# VELA:Visual Exploration and Analysis of large scale multimedia Archives

Bocoum Ousmane[a,*]

[a] *Computer Science and Technology, Southwest University of Science and Technology*

**Abstract**

We present a visual analytic framework for the exploration and analysis of large scale multimedia archives. By revealing different perspectives of a multimedia corpus, our framework gives high level overviews on each type of data and provides powerful mechanisms for detailed analysis. Using deep learning techniques applied to images, audio and text we designed a pipeline for the automatic indexation and classification of large archive of multimedia data. We demonstrate the applicability of our approach on a real world archive dataset.

*Keywords:* Visual analytics, Large Multimedia, Image Visualization, Text Visualization

## 1. Introduction

Large multimedia archives [1] are maintained by different services around the world. Although archivists have always built indexes and research tools, the quantities of the documents are such that a large part of them remains difficult to access and explore [2]. Archivists therefore are facing a major challenge : how to make accessible to the public these millions of documents in a meaningful and easy to use way.

Considering these challenges we designed a web based visualization framework for large scale multimedia archives exploration and analysis. Our system was developed in an iterative design process with continuous refinement guided by archivists. In collaboration with the archivists we identified four major requirements for the effective support of the analysis and exploration process of large multimedia corpora, these are:

(1) Being able to visualize the corpus in a single interactive view

(2)Being able to interact with the items displayed

(3) Being able to query the corpus and get relevant answers

(4)Provide a pipeline to automatically index and classify the data in the corpus.

To meet these requirements we combined visualization techniques and deep learning techniques. Deep learning methods were used to classify automatically multimedia data, witch include text, audio, image and video. The classified corpus was then visualized using multiple visual components, allowing the user to explore, query and interact with the corpus.

[*] Corresponding author.

*Email address:* 3036521131@qq.com (Bocoum Ousmane)

*January 17, 2019*

The main contributions of this paper are:

- We present a visual framework for the exploratory analysis of large scale multimedia archives.
- We introduce a novel method to model a large archive corpus in a directed acyclic graph
- We demonstrate how the combination of deep learning models and visualization techniques can be used to design powerful visualization systems.

This paper is organized as follows: Section 2 introduces a state of the art of information visualization and other related works, in section 3 we describe our framework and the methods used for the automatic indexation. Section 4 presents the visualization components and the different visualization tasks, we finally conclude in section 5 and provide perspectives for future improvements.

## 2. Related Works

Information visualization is a very active research area [3] , numerous visualization systems has been designed to deal with large collections of data. Our framework is informed by related research from the fields of visual content analysis and deep learning techniques applied to text, audio, and image classification.

### *2.1. Visualization of large multimedia data collections:*

#### *2.1.1. Image collection visualization*

Large image collection data visualization has been increasingly used in fields such as medicine [4] , security, and personal album [5] management. Numerous works have successfully used visualization techniques for the exploration of large image collections,Photoland [6] is a system that visualizes hundreds of photos on a 2D grid space to help users manage their photos,Tan et al. presented imageHive [7] an Interactive content-aware image summarization system, tipiX [8] is a system that allows a rapid visualization of large medical image collections. Xie et al. [9] proposed a semantic based method for visualizing large image collections using Convolutionnal neural networks.

Different visualization techniques such as scatter plots [10] , tree based methods [11] and directed graphs [12] has been used to facilitate large image collection analysis.

#### *2.1.2. Text Visualization*

The main goal of text visualization is to allow the discovery of useful knowledge from large document collections effectively without completely going through the details of each document in the collection. To reach this goal different techniques [13] has been proposed such as document similarity visualization [14] , text flowbased methods , Word-based methods, radial based methods, tree-based, and Semantic Oriented Techniques [13].

Using these methods researchers have designed numerous systems, the Text Visualization browser [15] present a survey of text visualization techniques, the Stanford Dissertation Browser [16] is a visualization system for document collections that enables richer interaction, it is an abstraction of Stanford's PhD dissertation from 1993-2008,the documents are presented through the lens of a text model that distills high-level similarity and word usage patterns in the data, they then present those patterns using visualization methods. Our system uses the same approach by extracting underlying patterns from raw documents and presenting them in a more clearly way using visual clues to the archivist.

The approaches presented above all focuses on a specific type of data collection or analysis task, although they allow the understanding of these single data types, such methods does not make use of the main structure of the collection as a whole. In contrast, our approach is more general and allows the user to visualize different data types in the same visualization component, and furthermore provides individual visual components for each type of document.

### *2.2. Deep Learning methods for Unstructured data classification*

The recent advancements in deep learning has made deep convolutionnal networks based methods the go to for unstructured data classification [17] .Convolutionnal neural networks have shown tremendous results in classifying images [18] with human level accuracy. They have revolutionized computer vision, achieving state-of-the-art results in many fundamental tasks, as well as making strong progress in natural language processing, computer audition, reinforcement learning, and many other areas [19].

#### *2.2.1. Image classification using CNNs*

Since the 2012 milestone, researchers showcased various cnn based architectures at the Imagenet Challenge [20].

The most notable models presented at imagenet are:

- AlexNet [21] imagenet 2012 with a top-5 error rate of 15.3%
- VGG16 [22] model imagenet 2014 with top-5 error rate of 7.3%
- GoogLeNet [23] model imagenet 2014 with top-5 error rate of 6.7%
- Inception V2 [24] model imagenet 2015 with a top-5 error rate of 3.58%
- Inception V4 [25] model imagenet 2016 with a top-5 error rate of 3.08%
- SE-Resnet [26] imagenet 2017 with a top-5 error rate of 2.25%.

These models are open source and can be used freely to perform custom image classification tasks. We used the inception V4 model as our image classifier.

### *2.2.2. Audio documents classification using CNNs*

With the success of deep neural networks, a number of studies applied them to speech and other forms of audio data [27] [28]. Representing audio in time is a challenging task, however Van Den Oord et al [29]. Addressed this challenge by introducing wavenet a deep neural network that generates waveforms from raw audio data to train custom classifiers. An alternative to their method is the spectogram of a signal which can represent both time and frequency [30] [31] .Spectograms are images and can be used to extract features form audio data and train a convolutionnal neural net.We used Spectograms to represent our audio data and trained a convolutionnal neural net to classify the audio files.

## 3. System Overview

### *3.1. Dataset*

The dataset used in this work is the compilation of multimedia archives data from the Malian National Archives service. The dataset comport images, audios,videos and text documents collected from 1960 to 2018.

### *3.2. Tasks analysis*

To fully understand the requirements and the tasks of our system we conducted multiple brainstorming sessions with professional archivists, In collaboration with them we identified four major tasks for the effective support of the analysis and exploration process of a large multimedia corpora, these tasks are:

T1. **Summarize a large corpus of multimedia data:**

A corpus of archives can contain a large number of items making it difficult to explore and analyze. Thus being able to visualize the corpus in a single interactive view is critical to conduct a smooth and meaningful exploration of its content.

T2. **Interact with the corpus:**

Interaction is an important task in a visualization system. The user should be able to interact with the content and operate analysis operations such as filtering or searching.

T3. **Query the corpus:**

The system should support various query methods. The user should be able to query the corpus and get relevant answers.

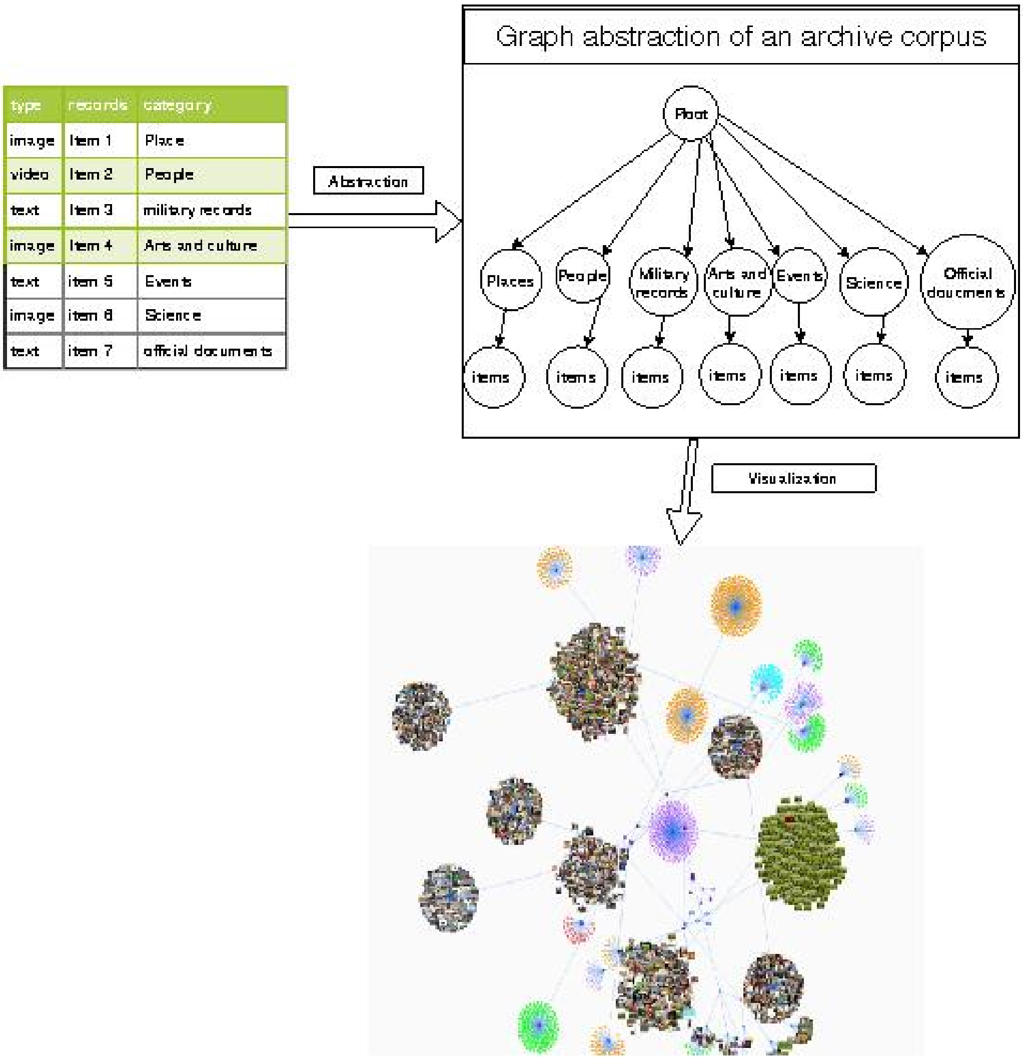
T4. **Automatic Indexation and classification:**

Indexing and classifying a large database of multimedia is a tedious task. Therefore, the necessity of an automatic indexation and classification pipeline.

### *3.3. System Design*

To cover the tasks described above we adopted the following design principles.

- Modeling a large archive corpus in a directed acyclic graph:



| type | records | category |
|---|---|---|
| image | Item 1 | Place |
| video | Item 2 | People |
| text | Item 3 | military records |
| image | Item 4 | Arts and culture |
| text | item 5 | Events |
| image | item 6 | Science |
| text | item 7 | official documents |

Figure 1: Process of modeling a large database of archive to an acyclic graph

A large archive corpus can be represented in an acyclic graph to abstract the hierarchy and the relations between the documents. The graph representation offers a multi-level overview(T1) of the corpus allowing the users to identify groups of interrelated documents. Such representation make it easy to implement various interaction and filtering methods(T2)

- Multiple Querying methods:

We implemented multiple querying methods such as filtering and searching, we furthermore implemented a document retrieval strategy allowing the user to query the corpus from an image or audio file(T3).

- Automatic indexation and classification pipeline:

Using deep learning methods we trained various classifiers for each data type(T4), section 4 describe the details of each model.

### *3.4. System Architecture*

Our system is a client-server application.We used Reactjs as front end framework,Flask as back end and MongoDB database. We trained our models using Tensorflow. The models are deployed on the server where the input data is prepossessed and classified. The classified data is then sent to the frontend for visualization.

## 4. Models

In this section we describe the models used for each type of data.

### *4.1. Image classification model*

Our image classification pipeline is a combination of the Inception V2 [24] trained on imagenet dataset [20] used as an entry classifier and a custom face recognition model. The Inception V2 takes the raw input images and classify the images to the 1000 classes of imagenet.

We then feed the pre-classified images containing people to a custom face classifier to index images pertaining to notable people present in the dataset.

Our face classifier was trained using transfer learning on facenet [32] , the classes are famous people and important people present in the corpus.

### *4.2. Audio Classification model*

We trained an audio classification model to recognize the genre of each file. Our model is a convolutionnal neural network based on the VGG16 [22] architecture and trained with our own datasets. For our training data we extracted the features for each category using melspectograms.A spectrogram is a visual representation of the spectrum of frequencies of sound or other signal as they vary with time. We then used the spectograms to train our model to map each data to one of the following classes:liste des categories

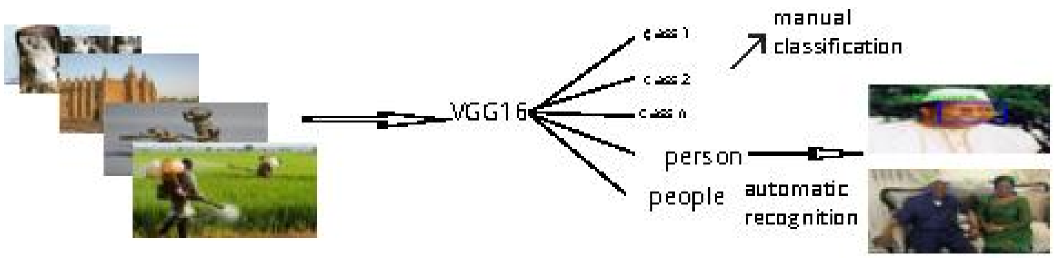


Figure 2: The image classification pipeline takes raw input images,and the VGG16 model trained on imagenet dataset gives a class to each image,the images containing people or a person are then sent to a custom face recognizer,the the face recognizer is trained to recognize famous and key people present in the corpus.

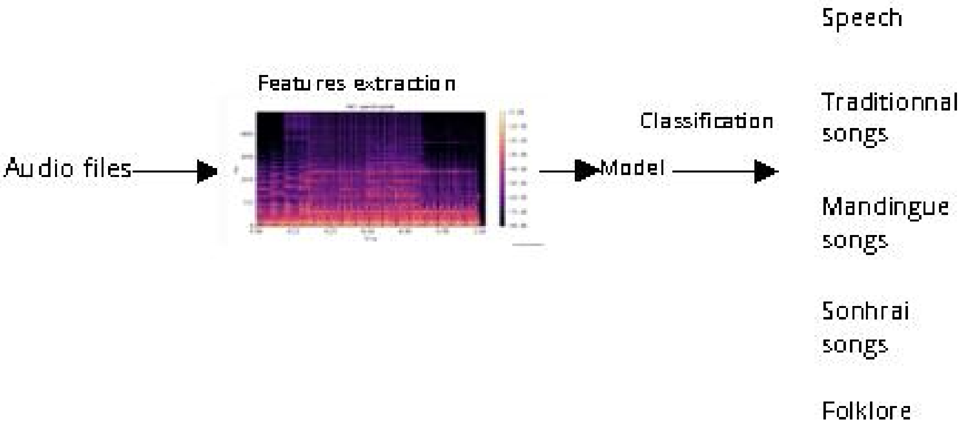


Figure 3: The audio classification model takes in the raw audio files and extracts melspectograms from them, we then uses the spectograms to train a convolutionnal neural network classifier, the output of the model are files classified per category.

Our goal with these models is to help the archivists to automatically classify and index raw multimedia data, however images and audio can be misclassified and have to be manually corrected by the archivists.

### *4.3. Text Similarity model:*

For each text document we conducted standard data processing,cleaning,n-gram extraction. We then used the Multi-Perspective Sentence Similarity model proposed by He et al [33] to measure the similarity between documents. The method uses convolutionnal neural nets to extract features from sentences and compare sentence similarity.

We also used the open source Stanford named entity recognizer [34] to extract the named entities present in each document. Named entities are very important in understanding the content of a text document, they help to abstract the context of a large document. We then visualized the entities in a named entity graph [35].

## 5. Visual components:

In this section we describe the visual components of our system. The system is composed of multiple views each of which displays an aspect of the corpus.

In the following we describe each visual components and its interactions.

### *5.1. Corpus view*

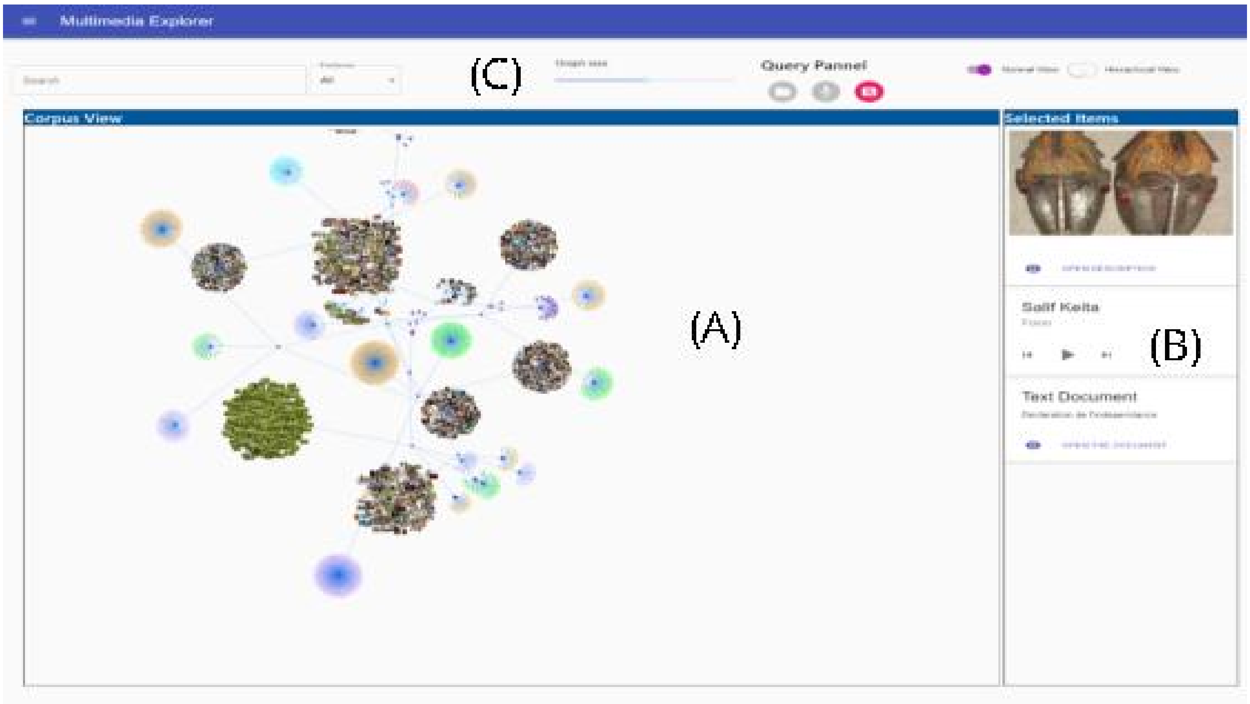


Figure 4: The corpus view summarizes the corpus in an interactive graph,the user can select individual elements for detailed analysis,the view supports zooming,filtering and searching.

The corpus view represents the complete corpus with entities highlighted in their respective colors. We choose to map each data type to a discriminative visual variable such as color or shape.

This visualization is composed of three components, the main layout(A) is a graph displaying the corpus, users can interact with the items on the graph,the graph is zoomable and draggable. The selected items are displayed on the selections panel(B), the selection panel allow the user to analyze the selected items with more detail.

At the top of the view is a query panel(C). The view supports searching and filtering.

### *5.2. Image view*

The image view is an abstraction of the images in the corpus as a single entity. The full corpus displayed in the corpus view can be difficult to comprehend directly, therefore this visualization focuses on displaying the image data in an interactive view. The main purpose of this view is to give the user the possibility to browse and explore the images in the corpus.

The view provides several convenience methods to navigate through the items such as zooming and selection. The selected items panel can be used to get more details about an item.The query panel offers multiple querying options such as searching, and image retrieval using an image file to retrieve similar images.

### *5.3. Audio View*

The audio view displays the classified audio data.A treemap is used to visualize different clusters of data per category. The user can select a given item and listen to its content. The purpose of this view is to summarize the audio data in the corpus in an organized component.

### *5.4. Video View*

The video view is an interactive visualization of the video files in the corpus. It is composed of three main components, a graph displays thumbnails and titles of the videos. The selected videos are displayed in the selection panel, and keyframes of each selected video are displayed on the keyframe panel. The view support searching and filtering.

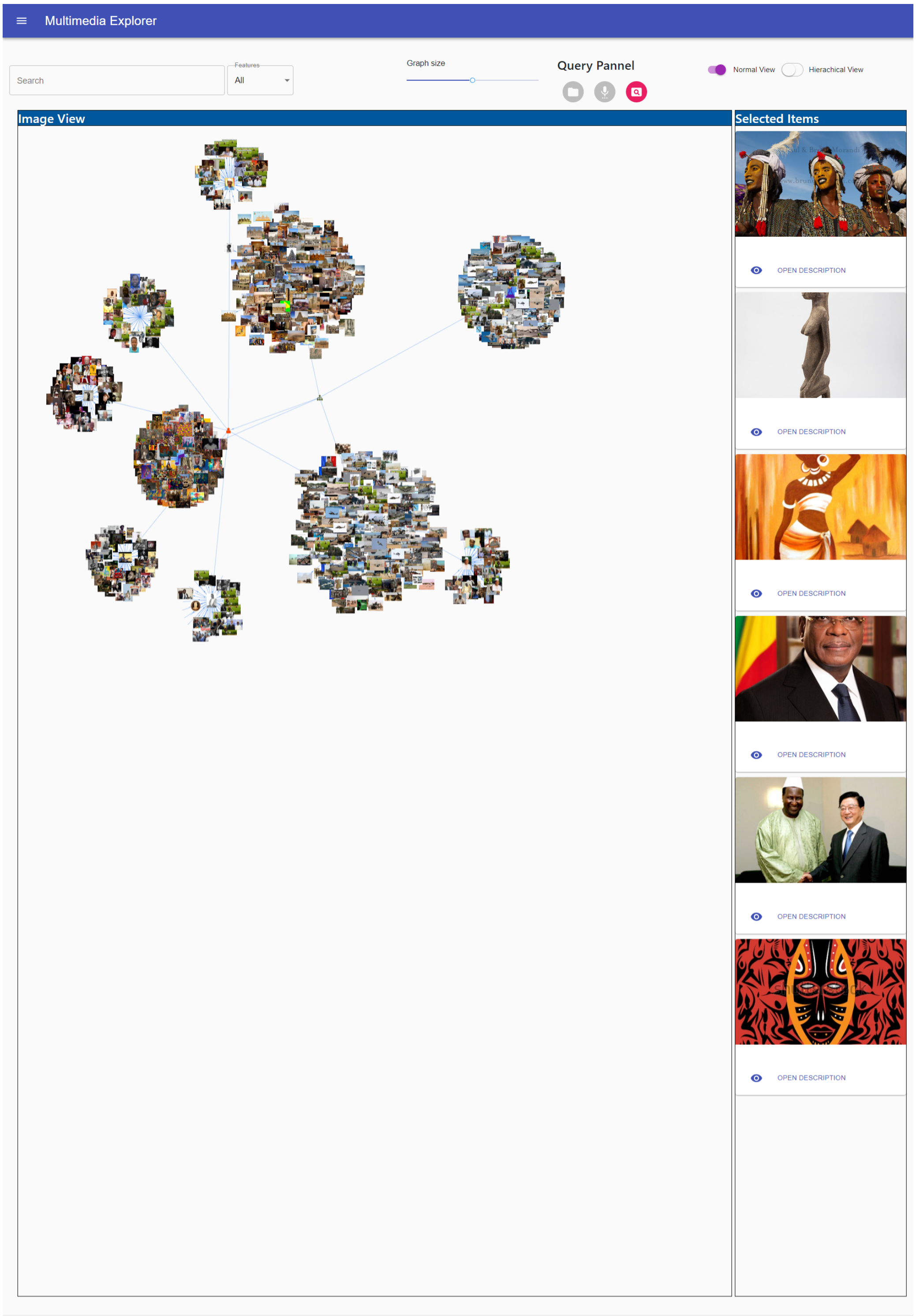
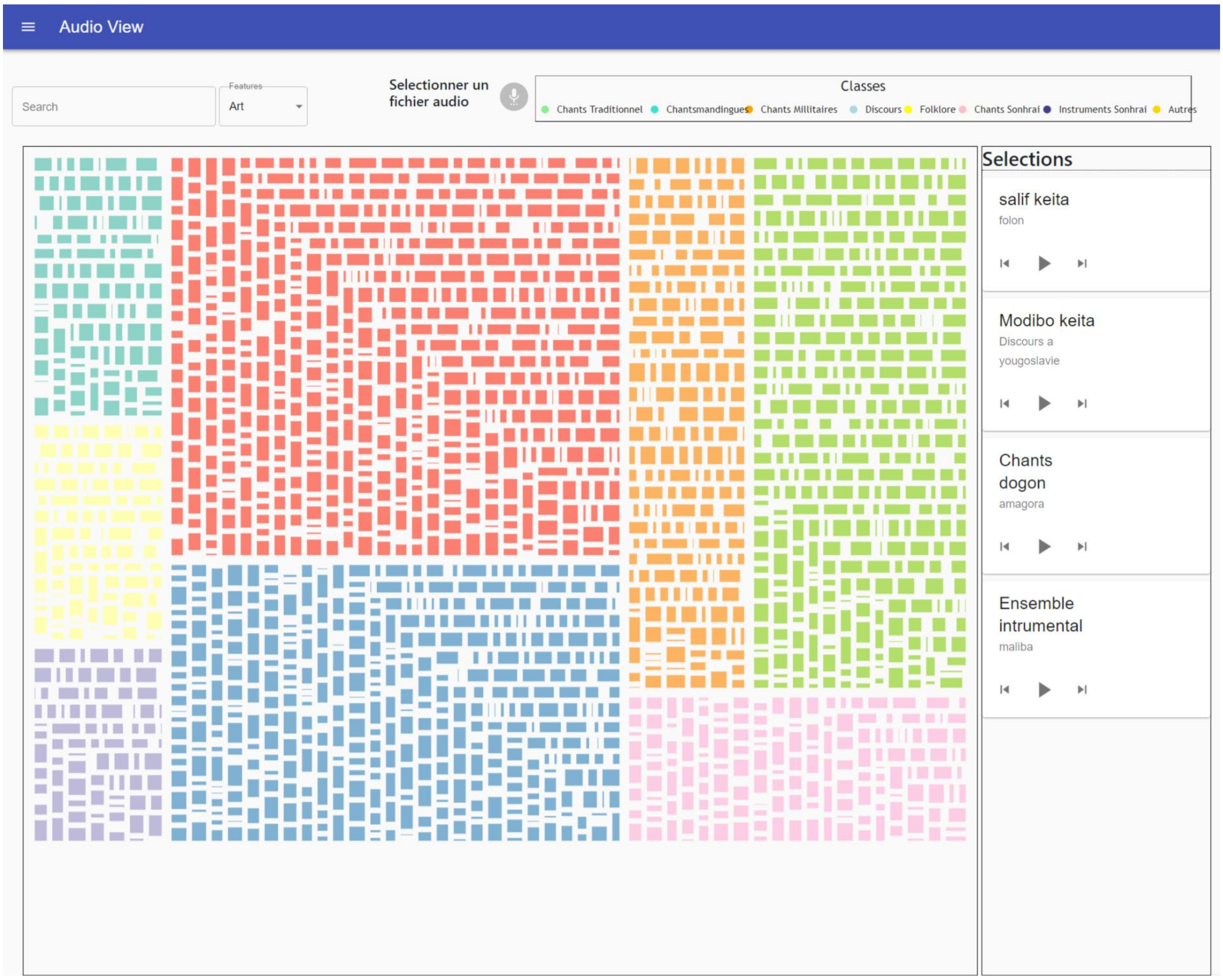


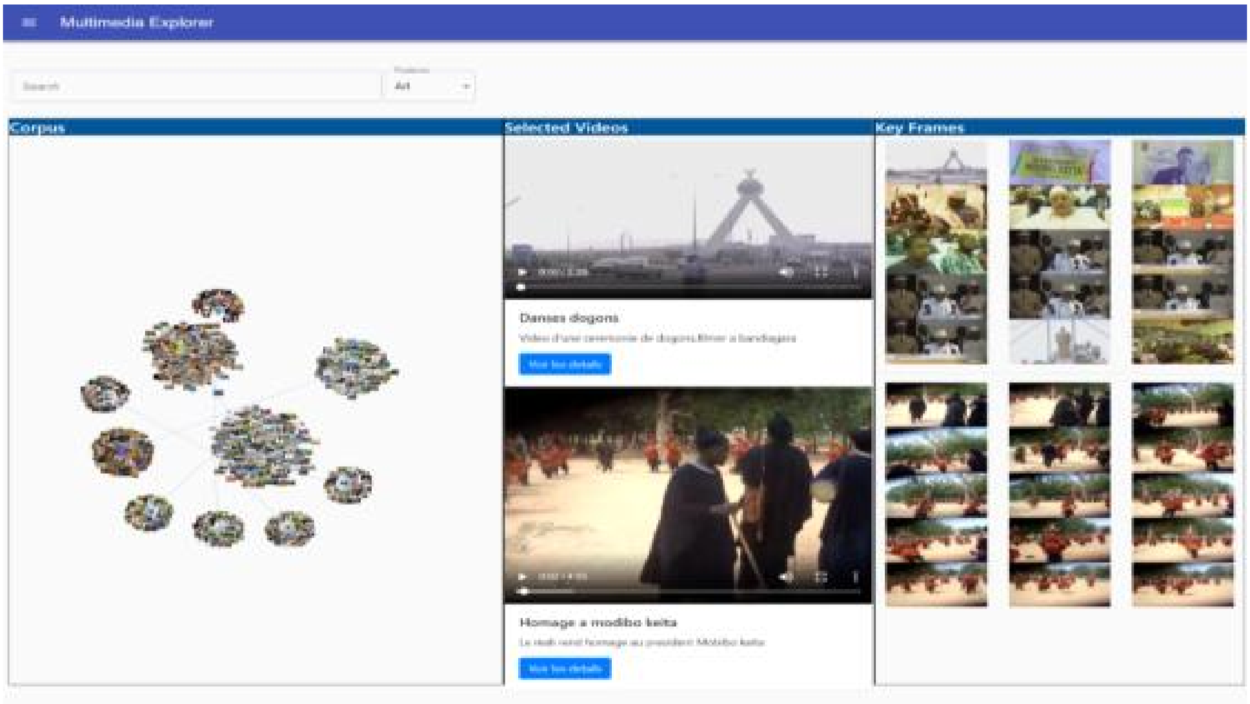
Figure 5: The image view visualizes the images in the corpus, its main purpose is to give the user a smooth exploration experience of the images without being distracted by other documents.



Features
Art

≡ Audio View
Search
Selectionner un fichier audio
Classes
● Chants Traditionnel ● Chantsmandingues ● Chants Millitaires ● Discours ● Folklore ● Chants Sonhrai ● Instruments Sonhrai ● Autres
Selections
salif keita
folon
Modibo keita
Discours a yougoslavie
Chants dogon
amagora
Ensemble intrumental
maliba

Figure 6: The audio view is a visualization of the output of our audio classification model, it displays a treemap containing clusters of items classified per category, the user can select an item and listen.



Danses dogons

Video d'une ceremonie de dogons,filmer a bandiagara

Homage a modibo keita

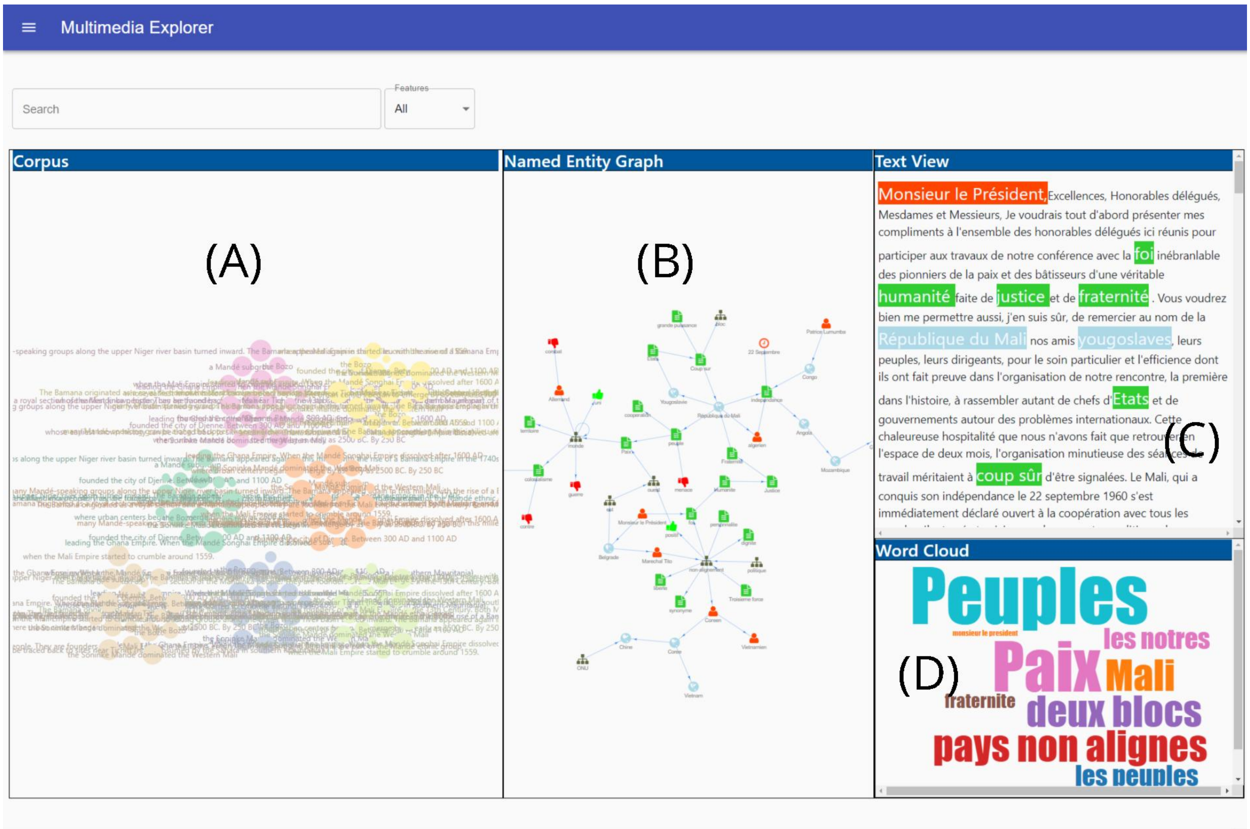
Le mali rend hommage au president Mobibo keita

Figure 7: The video view gives the user the possibility to explore the videos in the corpus, A graph displays thumbnails of the videos, the user can select a thumbnail and view the video, key frames of the selected videos are displayed in the key frame panel.

### 5.5. *Text View*

The text view visualizes the text documents of the corpus. It is composed of:

- A bubble graphFigure 8 (A) displaying the documents classified per category, the distance between each item is proportional to the similarity of the documents. The user clicks on a document and trigger the visualization of the content on the other components of the view.
- A named entity graphFigure 8 (B) shows the named entities contained in the document
- The text viewFigure 8 (C) displays the document with the named entities highlighted
- A Word cloud viewFigure 8 (D) displays the important words in the document per frequency



Monsieur le Président, Excellences, Honorables délégués, Mesdames et Messieurs, Je voudrais tout d'abord présenter mes compliments à l'ensemble des honorables délégués ici réunis pour participer aux travaux de notre conférence avec la foi inébranlable des pionniers de la paix et des bâtisseurs d'une véritable humanité faite de justice et de fraternité . Vous voudrez bien me permettre aussi, j'en suis sûr, de remercier au nom de la République du Mali nos amis yougoslaves, leurs peuples, leurs dirigeants, pour le soin particulier et l'efficience dont ils ont fait preuve dans l'organisation de notre rencontre, la première dans l'histoire, à rassembler autant de chefs d'Etats et de gouvernements autour des problèmes internationaux. Cette chaleureuse hospitalité que nous n'avons fait que retrouver en l'espace de deux mois, l'organisation minutieuse des séances de travail méritaient à coup sûr d'être signalées. Le Mali, qui a conquis son indépendance le 22 septembre 1960 s'est immédiatement déclaré ouvert à la coopération avec tous les

Figure 8: The text view is an interactive document browser, the documents in the corpus are displayed in the corpus panel(A), when the user selects a document its content is displayed in the text view(C) with named entities highlighted, a named entity graph displays(B) the named entities present in the document, a word cloud is used to display the most frequent words in the document.

## 6. Conclusion

We presented a visual analytic framework for large multimedia archives data visualization. Our approach represents a large corpus of archives in an acyclic directed graph. The graph model was then visualized using multiple interactive views, each view displayed an aspect of the corpus. We also used deep learning techniques to facilitate the tedious task of indexing and classifying a large corpus of multimedia data.

In future works we would like to improve the automatic indexation pipeline and make it more accurate.

## References

[1] C. G. Snoek, M. Worring, J. C. V. Gemert, J.-M. Geusebroek, A. W. Smeulders, The challenge problem for automated detection of 101 semantic concepts in multimedia, in: Proceedings of the 14th ACM international conference on Multimedia, 2006, pp. 421–430.

[2] M. R. Henley, Method and system for managing movement of large multi-media data files from an archival storage to an active storage within a multi-media server computer system (1998).

[3] C. North, Information visualization, Handbook of human factors and ergonomics (2012) 1209–1236.

[4] W. Plant, G. Schaefer, Visualisation and browsing of image databases (2011).

[5] E. Cambria, A. Hussain, Sentic album: content-, concept-, and context-based online personal photo management system, Cognitive Computation 4 (4) (2012) 477–496.

[6] D.-S. Ryu, W.-K. Chung, H.-G. Cho, Photoland: a new image layout system using spatio-temporal information in digital photos, in: Proceedings of the 2010 ACM Symposium on Applied Computing, 2010, pp. 1884–1891.

[7] L. Tan, Y. Song, S. Liu, L. Xie, Imagehive: Interactive content-aware image summarization, IEEE computer graphics and applications 32 (1) (2012) 46–55.

[8] A. V. Dalca, R. Sridharan, N. Rost, P. Golland, tipiX: Rapid Visualization of Large Image Collections.

[9] X. Xie, X. Cai, J. Zhou, N. Cao, Y. Wu, A Semantic-based Method for Visualizing Large Image Collections, IEEE Transactions on Visualization and Computer Graphics.

[10] G. P. Nguyen, M. Worring, Interactive access to large image collections using similarity-based visualization, Journal of Visual Languages & Computing 19 (2) (2008) 203–224.

[11] B. B. Bederson, PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps, in: Proceedings of the 14th annual ACM symposium on User interface software and technology, 2001, pp. 71–80.

[12] Y. Gu, C. Wang, J. Ma, R. J. Nemiroff, D. L. Kao, iGraph: a graph-based technique for visual analytics of image and text collections, in: Visualization and Data Analysis 2015, Vol. 9397, 2015, p. 939708.

[13] K. Kucher, A. Kerren, Text visualization techniques: Taxonomy, visual survey, and community insights, in: Visualization Symposium (PacificVis), 2015 IEEE Pacific, 2015, pp. 117–121.

[14] D. Baker, The Document Similarity Network: A Novel Technique for Visualizing Relationships in Text Corpora.

[15] K. Kucher, A. Kerren, Text visualization browser: A visual survey of text visualization techniques, Poster Abstracts of IEEE VIS 2014.

[16] J. Chuang, D. Ramage, C. Manning, J. Heer, Interpretation and trust: Designing model-driven visualizations for text analysis, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012, pp. 443–452.

[17] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, M. Hasan, B. C. V. Esesn, A. A. S. Awwal, V. K. Asari, The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches, arXiv preprint arXiv:1803.01164.

[18] C. Wang, Y. Xi, Convolutional Neural Network for Image Classification, Johns Hopkins University Baltimore, MD 21218.

[19] A. Voulodimos, N. Doulamis, G. Bebis, T. Stathaki, Recent Developments in Deep Learning for Engineering Applications, Computational intelligence and neuroscience 2018.

[20] title={Imagenet large scale visual recognition challenge}.

[21] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems (2012).

[22] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition., arXiv preprint 1409 (1556).

[23] C. Szegedy, Going deeper with convolutions., 2015.

[24] C. Szegedy, Rethinking the inception architecture for computer vision., 2016.

[25] C. Szegedy, Inception-v4, inception-resnet and the impact of residual connections on learning., AAAI 4.

[26] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks. arXiv. (2017).

[27] S. Hershey, "CNN architectures for large-scale audio classification." Acoustics, Speech and, 2017.

[28] J. Lee, Raw Waveform-based Audio Classification Using Sample-level CNN Architectures., arXiv preprint 1712 (866).

[29] V. D. Oord, Aäron, Wavenet: A generative model for raw audio., CoRR 1609 (3499).

[30] T. J. Lynn, A. Z. Sha'ameri, Automatic analysis and classification of digital modulation signals using spectogram time frequency analysis., Communications and Information Technologies.

[31] H. Lee, "Unsupervised feature learning for audio classification using convolutional deep belief networks." Advances in neural information processing systems (2009).

[32] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering., 2015.

[33] H. He, K. Gimpel, J. Lin, Multi-perspective sentence similarity modeling with convolutional neural networks., 2015.

[34] J. R. Finkel, C. D. Manning, Nested named entity recognition., Association for Computational Linguistics, 2009.

[35] M. El-Assady, NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations., Computer Graphics Forum 36 (3).