École Polytechnique de l'Université de Tours
64, Avenue Jean Portalis 37200 TOURS, FRANCE
Tél. (33)2-47-36-14-14
Fax (33)2-47-36-14-22
www.polytech.univ-tours.fr

Parcours des écoles d'ingénieurs Polytech

The Exact Duplicate Image Detection for TV Guide Image Crawling

Computer science

Graduation thesis

Author

Naifeng GAN

[naifeng.gan@etu.univ-tours.fr]

Supervisor

Mathieu Delalandre

[mathieu.delalandre@univ-tours.fr]

Polytech Tours Département Informatique

Abstract

As the internet grows, we can find many duplicate images on the site. They have the same content, but some images may be downscaling, changed the illumination, scaling or compressed. Therefore, it is necessary to use exact duplicate image detection to find unauthorized images on the Internet to avoid copyright infringement. At the same time, on the multimedia playback application platform, images can be detected to obtain the related web pages, videos and images which are linked according to duplicate images.

In this report, I use web image crawler to get episode images on TV guides on almost ten French video sites, creating a database that can be updated once a week. Find different images and exact duplicate images in the database, and use three image matching methods SAD, SSD and NCC to obtain the most robust image detection algorithm. The images detection algorithm is optimized by changing the normalized image size and image color space. According to their P-R curve and F-measure curve, it is clear that the accuracy of NCC is the highest; the larger the normalized image size, the higher the detection accuracy; the detection algorithm in color space is more robust than that in grayscale space.

Keywords: image processing, exact duplicate image detection, image crawler, SAD, SSD, NCC, P-R curve, F-measure curve.

Acknowledgments

I would like to express my gratitude to all those who helped me during the project.

My deepest gratitude goes first and foremost to Dr. Mathieu Delalandre, my supervisor, for his instructive advice and useful suggestions on my project and report. Without his patient guidance and illuminating instruction, this project could not have reached its present result.

Secondly, I would like to express my school Beijing Institute of Technology and Polytech, Tours. I wouldn’t have the chance to do the project in France without the cooperation of the two schools.

Last, my thanks would go to my beloved family for their loving considerations and great confidence in me all through these months. I also owe my sincere gratitude to my friends who gave me their help and time in listening to me and helping my work out my problems during the difficult course of the project. In addition, I would like to express my gratitude to Louis Babuchon who helped me to obtain image crawler and build image database.

Table of contents

Table of figures

Introduction

1. What is image processing?

Image processing is a method of performing some operations on an image in order to obtain an enhanced image or extract some useful information. There are two types of methods for image processing namely, called analogue and digital image processing. Analogue image processing can be used for hard copies like printouts and photos. Digital image processing helps manipulate digital images by using a computer. In this paper, we only discuss the narrowly defined image processing: digital image processing.

An image can be defined as a two-dimensional function, $s = f(x, y)$ (x, y are $spatial$ coordinates, s is the two-dimensional space), and the amplitude of f at any pair of coordinates (x, y) is called the $intensity$ or $gray\ level$. When x, y and the intensity values of f are all finite, discrete quantities, we call the image a digital image. Digital image processing refers to processing digital images by using computer algorithms.

Regarding the definition of digital image processing, due to its high correlation with computer vision and image analysis, we usually consider three types of computerized processes: $\text{low}-, \text{mid}-,$ and $\text{high}-\text{level processes}$.Low $-$ level processes involve primitive operations. The inputs and outputs of $\text{Low}-\text{level processes}$ are both images. $\text{Mid}-\text{level processes}$ involves tasks such as segmentation and recognition of individual objects. A $\text{Mid}-\text{level process}$ is characterized by the fact that its inputs are images, but its outputs are attributes extracted from those images like features or targets. $\text{High}-\text{level processing}$ involves "making sense" of an ensemble of recognized objects and performing the cognitive functions associated, which typically associated with computer vision. Besides, A $\text{High}-\text{level processes}$ is characterized by the fact that its inputs are images and then it carries out an action. Therefore, in general, we define image processing as: its inputs are images, outputs are also images, or attributes extracted from images, including the recognition of individual objects. At present, digital image processing has been applied in the fields of Communication Engineering, Aerospace Engineering, Biomedicine Engineering and machine vision. Some of these application areas are illustrated in the following section.

2. Examples of fields using image processing

Nowadays, image processing becomes an indispensable part of our daily lives. So, in this section, I will briefly introduce some application areas of image processing. Of

course, this discussion doesn't cover all of image processing's application fields, I will only introduce a small part of them.

(1) Communication Engineering

Multimedia communication combining sound, text, image and data is the main development direction of current communication engineering. Specifically, it uses the way called tri-network which combining mobile communication network, television network, and worldwide internet to spread on digital communication networks, such as FaceTime developed by Apple. (Figure 1)

Figure 1: People using FaceTime with their phone and computer

Among them, image communication is the most complicated and difficult part, because the amount of the image data is huge, such as the transmission rate of color television signals is more than $100Mbit/s$. Therefore, in order to transmit high-speed image data in real time, a new coding technique must be used to compress the number of bit of information.

In addition to the widely used entropy encoding, DPCM coding, and transform coding, new coding methods such as branch coding, adaptive modulation coding, and the image compression coding based on wavelet transform are being developed internationally. And, of course, the digital image processing runs through the entire encoding process.

(2) Aerospace Engineering

Digital image processing is widely used in the aerospace engineering. In addition to the image processing of outer space photographs such as the Moon and Mars, the image processing is also used in aircraft and satellite remote sensing technology. Investigate images taken by aircraft or satellites are digitally encoded in the air, then transmitted to the ground at high speed, and analyzed by the processing center. From the imaging, storage and transmission processes to analysis process, digital image processing methods must be used. Currently, the technology is mainly used in the world for satellite positioning, resource survey, city survey, etc. For example, the GPS (Global Positioning System) developed by the US military, the Galileo system of the European civil satellite navigation system, and the BeiDou Navigation Satellite

System (BDS) developed by China, more and more map software is extremely convenient for our travel, which is inseparable to digital image processing. (Figure 2)

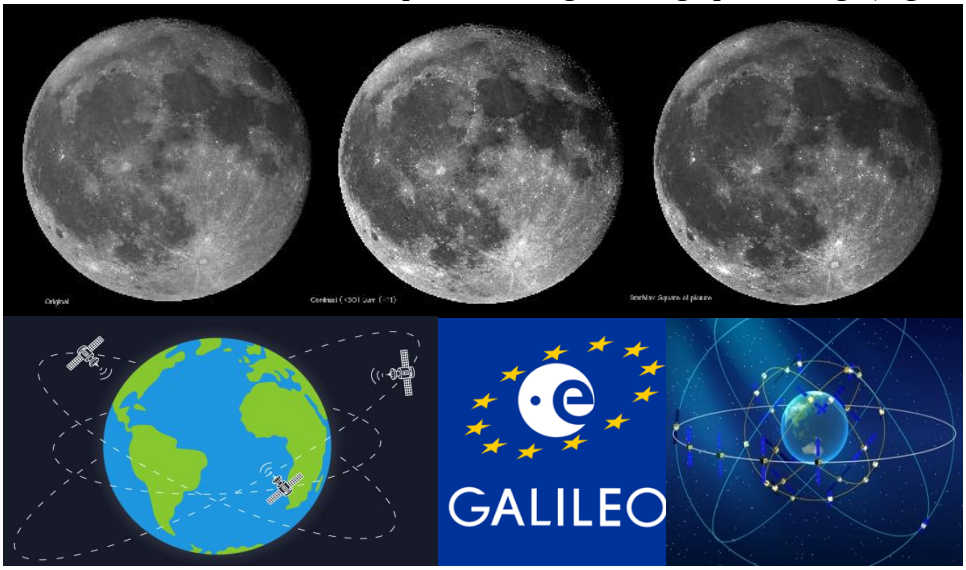
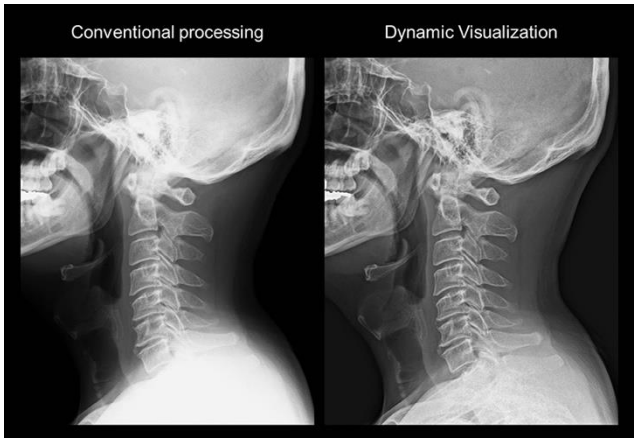


Figure 2: 1) Moon pictures (using image processing); 2) GPS system; 3) Galileo system; 4) Bei-Dou Navigation Satellite System

(3) Biomedicine Engineering

The internal structure and intravascular substances of the human body can be presented in the form of images using the digital image processing, and can be compared. Digital image processing provides a more scientific diagnostic approach to biomedical engineering. For example, gamma ray is used for localization of tumors or infections, X-rays for angiography (Figure 3), and computerized axial tomography (CAT) for three-dimensional images for blood vessel and human internal description.



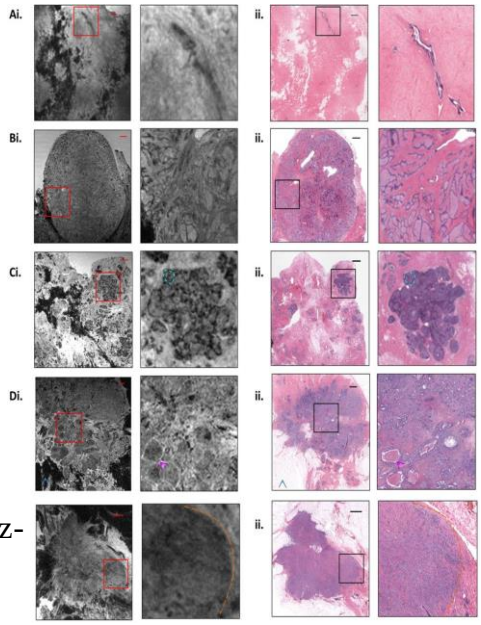


Figure 3: 1) The internal structure of human body using different kind of image processing. 2) Analyzing tumor by comparing the microscopic images.

There is also a class of processing and analysis of medical microscopic images, such as red blood cells, white blood cell classification, chromosome analysis, cancer cell recognition, etc.

(4) Machine Vision

Machine vision can accept and process an image of a real object automatically by optical means and non-contact sensors, by analyzing the image to obtain the information or controlling the motion of the machine. Therefore, the realization of machine vision is inseparable from image processing. For example, face and fingerprint recognition functions are widely used in smart phones, smart door locks and crime tracing. (Figure 4)

At the same time, machine vision is also used in military reconnaissance, logistics, hospital and home service intelligent robots, automated production line equipment and space robots, etc. Among them, the most popular technology at present is face recognition technology, including face tracking detection, automatic adjustment of image magnification, nighttime infrared detection, automatic adjustment of exposure intensity and other technologies. Face recognition technology consists of three parts:

1) Face detection: Face detection is to determine whether there are images in a dynamic scene or complex background, and to separate these images;

2) Face tracking: Face Tracking is a dynamic tracking of the detected face image;

3) Face comparison: Face comparison is to identify the detected face image or perform a target search in the face gallery to find the best match. [2]

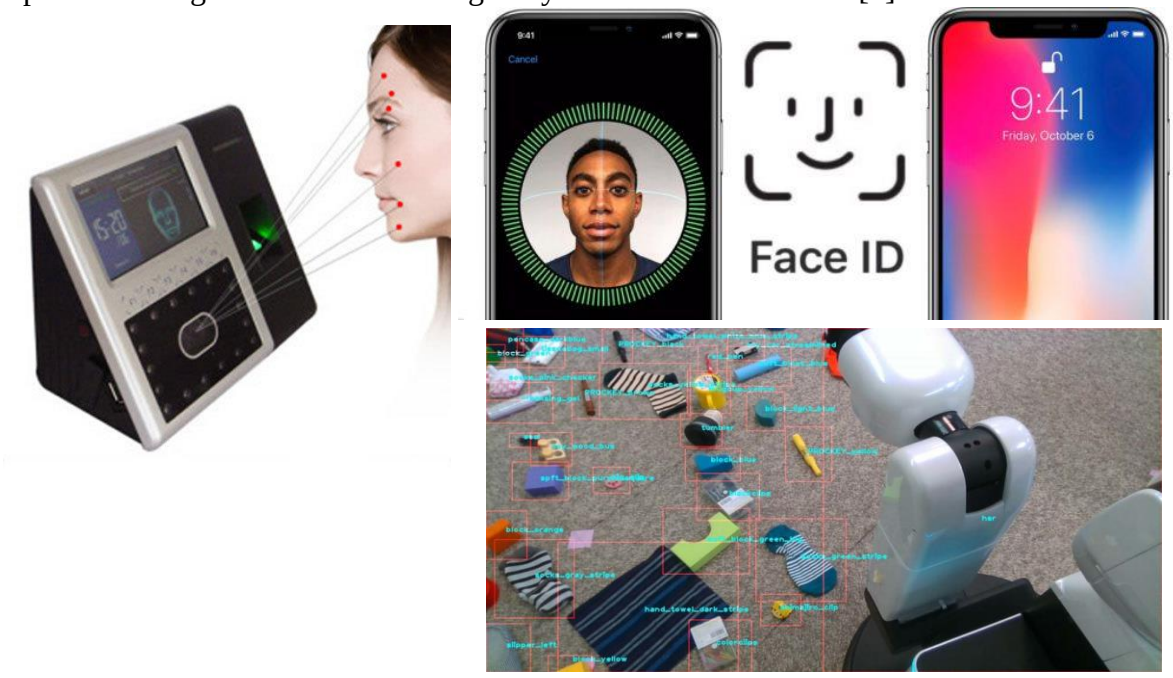


Figure 4: 1) Fingerprint and face recognition 2) Face recognition in iPhone 3) Machine vision in home service intelligent robots

In a conclusion, image processing has become an indispensable part of our daily lives. With the development of image processing technology, more and more technical fields will apply image processing technology to make our lives more intelligent and automated.

Chapter 1

General Description

1.1 Introduction of duplicate image detection

The definition of exact duplicate image detection is to detect exact duplicate image using computer algorithms and to drive a performance characterization. So, what is the duplicate image? I will introduce two types of duplicate images and compare them in the following part of this section.

The duplicate image are images that are visually similar, and can be divided into near duplicate image and exact duplicate image. Near Duplicate Image is a pair of images, one of which is close to the exact copy of the other, but is slightly different due to changes in shooting conditions (camera model, perspective, time, rendering conditions, editing operations, etc.). From the perspective of image encoding, near duplicate images are images in which the binary form are not completely repeated, but are similarly imaged by visual recognition.[3] The application ranges from exact duplicate detection where no changes are allowed to a more general definition that requires the images to be of the same scene, but with possibly different viewpoints and illumination.[4]At present, it is commonly used in the multimedia playback application platform to detect pictures by Near duplicate image detection, and then link related web pages, videos and pictures according to near duplicate images. After that, the website can use machine learning to obtain the interest preferences of their users. Finally, the website builds the user's persona based on user profile.

According to the article [4], we divide the near duplicate image into the following categories:

1. Scene changes: objects, absence or presence of foreground objects, background change, post editing of images, etc.
2. Camera parameter changes: camera perspective change, camera tilting, panning, zooming, etc.
3. Photometric changes: change of exposure or lighting condition, etc.
4. Digitization changes: hue shift, contrast change, resolution change, etc.

Therefore, each image may have many different types of copies. So, the near duplicate image detection becomes very difficult but important. (Figure 5)

The exact duplicate images refer to some images that have the same visual content (for the human eyes), but they present main difference in the way they are coded due to degradation such as downscaling, illumination changes,

Figure 5: Near duplicate image

shifting and compression, etc. Exact duplicate image detection is widely used in our daily lives. For example, using exact duplicate image detection to find unauthorized images on the internet to avoid copyright infringement.

For exact duplicate image, we divide it into several categories:

1. Downscaling: When the original image doesn't meet the required size, then we can downscale it to obtain a new image that fits the size. These two images are exact duplicate images.
2. Color space model change: Color space is the way in which colors are organized. A color model is a color that is described by a limited set of numbers. For example, the color space RGB is defined by three primary colors of red, green, and blue, three colors form three coordinate axes, and each possible color has a unique position in the three-dimensional space. The color space is not unique, HSV color space is produced by hue, saturation and value, and CMYK color space is produced by cyan, magenta, yellow and black. Because of the changes between these color space models, there are some images have the same content with different expression colors. These images are exact duplicate images. (Figure 6)

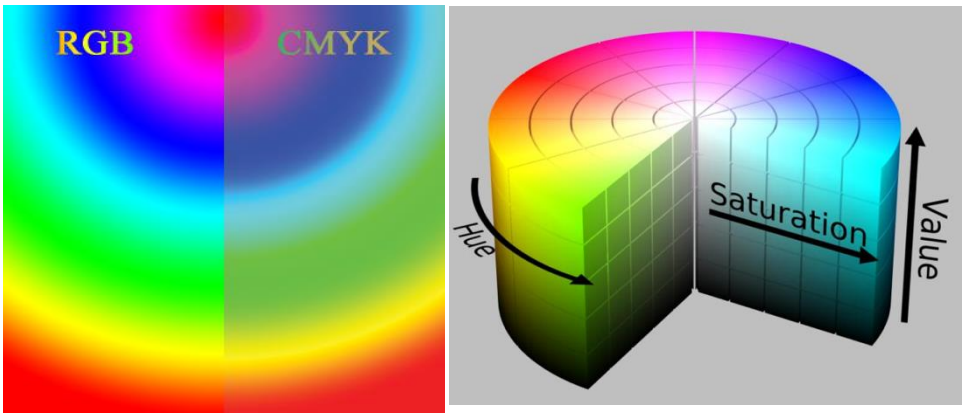


Figure 6: 1) Two different color spaces----RGB and CMYK; 2) HSV color

3. Illumination change: For some image acquisition systems, it is common that light changes from morning to night, so images taken at the same location

with the same content may have difference between light and dark. At the same time, in photography, we always adjust the lighting of the photos to make the main content clearer and brighter.

4. Compression: The purpose of image compression is to reduce redundant information in image data, so that data can be stored and transmitted in a more efficient format, and at the same time still ensuring the quality of the image. The compressed image and the original image are exact duplicate images.

Here, I use an RGB image about a cute cat to indicate the difference between these types of changes and the original image. (Figure 7)

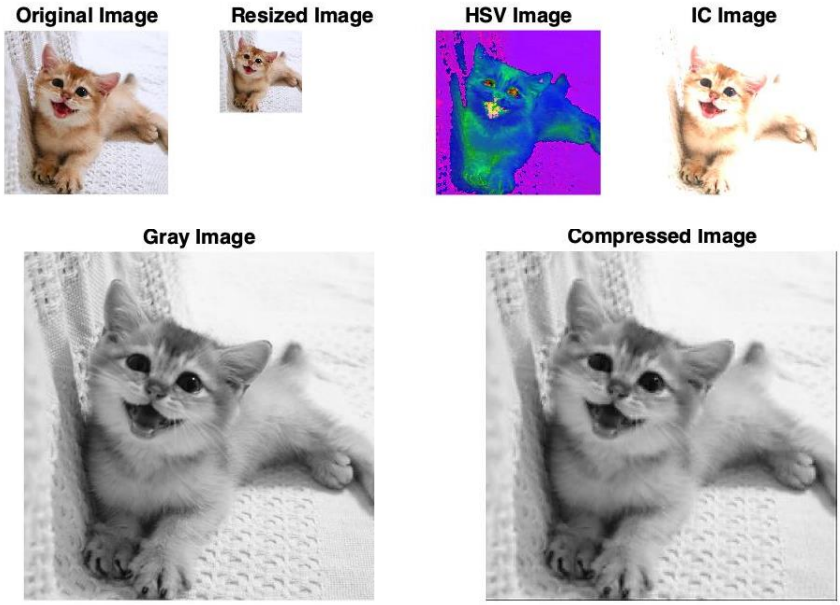


Figure 7: 1) The original image; 2) Scale the original image by 0.5 times; 3) Change the original image from RGB to HSV; 4) Change the illumination of the original image; 5) Change the original image to grayscale image; 6) Compress the grayscale image of the original image. (from left to right, top to bottom)

Therefore, the difference between exact duplicate image and near duplicate image is: in terms of vision, the exact duplicate image has the same visual content, but the near duplicate image has some differences; in terms of image encoding, there is a small difference between the exact duplicate images, and near duplicate image is image in which the binary form is not completely repeated. And in the following we mainly discuss the exact duplicate image and how to detect exact duplicate image.

1.2 Objective

Nowadays, duplicate images can be found widely in various websites. For example, the famous cartoon *"zip zip"* has only three different images on the TV guides of more than a dozen video websites in France. (Figure 8)

These images may have lower quality than the original one, or may be compressed, adjusted hues to change the size and illumination of the images. They are exact duplicate images, but those with aspect ratio changes or cropped images do not fall within this range. Based on this phenomenon, my goals are as follows (Figure 9):

Figure 8: Three different images in more than ten video websites

1. Establish a database: Firstly, I will create a huge image database through XMLTV, so that I can get enough images and extract exact duplicate images; Secondly, I will use image crawlers to obtain images of several video websites and save them in the database; (Using the algorithm of Louis, it will be introduced in the next section)
2. Detect exact duplicate images (Hand Made): In the image database, the images in the same folder will be detected to obtain the exact duplicate images. (Select a part of folders in the database for detection by hand) Finally, save the result in the

file named duplicate, and the different images will be saved in the file named different.

3. Three image matching methods: Use three image matching methods: SAD, SSD, NCC to obtain the distance between different images and the distance between exact duplicate images. And the results will be represented by frequency-distance curve.
4. Optimal robust method: By changing the three variables of the image attributes, image matching method and normalized image size, the precision and recall curves and F-measure curves will be drawn to obtain the most robust optimal algorithm.

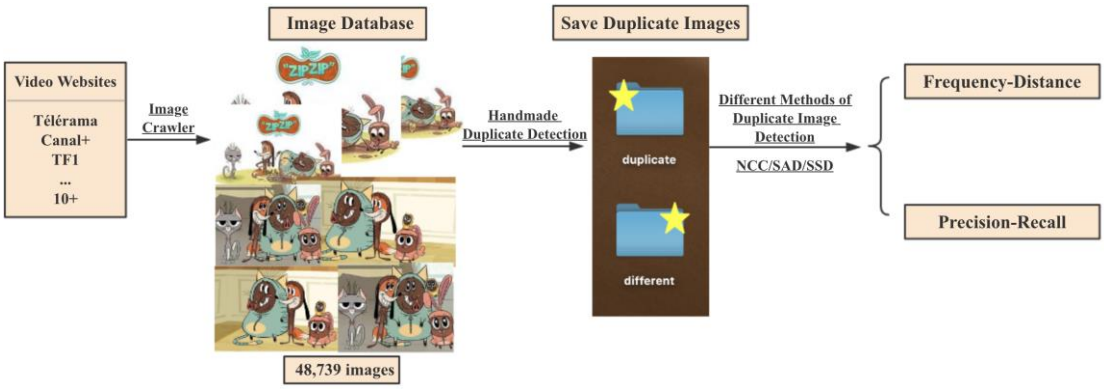


Figure 9: Project process

1.3 Image database from Louis

According the report: Smart image scraping for media guide TV / video portal from Louis Babuchon I build a image database to get images from websites.

1.3.1 Introduction of Louis's report

In this era, we are all consumers of the media. Nowadays, Internet TV has become the first choice for our consumption of media. Therefore, there are many TV guides on the Internet to classify TV program content, but the number of different guides is huge and there is data redundancy. We can find many duplicate images on the website. They have the same content, but some images may be downscaling, changed color space model, illumination changed or compressed. Louis formed an image database in XMLTV for analyzing the image features available on the TV guide; then used the web crawler to obtain the image of the website, and used duplicate image detection to extract the metadata in the image and store it in the XMLTV database. He implemented a program in Java which can automatically retrieve images from the

Télérama website to obtain an image database with sufficient data volume. The entire project can be divided into two parts including five components:

- Image crawler:
  1. Program extraction: Restore the information of the program according to the source (XMLTV or Web) and extract the links of different images, and send the information list.
  2. Download Image: Download all images from the information list and store them locally according to the defined file system.
- Selector:
  3. Duplicate image detection: Read the images stored on the previous step of the disk, detect duplicates and delete duplicates with the lowest quality(resolution/compression);
  4. Metadata extraction: Get the previously filtered images and extract amount of metadata (copyright, dimension, author);
  5. Database Record: Restore filtered and recorded images and record them in the XMLTV database.

1.3.2 The establishment of database

The XMLTV structure is used to represent television programs in an XML format. Louis chose the website *xmltv france* to use *telemara* API 2, which allows TV programs to be restored in a few weeks and provides amount of information on the programs, which is updated regularly. The XMLTV structure can describe a list of TV channels by name and icon. It is a list of programs which contains a start time, an end time, and a list that describe other information about the program. The most important is the <icon> tag, which is the link to the image. Finally, get these links to create a database and get a list of related programs and images.

1.4 Assumptions and constraints

In this project, I will obtain an image database which needs to update the images data in real time so that I can detect the exact duplicate images. I will use image crawler to get the latest image data on the websites and save them in the image database, then keep it updated every week. In order to achieve the image crawler, I chose Java of ImageJ library as the programming language.

I will use MATLAB to design a program which will automatically calculate the distance between exact duplicate images in the same folder in the image database and the distance between different images in the same folder. If it doesn't work, I will use Python or Opencv again to achieve the ultimate goal.

My ultimate goal is the detection of exact duplicate image, not including the near duplicate image. In addition, when the aspect ratios of the two images are different, it can be considered that two images are not exact duplicate images (may be cropped or scaled only in one direction).

Chapter 2

Analysis and design

2.1 Image crawler

The image crawler will be used to capture images from the TV guide of the video websites, and the crawled images will form my image database. Therefore, the image database has the following features: Real-time updates, large amount of data, and more data redundancy. Here, I used Louis's Java program to finish this part.

This program is done with Java tools, so I need to run the JAR file on my computer.

2.1.1 Running JAR file on windows

I have two methods to run JAR files on windows:

1. Installing the Java Runtime Environment (JRE); and then enter the CMD windows Switch to the directory where the jar is located.
2. When you don't have JRE:

- Open the Windows Explorer, from the Tools select 'Folder Options...'
- Click the File Types tab, scroll down and select JAR File type.
- Press the Advanced button.
- In the Edit File Type dialog box, select open in Actions box and click Edit...
- Press the Browse button and navigate to the location the Java interpreter javaw.exe.
- In the Application used to perform action field, needs to display something similar to C:\Program Files\Java\j2re1.4.2_04\bin\javaw.exe" -jar "%1" % (Note: the part starting with 'javaw' must be exactly like that; the other part of the path name can vary depending on which version of Java you're using) then press the OK buttons until all the dialogs are closed.

[from http://windowstipoftheday.blogspot.com/2005/10/setting-jar-file-association.html]

2.2 Image database

Based on the image crawler of Louis, after crawling the images on the TV guide of 5 video sites, I got an image database. (Figure 10)

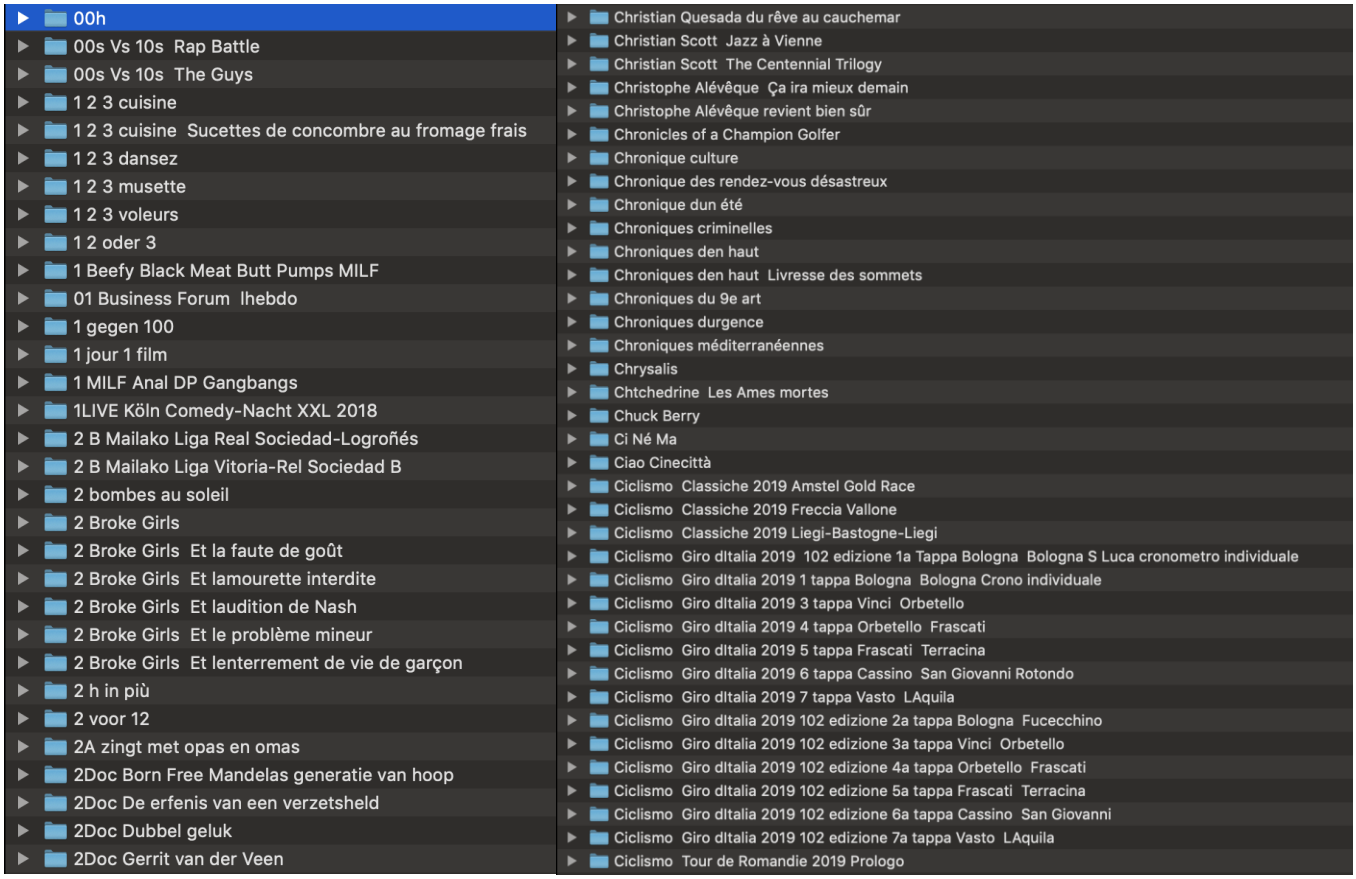


Figure 10: Entire image database

The database contents about 48,739 images, 17,391 programs, and the entire file size is 4.48 GB. Among them, each folder contains the same TV program images crawled from different video sites. Some of the images from the same folder are different images, some are near duplicate images and some are exact duplicate images.

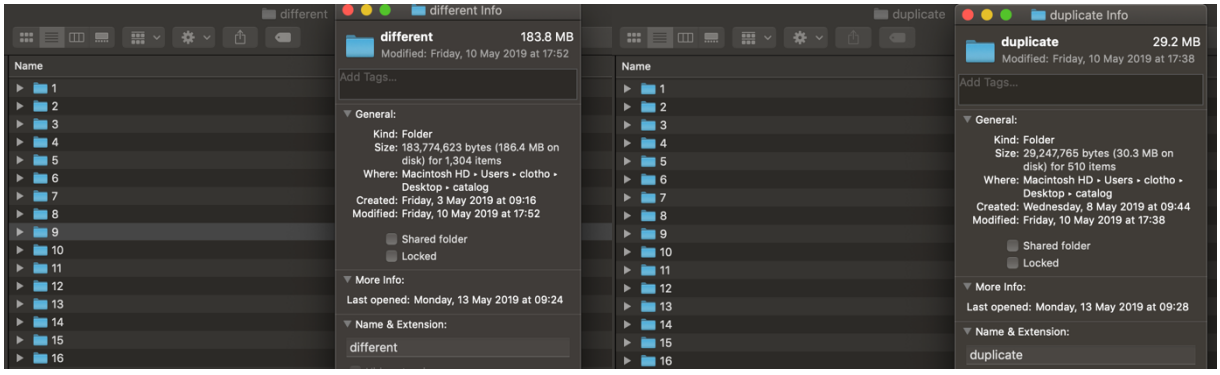


Figure 11: Duplicate and different files

Then, I performed image detection on these files. The exact duplicate images are stored in the duplicate folder, and the different images are stored in the different folder by manually folder. So, I got a classified database of images including 1,814 images. (Figure 11)

2.3 Exact duplicate image detection

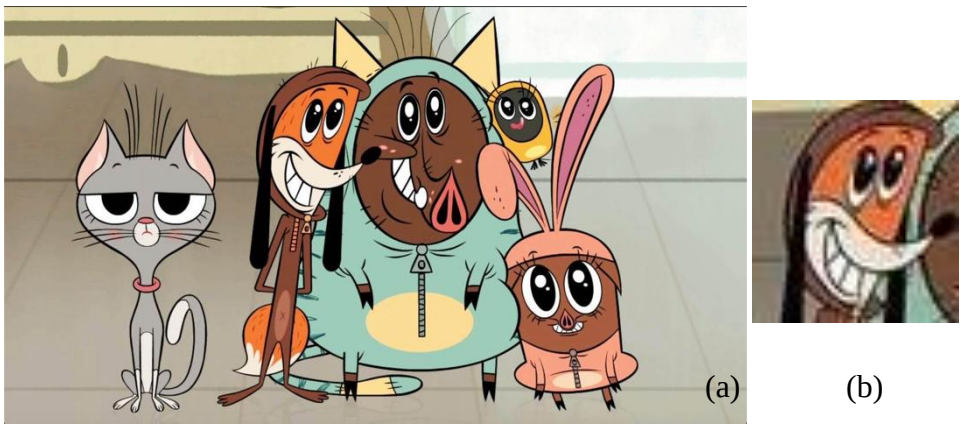
2.3.1 Introduction to detection methods

2.3.1.1 Mean Absolute Differences (MAD)

Mean Absolute Differences algorithm (MAD) is an image matching algorithm proposed by Leese in 1971. It is in commonly used in image pattern recognition. This algorithm has high matching precision and is widely used in image matching.

1. Interpretation

Suppose we have $S(x, y)$ which we called the *search image* of size $M \times N$, and (x, y) represents the coordinates of each pixel in the *search image*. $T(x_t, y_t)$ is a *template image* of $m \times n$, as shown in the figure 2.3.1.1-1, figure (a) is the *search image*, and figure (b) is the *template image*. our purpose is to find the area matching (b) in (a).



(a) (b)

Figure 12: (a) Search image; (b) Template image;

In the *search image* S, take (i, j) as the upper left corner, take the subgraph of $m \times n$ size, calculate its similarity with the *template image*; traverse the entire search image, find the most similar subgraph of the template image as the final match result in all the subgraphs that can be found.

The similarity formula of the MAD algorithm is as follows:

$$D(i,j) = \frac{1}{m \times n}\sum_{s=1}^{m}\sum_{t=1}^{n}|S(i+s-1, j+t-1) - T(s,t)|$$

Among them: $1 \le i \le M-m+1, 1 \le j \le N-n+1$

The smaller $D(i,j)$, the more similar the two images are, so as long as the smallest D(I,j) is found, the best matching position can be determined and the matching degree of the two images can be obtained. [5]

2. Algorithm implementation in MATLAB

```
%%MAD
clear all;
close all;
%%
src=imread('/Users/clotho/Desktop/image_1.jpg');
[M N d]=size(src);
if d==3
    src=rgb2gray(src);
end
mask=imread('/Users/clotho/Desktop/image_2.jpg');
[m n d]=size(mask);
if d==3
    mask=rgb2gray(mask);
end
%src = imresize(src,[a,a]);
%mask = imresize(mask,[n,n]);

%%
dst=zeros(M-m+1,N-n+1);
for i=1:M-m+1
    for j=1:N-n+1
        temp=src(i:i+m-1,j:j+n-1);
        dst(i,j)=sum(sum(abs(temp-mask)))/(m*n);
    end
end
abs_min=min(min(dst));
[x,y]=find(dst==abs_min);
figure;
subplot(121);
imshow(mask);title('template image');
subplot(122);
imshow(src);
hold on;
rectangle('position',[y,x,m,n],'edgecolor','r');
hold off;title('search image');
```

Figure 13: The program in MATLAB of MAD

First, two pictures are sequentially read to obtain a search image and a template image. Convert the RGB image to a grayscale image and scale the two images to

Figure 14: The result of the MAD program

squares. After performing the above steps, the subgraph that are most similar to the template image in the search image are obtained and framed by a square frame. (Figure 13)
Finally, we get the results of the image matching using MAD. (Figure 14)

3. Advantages and disadvantages

The MAD is simple and easy to understand. The essence of MAD is to calculate the average of the L1 distance between the subgraph and the template image. Besides the calculation process is simple and the matching precision is high.

However, the amount of calculation of MAD algorithm is too large and it has a bad anti-noise ability. Since the MAD is basically the same as SAD, only the SAD algorithm will be used for image processing in the following.

2.3.1.2 Sum of Absolute Differences (SAD)

1. Interpretation

In fact, the SAD algorithm is almost identical to the MAD algorithm, except that the similarity measurement formula has a slight change (calculating the square of the L1 distance between the subgraph and template image), so it will not be described again here.

The similarity formula of the SAD algorithm is as follows:

$$D(i,j) = \sum_{s=1}^{m}\sum_{t=1}^{n}|S(i+s-1, j+t-1) - T(s,t)|$$

Among them: $1 \le i \le M-m+1, 1 \le j \le N-n+1$

2. Algorithm implementation in MATLAB

```
%SAD
clear all;
close all;
%%
src=imread('/Users/clotho/Desktop/image_1.jpg');
[M N d]=size(src);
if d==3
    src=rgb2gray(src);
end
mask=imread('/Users/clotho/Desktop/image_2.jpg');
[m n d]=size(mask);
if d==3
    mask=rgb2gray(mask);
end
%src = imresize(src,[a,a]);
%mask = imresize(mask,[n,n]);

%%
dst=zeros(M-m+1,N-n+1);
for i=1:M-m+1
    for j=1:N-n+1
        temp=src(i:i+m-1,j:j+n-1);
        dst(i,j)=sum(sum(abs(temp-mask)));
    end
end
abs_min=min(min(dst));
[x,y]=find(dst==abs_min);
figure;
subplot(121);
imshow(mask);title('template image');
subplot(122);
imshow(src);
hold on;
daspect([1,1,1]);
rectangle('position',[y,x,m,n],'edgecolor','g');
hold off;title('search image');
```

Figure 15: The program in MATLAB about SAD

The SAD algorithm process is basically the same as MAD, but it is different when calculating similarity. (Figure 15)

Finally, we get the results of the image matching using SAD. (Figure 16)

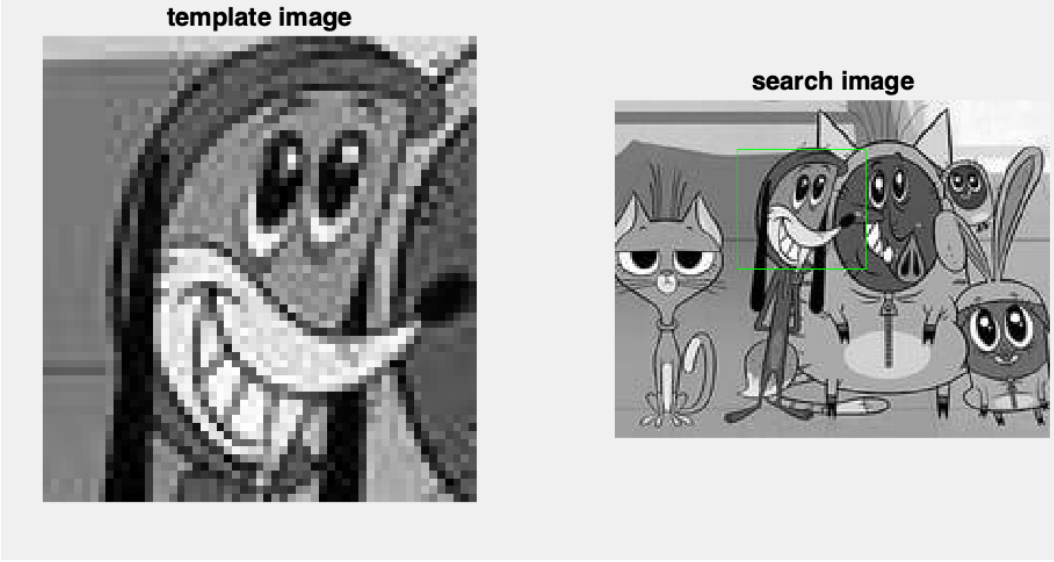


Figure 16: The result of the SAD program

3. Advantages and disadvantages

The SAD is a commonly used similarity metric. The same as MAD, the amount of SAD algorithm is too large and it will take a lot of time to complete the program. Although there are many approaches whose objective is to speed up the process of SAD

matching, they can only give the position of the minimum. When the distance SAD should be calculated for every location in the image, the direct SAD computation requires more much time. So, it needs to improve the computing time and computing method. [6]

2.3.1.3 Normalized Cross Correlation

Normalized cross correlation (NCC) is an algorithm for calculating the correlation between two sets of sample data based on statistics. Its value range is [-1, 1], and for image, each pixel can be regarded as RGB value. The whole image can be regarded as a collection of sample data. If it has a subset that matches another sample data, its NCC value is "1", indicating a high correlation. If its NCC value is "-1", it is completely irrelevant. Based on this principle, the first step of the image detection based on the template matching is to normalize the data.

1. Interpretation

The implementation of the NCC algorithm is similar to the above algorithms. It uses the normalized correlation formula to calculate the matching degree (NCC value) between the subgraph of the search image and the template image. [7]

The following is the formula of the NCC:

$$R(i,j) = \frac{1}{m \times n} \frac{\sum_{s=1}^{m}\sum_{t=1}^{n}\left[S^{i,j}(s,t) - E\left(S^{i,j}\right)\right] \times [T(s,t) - E(T)]}{\sigma_S \sigma_T}$$

Among them: $E\left(S^{i,j}\right), E(T)$ represent the average gray value of the subgraph and template image respectively. σ_S, σ_T represent the standard deviation of all pixels of subgraph and template image respectively.

$$\sigma_S = \sqrt{\frac{1}{m \times n}\sum_{s=1}^{m}\sum_{t=1}^{n}[S^{i,j}(s,t) - E(S^{i,j})]^2}$$

$$\sigma_T = \sqrt{\frac{1}{m \times n}\sum_{s=1}^{m}\sum_{t=1}^{n}[T(s,t) - E(T)]^2}$$

2. Algorithm implementation in MATLAB

Firstly, a part of NCC algorithm is the same as the above algorithm. Secondly, obtains the template image pixels and calculates the data samples of the mean and standard deviation; Thirdly, according to the template image size, moves the window from left to right and top to bottom on the search image, and calculates the NCC value(result) between the pixels of template image and the pixels in the window of the search

```
%ncc
clear;
src=imread('/Users/clotho/Desktop/image_1.jpg');
mask=imread('/Users/clotho/Desktop/image_2.jpg');
[M,N,b]=size(src);
[m,n,c]=size(mask);
src=rgb2gray(src);
mask=rgb2gray(mask);
result=zeros(M-m+1,N-n+1);
vec_sub = double(mask(:));
norm_sub = norm(vec_sub);
for i=1:M-m+1
    for j=1:N-n+1
        subMatr=src(i:i+m-1,j:j+n-1);
        vec=double( subMatr(:) );
        result(i,j)=vec'*vec_sub / (norm(vec)*norm_sub+eps);
    end
end
%find the max position
[iMaxPos,jMaxPos]=find( result==max( result(:)));
figure,
subplot(121);imshow(mask),title('template image');
subplot(122);
imshow(src);
title('search image'),
hold on
plot(jMaxPos,iMaxPos,'*');%plot the max position
 %create a square box
plot([jMaxPos,jMaxPos+n-1],[iMaxPos,iMaxPos]);
plot([jMaxPos+n-1,jMaxPos+n-1],[iMaxPos,iMaxPos+m-1]);
plot([jMaxPos,jMaxPos+n-1],[iMaxPos+m-1,iMaxPos+m-1]);
plot([jMaxPos,jMaxPos],[iMaxPos,iMaxPos+m-1]);
```

Figure 18: The program in MATLAB about NCC

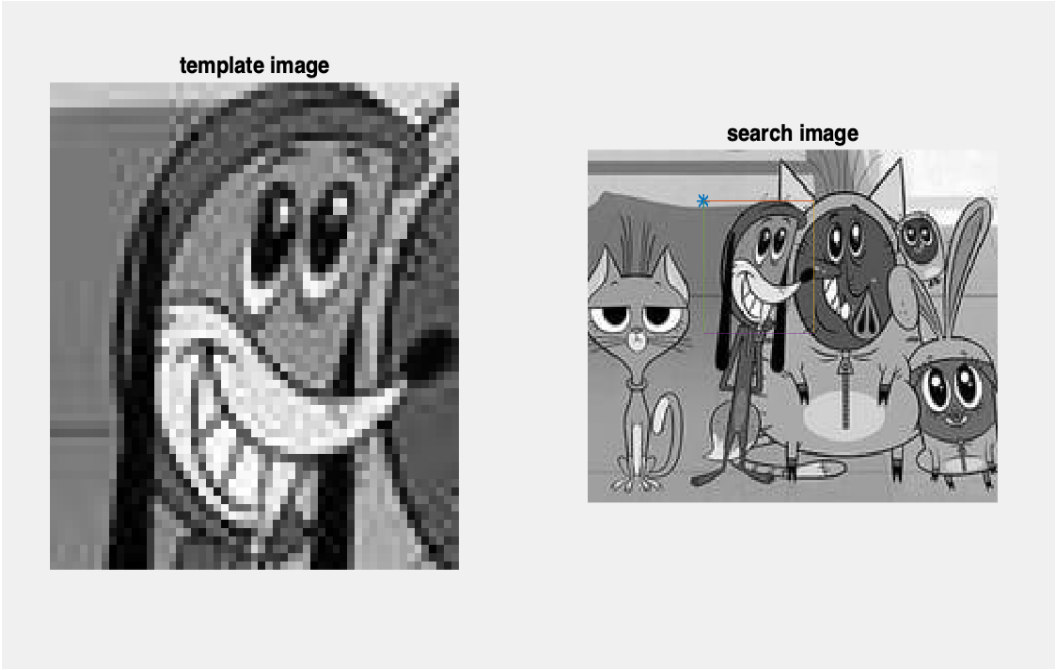


Figure 17: The result of the NCC program

image; Finally, find the position of the pixel which has the largest NCC value, and frame the matching part by a square frame.(Figure 18)

Here is the result of the image matching I got: (Figure 17)

3. Advantages and disadvantages

The feature point matching based on the normalized cross correlation is well employed because of its good anti-noise ability. However, large amount of calculations are needed by more feature points. Many researches mainly focusing on the template-based matching have been carried out to solve the problem of low efficiency of the NCC algorithm.

2.3.1.3 Sum of Squared Differences

Image matching between search image and template image, which is carried out using sum of square differences (SSD), has been widely used in various computer vision applications such as stereo measurements and superresolution image syntheses.[8] Describe SSD briefly 1) Sum of square differences between entries of the two descriptors; 2) Does not provide a way to discard ambiguous (bad) matches.

1. Interpretation

Sum of Squared Differences (SSD algorithm), also called difference and algorithm. In fact, the SSD algorithm is exactly the same as the SAD algorithm, except that the similarity measurement formula has a slight change (calculating the L2 distance between the subgraph of the search image and the template image).

The similarity formula of the SAD algorithm is as follows:

$$D(i,j) = \sum_{s=1}^{m}\sum_{t=1}^{n}[S(i+s-1, j+t-1) - T(s,t)]^2$$

Among them: $1 \le i \le M-m+1, 1 \le j \le N-n+1$

2. Algorithm implementation in MATLAB

```
%SSD
%%
clear all;
src=imread('/Users/clotho/Desktop/image_1.jpg');
[M N d]=size(src);
if d==3
    src=rgb2gray(src);
end
mask=imread('/Users/clotho/Desktop/image_2.jpg');
[m n d]=size(mask);
if d==3
    mask=rgb2gray(mask);
end
%src = imresize(src,[a,a]);
%mask = imresize(mask,[n,n]);

%%
dst=zeros(M-m+1,N-n+1);
for i=1:M-m+1
    for j=1:N-n+1
        temp=src(i:i+m-1,j:j+n-1);
        dst(i,j)=sum(sum((temp-mask).^2));
    end
end
abs_min=min(min(dst));
[x,y]=find(dst==abs_min);
figure;
subplot(121);
imshow(mask);title('template image');
subplot(122);
imshow(src);
hold on;
daspect([1,1,1]);
rectangle('position',[y,x,m,n],'edgecolor','y');
hold off;title('search image');
```

Figure 19: The program in MATLAB of SSD

The SSD algorithm process is basically the same as SAD, but it is different when calculating similarity. (Figure 19)

Finally, the results of the matching program are as follow: (Figure 20)

Figure 20: The results of SSD program

3. Advantages and disadvantages

Theoretically, SSD is simple among similarity measures method and has less computation cost since it only involves square operation and pixels subtraction between

template image and the search image compared with NCC. However, NCC is more robust than SSD in illumination change. [9]

2.3.2 Duplicate image detection algorithm

2.3.2.1 Software support

MATLAB is a high-performance language for technical computing. It integrates computing, visualization and programming in an easy environment. In this environment, the problem and its solution are expressed in the mathematical notation we are familiar with. Typical applications include the following:

- ✧ Mathematics and calculation；
- ✧ Algorithm development;
- ✧ Data acquisition;
- ✧ Modeling, simulation and prototyping;
- ✧ Data analysis, research and visualization;
- ✧ Science and engineering graphics;
- ✧ Application development, including image user interface construction.

MATLAB is an interactive system. Its basic data elements are an array that does not require the determination of dimensions. This allows people to solve many technical calculation problems by formulating methods, especially those involving matrix representation. [10]

The Image Processing Toolbox is a MATLAB function set that extends its ability to solve image processing problems. So, I used MATLAB to solve the duplicate image detection part.

2.3.2.2 The types of exact duplicate image

After studying the duplicate images in the database, I obtained the following types of duplicate image:

- ✧ Scaling;
- ✧ Downscaling;
- ✧ Illumination change;
- ✧ Compression;
- ✧ Image shifting;

It should be noted that when the aspect ratios of the two images are different, it can be considered that the two images are not exact duplicate images. One of the two images may be cropped or only scaled in a certain direction. At the same time, the near duplicate image is not within the scope of my discussion.

2.3.2.3 Algorithm process

I will use MATLAB to complete the following steps to get the distance between images (including duplicate and different images) using six methods mentioned above:

1. Extract images: Extract two images A, B from the same folder (duplicate or different) in the image database sequentially;
2. Image grayscale: Convert image A, B into grayscale images
3. Normalized size: Obtain the size of two images A and B, compare the number of row of pixels. Then for example make the image A with small pixel rows as the benchmark, and reduce B to the same size as A; According to the size of the two images, extract a part of the image of the center of the two images, get the normalized image size 128×128;
4. Calculating distance: Calculate the distance of the images using the above three methods;
5. Robustness analysis: By changing the variables – grayscale or color, image matching methods, normalized image size, drawing precision and recall curve and F-measure curve, analyzing algorithms' robustness, obtaining optimal image attribute, image matching method and normalized image size. (Figure 21)

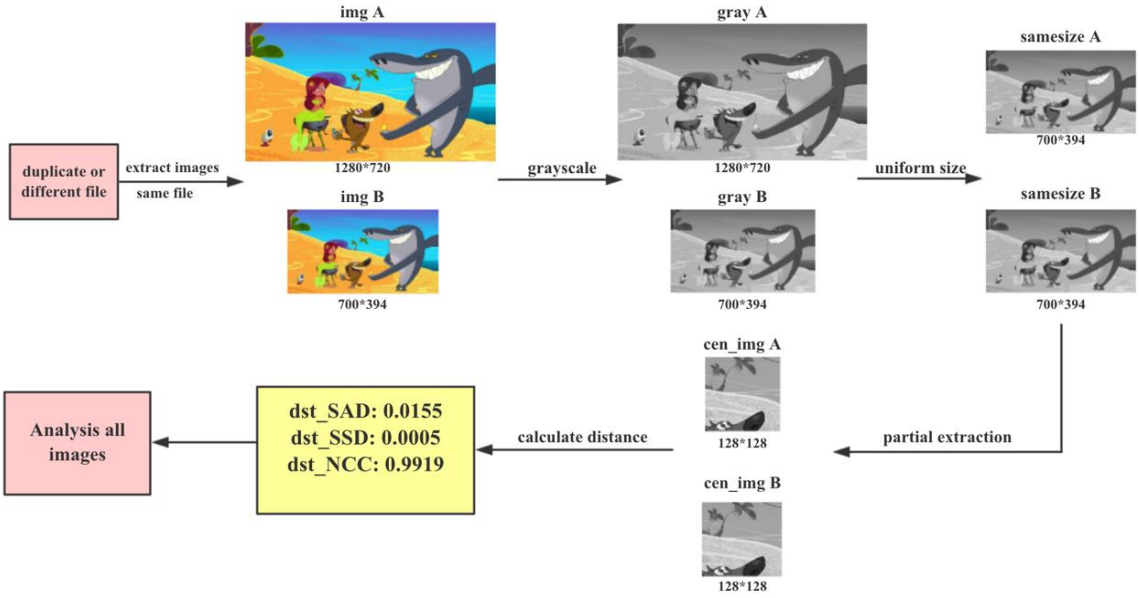
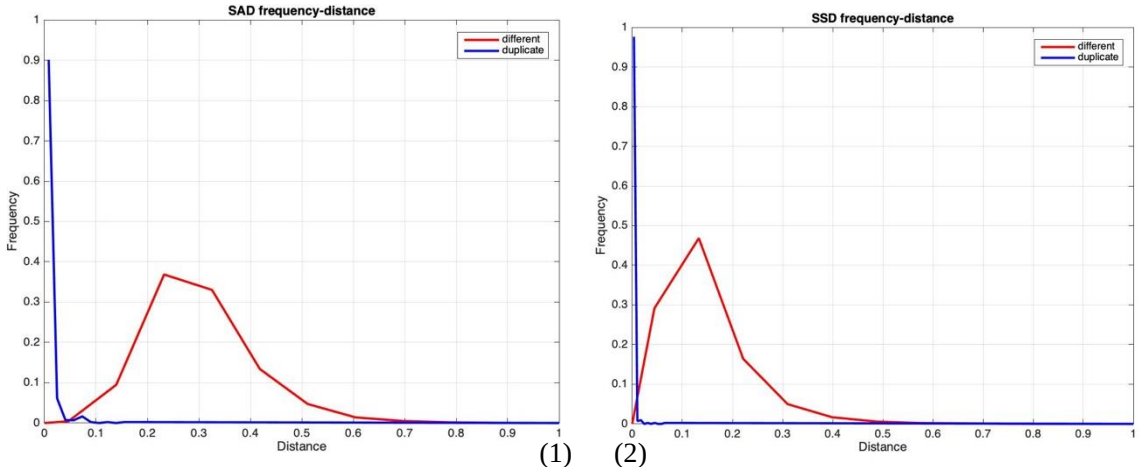


Figure 21: Algorithm flow map

2.4 Analysis of results

By performing these three image matching algorithms, I obtained the frequency-distance curves to describe the image distance and the frequency they appeared. In the figure, the red curve represents the distance between different images in the same

files, and the blue curve represents the distance between exact duplicate images in the same files. Among them, there are 1,304 samples of different images and 510 of the exact duplicate images. (Figure 22, Figure 23)



(1) (2)

Figure 22: 1) Frequency-Distance curves of SAD; 2) Frequency-Distance curves of SSD;

According to the three figures, we can find that the red and blue curves are separated when using NCC method. So, NCC's reliability is higher than the SAD and SSD methods. At the same time, when I use the SAD and SSD methods to detect exact duplicate images, some of the different images will be mistakenly identified as exact duplicate images during detection and some of the exact duplicate images will be mistakenly treated as different images because the different and duplicate curves of the both two methods have the overlapping area. Here are some error images in figure 24. Therefore, we need to set a distance threshold. For the SAD and SSD methods, when the two image distances are less than this threshold, the two images are exact duplicate images. When the two image distances are bigger than this threshold, the two images are different images. For the NCC method, when the two image correlations are less than this threshold, the two images are different images, and when the two image

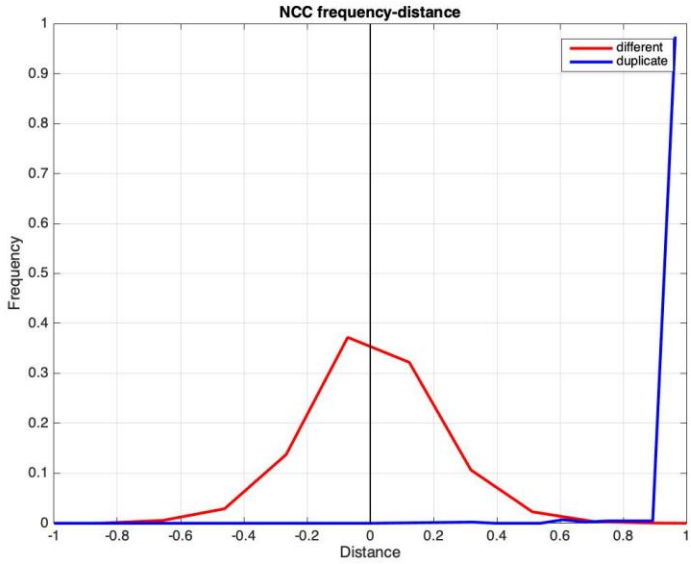


Figure 23: Frequency-Distance curves of NCC

distance correlations are bigger than this threshold, the two images are exact duplicate images. In order to show the accuracy of the three methods, I chose the precision-recall curves to indicate the accuracy of the three methods.

The P-R curve is often used for information retrieval. [11] It characterizes the re-

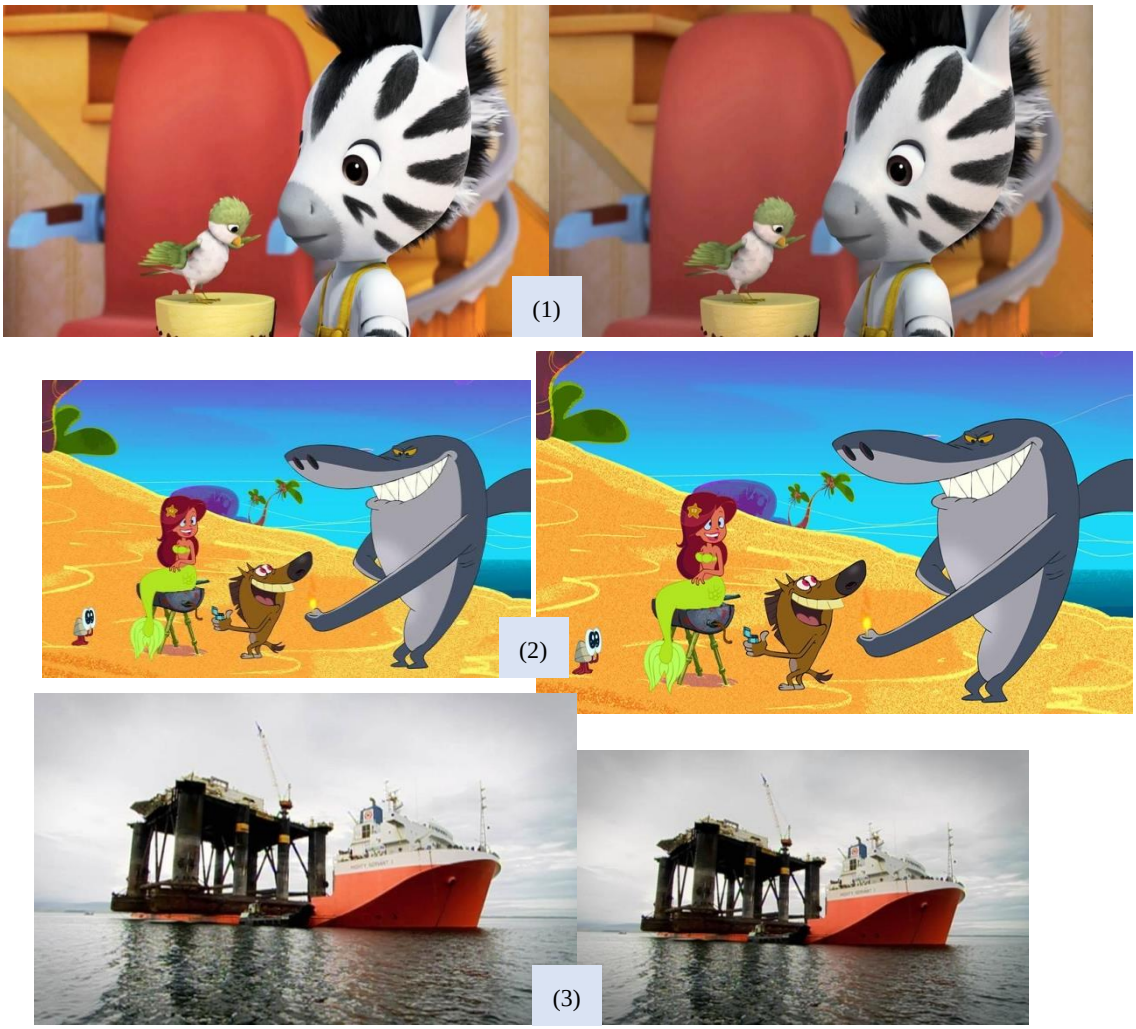


Figure 24: (1) Illumination changes; (2) Scaling; (3) Shifting

lationship between the precision and the recall rate. The precision refers to the proportion of true positive cases in all the data that are predicted as positive cases. The recall refers to the ratio of true positive predictions to all positive cases:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Among them:

TP (true positives) means the prediction is a positives sample, and the sample is actually positive;

FP (false positives) means the prediction is a positives sample, and the sample is actually negative;

TN (true negatives) means the prediction is a negatives sample, and the sample is actually negative;

FN (false negatives) means the prediction is a negatives sample, and the sample is actually positives;

The precision and recall are a pair of contradictory measures. Generally speaking, when the precision is high, the recall rate is often low. When the recall rate is high, the precision is often low. Therefore, when compared the matching algorithm, we can

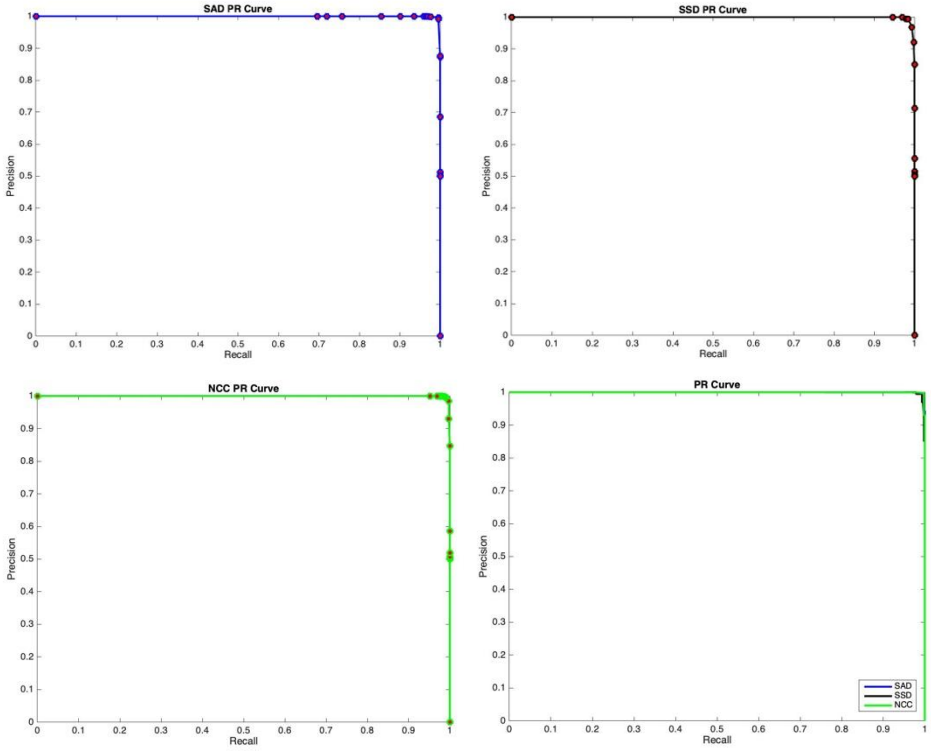


Figure 25: P-R curves of three methods and the comparation of them

sort the samples according to the prediction results of the matching algorithm. In the front of the sample is the positive example that the algorithm considers to be the most

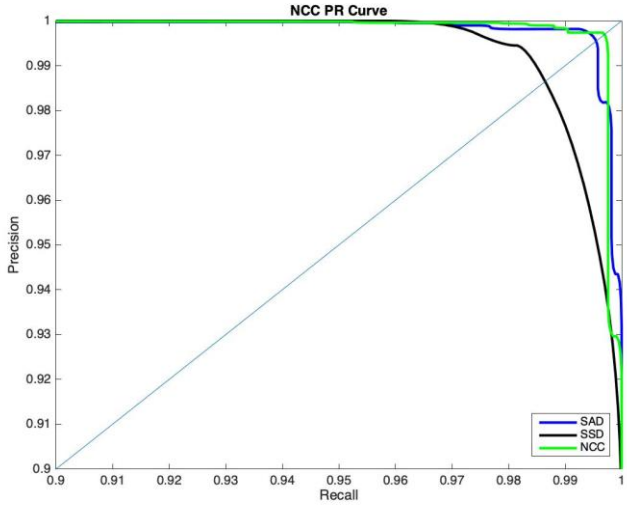


Figure 26: The detail of P-R curves

likely, and the one that follows is the sample that the algorithm considers to be the

least likely to be a positive example. Use the precision as the y-axis, and the recall as the x-axis. The P-R curve can be plotted. According to the property of precision and recall, the closer the P-R curve is to the upper right corner, the more accurate the algorithm.

Therefore, in order to show the precision of the three methods, I chose P-R curves to indicate the accuracy of the three methods. By changing the distance threshold, the P-R curves of the three methods at different thresholds were obtained. (Figure 25, Figure 26)

When using the P-R curve to judge the property of the matching method, if the P-R curve of one method is completely covered by the P-R curve of the other method, it can be concluded that the property of the latter is better than the former. When the method above cannot be applied, the Break Even Point and the F1 value can be used to judge the property of the balance. The Break Even Point (BEP) is the value when the precision is equal to the recall. If the value is larger, the property of the method is better. The F1 means F1-Score. [12] F-Score is a weight harmonic average of precision and recall. It is commonly used evaluation standard in the field of IR (Information Retrieval). Meanwhile, it is often used to evaluate the quality of classification models.

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Among them:

The P represents the precision;

The R represents the recall;

The β is a parameter;

When $\beta = 1$, it becomes $F_1 - Score$:

$$F_1 = \frac{2 \times PR}{P + R}$$

The precision, recall and the F-Score methods all concentrate on one class (positive examples). Recall is function of its correctly classified examples (true positives) and its misclassified examples (false negatives). Precision is a function of true positives and examples misclassified as positives (false positives). The $F - Score$ is evenly balanced when $\beta = 1$.[13] So I use $F_1 - Score$ to represent the result of the balance of the two parameters.

It can be seen from the figure 26 that the BEP of NCC is larger than SAD and SSD. And from the figure 27 that the $F_1 - Score$ of NCC is higher than SAD and SSD, indicating that the NCC's robust is the best. The maximum value of NCC's F-measure is 0.99749.

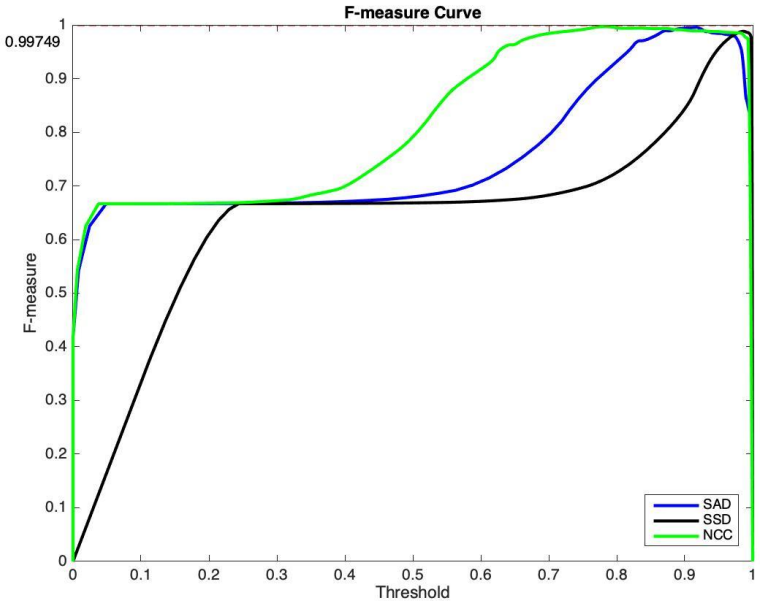


Figure 27: the F1-Score curves of three methods

2.5 Algorithm optimization and comparison

From the conclusions in the previous section, it can be seen that among the three matching methods of SAD, SSD and NCC, the robustness of NCC is the highest. To get a more optimized algorithm, I decided to use Control Variates Method to change the image matching method (above), change the image normalization size, and change the image color space to get the most optimal algorithm.

2.5.1 Change the normalized image size

In the above, I used the normalized size of 128×128. Here, I chose 64×64,

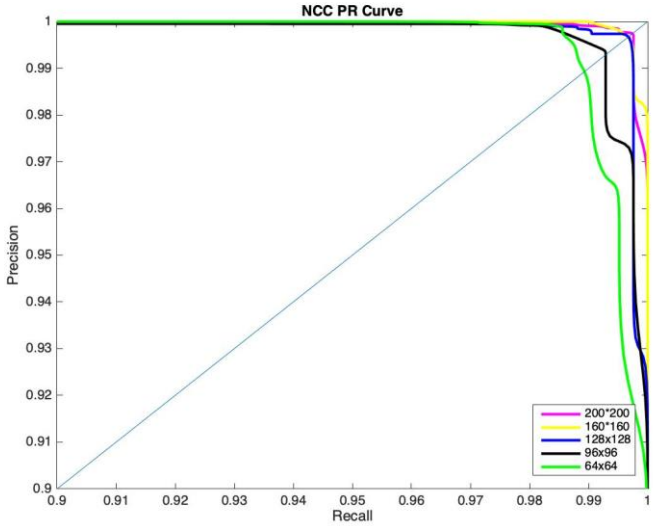


Figure 29: P-R curve for different sizes

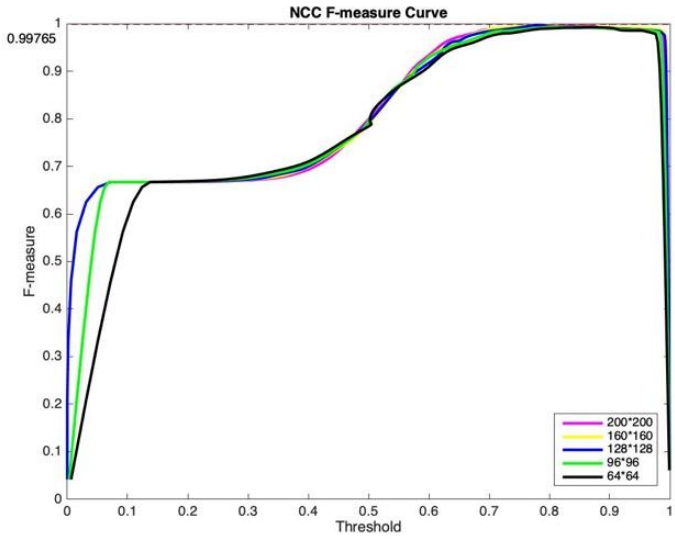


Figure 28: F-measure curve for different sizes

96×96, 160×160 and 200×200 as the normalized sizes. Then, I used the above three image matching methods to obtain the frequency-distance curves. Finally, I plotted the P-R curves and $F_1 - Score$ curves to compare the five sizes. (take NCC as an example) Finally, I use the Max F-measure-size figure to show how the maximum value of F-measure changes with normalized image size. (Figure 30)

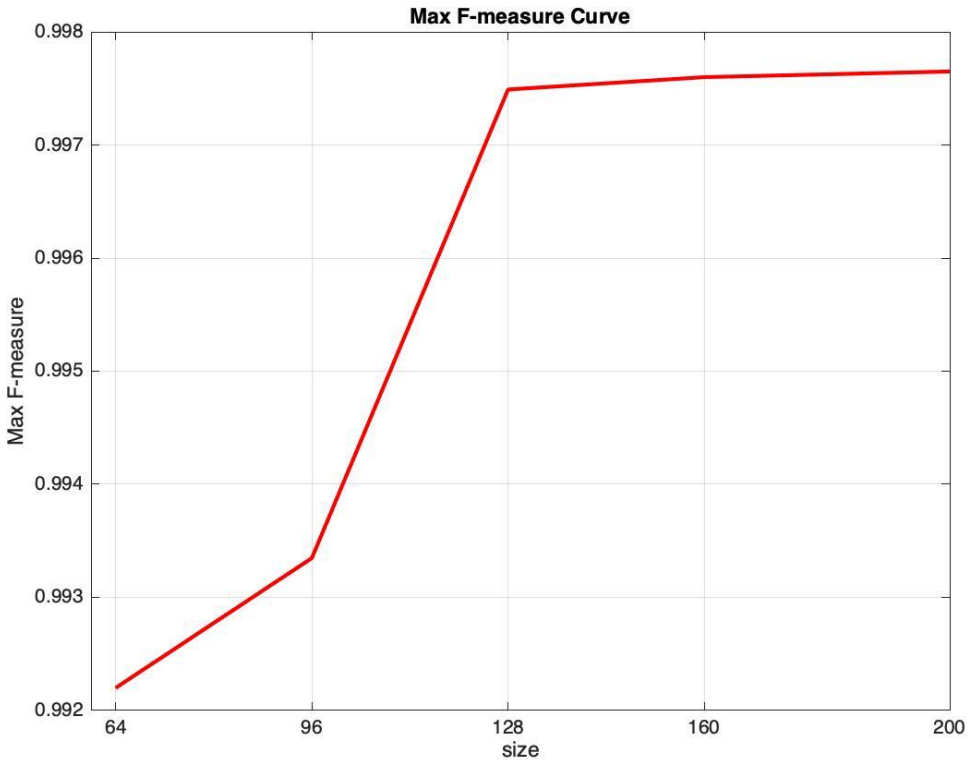


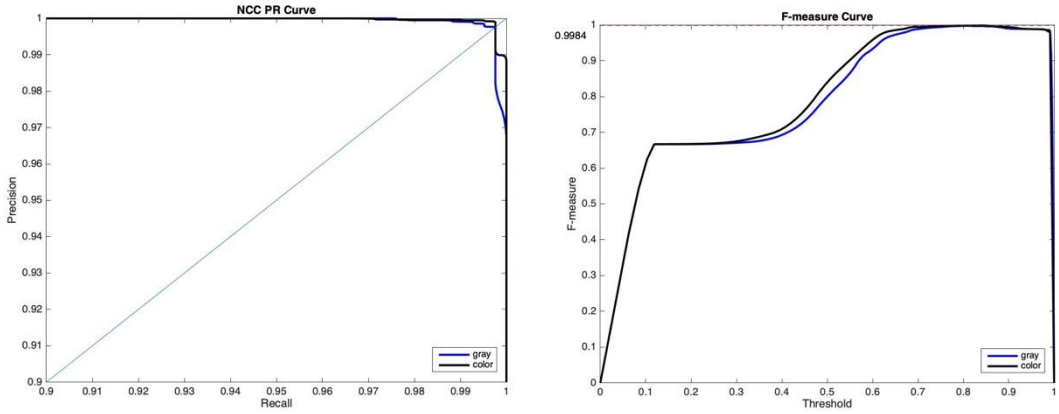
Figure 30: Max F-measure curve for different sizes

According to the figure 28 and figure 29, the algorithm with 200×200 as the normalized image size has the highest robustness, so that when the normalized size is larger, indicating that the more pixels are selected, and the performance of the algorithm is better.

2.5.2 Change the image to gray or color

In the above, I converted the RGB image to the grayscale image and performed the calculation between the pixels. Here, I decided to use the RGB image to test the above three image matching methods, and compare the results of the RGB image with the results of the grayscale image to obtain the relationship between image attributes and algorithm robustness. (taking NCC and 200×200 as an example)

As we can see from the above figure 31, the max F-measure is 0.99775 of the RGB space. So, the RGB image is more robust than the grayscale image algorithm. There are certain errors in the conversion of color images to grayscale images, which may cause variations in the distance between the pixels of the two images. Besides, we will lose the information of color after grayscale so the algorithm will ignore some color-related difference. However, using color image for detection will slow down the algorithm and increase the amount of computation. In conclusion, when we use color images for detection, the algorithm will be more reliable, but at the same time, the program running time will be slower.



(1) (2)

Figure 31: (1) P-R curve for different color spaces;
(2) F1-measure for different color spaces;

2.5.3 Best result

After the comparison above, I obtain the most reliable image matching method, image normalized size and image attribute. Therefore, I unify the RGB image to size 200×200, and calculate the distance between different images and the distance between duplicate images using the NCC image matching method. Then, I obtain the P-R curve and F-measure curve, as shown in figure 32 and figure 33. Among them, the maximum value of F-measure is 0.9984. It is the largest F-measure in all of the above cases. So, this algorithm is the most reliable one.

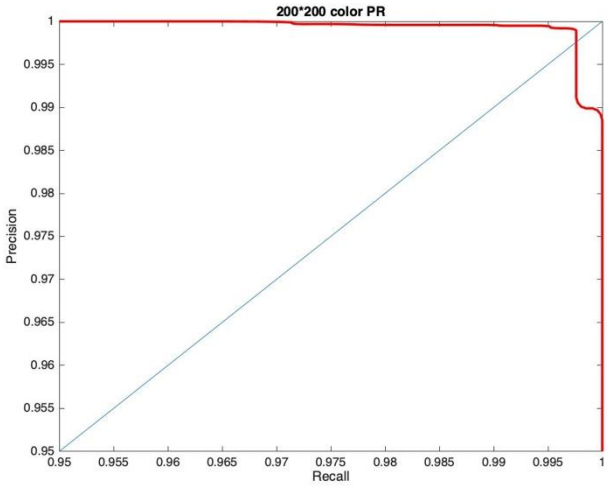


Figure 32: P-R curve for size 200*200 in color space using NCC

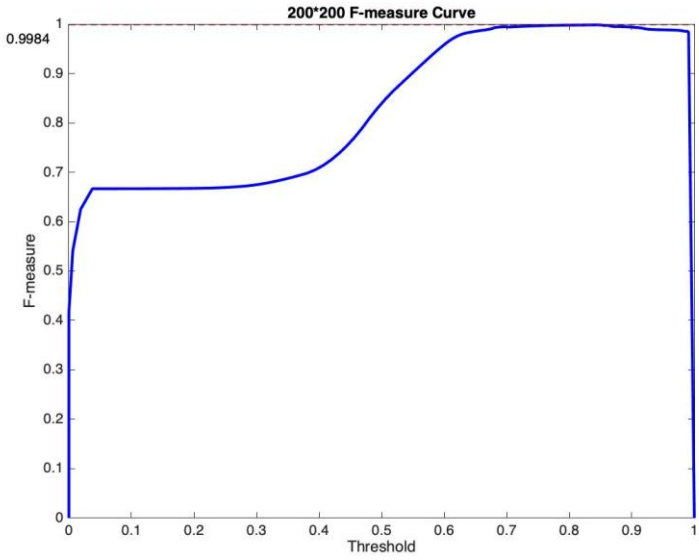


Figure 33: F-measure curve for size 200*200 in color space using NCC

Chapter 3

Conclusion

Image processing is very important in the Information Internet era. The detection of duplicate images is the application of image processing in the direction of image matching and image detecting.

In this article, I use image crawler to capture images from TV guide in more than 10 French video sites and store them as an image database. Then use Sum of Absolute Differences, Sum of Squared Differences, Normalized Cross Correlation image matching methods to calculate the distance between the images of the same TV program, and obtain the distance between the images under the same file, including duplicate image files and different image files. Finally, use these distances to get the Frequency-Distance curve.

From the above, the different and duplicate curves of NCC are separate, so it has 100% reliability. The different and duplicate curves in SAD and SSD are intersected, so when the image is judged using the SAD and SSD methods, the possibility of misclassification will be generated. After that, I use the Frequency-Distance curve to set a list of thresholds, calculate precision and recall at different thresholds, and then get the Precision-Recall curve to describe the reliability of the three matching methods. From the Precision-Recall curve, we can conclude that the NCC's PR curve is closer to the upper right corner and its Break-Even Point is larger. Therefore, the NCC has higher reliability and robustness. At the same time, I also use the F-measure curve to describe the reliability of the three methods and reach the same conclusion.

To get the more optimized algorithm, I chose 5 normalized sizes: 200×200, 160×160, 96×96 and 64×64, comparing the reliability of the algorithm when using these 5 sizes as the normalized image size. As we can see from the P-R curve and F-measure curve, when 200×200 is used as the normalized image size, the reliability of the algorithm becomes higher. In addition, I compared the reliability of the algorithm in RGB space and grayscale space. When the image is not converted to grayscale image, which means image still keep in RGB space, the algorithm will be more reliable.

From what has been discussed above, we may safely draw the conclusion that when exact duplicate image detection is performed on image in RGB space, using NCC as the image matching method and choosing a larger normalized size can obtain a higher robust algorithm.

In this report, the number of images I used to detect is 1,814. It’ s pretty small. Besides, I classify them by hand to obtain the image database for detection. The

duplicate images I can precisely detect are only compression and scaling. When there is illumination change, shifting and cropping between images, the reliability of my algorithm will be drop. Besides, I only test the five normalized sizes but not compare the running time of them, as we all know, if the size of images become bigger, the running time will be longer. So, more work can be done in the future.

In the future, we can further expand the database and detect the duplicate or different automatically. Besides, we can use more image matching methods such as scale invariant template matching to detect the duplicate of shifting and cropping. And for illumination change, we can use the contrast normalization to obtain the images that suit for my algorithm. What's more, I will record the program running time when changing the normalized size, and get the best size according to the proportional relationship between the algorithm optimization degree and the running time changes. Then we can find the most robust algorithm for exact duplicate image detection.

Reference

[1] Rafael C.Gonzalez, Richard E.Woods. Digital Image Processing Third Edition. London :Pearson Education, 2008:1-3
[2] StanZ.Li. Handbook of face recognition: Library of Congress Cataloging-in-Publication Data, 2004:52-53
[3] High-Confidence Near Duplicate Image Detection; Wei dong, Zhe wang, Moses Charikar, Kai Li
[4] Ond rej Chum, James Philbin, Andrew Zisserman, Near Duplicate Image Detection: min-Hash and tf-idf Weighting: BMVC: 2008
[5] D.I Barnea, H.F Silverman, A class of algorithms for fast digital image registration, IEEE: 1972
[6] Bhavika K. Desai, DR. M. B. Potdar, Manoj Pandya, Manish P. Patel, Paru Thakkar, Template Matching Technique using Enhanced SAD Technique: International Journal of Engineering Research & Technology, 2014:1371
[7] Bo Yi, Ping Yu, Guo-zhu He, Jing Chen, A Fast Matching Algorithm with Feature Points Based on NCC, International Academic Workshop on Social Science, 2013:955-958
[8] Hitoshi Hishiguchi, Yoshihiko Nomura, A study on SSD calculation between input image and subpixel-translated template images and its applications to a subpixel image matching problem, the international society for optical engineering, 2009
[9] Badrul hisham Mohamed, Shahrul Nizam Yaakob, Rafikha Aliana A.Raof, Template Matching using Sum of Squared Difference and Normalized Cross Correlation, IEEE: 2015
[10] Rafael C.Gonzalez, Richard E.Woods, Steven L.Eddins. Digital Image Processing using MATLAB. London :Pearson Education, 2004:4-5
[11] Jesse Davis, Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. ICML '06 Proceedings of the 23rd international conference on Machine learning:233-240
[12] Peter A.Flach, Meelis Kull. Precision-Recall-Gain Curves: PR Analysis Done Right, Advances in Neural Information Processing Systems 28 (NIPS 2015)
[13] Marina Sokolova, Nathalie Japkowica, Stan Szpakowicz. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence: 1015-1021