

# Visual Data Mining, Techniques and Tools for Data Visualization and Mining: Multimedia Datasets Case Study

Josiane Kayitesi and Papias Niyigena

*University of Lay Adventists of Kigali*

Submitted: 10-03-2022

Revised: 21-03-2022

Accepted: 23-03-2022

## ABSTRACT

Visual data processing and Machine Learning mechanically generate insights that alter you to spot the foremost common variables across all models, the foremost vital variables chosen across models, and assessment results for all models. Data visualization on images does not always need training machine learning algorithms to a closely related set of images. The current study proposed an approach to image mining that uses visual analytics under Python environment using NumPy, Matplotlib, and OpenCV libraries to combine interactive visualizations and automated data analysis without including machine learning. Three features, “Pixels analysis”, “Fourier transformation”, and then “Edge detection”, were studied and the outcomes are of that the intensity of red is an indication of the skin data point in the pixel, the intensity of the blue indicates a background in the image and the intensity of the green indicates the person's clothing in the image. The main goal behind detecting edges from an image is to visualize important events, trends, and changes in properties of the image. It is one of the basic steps in image processing, image patterns recognition, and computer vision techniques.

## I. INTRODUCTION

Whether you work as a data analyst, run a data-centric startup or want to learn more about data mining as a whole, visual data mining can be an interesting place to start (Tank, 2019). For data mining to be effective, it is important to include the human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today's computers (Daniel, A. K. and Mike, S., 2004).

The data analyst typically specifies first some parameters to restrict the search space. Visual Data Mining and Machine Learning automatically generates insights that enable you to identify the

most common variables across all models, the most important variables selected across models, and assessment results for all models (SAS, 2020).

Visual data mining discovers implicit and useful knowledge from large data sets using data and/or knowledge visualization techniques (Jiawei, H. and Jian, P., 2012). Data mining and data visualization is gradually becoming an indispensable organizational weapon for achieving competitive advantage in many data-driven industries (Yeh, 2015). Presenting data in an interactive, graphical form often fosters new insights, encouraging the formation and validation of new hypotheses to the end of better problem-solving and gaining deeper domain knowledge (Daniel, A. K. and Mike, S., 2004).

Data visualization provides an important suite of tools for identifying a qualitative understanding. This can be helpful when trying to explore the dataset and extract some information to know about a dataset and can help with identifying patterns, corrupt data, outliers, and much more (Aashi, 2021). Visualization is the key process in data mining; Visualization technique can provide a clear and more detailed view on different aspect of data as well as results of automated mining algorithms. Nowadays, the challenge is to mine data in the multimedia objects such that in one way is to extract relational data from multimedia data as an object databases directly (Karthikeyan, C. and Rajaraman, A., 2019).

It is obvious that the influence and significance of data visualization cannot be underestimated. Although data visualization has proven to be a powerful tool in data mining and knowledge discovery, its use in business and corporate world is still new and fledgling (Yeh, 2015).

This paper aims at briefly review related works produced previously. Next, we will present a multimedia dataset and use of Python libraries as visualization tools focusing on new analytical

visualization techniques like NumPy, OpenCV, and Matplotlib.

## II. RELATED WORKS

There are various visualization approaches for multimedia dataset mining over time in the literature, below are some of them reviewed for this paper objectives achievement.

In their paper, (Karthikeyan, C. and Rajaraman, A., 2019), cited that Visual data mining is a novel approach to data mining. The aim of their work was to combine traditional data mining algorithms with information visualization techniques to utilize the advantages of both approaches. The utilization of both automatic analysis methods and human perception/understanding promises better and more effective data exploration. Visualization is a key process in visual data mining and the focus is on the presentation of all aspect of multimedia objects, their identification, their analysis and relationships.

The second, (Vijayarani, S. and Sakila, A., 2015), researchers indicated that the basic concepts of multimedia mining and its essential characteristics. Multimedia mining architectures for structured and unstructured data, research issues in multimedia mining, data mining models used for multimedia mining and applications are also discussed in their paper to help the future researchers to get the knowledge about how to do their research in the field of multimedia mining.

The third work, the researchers (Chen, Y., 2014) proposed a new framework of visual mining of multimedia data and the key idea is to integrate data visualization and data mining. Overall, their visualization tool, allows users to not only easily examine and synthesize information into new ideas and hypotheses, but also quickly quantify and test the insights gained from visualization; the next step is to conduct a systematical evaluation of our prototype system. They used the experimental paradigm to test what kinds of new findings researchers can obtain by using that tool. The presented works are all related to data mining, the current research is about to emphasize the conducted researches and go deeper into the visual data mining techniques and tools for multimedia dataset mining.

## III. METHODOLOGIES

To carry out a scientific and analytical research such as this, we base on responding the research questions which also drives researchers to achieve the goals of research. It is in that regards, this works aims at briefly review related works produced previously. Next, we will present a

multimedia dataset and use of Python libraries as visualization tools focusing on new analytical visualization techniques like NumPy, OpenCV, and Matplotlib.

### 3.1 Research Design

For many studies, papers are the exceptional method for obtaining insights which are obvious in multimedia data, especially images. In this work, we are planning to enlighten the techniques of image visualization and help researchers on how the images can be readable with different ways to come up with its meaning in data mining concepts. A set of techniques, will be used to see how images can offers us various messages; we will use thoughts from previous publishers as secondary data after examine them to get new designs to fulfill the studies objectives.

### 3.2 Research approaches

We are learning image processing using Python with two libraries, Matplotlib, OpenCV and NumPy, for “Pixels analysis”, “Fourier transformation”, and then “Edge detection” features extraction. We will use these images manipulation techniques throughout the remainder of the study to achieve the goal of our research.

### 3.3 Features to extract

#### Pixels analysis

Computers store images as a mosaic of tiny squares, if these square tiles are too big, it's then hard to make smooth edges and curves. Vector graphics are a somewhat different method of storing images that aims to avoid pixel related issues. A simple way to describe each pixel is using a combination of three colors, namely Red, Green, and Blue. This is what we call an RGB image (Mohammed, 2018).

#### Fourier Transformation

Some images contain systematic noise that users may want to remove; if such noise is regular enough, employing Fourier Transformation adjustments may aid in image processing (Tonichi, 2021). This is what we are going to prove in this study.

#### Edge detection

Edge detection is an image processing technique for finding the boundaries of objects within images where mainly works by detecting discontinuities in brightness. One of the most popular and widely used algorithm is canny edge detector and this is the one to be used in this paper experimentation (Abdou, 2020).

#### IV. RESULTS AND ANALYSIS

This section presents the study experimentations, their results, and discuss the outcomes to get which of these techniques and tools for data visualization is efficient in visual data mining.

##### 4.1 Dataset loading

A dataset, with 3725 images of people with face masks, is used within this study to analyze the images by studying their “Pixels”, “Fourier transformation”, and then “Edge detection”. An image was selected to be used as sample image. We are visualizing their features, insights and message from it as listed in the previous sections.

##### 4.2 Features extraction

While extracting features from the selected images of the dataset, Python environment with both Matplotlib, OpenCV, and NumPy libraries are chosen to help us coming up with the expected outcomes.

##### 4.2.1 Experimentation with Python

###### a. Import libraries

As listed above, Python libraries for image visualization like Matplotlib, OpenCV, and NumPy are used in this experimentation and must be loaded into the Python environment for being used. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible (Matplotlib, 2022). NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python (W3Schools, 2022).

While, OpenCV (Open Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez (which was later acquired by Intel). The library is cross-platform and free for use under the open-source Apache 2 License (Wikipedia, 2012).

###### b. Image loading and processing

After loading the libraries, we loaded image and run the experimentation with it to visualize the features as expected.



Figure 1 - Image presented in x, y plane

After the image is loaded the next step is to inspect its type, shape, and size; Type of the image : <class 'imageio.core.util.Array'>, Shape of the image: (1500, 1500, 3), Image Height 1500, Image Width 1500, Dimension of Image 3. Image size 6750000bits, Maximum RGB value in this image 255, and the Minimum RGB value in this image 0.

Our image has three channels of colors: Red, Green, and Blue; now let's take a quick view of each channel in the whole image.

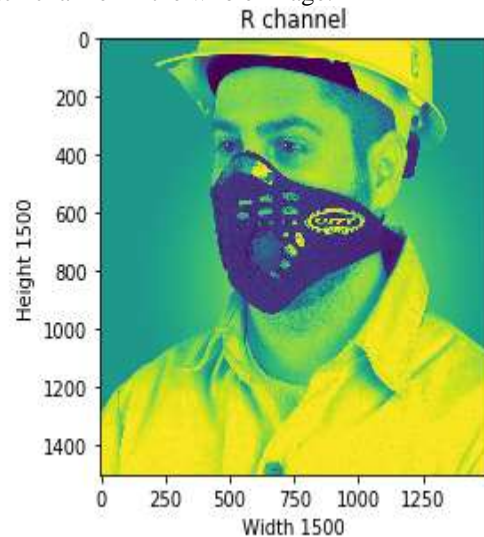


Figure 2 - Red channel image

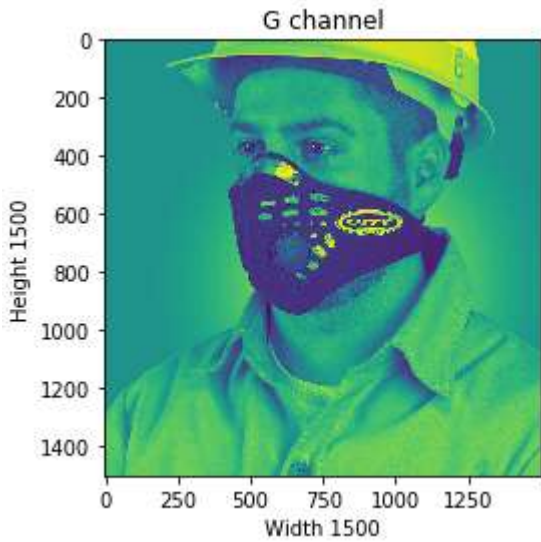


Figure 3 - Green channel image

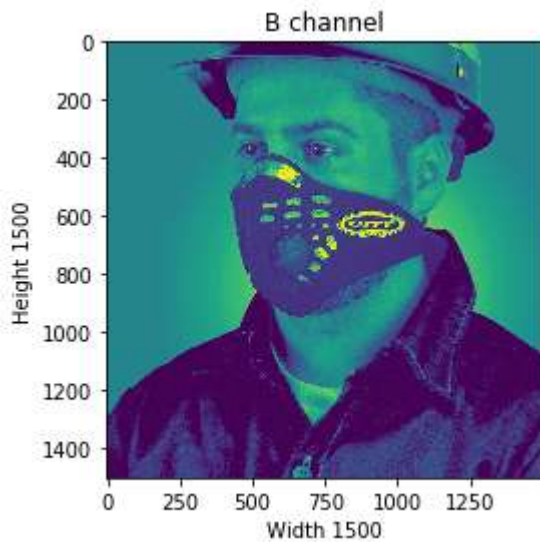


Figure 4 - Blue channel image

For the sake of clarity, it's reasonable to assume that color images would require three full color bands, but it turns out a system can produce high quality color images and then apply specialized image processing algorithms to the resulting data to thoroughly analyze the image.

Here we have something interesting, the colors in each RGB plane mean something. For example, the intensity of red is an indication of the skin data point in the pixel, the intensity of the blue indicates a background in the image and the intensity of the green indicates the person's clothing in the image. Just by looking at the following colorful images, a trained eye can tell what is the skin, what is the background, and what is the clothing.

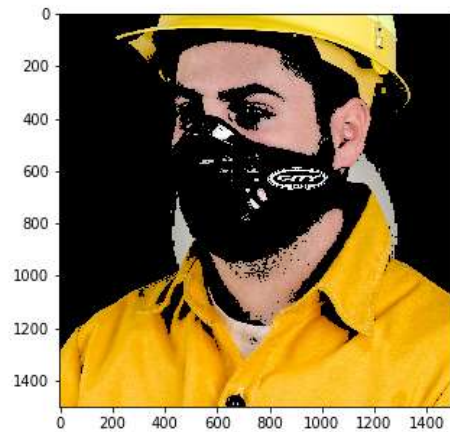


Figure 5 - Only Red Pixel value, higher than 180

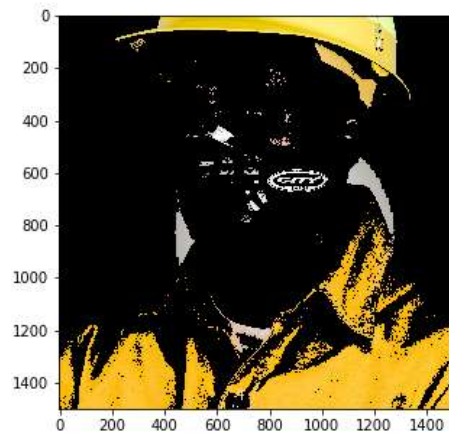


Figure 6 - Only Green Pixel value, higher than 180

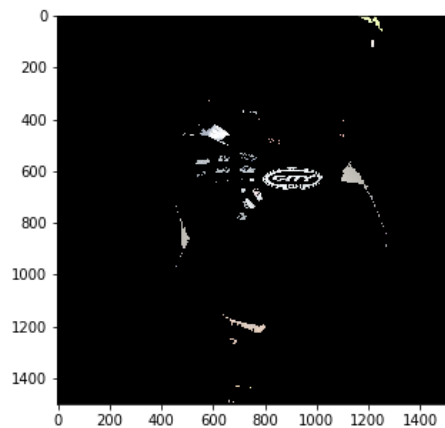


Figure 7 - Display only Green Pixel value, higher than 180

After the pixels analysis, the next step in our experimentations was about Fourier transformation for noise inspection and removal for the current image. We are going to convert our image into a grayscale image, remember as seen



above, some color image formats are 3D arrays as they have a red, green, and blue layers.



Figure 8 - Converted to grayscale image

A Fourier transformation for our image is calculated and two images are displayed as following; at the center of the Fourier transform, the constant term is brighter than the remaining points of the image.



Figure 9 - Fourier transformation image

As well as the dots are far away from the center, the higher is the frequency and the brighter these dots are, the more prominent that grating is in the image and confirms its higher amplitude.



Figure 10 - Smudged image

We can see that decreasing the worth has nearly no result on the first image, however, increasing the worth appears to darken the above image. Finally, allow us to indorse the Fourier Transformation adjustment whereas holding the colors of the first image as it is shown here below.



Figure 11 - Transformed image

In this way in visual data mining, the Fourier Transformation is an important tool that can be used to minimize the noise in images. This experience showcases how we can make understated changes to an image via Fourier Transformation.

The last feature to be extracted from our image was the edge detection. Also, we started by loading the image, converting it into the grayscale image after setting the image size appropriate for visualization of the image. The final step for edge detection is applying the Canny Algorithm on the image as converted into grayscale, and from this step we have the following output image.



Figure 12 - Edge detection image

The main goal behind detecting edges from an image is to visualize important events, trends, and changes in properties of the image. It is one of the basic steps in image processing, image patterns recognition, and computer vision techniques. If you see clearly the image above, we have the lines that are defining the limitation of objects that constitute our image.

## V. CONCLUSION

Furthermore, to increase the standardization of data-mining methods, especially visual data mining, as well as novel methods capable of addressing unstructured multimedia data, such as graphics, audio, video, and images grouped into a dataset. Data-mining technology can improve user perception towards images data, detect colors and edges, and then analyze the Fourier transformation of the images to remove the noise.

In this study, an approach for visual data mining using NumPy, Matplotlib, and OpenCV libraries from Python which combines cooperating visualizations and automated data analysis without including machine learning was proposed. With this study, three image features were targeted, "Pixels analysis", "Fourier transformation", and then "Edge detection", and the results from Pixel analysis are of that the intensity of red color channel is an indication of the skin data point in the pixel, the intensity of the blue color channel indicates a background in the image and the intensity of the green color channel indicates the person's clothing in the image. On the edge detection, it is shown that detection of edges is one of the basic steps in image processing, image patterns recognition, and computer vision techniques main goal behind detecting edges from

an image is to visualize important events, trends, and changes in properties of the image and for the current image in this study you see clearly the image above, we have the lines that are defining the limitation of objects that constitute our image.

## REFERENCES

- [1]. Aashi, G. (2021, June 7). Must Known Data Visualization Techniques for Data Science. Retrieved from Analytics Vidhya: <https://www.analyticsvidhya.com/blog/2021/06/must-known-data-visualization-techniques-for-data-science/>
- [2]. Abdou, R. (2020, Oct). How to Perform Edge Detection in Python using OpenCV. Retrieved from ThePythonCode.com: <https://www.thepythoncode.com/article/canny-edge-detection-opencv-python>
- [3]. Chen, Y. (2014). Visual Mining of Multimedia Data for Social and Behavioral Studies. Fujian Agriculture and Forestry University, 1 - 8.
- [4]. Daniel, A. K. and Mike, S. (2004). Visual Data-Mining Techniques. Visualization Handbook. Amsterdam:Elsevier, 831-843.
- [5]. Jiawei, H. and Jian, P. (2012). 13 - Data Mining Trends and Research Frontiers. Retrieved from Science Direct: <https://www.sciencedirect.com/topics/computer-science/visual-data-mining>
- [6]. Karthikeyan, C. and Rajaraman, A. (2019). Visual Data Mining for Multimedia databases. International Journal of Recent Technology and Engineering, 277 - 279.
- [7]. Matplotlib. (2022, March 22). Matplotlib: Visualization with Python. Retrieved from Matplotlib.com: <https://matplotlib.org/#:~:text=Matplotlib%20is%20a%20comprehensive%20library,easy%20and%20hard%20things%20possible.>
- [8]. Mohammed, I. (2018, June 26). Basic Image Data Analysis Using Python: Part 1. Retrieved from Dzone.com: <https://dzone.com/articles/image-data-analysis-using-numpy-amp-opencv>
- [9]. SAS. (2020). SAS® VISUAL DATA MINING AND MACHINE LEARNING. Retrieved from SAS: [https://www.sas.com/en\\_us/software/visual-data-mining-machine-learning.html](https://www.sas.com/en_us/software/visual-data-mining-machine-learning.html)
- [10]. Tank, P. (2019, 11 22). What is Visual Data Mining? Retrieved from Planning Tank: <https://planningtank.com/computer-applications/visual-data-mining>
- [11]. Tonichi, E. (2021, Jan 28). Image Processing with Python — Application of Fourier

- Transformation. Retrieved from towardsdatascience.com:  
<https://towardsdatascience.com/image-processing-with-python-application-of-fourier-transformation-5a8584dc175b>
- [12]. Vijayarani, S. and Sakila, A. (2015). MULTIMEDIA MINING RESEARCH – AN OVERVIEW. International Journal of Computer Graphics & Animation (IJCGA), 69 - 77.
- [13]. W3Schools. (2022, March 12). NumPy Introduction. Retrieved from W3Schools.com:  
[https://www.w3schools.com/python/numpy/numpy\\_intro.asp#:~:text=NumPy%20is%20a%20Python%20library,NumPy%20stands%20for%20Numerical%20Python](https://www.w3schools.com/python/numpy/numpy_intro.asp#:~:text=NumPy%20is%20a%20Python%20library,NumPy%20stands%20for%20Numerical%20Python).
- [14]. Wikipedia. (2012, November). OpenCV. Retrieved from Wikipedia.org:  
<https://en.wikipedia.org/wiki/OpenCV>
- [15]. Yeh, R. K. (2015). Visualization Techniques for Data Mining in Business Context: A Comparative Analysis . University of Texas at Arlington, 310 - 320.