UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Academic Year 2020/2021

Master Degree in ICT for Internet and Multimedia

02 · 2021

3D AUGMENTED REALITY

PROJECT REPORT - B.2

Casagrande Nicola - De Toni Alberto
1242363 - 1236657

Local feature compression using autoencoders - SfM tests
Design a compression strategy for local SURF descriptors using autoencoders. Training data can be generated using the images of dataset Portello and Castle. Testing must be done on dataset FountainP-11 and Tiso (available at https://github.com/openMVG/SfM_quality_evaluation/tree/master/Benchmarking_Camera_Calibration_2008 and http://www.dei.unipd.it/~sim1mil/materiale/3Drecon/). Software must be implemented in MATLAB, Keras or Pytorch.
Testing on 3D reconstruction using SfM: The reconstructed descriptors (only for the test set) are used to perform a SfM reconstruction using COLMAP (using the two test dataset).
Programming languages: MATLAB/Python/C++.

1 Introduction

1.1 What are the descriptors

In computer vision, visual descriptors or image descriptors are descriptions of the visual features of the contents in images, videos, or algorithms or applications that produce such descriptions. They describe elementary characteristics such as the shape, the color, the texture or the motion, among others[1]. A feature detector (*extractor*) is an algorithm taking an image as input and outputting a set of regions ("local features" = "Interest Points" = "Keypoints" = "Feature Points"). A descriptor is computed on an image region defined by a detector. The descriptor is a representation of the image function (colour, shape, ...) in the region (typically an array). Two key operations are related to feature extraction:

- Feature detection: extract the features of interest
- Feature description: associate a descriptor to each feature in order to distinguish from the others

There are many pre-defined descriptors available to the users. Some of them are reported in fig.1.

1.2 Speeded Up Robust Features (SURF)

In computer vision, Speeded Up Robust Features (SURF) is a local feature detector and descriptor. It can be used for tasks such as object recognition, image registration, classification, or 3D reconstruction. It is partly inspired by the Scale-Invariant Feature Transform (SIFT) descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT[2]. The SURF algorithm

[1] https://en.wikipedia.org/wiki/Visual_descriptor, 01/02/21
[2] https://https://https://en.wikipedia.org/wiki/Speeded_up_robust_features, 01/02/21

Desc.	Binary	Invariant	Size	Complexity	Robustness	Patents
SIFT	no	Scale, rotation, mild comp.	128 float= 512 B (128B in some impl)	High	High	Yes
SURF	no	Scale, rotation, mild comp.	64 float = 256 B or 128 float = 512 B (conv. to bytes sometimes)	Medium/High	High	No
BRIEF	yes	Scale, rotation	256 or 128 bits	low	Low	no
BRISK	yes	Scale, rotation	512 bits	low	low	no
FREAK	yes	Scale, rotation	512 bits	very low	low	no

Figure 1: Common algorithms for feature description.

is based on three main parts:

- Detection: using a blob detector method based on the Hessian matrix to find feature points
- Description: the goal of a descriptor is to provide a unique and robust description of an image feature
- Matching: comparing the descriptors obtained from different images to find the matching pairs

1.3 What is an autoencoder

An autoencoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal “noise”. Along with the reduction side, a reconstructing side is learnt, where the autoencoder tries to generate from the reduced encoding a representation as close as possible to its original input[3]. There are two categories of autoencoder:

- Regularized autoencoders: they are divided into three types such as Sparse autoencoder (SAE), Denoising autoencoder (DAE) and Contractive autoencoder (CAE)
- Variational autoencoder (VAE): generative model akin to generative adversarial networks (GAN)

[3]https://https://en.wikipedia.org/wiki/Autoencoder, 01/02/21

A basic autoencoder as shown in fig.2 has an internal hidden layer and it's made of two main parts which are the encoder that turns a high-dimensional input into a latent low-dimensional code, and the decoder that performs a reconstruction of the input with this latent code. It's also possible to train the autoencoder with multiple layers to exponentially reduce the computational cost of representing some functions and to exponentially decrease the amount of training data needed to learn some functions. The two steps of using an autoencoder are:

- Training
- Testing

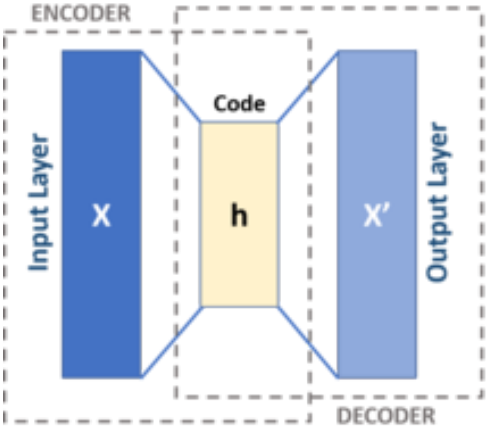

Figure 2: Schema of a basic autoencoder.

1.4 Development software

MATLAB is a proprietary multi-paradigm programming language and numeric computing environment developed by MathWorks. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages[4]. We decided to use Matlab as Machine Learning tool because we felt more confident, it has an extensive documentation and it offers different advantages:

- Implement and test algorithms easily
- Develop the computational codes easily
- Debug easily
- Use a large database of built in algorithms

[4]https://https://https://https://en.wikipedia.org/wiki/MATLAB, 02/02/21

- Process still images and create simulation videos easily
- Symbolic computation can be easily done
- Call external libraries
- Perform extensive data analysis and visualization
- Develop application with graphics user interface

1.5 Goal of the project

The objective of the project is to develop a compression strategy for local SURF descriptors using an autoencoder and perform a SfM reconstruction using COLMAP software with the obtained data. The matchings and the features must be saved in a .txt file. The autoencoder has to be trained with training data generated from a given dataset and has to be tested on another dataset. The final result should show a worse reconstruction due the compression of the data and the goal is to find a good setting for the autoencoder.

2 Data

2.1 Dataset

The data of the experiment are divided into two groups of images, one used for the training of the autoencoder and another one for the testing. The images represent four buildings from different points of view. In the training group there are images of Portello and Castle (some of them are reported in fig. 3 and fig. 4), while in the group of the testing images there are images of fountain and Tiso, some of them reported in fig. 5 and fig. 6.

Figure 3: samples from Portello dataset.

Figure 4: samples from castle dataset.

Figure 5: samples from Fountain-P11 dataset.

Figure 6: samples from Tiso dataset.

3 Development

3.1 Training

The images from the training set are loaded into a single image datastore. After that, a loop for each image is designed in order to:

1. Convert the image to grayscale;
2. Detect the SURF features (using the MATLAB function “detectSURFFeatures”);
3. Extract the features (descriptors) using the function “extractFeatures”. In this case they are represented as a vector of 64 columns and N rows;
4. Stack vertically the descriptors of each image in a single matrix.

The result is a big matrix containing all the descriptors for each image in column. The autoencoder has been built using the function “layerGraph”, which requires in input a vector “layers” containing the structure of the Neural Network. For this purpose, we chose the following three-layer structure:

- an input layer that receives features of dimension 64 (using “featureInputLayer(64)”);
- two fully connected layers of dimension 6 and 64 (using “fullyConnectedLayer(6)” and “fullyConnectedLayer(64)”), the second one having dimension 64 as will work as output layer;
- a regression layer that will predict the reconstructed values (using “regressionLayer”).

After some testing, we found out that for the inside layer a dimension of 4 was too strict (The reconstructed values were too compressed, i.e. too far from the input ones) while for 8 they were too close to the input ones, so we opted for the compromise of 6. Regarding the training options, we decided to use an *adam* (Adaptive Moment Estimation) optimizer and a maximum number of epochs of 2, since even after one epoch the training loss was pretty stable around 5% (fig. 7).

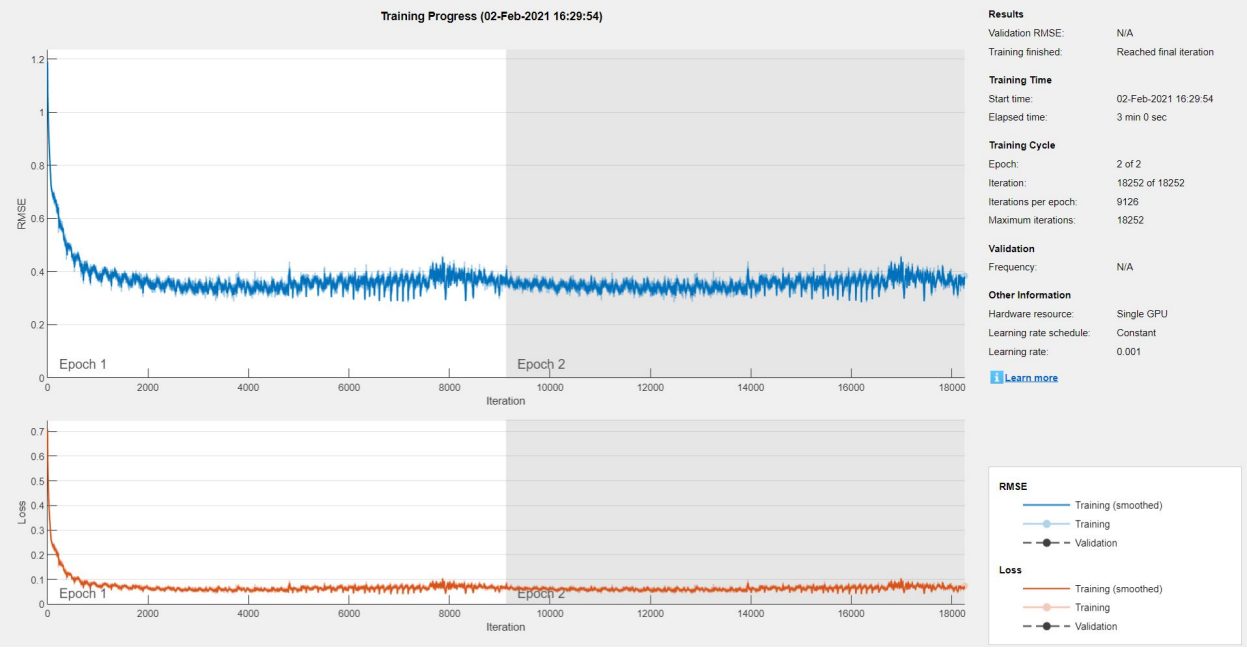

Figure 7: Training results after 2 epochs (18252 iterations).

The autoencoder is then succesfully trained (using the function “trainNetwork”) in order to be ready to be used for the testing phase. A useful trick in order to fasten up the operation of testing and not doing everytime the training (which could require some time if the number of epochs is increased) is to save the

workspace at the end of the training section and load it at the beginning of the testing one, so that everytime that section is called the autoencoder will remain trained just like it was the first time.

3.2 Testing

For the phase of testing the image datasets must be loaded separately (one for each building). To be clearer, this section simply called two functions, one for the features extraction, and one for the matchings calculations. This was done because COLMAP requires specific type of files in order to perform a reconstruction (will be better explained in sec. 4).

3.2.1 function FEATURES

This function aims to reconstruct the features of each image with the autoencoder and the original ones in a .txt file for each image in the form "location_x location_y scale orientation" followed by 128 zeros since COLMAP works with SIFT features, which are of that size. The workflow of the function is the following:

- The path containing the features is cleaned from older files containing the features;
- For each image in the image datastore considered, which is given in input along with the autoencoder:
  1. The image is converted to grayscale;
  2. The features are detected and then extracted like for the phase of training, only this time they won't be stacked in a single matrix but the features **extracted** are going to be fed into the autoencoder and be returned from the function while the **detected** ones are going to be divided in Location, Scale, Orientation in order to be printed in a .txt file of the form

     N 128
     $location_{x0}$ $location_{y0}$ $scale_0$ $orientation_0$ 0 ...(x128) 0
     $location_{x1}$ $location_{y1}$ $scale_1$ $orientation_1$ 0 ...(x128) 0
     ...
     $location_{xN}$ $location_{yN}$ $scale_N$ $orientation_N$ 0 ...(x128) 0

     where N is the number of features.

3.2.2 function MATCHINGS

The function, given in input the image datastore considered and the features (compressed or not) matches the features of each couple of image i, j. To avoid repetitions (i, j and j, i have the same matchings) the MATLAB function "matchFeatures" is called for every i, j for i going from 1 to the last feature -1 and j going from $i + 1$ to the last feature. The matchings are all stacked in a single .txt file of the form:

$$
\begin{aligned}
&image0.jpg\ image1.jpg \\
&matching1_0\ matching1_1 \\
&matching2_0\ matching2_1 \\
&... \\
&matchingN_0\ matchingN_1 \\
& \\
&image0.jpg\ image2.jpg \\
&matching1_0\ matching1_2 \\
&...
\end{aligned}
$$

4 COLMAP

Having all the required files the performances of the compression system can be tested on a 3D-Reconstruction software, like COLMAP (https://colmap.github.io/). This software was meant for using SIFT features (see fig. 1), which are heavier and more performing in terms of quality but also less robust to image transformations (which must be accounted since the images are taken with a mobile phone) and patented. A trick that can be used is to use SURF features (64 float) and faking a larger dimensionality (128 float) by adding zeros. This way COLMAP recognizes the right dimensionality of the files while using SURF features instead of SIFT. In order to start the reconstruction, COLMAP will need three things:

1. A set of images of the same scenario in different perspectives;
2. A .txt file for each image containing the features, in the format specified in sec. 3.2.1. The title of the file must be the name of the image followed by .txt (*e.g: image001.jpg.txt*);
3. A .txt file containing the matchings between each image in the format specified in sec. 3.2.2.

Not having these will result in an error inside COLMAP which won't start the reconstruction.

5 Conclusions

The plots and the momenta analysis display a strong similarity between the data collected and the theoretical models. The Arecchi experiment exhibits the difference between a coherent and a thermal source: the statistical description of a coherent source follows a Poisson distribution of mean λ, while a thermal source follows a Bose-Einstein distribution of parameter λ. This can be verified empirically because by reducing the time bin the less photons are collected and the Bose-Einstein distribution follows this kind of behaviour (strong peak around 0), while the Poisson distribution tends to keep a bell-shape around the mean (like a coherent light source) and by changing T all it does is shifting and spreading the “bell”. Note that the differences between the empirical and theoretical momenta are really small: the twos can be considered approximately equal despite some error due to the small set of n.

6 MATLAB Code

```
%% STILL

clear all;
close all;
clc;


files = dir('Wheel_still_2020\*.txt');

datasets=[];
for i=1:length(files)

    data=importdata(strcat(files(i).folder, '\', files(i).name));
    if i>2 && length(data(:,1))<length(datasets(:,i-1))
        datasets(length(data(:,1))+1:length(datasets(:,i-1)),:)=[];
    end
    datasets(:,i)=data(:,1);

    clear data;
end
clear files i;

%%%%%%%%%%%%%%%%%%%%%%%%
        T=15; %us
%%%%%%%%%%%%%%%%%%%%%%%%


tags=datasets.*81e-12*1e6; %tags in us (microseconds)

clicks=tags(:,1);
for i=2:10
    clicks=[clicks; tags(2:length(tags(:,i)),i)+clicks(length(clicks))];
end
tags=clicks; clear clicks i;



m=discretize(tags,0:T:max(tags));
figure('Renderer', 'painters', 'Position', [500 300 900 600]);
hold on; grid on;

    [GC,GR] = groupcounts(m);
    GC(length(GC)+1:max(GR))=0; %zero padding

    [vals_y,~]=histcounts(GC, 'BinMethod', 'Integers');

    vals_x=0:length(vals_y)-1;

    N=sum(vals_y);

    vals_x=vals_x';
    vals_y=vals_y';

    stem(vals_x,vals_y/N, 'filled')

    lam=length(tags)/tags(length(tags))*T;


n=0:length(vals_x); n=n';
poisson=exp(-lam).*(lam.^n)./factorial(n);

plot(n,poisson, 'r', 'LineWidth', 1.5)

```

```
xlim([0 length(vals_x)])
title(strcat('Wheel still - T=', string(T), ...
    '\mus - empirical distribution'), 'FontSize', 15)
ylabel('number of occurrences normalized', 'FontSize', 15);
xlabel('number of photons in the interval T', 'FontSize', 15);

pbaspect([1.5 1 1])
clear i j initial final datasets;


%%%%%%%%%% MOMENTA %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

mu=lam;
m1P=sum((n.*exp(-mu).*mu.^n)./(factorial(n)));
diff1P=m1P-mu;
m2P=sum(((n-mu).^2.*exp(-mu).*mu.^n)./(factorial(n)));
diff2P=m2P-mu;
m3P=sum(((n-mu).^3.*exp(-mu).*mu.^n)./(factorial(n)));
diff3P=m3P-mu;
m4P=sum(((n-mu).^4.*exp(-mu).*mu.^n)./(factorial(n)));
diff4P=m4P-(mu+3*mu^2);

errorbar(n,poisson,...
    std(vals_y/N-poisson(1:length(vals_y)))*ones(size(poisson)))

legend('occurrences',strcat('Poisson distrib. (\lambda=', string(lam),...
    ')'), 'std dev', 'FontSize', 15);

%% SPINNING

clearvars -except T;
clc;


files = dir('Wheel_spinning_2020\*.txt');

datasets=[];
for i=1:length(files)

    data=importdata(strcat(files(i).folder, '\', files(i).name));
    if i>2 && length(data(:,1))<length(datasets(:,i-1))
        datasets(length(data(:,1))+1:length(datasets(:,i-1)),:)=[];
    end
    datasets(:,i)=data(:,1);

    clear data;
end
clear files i;

tags=datasets.*81e-12*1e6; %tags in us (microseconds)

clicks=tags(:,1);
for i=2:10
    clicks=[clicks; tags(2:length(tags(:,i)),i)+clicks(length(clicks))];
end
tags=clicks; clear clicks i;

m=discretize(tags,0:T:max(tags));
figure('Renderer', 'painters', 'Position', [500 300 900 600]);
hold on; grid on;

    [GC,GR] = groupcounts(m);
    GC(length(GC)+1:max(GR))=0; %zero padding


    [vals_y,¬]=histcounts(GC, 'BinMethod', 'Integers');

    N=sum(vals_y);
```

```

    vals_x=0:length(vals_y)-1;

    vals_x=vals_x';
    vals_y=vals_y';

    stem(vals_x,vals_y/N, 'filled')

     lam=length(tags)/tags(length(tags))*T;

n=0:length(vals_x); n=n';
BE=(1/(lam+1)).*(lam/(lam+1)).^n;

plot(n,BE, 'r', 'LineWidth', 1.5)

title(strcat('Wheel spinning - T=', string(T), ...
    '\mus - empirical distribution'), 'FontSize', 15)
ylabel('number of occurrences normalized', 'FontSize', 15);
xlabel('number of photons in the interval T', 'FontSize', 15);
xlim([0 30]); ylim([-inf inf]);

pbaspect([1.5 1 1])
clear i j initial final datasets h;

%%%%%%%%% MOMENTA %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

mu=lam;
m1BE=sum(n.*1./(1+mu).*(mu./(1+mu)).^n);
diff1BE=m1BE-mu;
m2BE=sum((n-mu).^2.*1./(1+mu).*(mu./(1+mu)).^n);
diff2BE=m2BE-(mu+mu^2);
m3BE=sum((n-mu).^3.*1./(1+mu).*(mu./(1+mu)).^n);
diff3BE=m3BE-(mu+3*mu^2+2*mu^3);
m4BE=sum((n-mu).^4.*1./(1+mu).*(mu./(1+mu)).^n);
diff4BE=m4BE-(mu + 10*mu^2 + 18*mu^3 + 9*mu^4);

errorbar(n,BE,std(vals_y/N-BE(1:length(vals_y)))*ones(size(BE)))

legend('occurrences',strcat('Bose-Einstein distrib. (\lambda=', ...
    string(lam), ')'), 'std dev', 'FontSize', 15)
```