UNIVERSITY of Prince Edward ISLAND

CAIRO Campus

Data Science

Extracting Bag-of-Visual-Words for Hybrid Optical Character Recognition

---

Authors' Name	Student ID
Islam Anwar	201900337
Karim Elbowety	201800392

Faculty of Science & Innovation,
The Universities of Canada

May 22, 2022

Abstract

This paper showcases innovative methodology at extracting and clustering a bag of visual words from text images. The main purpose of this paper is to present the methodology applied at extracting a series of text characters from text images as well as using various clustering techniques to separate them. Several machine learning strategies are employed such as image thresholding, extracting connected components from images using an image processing tool, individual component extraction and more. We go into further detail regarding topics such as the hyperparameter optimization of clustering techniques and feature extraction methods employed. Performances indicate that applying PCA with agglomerative clustering visually produces tighter clusters with the same character or symbol.

Contents

1 Approach

1.1 Chosen Images

Prior to performing clustering, 3 images were to be chosen. Every image is of a different language and is composed of a multitude of unique characters and symbols. The 3 languages we chose were English, Arabic, and Russian. Figures (3-5) show the contents of the chosen documents.

1.2 Image Properties

Upon reading the images, we converted the pixel colours to gray scale to reduce image noise and because we will then apply image thresholding. Thresholding is the simplest method of image segmentation. From a gray scale image, thresholding can be used to create binary images [Guruprasad, 2020]. An example of a gray scaled image is in figure (1), as well as figures (6, 7) in the appendix.

After applying the threshold, we insert it into an method that identifies the connected components of the image using OpenCV. This returns statistics about the connected components as well as the number of labels within the image. Examples are in figures (2, 8, 9).

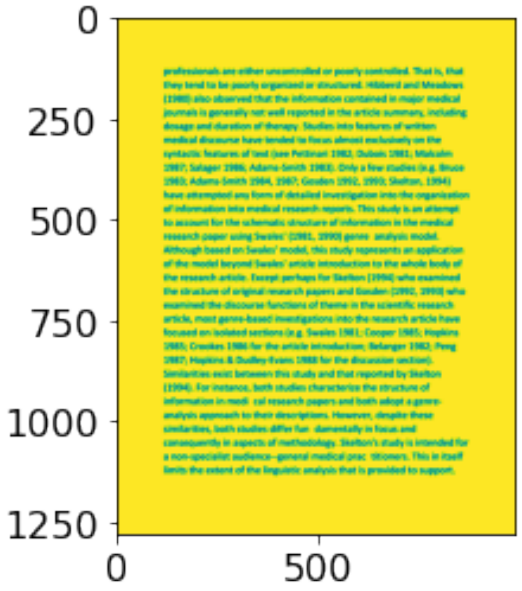


Figure 1: English Grayscale

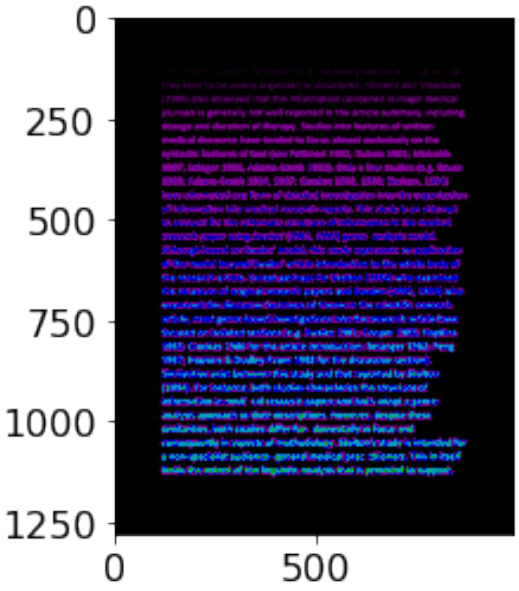


Figure 2: English Labels

1.3 Component Extraction

After component identification, we perform component extraction to distinguish between unique characters in a set language. This was done by iterating through the collection of labels we obtained the previous step. Furthermore, we went through every single pixel in the image and used its corresponding stats measure we also identified to separate the connected components into separated components. Once they we're all separated, they we're added to a list containing all possible components.

1.4 Resizing

Concerning resizing, we are performing resizing to maintain image size consistency before feeding it into the cluster as well as the feature extraction process. Moreover, we resized by padding to minimize the loss of information from interfering with the image's original aspect ratio.

1.5 Feature Extraction

Regarding the feature extraction, we opted to try two feature extraction techniques:

- HOG (Histogram of Oriented Gradients)
- PCA (Principal Component Analysis)

1. Histogram of Oriented Gradients:
HOG is a feature descriptor that is often used in computer vision to identify features in an image. It partitions the image into smaller portions where it can calculate the orientation and gradient of each portion and represent them in the form of histograms.

2. Principal Component Analysis: PCA is one of the most commonly used feature extractors as it can be used to identify the most important features within a message while preserving a certain threshold of information. However, rather than choosing the specific number of principal components, we specified how much information we want to be retrieved post-extraction. Our best case was when we applied the hyperparameter of PCA to retrieve 99% of the information.

Figures (10, 11) in the appendix showcase examples of applying HOG to extract the orientation and gradients of sub-portions of an image as well as PCA to extract the most significant features while retaining valuable information.

1.6 Clustering & Evaluation

After applying feature extraction, we used clustimage, which is a computer vision package dedicated to detecting natural groups or clusters of images. The clustering techniques we deployed are DBSCAN and Agglomerative clustering. DBSCAN seemed like an intriguing model as it relies on similarly dense data and is robust to outliers. Agglomerative clustering was another promising prospect as the number of clusters was unknown to us. Therefore, by using agglomerative clustering we had flexibility regarding the variability of the number of clusters throughout each language.

For hyperparameter optimization, the table below highlights the chosen hyperparameters for PCA and Agglomerative clustering which would end up being our most effective combination after exhaustive testing between PCA, HOG, DBSCAN, and Agglomerative.

Technique	Hyperparameter	Value
PCA	N Components	0.99%
HOG	Orientation	8
	Pixels Per Cell	(2,2)
	Cells Per Block	(1,1)
DBSCAN	Eps	0.5
	Min Samples	5
	Cluster Space	Low
Agglomerative	Metric	Euclidean
	Min Clusters	20
	Max Clusters	80
	Cluster Space	High

Our evaluation metric is the silhouette score. The silhouette value is a measure of how similar a sample is to its own cluster in comparison to other clusters. Accordingly, the higher the score the more a sample resembles its cluster correctly.

1.7 Results

As shown in Figures (12-14), we managed to cluster the English and Russian languages somewhat effectively. However, Arabic has proven to be a much more difficult challenge as the characters as not as separable as in English and Russian. The grammatical structure of Arabic makes it more difficult to separate the connected components into singular symbols.

2 Challenges & Solutions

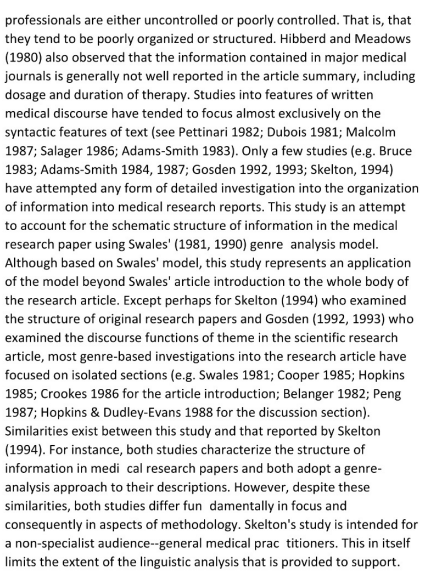
2.1 Challenges

We faced several challenges during this project. This includes the difficulty of getting our extracted components as separate characters, as well as the problem of the image background. The image background is considered the first component of any image, therefore we had to account for that issue. Furthermore, resizing our images without significant loss of information was another problem we needed to mitigate. However, the most frustrating aspect was finding a similarity measure to compare the components before clustering them.

2.2 Solution to Challenges

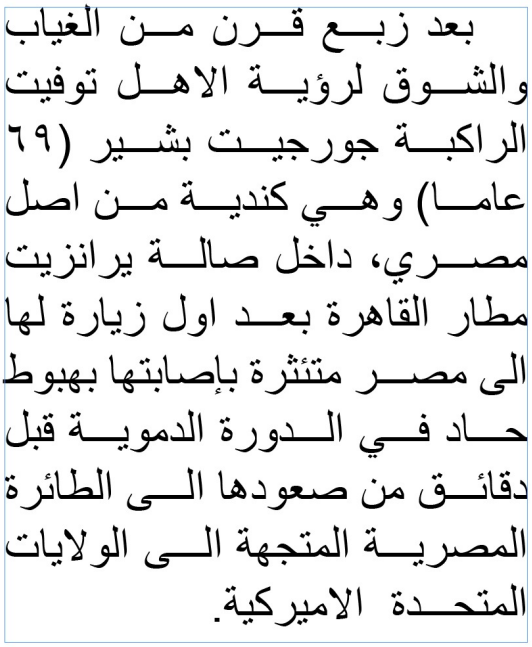
We attempted to solve as many of our problems as possible. We separated our extracted components using the labels that we obtained from the image processing tool OpenCV, with which we operated. In addition to that, applying padding resizing was integral to avoid losing information from our components. Regarding our similarity measures, we used HOG as well as PCA to attempt to extract the most meaningful features while retaining information. It provided a strong base by which we were able to then cluster our extracted features.

3 Appendix



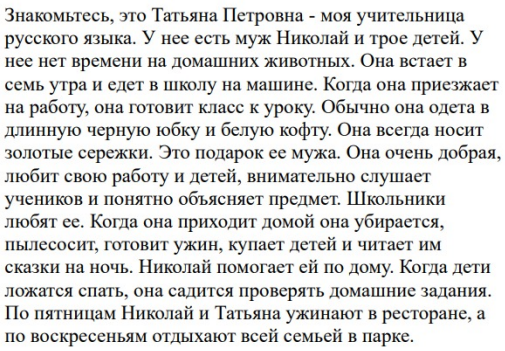
professionals are either uncontrolled or poorly controlled. That is, that they tend to be poorly organized or structured. Hibberd and Meadows (1980) also observed that the information contained in major medical journals is generally not well reported in the article summary, including dosage and duration of therapy. Studies into features of written medical discourse have tended to focus almost exclusively on the syntactic features of text (see Pettinari 1982; Dubois 1981; Malcolm 1987; Salager 1986; Adams-Smith 1983). Only a few studies (e.g. Bruce 1983; Adams-Smith 1984, 1987; Gosden 1992, 1993; Skelton, 1994) have attempted any form of detailed investigation into the organization of information into medical research reports. This study is an attempt to account for the schematic structure of information in the medical research paper using Swales' (1981, 1990) genre analysis model. Although based on Swales' model, this study represents an application of the model beyond Swales' article introduction to the whole body of the research article. Except perhaps for Skelton (1994) who examined the structure of original research papers and Gosden (1992, 1993) who examined the discourse functions of theme in the scientific research article, most genre-based investigations into the research article have focused on isolated sections (e.g. Swales 1981; Cooper 1985; Hopkins 1985; Crookes 1986 for the article introduction; Belanger 1982; Peng 1987; Hopkins & Dudley-Evans 1988 for the discussion section). Similarities exist between this study and that reported by Skelton (1994). For instance, both studies characterize the structure of information in medi cal research papers and both adopt a genre-analysis approach to their descriptions. However, despite these similarities, both studies differ fun damentally in focus and consequently in aspects of methodology. Skelton's study is intended for a non-specialist audience--general medical prac titioners. This in itself limits the extent of the linguistic analysis that is provided to support.

Figure 3: English Document



بعد زبـــع قـــرن مـــن الغياب والشـــوق لرؤيـــة الاهـــل توفيت الراكبـــة جورجيـــت بشـــير (٦٩ عامـــا) وهـــي كنديـــة مـــن اصل مصـــري، داخل صالـــة يرانزيت مطار القاهرة بعـــد اول زيارة لها الى مصـــر متئثرة بإصابتها بهبوط حـــاد فـــي الـــدورة الدمويـــة قبل دقائـــق من صعودها الـــى الطائرة المصريـــة المتجهة الـــى الولايات المتحـــدة الاميركية.

Figure 4: Arabic Document



Знакомьтесь, это Татьяна Петровна - моя учительница русского языка. У нее есть муж Николай и трое детей. У нее нет времени на домашних животных. Она встает в семь утра и едет в школу на машине. Когда она приезжает на работу, она готовит класс к уроку. Обычно она одета в длинную черную юбку и белую кофту. Она всегда носит золотые сережки. Это подарок ее мужа. Она очень добрая, любит свою работу и детей, внимательно слушает учеников и понятно объясняет предмет. Школьники любят ее. Когда она приходит домой она убирается, пылесосит, готовит ужин, купает детей и читает им сказки на ночь. Николай помогает ей по дому. Когда дети ложатся спать, она садится проверять домашние задания. По пятницам Николай и Татьяна ужинают в ресторане, а по воскресеньям отдыхают всей семьей в парке.

Figure 5: Russian Document

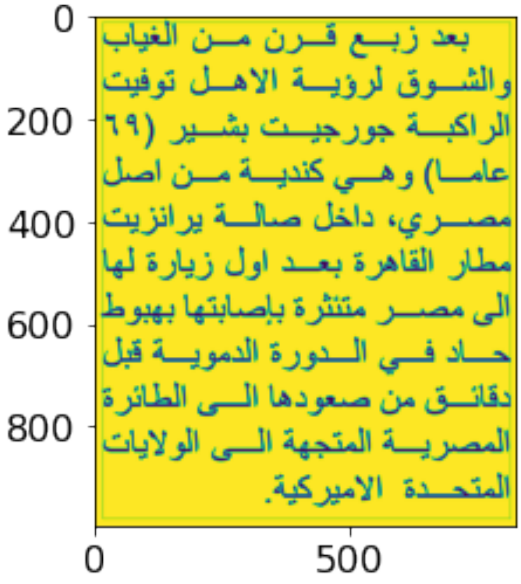


Figure 6: Arabic Grayscale

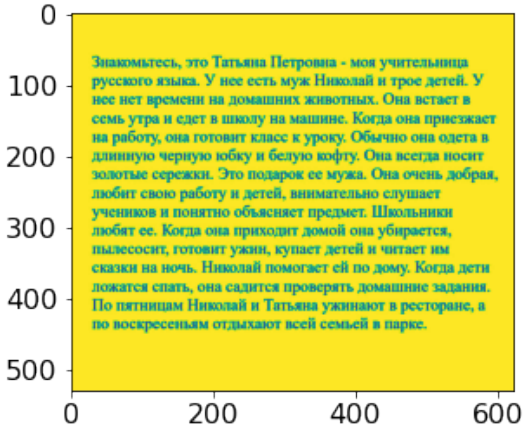


Figure 7: Russian Grayscale

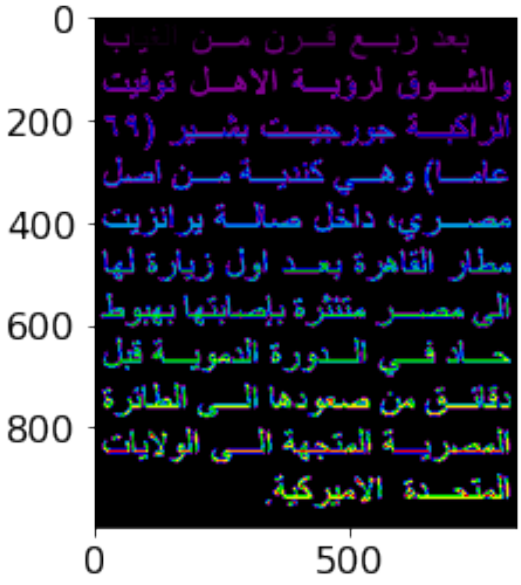


Figure 8: Arabic Labels

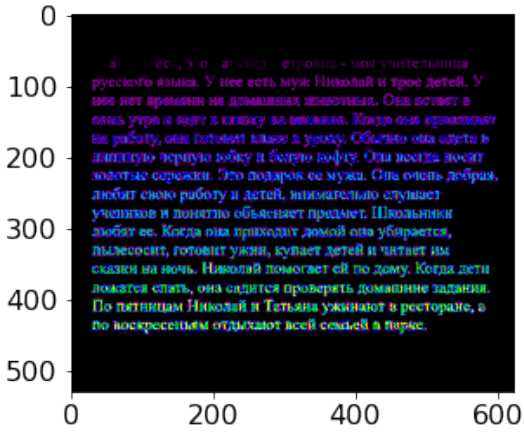


Figure 9: Russian Labels

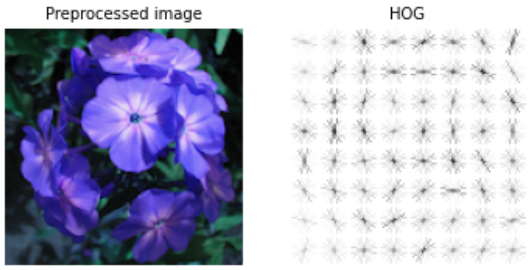


Figure 10: HOG Example

Figure 11: PCA Example

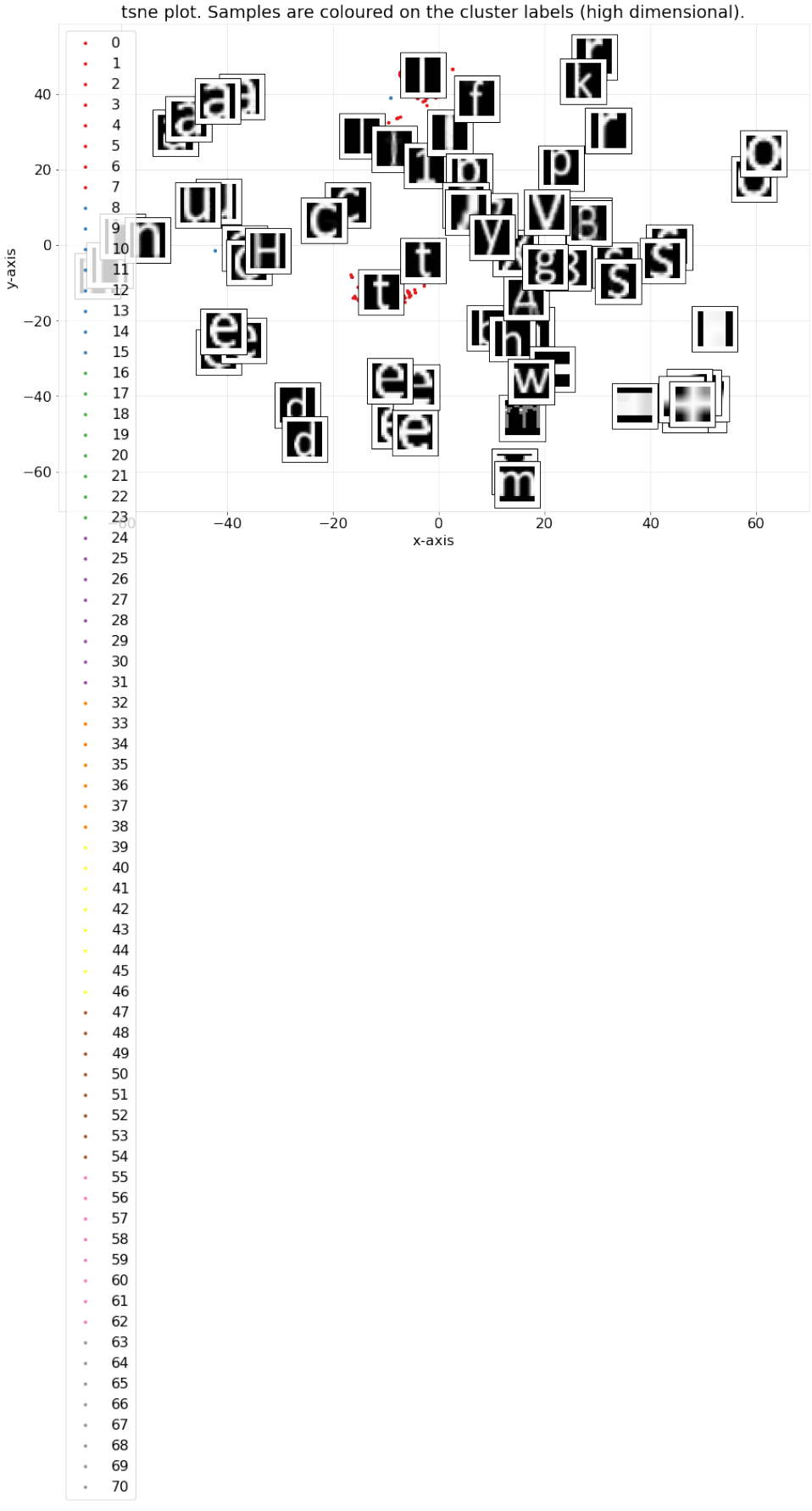


Figure 12: English Cluster

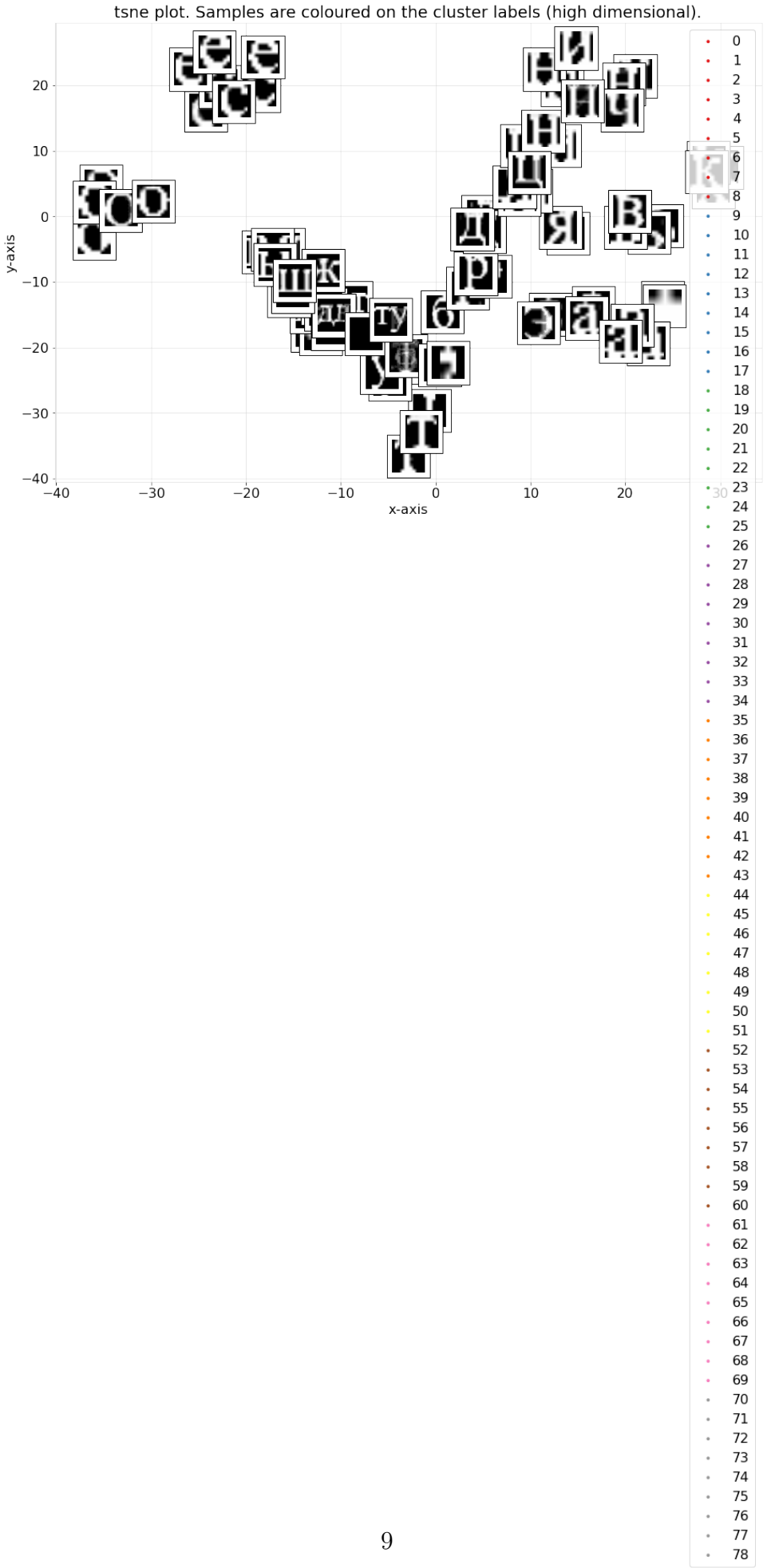


Figure 13: Russian Cluster

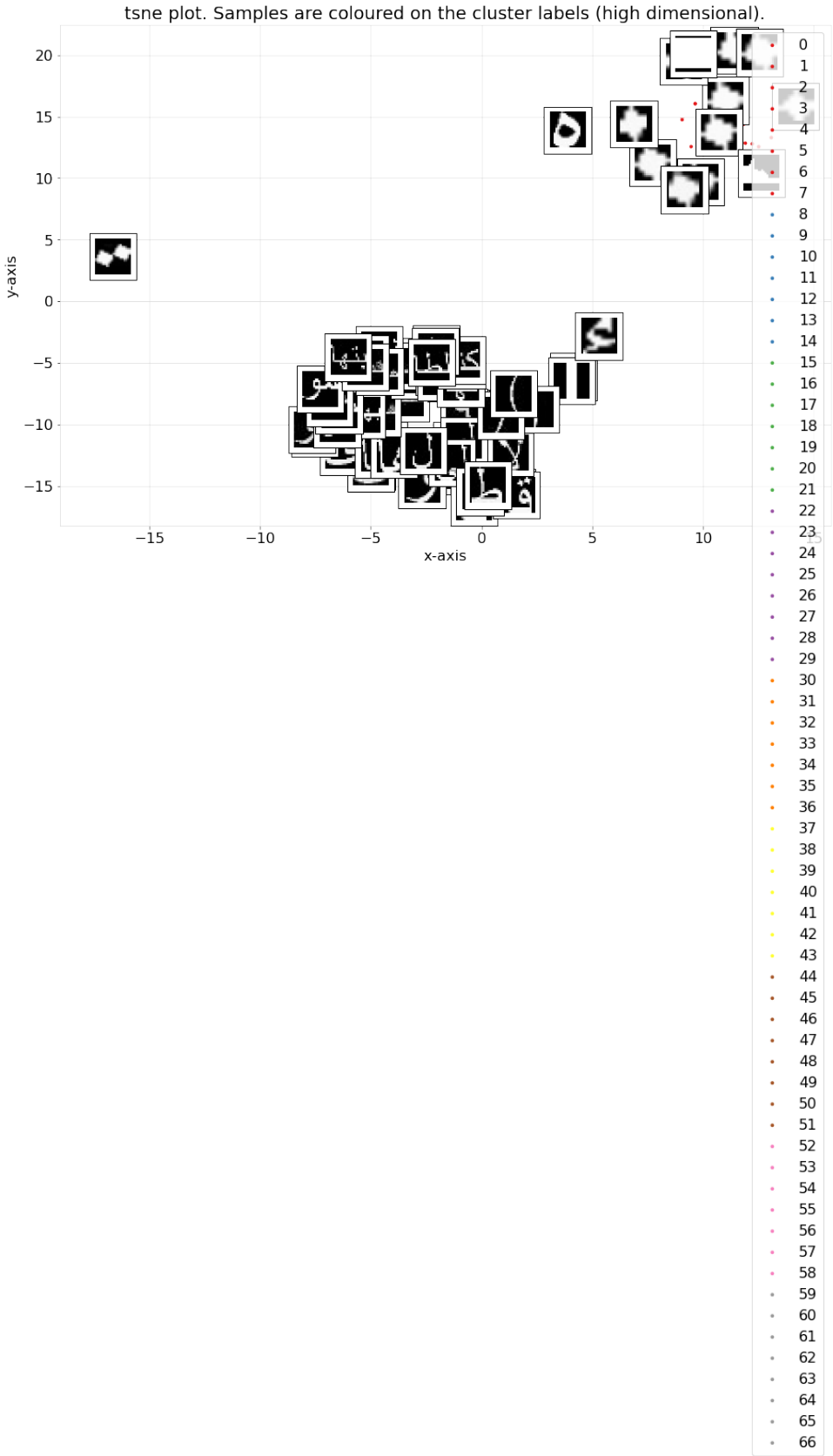


Figure 14: Arabic Cluster

References

[Guruprasad, 2020] Guruprasad, P. (2020). Overview of different thresholding methods in image processing.