Master's degree in Computer Engineering for Robotics and Smart Industry - a.y. 2022/2023
Computer Vision

# Automatic 2D mosaicing

Lorenzo Busellato - VR472249 - lorenzo.busellato_02@studenti.univr.it

CONTENTS

Master's degree in Computer Engineering for Robotics and Smart Industry - a.y. 2022/2023
Computer Vision

# Automatic 2D mosaicing

## I. INTRODUCTION

Image mosaicing is a method of combining multiple images of the same scene into a single, larger image. The mosaicing process can be broadly divided in five steps:

- Feature point extraction
- Feature point matching
- Robust homography computation
- Image warping
- Image blending

Feature points are be found using Scale Invariant Feature Transform (SIFT). Their matching is obtained by comparing the descriptors resulting from SIFT. The homography computation is made statistically robust to outliers using RANdom SAmple Consensus (RANSAC). The images are correctly aligned using image warping, i.e. by using the homography between them. The quality of the resulting mosaic is improved using image blending, which makes the colors more uniform near the seams between the images.

## II. OBJECTIVE

This project's objective is to develop a software application for the automatic generation of a 2D mosaic from a set of images of a planar scene.

## III. METHODOLOGY

### A. SIFT

Scale Invariant Feature Transformation (SIFT) is a method used to identify a set of feature points, or features, within an image.

Features are regions within an image that carry some information about the image's content. Features can often be associated to structures within the image, such as corners or edges.

To be able to match detected features in a set of images, the features should be invariant to:

- Scale
- Illumination
- Rotation

Given an input image I, the algorithm first finds candidate features as follows:

1) The image is downsampled four times, yielding a set of scaled images.
2) Each scaled image is convolved five times with a Gaussian kernel, yielding five sets of blurred images called **octaves**.
3) For each octave, four **difference-of-gaussians** (DoG) images are computed by taking the difference of adjacent images, yielding sixteen DoG images.
4) For each pixel in each DoG image, features are defined as the local maxima and minima in the 3x3x3 region surrounding the pixel in the previous, current and next DoG image in the octave.

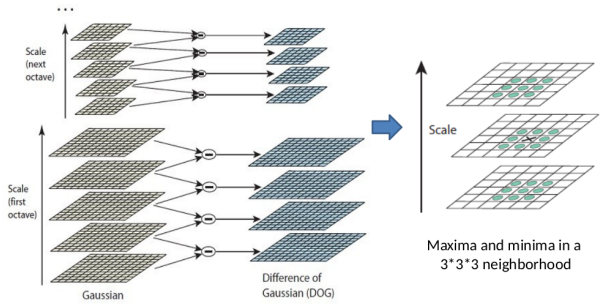


Fig. 1. From image pyramids to candidate feature points (image source [1])

The scaling and blurring introduce invariance to scale and illumination conditions. To introduce invariance to rotation, the orientation associated to the feature is estimated by computing the histogram of gradients for the 16x16 pixel region on the Gaussian pyramid scale the feature was found at. The orientation associated to the feature is the histogram bin corresponding to the highest peak in the histogram.

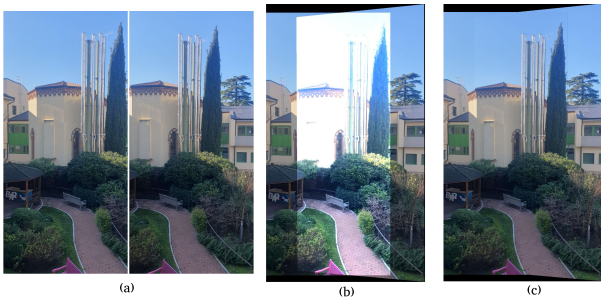


Fig. 2. SIFT descriptor (image source [1])

The feature descriptor finally is computed by considering the 16x16 region around the feature, which is then divided into 16 4x4 sub-regions for each of which a 8-bin histogram is computed. The feature descriptor is then the concatenation of the 16 resulting histograms, i.e. a 128 by 1 vector.

Feature matching is done with an heuristic based on distance. Given two images and the corresponding features

extracted with SIFT, for each feature of the first image the distance to each feature of the second is computed. The candidate match is the feature pair that results in the smallest distance. To reduce the influence of outliers, the candidate match is accepted only if the distance is smaller than an arbitrary threshold.

### B. Homography

Any two images of the same planar scene can be related through a linear relationship called homography.

Let $M$ be a 3D point in the reference frame centred on the left camera and let $M'$ be the same point in the reference frame centred on the right camera. The perspective matrices that describe the cameras are then:

$$P = K[I \mid 0] = [K \mid 0] \qquad P' = K'[R \mid T] = K'G$$

The two points are linked by the rototranslation matrix $G$:

$$M' = GM = RM + T \tag{1}$$

Let $n^T \tilde{M} = d$ be the equation of the plane containing the scene, where $n$ is the plane normal and $d$ is some scalar representing the distance of the plane from the origin. The projections of $M$ and $M'$ on the image planes are:

$$m \simeq KM \qquad m' \simeq K'M'$$

Since $M$, by construction, belongs to the plane we have:

$$n^T M = d \implies \frac{n^T M}{d} = 1$$

Plugging the fraction into the equation 1 (treating $T$ as $1 \cdot T$):

$$M' = RM + \frac{n^T M}{d} T$$

Therefore:

$$M' = K'^{-1} m' = \left( R + \frac{n^T}{d} T \right) M = \left( R + \frac{n^T}{d} T \right) K^{-1} m$$

Finally:

$$m' = K' \left( R + \frac{n^T}{d} T \right) K^{-1} m = H_\pi m$$

$H_\pi$ is the homography, i.e. the linear relation between the pixels in the two images.

To compute the homography, we start from a set of $n$ conjugate points $(m_i, m'_i)$, and we want to estimate the matrix $H$ such that:

$$m'_i = H m_i \quad i = 1, \ldots, n$$

Taking the cross-product of both sides with $Hm_i$ yields:

$$m'_i \times H m_i = 0 \implies [m'_i]_x H m_i = 0$$

where $[m'_i]_x$ denotes the skew symmetric matrix of $m'_i$. To get a linear system, we need multiple instances of this equations:

$$vec([m'_i]_x H m_i) = (m_i^T \otimes [m'_i]_x) vec(H) = 0$$

where $\otimes$ denotes the Kronecker product and $vec()$ the vectorization transformation.

The system can be solved for $H$ using singular value decomposition (SVD), for which at least four conjugate point pairs are needed, since each pair gives two linearly independent equations and $H$ has nine unknowns.

### C. RANSAC

RANdom SAmple Consensus (RANSAC) is an iterative algorithm which aims to improve the estimation of a model given a set of observations.

The algorithm randomly samples the observations and creates a model estimate on this subset. The amount of data points that are closely explained by the computed model is called its consensus. This procedure is repeated a number of times, and the resulting best-estimate for the model is the one with the highest consensus (i.e. with the least amount of outliers).

The algorithm is defined as follows:

- Input: a set of observations $(y_i, x_i)$, $i = 1, \ldots, n$, a threshold $\varepsilon$ and a number of iterations $k$.
- Algorithm:
  1) Repeat $k$ times:
     a) Take a random sample of $p$ elements from the observation set.
     b) Use the subset to estimate a probe model $\hat{\theta}_j$ (e.g. by regression).
     c) Compute the consensus set of the probe model:
     $$C = \{y \mid (y_i - f(x_i, \hat{\theta}_j))^2 < \varepsilon\}$$
     where $f$ is the function that relates $x$ to $y$ given the model $\hat{\theta}_j$.
  2) Among all the consensus sets, pick the one with the most elements.
  3) The probe model corresponding to the most numerous set is the best estimation. Its consensus set is the set of inliers of the data set. The remaining observations are considered outliers.

RANSAC will be used to refine the set of detected features in each image. This means that features of a given image that are unlikely to correspond to features in a subsequent image will be treated as outliers and therefore removed.

The algorithm will also be used to improve the computation of the homography, resulting in the transformation that links as many feature matches as possible.

### D. Image warping

To obtain the mosaic, the images are treated sequentially. Given two images, the first is treated as the reference image, while the second image undergoes a procedure of warping. Warping means that the homography that relates the two images is used to project the second image onto the plane of the reference.

### E. Image blending

Once the second image has been warped, a simple superposition of it with the reference image yields the mosaic. The main issue is that the color blending is not uniform in the overlapping region and especially near the seams between the images. A first approach is to create a binary mask for the overlapping region, and use the averaged values of pixel intensities of the two images in the overlap, thus correcting the intensities in the stitch. This simple approach does not fix however the noticeable seams there are when the images are too misaligned.

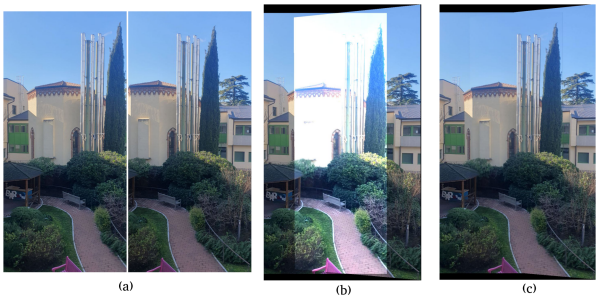


Fig. 3. Reference images (a) and the resulting mosaic without blending (b) and with blending (c).

## IV. IMPLEMENTATION

## V. TESTS AND RESULTS

The following tests are performed on two sets of images taken with a smartphone camera.

The following tests were performed:

- Test 1: 10 pictures of set 1, 1000 RANSAC iterations with tolerance of 1 pixel.
- Test 2: 10 pictures of set 1, 1000 RANSAC iterations with tolerance of 5 pixels.
- Test 3: 10 pictures of set 1, 1000 RANSAC iterations with tolerance of 15 pixels.
- Test 4: 10 pictures of set 1, without RANSAC.
- Test 5: 10 pictures of set 2, 1000 RANSAC iterations with tolerance of 5 pixels.
- Test 6: 4 pictures of set 2, 1000 RANSAC iterations with tolerance of 5 pixels.

Tests 1 through 3 highlight the influence of the pixel tolerance used in RANSAC. Test 4 was done to show the importance of using RANSAC. Tests 5 and 6 show the functionality of the pipeline on other image sets as well as the increased noise resulting from the usage of multiple images.

### A. Test 1

### B. Test 2

### C. Test 3

### D. Test 4

### E. Test 5

### F. Test 6

## VI. CONCLUSIONS

The pipeline is capable of producing convincing mosaics from different sets of input images, as shown by tests 2 and 5.

The dimensionality of the image sets is important because increasing the number of input images results in an increase in noise in the mosaic due to the subsequent warping.

The usage of the RANSAC algorithm has been justified by test 4. Having a statistically robust method for the estimation of the homography is important to ensure that the produced mosaic has no evident misalignments. Tests 1 through 3 showed the importance in picking the right tolerance for the algorithm, because the tolerance directly influences the quality of the resulting homography.

## REFERENCES

[1] U. Castellani. Lecture slides of the computer vision course, master's degree in computer engineering for robotics and smart industry, 2022.