by annotating a few reference points. However the ratio of recovered poses could be increased by investigating pre-processing techniques or extracting features that would be more robust to the low resolution of the images.

References

[1] F. C. et al., *Isr*, https://github.com/idealo/image-super-resolution, 2018 (cit. on p. 8).

[2] AliceVision. (). Meshroom, [Online]. Available: https://alicevision.org/#meshroom. (accessed: 11.06.2021) (cit. on p. 7).

[3] S. Ares Oliveira, B. Seguin and F. Kaplan, 'Dhsegment: A generic deep-learning approach for document segmentation', in *Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on*, IEEE, 2018, pp. 7–12 (cit. on p. 4).

[4] F. Maiwald, T. Vietze, D. Schneider, F. Henze, S. Münster and F. Niebling, 'Photogrammetric analysis of historical image repositories for virtual reconstruction in the field of digital humanities', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2/W3, pp. 447–452, Feb. 2017. DOI: 10.5194/isprs-archives-XLII-2-W3-447-2017 (cit. on p. 2).

[5] N. Snavely, S. Seitz and R. Szeliski, 'Photo tourism: Exploring photo collections in 3d. acm trans graph 25(3):835-846', *ACM Trans. Graph.*, vol. 25, pp. 835–846, Jul. 2006. DOI: 10.1145/1141911.1141964 (cit. on p. 2).