ABSTRACT

We present the first step of the Rheims city (France) 3D reconstruction based on old postcards. Rheims endured drastic destruction during the First and then the Second World Wars. Old postcards, embodying the period from the beginning to the end of the 20th century, establish a testimony of the evolution in time and space of the most important places and monuments in the city. We aim to localize, render and visualize in 3D the buildings of this city, by using this sparse dataset and incomplete cadastral surveys. Image matching is initially required for locating a building in the city space. In this paper we propose a performance comparison of state-of-the-art detector-descriptor couples. We focus on their robustness and their transformations invariance when low resolution data is subject to temporal and space modifications.

Keywords
Sparse data, Spatio-temporal 3D reconstruction, Image matching, Old postcards, Rheims

1. INTRODUCTION

Terms like “urban planning”, “virtual and photo tourism” and “3D modeling” refer to common applications of 3D reconstruction. This mainly involves photographic data analysis and understanding in an overall reconstruction process. We witness various works that aim to the conception of low cost automatic and semi-automatic systems for modeling urban areas. Many of these works rely on meta-data [15] from Geographic Information System(GIS) combined with photographic data to get a good accuracy in 3D modeling of urban monuments. A data set of images is used for 3D scene reconstruction. A sequence of images taken by a moving camera is used in [13] while a panoramic street view sequence serves in [8] to robustly estimate the camera poses. Advanced works [14, 1] exploit a more complicated and large dataset of existing photos collected from the Internet. Those correspond to photos taken by tourists for the same urban sites.

Aforementioned image-based automatic systems [14, 1] reconstruct 3D urban sites as individual unconnected models of important historical monuments. They inquire a data set of a monument using a keyword-based search over image hosting web services such as flicker. The images are unstructured, uncalibrated and do not hold any Geo-referencing tags. They can be taken at different scales and points of view.

Structure from motion is used for the scene pose estimation. The first step of this computer vision technique is the extraction of the most distinctive feature points in the images. The understanding of the scenes illustrated in the images is processed by matching similar features across image pairs for further pose estimation and 3D reconstruction.

In our project, we consider the problem of recognition, extraction and 3D-reconstruction of geo-located buildings in Rheims city (France) under a “past-present” comparative approach. Rheims was founded 80 BC by Gauls and played a prominent role in French monarchical history as the traditional site where the kings of France were crowned. Because of this rich historical past, various documents testify to the evolution of this city and more particularly old postcards for the period from the beginning to the middle of the 20th century. We aim to propose a collaborative tool for modeling and visualizing the buildings of the city across time. This tool should give the citizens the opportunity to navigate in the city and track the temporal evolution of the sites.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Figure 1: Postcards of Saint-Jacques church: in the 19th century and after 1st World War destructions.

The reconstruction of Rheims from old postcards is a chal-
lenging project. The dataset is sparse and illustrates important spatial and temporal evolution of Rheims City. Postcards correspond to the most significant places and historical moments of Rheims. We have several postcards at our disposal which represent the same place taken at different times. As Rheims was heavily damaged during the First World War and partially during the Second one, we have pictures taken before and after destruction, and during and after reconstructions. Figure 1 illustrates two very different aspects of the Saint-Jacques church zone of the Erlon square. Postcards may in addition contain text, stamps and postmarks. They are also often of low resolution. In order to match these images together the problem of finding most distinctive and repeatable features in the image is hard and crucial in our case.

In this paper we present the performance comparison of state-of-the-art detector-descriptor couples that are the most robust and reliable in the context of this application. We first give a brief overview of point of interest detectors and local descriptors. We then proceed to the evaluation process that consists in two main steps. The first one validates the robustness of each couple when the ground truth is known, testing their invariance to affine transformations and occlusions. In the second step, the ground truth is unknown. In this case we evaluate the use of homography transformation and epipolar geometry to estimate it. We finally discuss the results and conclude.

2. RELATED WORK

A feature point represents a salient region that is invariant to potential geometric and photometric transformations that may occur during the image acquisition process. Nowadays, the evaluation of detector-descriptor couples is the aim of numerous works. They consist in couple performance comparison when images are subject to geometric similarity, affine and illumination transformations under controlled variations of a scene. These controlled variations between two images are convenient to establish a ground truth, allowing an easy control of the obtained results.

A number of different detectors and descriptors are combined. Several comparative studies of local region detectors have been presented in the literature using scenes that geometrically relate to homographies. Mikolajczyk and al. [10] proposed a general study in which they extract affine invariant regions using different detectors and then compared different description methods. They characterized and compared several interest point detectors focusing mainly on the descriptor choice and considering scale and affine invariant regions.

There are two kinds of features. Features of the first kind contain texture information and are far away from object boundaries, while the second kind translate geometric properties of an object and are detected near corners and edges of the object. Even if the first kind exhibits a greater stability according to geometric transformations of an object (viewpoint changes) [10] most of the features belong to the second family. Moreels and Perona [11] extended the evaluation of detector-descriptor couples to 3D scenes. They compared several couples to match 3D objects features across different viewpoints and lighting conditions, proposing a novel and automatic approach to estimate ground truth for 3D scenes. Other authors compared different detector-descriptor couples in the context of visual SLAM [5], historic repeat pho-

tography [3] or real-time visual tracking [4].

3. EVALUATION PROCESS

The aim of our project is to localize, render and visualize in 3D the buildings of Rheims city, by using a sparse dataset based on old postcards and incomplete cadastral surveys. Image matching is the required initial step. Once several detector-descriptor couples are selected for testing, our evaluation process methodology consists in matching two images under three different conditions. As the matching process may produce false positives, it is crucial to know the ground truth. For every pixel in a reference image the ground truth allows for accurate determination of its real correspondence in the query image.

The first step of our evaluation process consists in testing detector-descriptor couples performance when the ground truth is certain. In this case the two images that are compared are the same, one of them being the result of applying a known transform to the other image. This stage allows us to select the most stable and robust state-of-the-art detector-descriptor couples in the context of our application. In the second step we compare two different postcards of the same location. The buildings represented in these image can be taken at different periods and be subject to architectural changes. Viewpoint, scale and image quality can also vary in these image couples. In this case the ground truth has to be estimated. We propose to estimate the transformation existing between the two images by a homography and then by epipolar geometry. The goal of this second step is to evaluate the invariance of the previously selected detector-descriptor couples when data is spatially and temporally sparse.

3.1 Interest point detectors

In order to extract visual landmarks in old postcards we evaluate six state-of-the-art interest points detectors. These detectors are scale or affine invariant.

Harris Laplace [9, 10] detector provides corner-like points that are invariant to scale and rotation changes. They are detected using a scale-adapted Harris function, then selected in scale-space by the Laplacian-of-Gaussian operator.

Hessian Laplace [9, 10] detector provides blob-like structures that are invariant to scale and rotation changes. Points are localized in space at the local maxima of the Hessian determinant and in scale at the local maxima of the Laplacian-of-Gaussian operator.

Harris Affine (resp. Hessian Affine) [9, 10] detector is invariant to affine transformations. The interest points are computed using the Harris Laplace detector (resp. Hessian Laplace detector) then an affine neighborhood is determined by the affine adaptation process based on the second moment matrix.

SIFT (Scale Invariant Feature Transform) [7, 16] detector is scale and rotation invariant and partially invariant to illumination and viewpoint changes. It detects distinctive points using a difference of Gaussian function (DoG) applied in scale space. Several points are selected as local extrema of the DoG function, rejecting low contrast points and points localized on low curvature contours.

ASIFT (Affine invariant extension of Scale Invariant Feature Transform) [17, 12] detector is fully invariant to affine transformations. This algorithm is based on the SIFT one. It consists in simulating all possible affine distortions caused by the change of camera optical axis ori-
entation from a frontal position. These rotations and tilts are performed for a finite and small number of latitude and longitude angles. A set of interest points is then computed on all the simulated images using the SIFT detector.

3.2 Local descriptors

We evaluate three different local descriptors in order to characterize previously detected visual landmarks.

**Steerable filters** [10]. Designing steerable filters consists in computing up to 4th order derivatives of a Gaussian function. Correlations (convolutions) between rotated versions of the filters with the image lead to a 14-dimensional descriptor.

**SIFT** [7, 16]. This descriptor is similarity invariant and assigns a dominant orientation to each feature point based on local image gradient directions. The descriptor is deduced from orientation histograms computed in sub-regions around the point. The resulting descriptor of dimension 128 is then normalized to ensure illumination invariance.

**PCA-SIFT** [6, 10]. This descriptor is based on a SIFT-like descriptor on which a PCA (Principal Component Analysis) is applied. To compute the 36 dimensional vector corresponding to this descriptor, \( x \) and \( y \) gradient images are computed in a support region, sampled at 39 \( \times \) 39 locations and then reduced by PCA.

In order to match two images we first determine the detector-descriptor couples we want to evaluate. We associate Hessian and Harris Laplacian and Affine detectors with steerable filters and PCA-SIFT descriptors. For the SIFT descriptor the four previous and the SIFT detectors are tested. As the ASIFT is a matching method derived from the SIFT, we will test it separately in order to compare it to the SIFT.

3.3 Matching process and ground truth

We conduct an evaluation process in order to analyze the quality of the feature matches. For every detector-descriptor combination, we compute the feature points for every test image. For an image pair, features are matched according to a metric distance that depends of the descriptor type: Euclidean for SIFT and PCA-SIFT, and Mahalanobis for steerable filters. A retained match corresponds to the nearest neighbor distant feature that satisfies a nearest neighbor distant ratio constraint as in [7]. A percentage of correct matches is computed according to a ground truth.

The ground truth knowledge enables the definition of the pixel localization corresponding to a pixel in a reference image depending on the transformation between the images. When a match is established, its is retained if it corresponds to the real match derived from the ground truth measurement. Otherwise the match is considered as a false match.

The percentage of correct matches is then computed. We describe in the following two types of ground truth measurement.

3.3.1 Certain ground truth

We refer to a known transformation that we impose to a reference image to get a query image. In this case the generated ground truth measurement is certain since we compare the image to its transformation copy. We test a scale change going from 0.4 to 2.0 increasing by 0.2 each time. A rotation through 11 angles going from -50° to 50° is also studied for the images of the experimental data set.

3.3.2 Ground truth estimation

To evaluate the detector-descriptor couples on two postcards of the same location at different time periods, we first generate the ground truth existing between the two images.

**Homography**

Since homography is considered for planar surfaces, we suppose that a building represents a planar scene viewed by the camera from two different points of view. Four couples of correspondences are used to compute the homography matrix. They correspond to the extreme points of the main facade of the building (figure 2).

**Epipolar geometry**

Homography relates only to any two images of the same planar surface in space, this transformation is thus not sufficient to represent the deformations of a 3D scene. On the opposite, epipolar geometry models the 3D deformation occurring when two cameras take a photo of the same 3D scene. It allows a point-to-line correspondence associating a point in a reference image and an epipolar line to which the real corresponding point belongs in the query image (figure 3). This can be obtained by computing the fundamental matrix corresponding to the epipolar geometry and requires a minimum of eight couples of correspondences between the two images.

In order to decide if a match is correct, we check if it belongs to the epipolar line. Figure 3 illustrates the way we proceed. The epipolar point is determined using the fundamental matrix. This epipolar point corresponds to the crossing of all the colored lines in the two images, and represents the same location in both images. Several red points are pointed in the reference image (top), their matches have to belong to the corresponding lines in the query image (bottom). In order to better see the lines to which the matches have to belong to, we draw the lines in both images by using the same color, even if the lines are not to be considered in the reference image.
4. EXPERIMENTAL RESULTS

4.1 Certain ground truth

In this subsection we present the results of the matching process of an image to its transform. Rotation, scale sampling and occlusion cases are presented. First the results are computed on one entire image then this image is duplicated, and the two images to be matched are partially occluded (left third part for the first image, right third part for the second one). The results are computed on the images composing our database. Several detector-descriptor couples are tested as described in sections 3.2 and 3.3.1.

Figures 4 illustrates the performances for every detector-descriptor combination, with respect to the transformation parameters. The graphs show the mean performance of Steerable filter, PCA-SIFT and SIFT descriptors each one being computed for the Hessian Laplacian, Hessian Affine, Harris Laplacian and Harris Affine detectors. We compute an additional mean performance estimation for the SIFT descriptor with the SIFT detector. The top row represents the performance behavior with respect to the 11 angles of rotation transformation. In the bottom row, the same performances are evaluated for the 9 scale factors. Each column is associated to a descriptor. We attest that the detectors are less sensitive to rotation than scale changes, independently of the descriptor. Fortunately, the SIFT descriptor performs invariantly to scale when combined to the SIFT detector. The SIFT descriptor over performs the other descriptors for all associated detectors. Composed with the SIFT detector, it gives the best performance.

The same tests are conducted for the image set with a partial occlusion constraint. The performance of the SIFT descriptor remains the same regardless of the occlusion. An instability is witnessed for the steerable filter and the PCA-SIFT descriptors in the scale change tests. The performance of those descriptors decreases dramatically at low scales. We notice a decrease of 10% of the performance of the overall detector-descriptor combination in the rotation test case. As the SIFT descriptor over performs the other descriptors, we chose to compare the SIFT descriptor performance when using the different detectors. Figure 5 present the results for rotation and scale tests. We represent the mean performances in the top row and the evolution of the number of correct matches in the bottom row. We observe good and stable results for the SIFT descriptor even if the images are partially occluded. A lower precision is noticed for the ASIFT detection method compared to the other detectors.
while the advantageous results remain for the SIFT detector. Besides, we observe that the ASIFT detector produces remarkably a large number of matches compared to other methods.

In addition to a homography ground truth transformation, we produced interactively an epipolar transformation as a ground truth reference. Eight pairs of corresponding points were chosen for the computation of the epipolar geometry. The results are presented in Table 2 and lead to similar conclusions as in the previous case. We observe slightly better percentages since an epipolar geometry better describes the transformation between the pairs of images used in this section.

Table 3: SIFT versus ASIFT performances: estimate ground truth by epipolar geometry (top two lines) and epipolar geometry (bottom two lines).

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Maximum (%)</th>
<th>Mean (%)</th>
<th>σ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT (Lowe)</td>
<td>99.1</td>
<td>30.7</td>
<td>32.5</td>
</tr>
<tr>
<td>ASIFT</td>
<td>98.5</td>
<td>14.3</td>
<td>25.6</td>
</tr>
<tr>
<td>SIFT (Lowe)</td>
<td>100.0</td>
<td>25.6</td>
<td>31.8</td>
</tr>
<tr>
<td>ASIFT</td>
<td>97.0</td>
<td>9.0</td>
<td>18</td>
</tr>
</tbody>
</table>

As the SIFT descriptor-detector is the most performant we compared it to the ASIFT, which provides a large number of correspondences as seen in the previous stage. As an example, Figure 6 presents a couple of images where ground truth is estimated by epipolar geometry. Table 3 compares the mean precision of both SIFT detector-descriptor couple and ASIFT. Once more the two ground truth estimates produce similar results, with an advantage to homography.

5. CONCLUSIONS

We proposed an evaluation methodology to analyze the image matching performance. We evaluated different combination of state-of-the-art detectors and descriptors that were presented as the most robust in the literature. The application data set represents sparse and low quality post cards of the Rheims city. The most important monuments of the city are illustrated in the post cards. Those can represent the same building at different epochs or taken from different perspectives. We studied the steerable filters, PCA-SIFT and SIFT descriptors, associated to the Harris (resp. Hessian) Laplacian and affine, as well as the SIFT and ASIFT

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Maximum (%)</th>
<th>Mean (%)</th>
<th>σ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steerable Filters</td>
<td>39.2</td>
<td>9.6</td>
<td>6.12</td>
</tr>
<tr>
<td>PCA-SIFT</td>
<td>94.7</td>
<td>14.6</td>
<td>19.6</td>
</tr>
<tr>
<td>SIFT</td>
<td>100.0</td>
<td>36.4</td>
<td>42.1</td>
</tr>
</tbody>
</table>

Figure 5: Performance of SIFT descriptor with all the detectors including ASIFT detector, according to rotation (left) and scale sampling (right) in terms of precision (top) and number of matches (bottom).

Figure 6: Valid matches obtained with the epipolar ground truth estimate, using the SIFT detector-descriptor (left) and the ASIFT algorithm (right).

Figure 7: Performance of SIFT descriptor with all the detectors including ASIFT detector, according to rotation (left) and scale sampling (right) in terms of precision (top) and number of matches (bottom).

Table 1: Matching precision of the descriptors: ground truth estimate by homography.

Table 2: Matching precision of the descriptors: ground truth estimated by epipolar geometry.
detectors. We presented the results in an hierarchical manner in order to reduce the amount of variables propagated along the analysis process.

On the first hand, we presented the simple case where we relied on a certain ground truth laying out affine transformation applied to an image. We could attest that the SIFT descriptor over performed other descriptors. Furthermore, this descriptor was the most robust when it was computed with the SIFT detector. It certified its invariance to scale changes, rotation as well as occlusion. On the other hand, we estimated a homography transformation as well as an epipolar geometry between pairs of dissimilar images. We used this estimation as ground truth validation criteria to evaluate the percentage of correct matches. The SIFT descriptor pointed out as the most efficient. While less precise, the ASIFT produced a larger number of matches.

Further work will lead to infer a temporal ordering of a series of images of the same urban site. The geometry of the corresponding models will incorporate the imperfection of spatiotemporal information [2]. The models represent the 3D elements to be integrated in a collaborative GIS for the city modeling.

6. ACKNOWLEDGMENTS

The authors are grateful to Olivier Rigaud for providing old Rheims postcards, Krystian Mikolajczyk, David Lowe, Guoshen Yu and Jean-Michel Morel for providing part of all of their detectors and descriptors implementations.

7. REFERENCES