Detection and Extraction of Faces and Text Lower Third Techniques for an Audiovisual Archive System using Machine Learning

Khalid El Fayq[1], Said Tkatek[2], Lahcen Idouglid[3], Jaafar Abouchabaka[4]
LaRIT, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

***Abstract*—As part of the audiovisual archive digitization project, which has become a complex field that requires human and material resources, and its automation and optimization have so far represented a center of interest for researchers and media manufacturers, in particular those linked to the integration of artificial intelligence tools in the industry, an elaborate work for the development of an optical character and face recognition model, to digitize the tasks of audiovisual archivist from the manuscript method in automation, from a TV news video. In this article, an approach to develop an example of lower third in Arabic language and facial detection and recognition for news presenter that provide accurate classification results as well as the presentation of different methods and algorithms for Arabic characters. Many studies have been presented in this area, however a satisfactory classification accuracy is yet to be achieved. The comparative state-of-the-art results adopt the latest approaches to study face recognition or OCR, but this model supports both at the same time. it will present the context of realization, the method proposed to extract the texts in the video, using machine learning, about the specificity of the Arabic language, and finally the reasons that govern the decisions taken in the steps of realization. The best results from this approach in real project at the media station was 90.60%. The dataset collected via presenters images and the character dataset via the Pytesseract library.**

Keywords—Image processing; OpenCV; Tesseract; video OCR; face detection

I. Introduction

In recent years, significant and rapid developments in the field of artificial intelligence (AI), the demand for smart applications has increased and found significant interest and use in many fields. As expected, audiovisual production is no exception to this, but radical transformations have taken place in the production process, facial recognition, text, and sound, and also video editing thanks to a set of tools that rely on AI technologies, as a support to the human element, most notably in tasks that require time and repetitive effort. This success in the use of AI is due to two main reasons:

1) Big Data Availability, such as photos and videos, in media stations and on the Internet.

2) Advances in digital computing manufacturing.

This article is the result of work carried out within Moroccan Television, for the purpose of researching a processing model that will allow the analysis of audiovisual streams of television news, by analyzing the news presenter faces and the writing in each news coverage.

Videos have become a great source of information; the text of the video contains a substantial amount of information in a non-editable form. If this text is converted into an editable form, it becomes easier and more efficient to store and redistribute [1].

It has been observed that media channels in general have a growing need for automatic facial and handwriting detection systems to integrate into their systems.

One only has to think of the need for the development of active systems to extract the data in the television news videos in order to help the archiving service to fill in the current file, which is currently manually filled in on a daily basis in a register (Fig. 1).

The user must first provide the video as the input from which he wants to extract text. The system will then process the video and generate the editable text output. the latter must ensure the reliability of the information to an acceptable percentage. The solution must extract information from people who appear on the screen, using two methods: facial and text recognition in the video.

It should be noted that there are now commercial automatic tools for processing texts, photos, and videos [2][3], with a complex background, alignment and color variants, etc.

This work will be limited to automating the task of gender recognition from faces only and optical character recognition from the Lower Third present in the video. Also proposed in this work is a video processing chain which includes parts of detection, and tracking of faces and text at the same time.

Effective and efficient text extraction has been a difficult topic in recent years, and the Arabic language is one of the most popular languages in the world. Hundreds of millions of people in many countries around the world speak Arabic as their native language. However, due to the complexity of the Arabic language due to its cursive nature, the recognition of printed and handwritten Arabic text remained untouched for a very long time and did not receive the same attention, compared to the Latin script [4].

The remaining of this paper is structured as follows:

related works developed for text and face recognition from scenes/videos, diagnosis and analysis of the implementation project and its difficulties, in the face detection and Arabic writing step are discussed in Section II, the presentation of Lower third and face extraction technique and the algorithm to follow for the deployment of this system, discusses the process of preparing the data set are mentioned in Section III and Section IV, the conclusion and some future work discussions are mentioned in Section V.

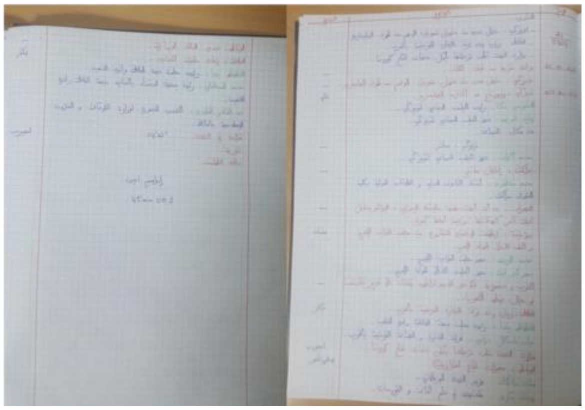


Fig. 1. Handwritten News Archive.

II. Related Work

Although a large number of printed Arabic character recognition approaches have been proposed in recent years, there is still a need to improve the recognition rate in Arabic OCR systems. This section presents some of these approaches.

Text recognition methods can be categorized into three broad categories, where some approaches recognize text using text segmentation and offering profiling learning with their own features. Methods in the second category recognize text without text segmentation, using a framework based on multiple hypotheses. The methods of the third category improve the text to increase the recognition rate by using binarization of scene images.

Each of these three categories has its own limitations. Approaches in the first category only work well for data from specific scripts, as they need training from their own samples and a classifier to recognize text based on that training. Methods in the second category require multiple hypotheses to set thresholds, but it is unclear how to derive different hypotheses to set specific thresholds. However, the methods of the third category do not need any classifier or hypotheses to define certain thresholds and they also improve the recognition rate through binarization. However, the approaches due to the third category do not provide satisfactory recognition performance for low resolution scene/video images [5], however, these methods perform well on horizontal scene text.

A. BACKGROUND: Text Recognition

Many systems have been developed to detect text in videos. Each system is based on a specific method and has its associated shortcomings. Some of the commonly used methods to detect text are:

1) Method based on the Sliding Window:

This approach uses a sliding window to search for specific text. It first takes a small rectangular block of a given image. The rectangular patch has a specific size. Drag this rectangular block over the entire area covered by the image to check if there is any text in this image block. Different sliding window classifiers are used to determine if there is text in the patch. The window is initially placed in the upper left corner of the image, and slides to different positions of the image starting from the first row, then slides through the other rows of the image. This method is slow because the image must be processed at several scales.

2) Method based on Connected Components:

In the connected component-based approach, it first extracts regions of pixels that have a similar color, edge strength, or texture and evaluate each of them to be text or not using techniques of machine learning.

The connected component method works well for caption text with single background images, but it doesn't work well for images with a grouped background.

Text recognition in video images is more difficult than that of natural scene images, difficulties video text analysis are due to complex background images, color variations, font size, camera movement, etc.

Aasim Zafar; Arshad Iqbal [6] compared K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) classifiers in the recognition of printed Arabic characters.

First, feature extraction techniques such as oriented gradient histogram (HOG) and local binary pattern (LBP) have been applied for feature extraction based on the structure of Arabic handwritten texts. SVM has been found to perform better than KNN. Amara et al. [7] presented Arabic OCR using Support Vector Machines (SVM). Although SVM has proven its effectiveness in different fields among other classification tools, SVM has not been effectively applied to recognize Arabic characters.

Saidane and Garcia [8] proposed a convolutional network-based binarization method for the color text area of video images and its performance depends on the amount of training samples used.

Ahmed H and Mahmoud [9] proposed a small-sized printed Arabic text recognition approach based on the estimation of the Hidden Markov Model (HMM). Although applying a hidden Markov model has some advantages (such as no pre-segmentation), the poor image quality of small printed Arabic text makes it difficult to find accurate model boundaries.

Sarfraz and al. [10] proposed an offline Arabic character recognition system. The proposed system has four steps. In the first step, the text preprocessing step removes the isolated pixel and corrects the drift. A pixel is considered isolated if it has no neighboring pixels. Drift is corrected by rotating the image according to the angle with the greatest number of occurrences between all angles of all line segments between any pair of black pixels in the image. In the second step, line and word segmentation is performed using horizontal and vertical projection. Words are segmented into individual characters by comparing the vertical projection profile with a fixed threshold. The feature space is constructed using the moment invariance technique [11].

Zagoris and al. [12] proposed an approach to differentiate handwritten text from machine-printed text. The text image is segmented into blocks. Each block is represented as a vector of words, which contains local features identified using scale-invariant feature transformation. Based on support vector machines, the proposed approach decides whether the text block is handwritten, machine-printed or noise by comparing its word vector with the codebook.

B. Background: Reconnaissance Facial

Facial recognition started about twenty years ago, and play an important role in various fields, the OpenCV library [13] is free, used in the field of research, because it provides a large number of important tools (more of 500 functions) in the field of computer vision.

The processing and description of audiovisual content presents several opportunities for description through automation and machine learning. Some methods are entirely based on image and sound content, such as facial recognition and audio indexing; other methods rely on text, either inherent in the digital file or extracted from audio or video.

It will discuss the state of the art of facial recognition then face detection methods have been classified by Yang [14], in four approaches:

- Approach based on recognition.
- Approach based on invariant characteristics.
- Approach based on template matching.
- Approach based on appearance

Image classification can be implemented using various supervised techniques such as Naive Bayes [15], K-Nearest Neighbor (KNN) [16], Support vector machines (SVM) [17] [18] [19], Decision trees [20], Random forests [21], Convolutional Neural Network (CNN) [22] [23]and Recurrent Neural Networks (RNN) [24]. These techniques process and classify images into different classes.

In Eidinger, Roee, & Tal [25], they proposed a method based on two tasks based on face representation with local binary patterns (LBP) and linear SVM with dropout. Dropout-SVM is based on the assimilation of a linear SVM to a single layer of a neural network.

Hassner, [26] used the same classification technique as Eidinger, Roee [25], on the front view, projecting 2D points of interest from the front to the 3D face as a reference. They showed that reconstruction of the face in the forward position can improve the performance of facial recognition tasks, in particular gender recognition.

Recently, AZZOPARDI, [27] also proposed a method based on artificial face extraction representation (cosfire-filters), similar to LBP. These representations are also inputs for the linear SVM.

Levi & Tal, [28] implementation of a convolutional neural network (CNN) with tree convolutional layers and three fully connected layers for the task of estimating age and sex; he was specially trained by Audience. To make predictions, they made several crops of different sizes around the face. The final forecast is the average forecast value of all these crops.

Moreover, Afifi & Abdelhamed, [29] also applied a method based on local facial features, dividing it into several parts (mouth, eyes, nose). They also include insightful strokes around the face while blurring it. Then use these images to train multiple CNNs.

Other papers have also presented a CNN as Ranjan, M. Patel, & Chellappa, [30] proposed two Hyperface models, a CNN network based on the AlexNet architecture and a residual network based on the ResNet 101 architecture. Models are multitasking and can be trained to perform face detection, POI coordinate prediction, pose estimation, and gender recognition simultaneously.

Wolfshaar, F. Karaaba, & A Wiering [31], use a CNN (ImageNet's BVLC) designed to recognize objects belonging to 20,000 categories. In a separate experiment, two datasets (ColorFeret and Adience) of the face were further trained. Then extract the visual features from the penultimate convolutional layer and use them to train the linear SVM.

Ozbulak, Aytar, & Hazim, [32], showed that transfer learning domain-specific models (such as VGG Face) can perform better than recycled CNN with limited data. They achieved this by comparing GilNet (a shallow benchmark CNN trained on the Adience dataset) with two enhanced deep CNNs (one for the VGGFace face and the other more general Alexnet). Then these two enhanced CNNs are used as descriptor extractors, and these descriptors will become the training data of SVM.

C. Problem Statement

For the moment, Laâyoune TV does not have a computerized archiving system, neither a desktop application nor a web application, and only uses handwritten recording.

On a daily basis, the archivist proceeds to view each video clip, and writes the visual information of each video, which drew my attention to the design of a model that uses two methods of machine learning, in particular the vision by computer, in order to help the television set up an automated system which does the same work of an archivist to extract the data from the video as it is written in the register, with the aim of minimizing the lead time of the archiving task which consumes about one hour of work daily which will free the archivist to use this time for other more important tasks.

Also depending on the global health situation due to the Corona 19 epidemic [33], the Moroccan national radio and television company complies with government laws that encourage remote work.

The main purpose of a text extraction system is to accept video files, detect the text, extract it, and produce an ASCII file including the text in a format that can be used by other applications.

Text detection is performed in each frame of the video. The rectangles representing the location of the text are followed during their period of appearance to associate the corresponding rectangles in the different frames.

This information is necessary to improve the content of the image, which can be achieved by integrating several rectangles containing the same text. This phase must produce sub-images of a quality in accordance with the prerequisites of an OCR process. Therefore, face recognition of news anchors is used to separate each story from the other.

This system makes it possible to extract several information of the news anchor (Fig. 2), the cities (Fig. 3), the people who speak in the microphone (Fig. 4), and the journalists who work in the reports (Fig. 5), in order to provide a complete document.

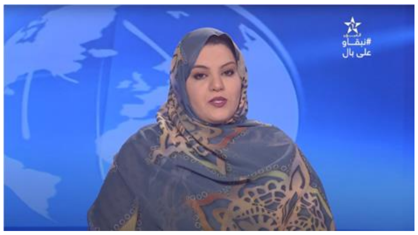


Fig. 2. Example of a TV News Presenter.

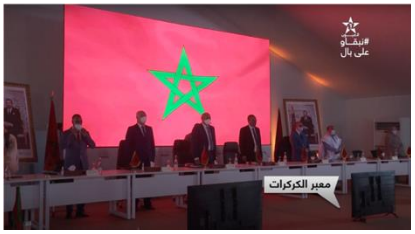


Fig. 3. Example of a City Name in a Report.

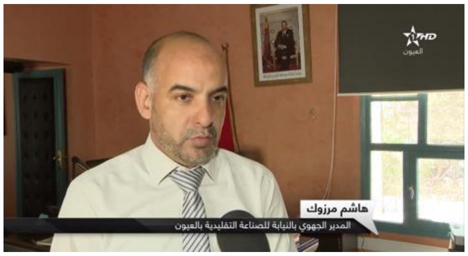


Fig. 4. Person Speaks with a Microphone.

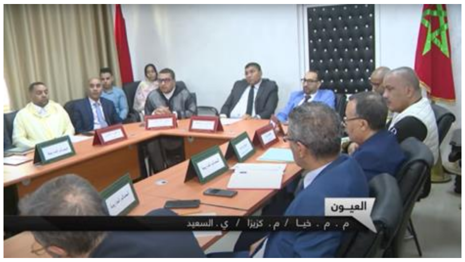


Fig. 5. Example of a Team that Produced a Report.

The detection and recognition system must be able to observe a scene. The acquisition conditions of each sequence of images obtained are checked. Usually, capturing the news anchor's face images will be done in front view and in best Full HD (1980x1080) image quality. "Table I" summarizes the video resolution specifications.

TABLE I. VIDEO TECHNICAL SPECIFICATIONS

Type	Video
Codec	MPEG-2 Video (mp2v)
Resolution	1920X1080
Frame rate	25
Planar decoded format	4:2:2 YUV

This system tries to extract and process much information to extract each text frame and recognize faces at the same time (Fig. 6).

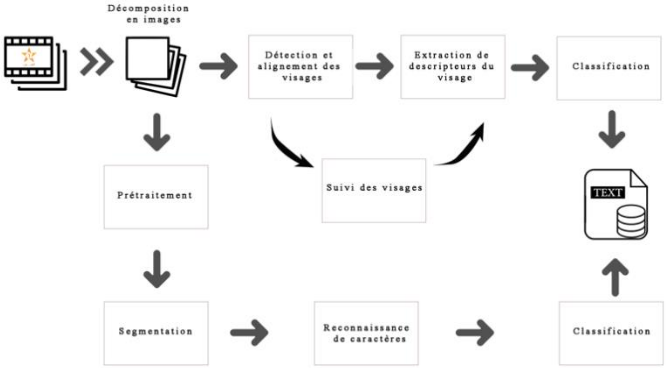

Décomposition en images
Détection et alignement des visages
Extraction de descripteurs du visage
Classification
Suivi des visages
Prétraitement
TEXT
Segmentation
Reconnaissance de caractères
Classification

Fig. 6. Processing Model Diagram.

There are two types of text in a video:

- Natural text.
- Overlay text.

1) Natural text is the text that appears in the video when it is recorded. These texts are part of the scene where the video is recorded. Example: nameplate number, text on a man's t-shirt, license plate number (Fig. 7).

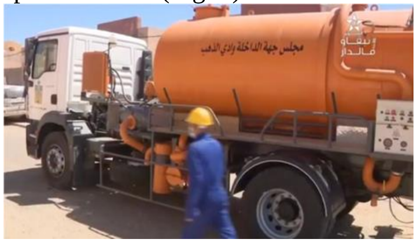


Fig. 7. The Name Appears in a Drinking Water Cistern.

2) Overlaid text is the text that was not part of the video when it was recorded, but is overlaid to give additional information about that particular scene. Example: text appearing in the Lower third of TV shows (Fig. 8).

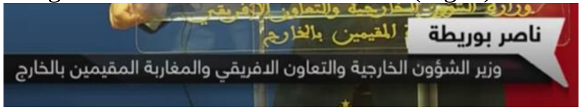

ناصر بوريطة
وزير الشؤون الخارجية والتعاون الافريقي والمغاربة المقيمين بالخارج

Fig. 8. Example of Lower Third.

Natural text is not of much use because it contains less important information, but overlay text contains information that is of great importance. Therefore, the main objective of the proposed system is to detect the superimposed text appearing in the video.

News Lower third is static or scrolling depending on the choice of news channel. Both variants have their own challenges. Static Lower third have a fixed space on the video frame to accommodate all text. News channels decrease font size to accommodate more text in some cases. Some channels also perform horizontal compression of text within the frame.

In the context of news videos, the space allocated to the Lower third is fixed and generally designed to accommodate most ligatures, but some are complex.

D. Material and Method

Lower third Extraction, has been refined an implementation of the efficient and accurate scene text detector (Tesseract) using the Python programming language [34], on a video to be able to return the bounding boxes of all text in a frame image (Fig. 9).

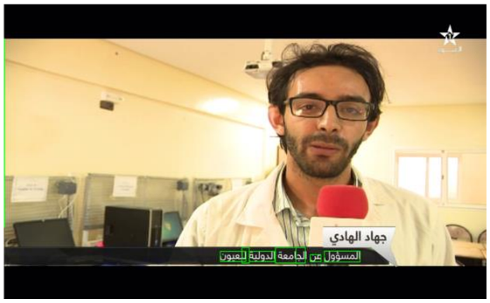


Fig. 9. Extraction using a Text Detector (Tesseract).

Python has very strong community support with many useful packages and libraries. It is also one of the most popular programming languages for data science, machine learning, artificial intelligence, and scientific computing in general.

The Tesseract engine supports multilingual text recognition [35] [36]. However, recognizing cursive scripts using Tesseract is a difficult task, the Tesseract engine is analyzed and modified for recognition of Arabic writing style. The original Tesseract system has accuracies of 65.59% and 65.84% for 14 and 16 font sizes respectively, while the modified system, with reduced search space, yields accuracies of 97.87% respectively, and 97.71%, this algorithm uses the characteristic of the densities of special symbols in each line of text, which is calculated using the built-in character classifier in Tesseract [37].

III. Text Recognition and Extraction

A. Segmentation

This step is applied to each Frame or image, using the OCR algorithm, the text box is detected. The detected text regions are then refined to increase the efficiency of text extraction. The effectiveness of text detection depends on the font color, text size, background color, and video resolution.

B. Classification

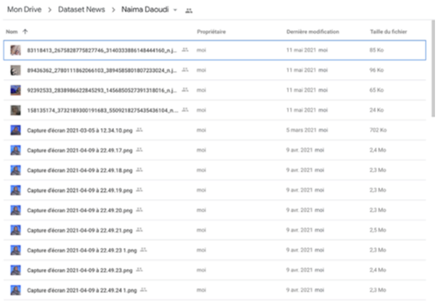
Classification is a form of data analysis that extracts patterns describing classes of data, these patterns are called classifiers, and they predict class labels. In this step, the system must make a decision based on the used algorithm. This analysis can help us better understand the data set.

C. Competition Dataset

In this section, he discusses in detail the process of preparing the dataset and the rigorous measures that were taken to ensure maximum diversity throughout in the dataset, through the collection of newsreel video recordings from the source of the National Radio and Television Company (SNRT), in High Definition (HD). At the face recognition level, there is a Database (Dataset) of 7 TV news anchors (Fig. 10), where the system will process these 7 faces to detect them in the video.

The storage of the extracted information for each video is ensured in a separate file bearing the title of the video in the form of a text file, with a small volume.

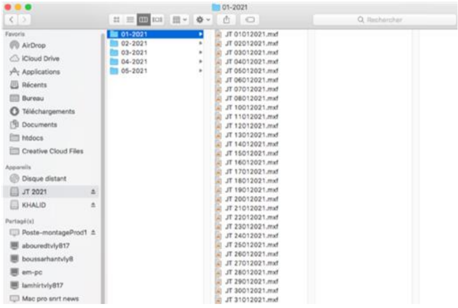
The videos are stored in a storage server, each video bears the name of the log "newsfeed" + the date (Fig. 11).



Mon Drive > Dataset News > Naima Daoudi

Nom ↑	Propriétaire	Dernière modification	Taille du fichier
83118413_2675828775827746_3140333886148444160_n.j...	moi	11 mai 2021 moi	85 Ko
89436362_2780111862066103_3894585801807233024_n.j...	moi	11 mai 2021 moi	96 Ko
92392533_2838986622845293_1456850527391318016_n.j...	moi	11 mai 2021 moi	65 Ko
158135174_3732189300191683_5509218275435436104_n...	moi	11 mai 2021 moi	24 Ko
Capture d'écran 2021-03-05 à 12.34.10.png	moi	5 mars 2021 moi	702 Ko
Capture d'écran 2021-04-09 à 22.49.17.png	moi	9 avr. 2021 moi	2,4 Mo
Capture d'écran 2021-04-09 à 22.49.18.png	moi	9 avr. 2021 moi	2,3 Mo
Capture d'écran 2021-04-09 à 22.49.19.png	moi	9 avr. 2021 moi	2,3 Mo
Capture d'écran 2021-04-09 à 22.49.20.png	moi	9 avr. 2021 moi	2,3 Mo
Capture d'écran 2021-04-09 à 22.49.21.png	moi	9 avr. 2021 moi	2,5 Mo
Capture d'écran 2021-04-09 à 22.49.23 1.png	moi	9 avr. 2021 moi	2,3 Mo
Capture d'écran 2021-04-09 à 22.49.23.png	moi	9 avr. 2021 moi	2,4 Mo
Capture d'écran 2021-04-09 à 22.49.24 1.png	moi	9 avr. 2021 moi	2,3 Mo

Fig. 10. Training Dataset for Face Detection.



01-2021
02-2021
03-2021
04-2021
05-2021

JT 01012021.mxf
JT 02012021.mxf
JT 03012021.mxf
JT 04012021.mxf
JT 05012021.mxf
JT 06012021.mxf
JT 07012021.mxf
JT 08012021.mxf
JT 10012021.mxf
JT 11012021.mxf
JT 12012021.mxf
JT 13012021.mxf
JT 14012021.mxf
JT 15012021.mxf
JT 16012021.mxf
JT 17012021.mxf
JT 18012021.mxf
JT 19012021.mxf
JT 20012021.mxf
JT 21012021.mxf
JT 22012021.mxf
JT 23012021.mxf
JT 24012021.mxf
JT 25012021.mxf
JT 26012021.mxf
JT 27012021.mxf
JT 28012021.mxf
JT 29012021.mxf
JT 30012021.mxf
JT 31012021.mxf

Fig. 11. News Video Clips Stored in Hard Drive.

The advantages of this system:

- Facilitates the task of the archivist and makes the data search and classification simpler.
- Generates a complete newscast information file for extraction. The disadvantage of this system
- Must have a robust computer capable of mass computing and processing information.
- Very slow extraction due to processing each frame for text and face recognition at the same time.

IV. Results and Discussion

The extraction of the Lower third was however more complex. It is useful to know the fixed position and speed of the Lower third output for each part. Using this information, the algorithm extracts the two parts of the recognition to prepare each frame received from the video file in MXF format for the news anchor face localization stage (Fig. 12).

Start of the video with the main presenter (start of the report 0) ------------- → report 0 --- ----------- → senior reporter start report 1 ---------- → report 1 ------------- → senior reporter start report 2 -------------- → report 2 ------------- →
.
.

.
.

senior reporter start report n -------------- → report n ------------- → start the conclusion (end of images) -------------- → conclusion ------------- → end

After entering the correct addresses (Fig. 13), you will have access to the reception platform which consists of:

- The "Choose a file" button to load the selected video into a disk.
- The "Download video" button to start the extraction.

In this Application, two techniques have been implemented; face detection and text detection.

Before launching data extraction, it is first necessary in the first time to use this system, to train the news presenters in order to generate a PICKLE type file.

In this case, 900 photos from these seven news presenters are used.

Once the news presenters to be recognized does not correspond to any person in the file (PICKLE), a message will be displayed to indicate that this person is unknown, this message will also appear in the final result file and also a copy of the images in the "image" folder. If the news anchor you want to recognize exists in database, the application will display his first and last name.

The objective of this final report is to filter useful information for the archiving service in order to recognize the news anchor, the people, the number of reports and the employees who produced the television news.

Given the amount of potential software in the audiovisual field that can be based on this type of application, this software must meet the requirements of speed and robustness of results.

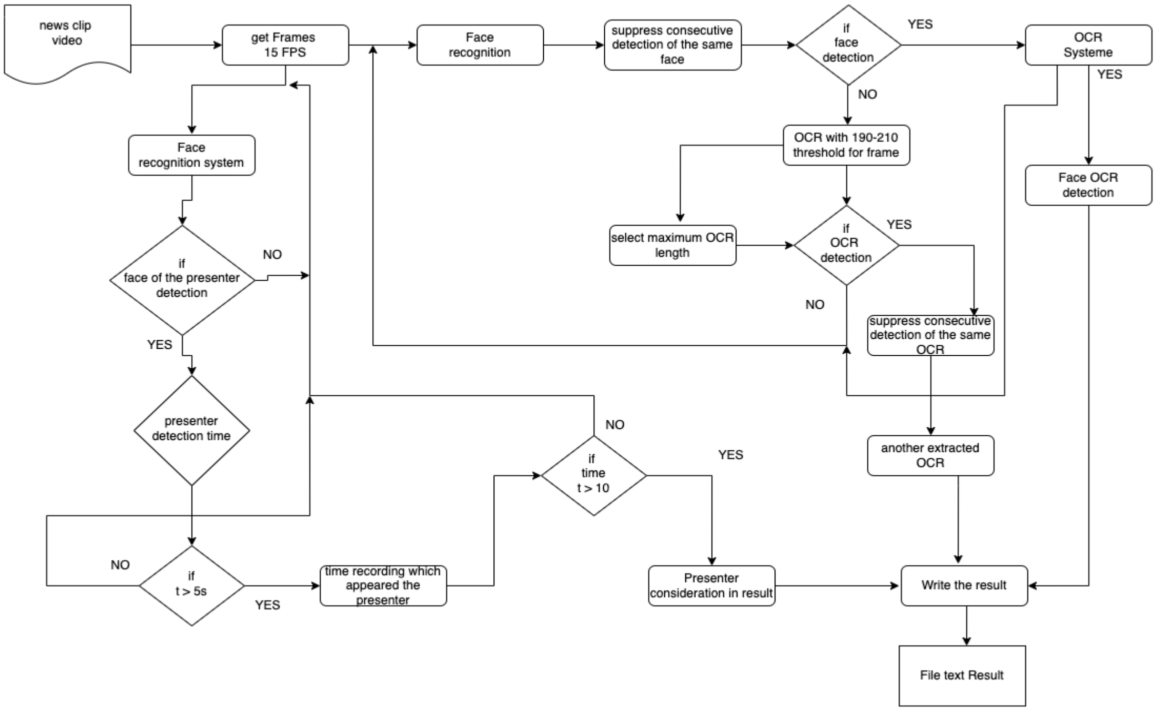

news clip video
get Frames 15 FPS
Face recognition
suppress consecutive detection of the same face
if face detection
YES
OCR Systeme
YES
NO
OCR with 190-210 threshold for frame
Face OCR detection
Face recognition system
select maximum OCR length
if OCR detection
YES
NO
if face of the presenter detection
NO
suppress consecutive detection of the same OCR
YES
presenter detection time
NO
if time t > 10
YES
another extracted OCR
NO
if t > 5s
YES
time recording which appeared the presenter
Presenter consideration in result
Write the result
File text Result

Fig. 12. Facial Recognition and OCR Home Interface.

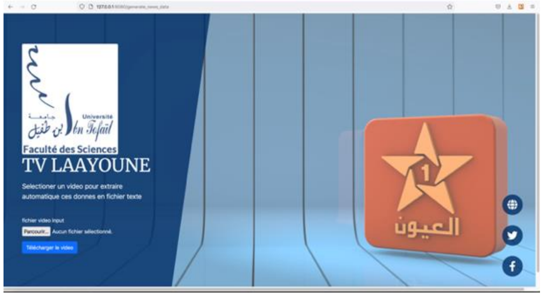
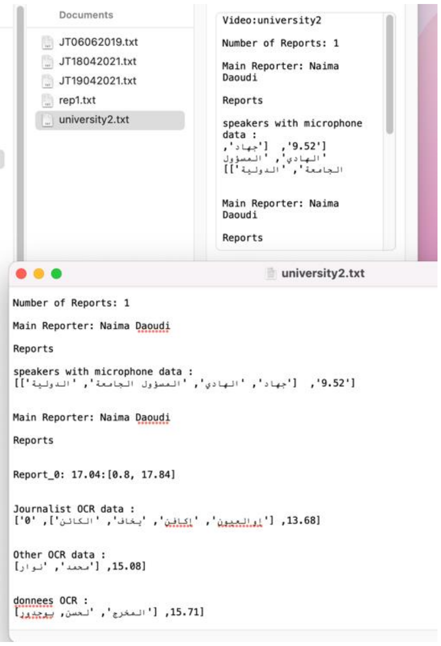

Université Ibn Tofail
جامعة ابن طفيل
Faculté des Sciences
TV LAAYOUNE
Selectioner un video pour extraire automatique ces donnes en fichier texte
fichier video input
Parcourir... Aucun fichier sélectionné.
Télécharger le video
العيون
1

Fig. 13. Facial Recognition and OCR Home Interface.

Finally, a custom web application is created (Fig. 14). The user interface (frontend) is implemented using HTML, CSS and JavaScript. The (backend) of this test application was built using Python with Flask [38], a library for building APIs, combined with the packages and scripts needed to implement OCR and facial recognition. The use of this web application is allowed to get an idea of the video treatments that can be processed successfully.



```
Video:university2

Number of Reports: 1

Main Reporter: Naima
Daoudi

Reports

speakers with microphone
data :
,'جهاد']  ,'9.52']
'الهادي' ,'المسؤول
الجامعة' ,'الدولية']]


Main Reporter: Naima
Daoudi

Reports
```

```
Number of Reports: 1

Main Reporter: Naima Daoudi

Reports

speakers with microphone data :
[['الدولية' ,'المسؤول الجامعة' ,'الهادي' ,'جهاد']  ,'9.52']


Main Reporter: Naima Daoudi

Reports


Report_0: 17.04:[0.8, 17.84]


Journalist OCR data :
['0' ,['الكائن' ,'بخاف' ,'إكافن' ,'إوالعيون'] ,13.68]


Other OCR data :
[نوار' ,'محمد'] ,15.08]


donnees OCR :
[بوجدور, لحسن' ,'المخرج'] ,15.71]
```

Fig. 14. Extract Result in a System Report.

V. Conclusion

In this paper, a prototype audio-visual archiving system based on face and text detection in Arabic language is designed and implemented. Experimentation shows that the overall recognition accuracy is greater than 88%.

On the other hand, the majority of research works related to video presents a variety of approaches concerning different domains and processes video in general, but does not consider integrating both face and text recognition into a single model. This is therefore the first objective to be attained through this paper.

This approach ensures that human resources help achieve the desired objective by automating the required filing system fixed by managers. This study shows that this problem is very complex. For this, it was used the algorithm of SVM as well as CNN and library Pytesseract at the same time to solve this problem by the good choice of the parameters and the tolerance coefficient. The optimal solution obtained makes it possible to validate the proposed approach but has a problem of program execution time given the complexity of video processing and computer processors. In future work, he will study similarity semantics for image video text classification. Additionally, it will try to build a bigger model for different languages.

Acknowledgment

The authors would like to acknowledge the support of the General Administration of the National Society of Moroccan Radio and Television (http://www.snrt.ma/), as well as the Faculty of Sciences of Ibn Tofail University (https: //fs.uit.ac.ma/).

References

[1] A. A. Shahin, “Printed Arabic Text Recognition using Linear and Nonlinear Regression,” 2017. [Online]. Available: www.ijacsa.thesai.org

[2] A. Mittal, P. P. Roy, P. Singh, and B. Raman, “Rotation and script independent text detection from video frames using sub pixel mapping,” J Vis Commun Image Represent, vol. 46, pp. 187–198, Jul. 2017, doi: 10.1016/j.jvcir.2017.03.002.

[3] Josef Chaloupka, A prototype of Audio-Visual Broadcast Transcription System. 2019.

[4] M. Rashad and N. A. Semary, “CCIS 488 - Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers,” 2014.

[5] A. Kumar Bhunia, G. Kumar, P. Pratim Roy, and R. Balasubramanian, “Text Recognition in Scene Image and Video Frame using Color Channel Selection”, doi: 10.1007/s11042-017.

[6] Institute of Electrical and Electronics Engineers, Institute of Electrical and Electronics Engineers. Delhi Section, and I. INDIAcom (Conference) (14th : 2020 : New Delhi, Machine Reading of Arabic Manuscripts using KNN and SVM Classifiers. 2020.

[7] M. Amara, K. Zidi, S. Zidi, and K. Ghedira, “CCIS 488 - Arabic Character Recognition Based M-SVM: Review,” 2014.

[8] C. Garcia and Z. Saidane, “Automatic Scene Text Recognition using a Convolutional Neural Network Metric Learning and Siamese Neural Networks View project PhD Thesis-Toward unsupervised activity monitoring with sequence metric learning View project Automatic Scene Text Recognition using a Convolutional Neural Network,” 2007. [Online].Available:https://www.researchgate.net/publication/251423608

[9] Ahmed H. Metwally, Mahmoud I. Khalil, and Hazem M. Abbas, “Offline Arabic handwriting recognition using Hidden Markov Models and Post-Recognition Lexicon Matching,” 2017.

[10] “Optical Character Recognition (OCR) of Arabic Characters,” https://ukdiss.com/examples/optical-character-recognition.php, 2018.

[11] J. Chaloupka, “Audio-Visual TV Broadcast Signal Segmentation,” in Advances in Intelligent Systems and Computing, 2020, vol. 1061, pp. 221–228. doi: 10.1007/978-3-030-31964-9_21.

[12] K. Zagoris, I. Pratikakis, A. Antonacopoulos, B. Gatos, and N. Papamarkos, “Distinction between handwritten and machine-printed text based on the bag of visual words model,” in Pattern Recognition, Mar. 2014, vol. 47, no. 3, pp. 1051–1062. doi: 10.1016/j.patcog.2013.09.005.

[13] Intel corporation, “Open Source Computer Vision Library,” 2021.

[14] M.-H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting Faces in Images: A Survey,” 2002.

[15] S.-C. Hsu, I.-C. Chen, and C.-L. Huang, “Image Classification Using Naive Bayes Classifier With Pairwise Local Observations,” XXXX-XXXX, 2017.

[16] A. Štulienė and A. Paulauskaitė-Tarasevičienė, “Research on human activity recognition based on image classification methods,” 2017.

[17] C. H. Qian, H. Q. Qiang, and S. R. Gong, "An Image Classification Algorithm Based on SVM," Applied Mechanics and Materials, vol. 738–739, pp. 542–545, Mar. 2015, doi: 10.4028/www.scientific.net/amm.738-739.542.

[18] S. Amassmir, S. Tkatek, O. Abdoun, and J. Abouchabaka, "An intelligent irrigation system based on internet of things to minimize water loss," Indonesian Journal of Electrical Engineering and Computer Science, vol. 25, no. 1, pp. 504–510, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp504-510.

[19] R. Dahmani, A. Belmzoukia, S. Tkatek, and A. Aït Fora, "Automatic slums identification around normal and smart cities: using Machine-learning on VHR Satellite Imagery," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 5, p. 9071-9079, Oct. 2020, doi: 10.30534/ijatcse/2020/312952020.

[20] E. Gyimah and D. K. Dake, "Using Decision Tree Classification Algorithm to Predict Learner Typologies for Project-Based Learning," in Proceedings - 2019 International Conference on Computing, Computational Modelling and Applications, ICCMA 2019, Mar. 2019, pp. 130–134. doi: 10.1109/ICCMA.2019.00029.

[21] H. Guan, J. Yu, J. Li, and L. Luo, "RANDOM FORESTS-BASED FEATURE SELECTION FOR LAND-USE CLASSIFICATION USING LIDAR DATA AND ORTHOIMAGERY," 2012.

[22] A. Sharif, R. H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features off-the-shelf: an Astounding Baseline for Recognition," 2014.

[23] S. Dey, R. Bhattacharya, F. Schwenker, and R. Sarkar, "Median filter aided CNN based image Denoising: An ensemble Aprroach," Algorithms, vol. 14, no. 4, Apr. 2021, doi: 10.3390/a14040109.

[24] J. Chaloupka, K. Palecek, P. Cerva, and J. Zdansky, "Optical character recognition for audio-visual broadcast transcription system," in 11th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2020 - Proceedings, Sep. 2020, pp. 229–232. doi: 10.1109/CogInfoCom50765.2020.9237867.

[25] E. Eidinger, R. Enbar, and T. Hassner, "Age and Gender Estimation of Unfiltered Faces," 2013. [Online]. Available: http://www.adience.com

[26] T. Hassner, S. Harel, E. Paz, and † Roee Enbar, "Effective Face Frontalization in Unconstrained Images," 2015. [Online]. Available: www.openu.ac.il/home/hassner/projects/frontalize

[27] G. Azzopardi, A. Greco, A. Saggese, and M. Vento, "Fusion of Domain-Specific and Trainable Features for Gender Recognition from Face Images," IEEE Access, vol. 6, pp. 24171–24183, Apr. 2018, doi: 10.1109/ACCESS.2018.2823378.

[28] G. Levi and T. Hassner, "Age and Gender Classification using Convolutional Neural Networks," 2015. [Online]. Available: www.openu.ac.

[29] M. Afifi and A. Abdelhamed, "AFIF4: Deep Gender Classification based on AdaBoost-based Fusion of Isolated Facial Features and Foggy Faces," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.04277

[30] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition," Mar. 2016, [Online]. Available: http://arxiv.org/abs/1603.01249

[31] Jos van de Wolfshaar, Mahir F. Karaaba, and Marco A. Wiering, Deep Convolutional Neural Networks and Support Vector Machines for Gender Recognition. 2015.

[32] G. Özbulak, Y. Aytar, and H. K. Ekenel, "How transferable are CNN-based features for age and gender classification?," 2016.

[33] S. Tkatek, A. Belmzoukia, S. Nafai, J. Abouchabaka, and Y. Ibnou-Ratib, "Putting the world back to work: An expert system using big data and artificial intelligence in combating the spread of COVID-19 and similar contagious diseases," Work, vol. 67, no. 3, pp. 557–572, 2020, doi: 10.3233/wor-203309.

[34] https://www.python.org/, "Python programming language," [Online], 2021.

[35] Q. U. A. Akram, S. Hussain, A. Niazi, U. Anjum, and F. Irfan, "Adapting tesseract for complex scripts: An example for Urdu Nastalique," in Proceedings - 11th IAPR International Workshop on Document Analysis Systems, DAS 2014, 2014, pp. 191–195. doi: 10.1109/DAS.2014.45.

[36] R. Smith, "An Overview of the Tesseract OCR Engine," 2007. [Online]. Available: http://code.google.com/p/tesseract-ocr.

[37] Z. Liu and R. Smith, "A simple equation region detector for printed document images in tesseract," in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2013, pp. 245–249. doi: 10.1109/ICDAR.2013.56.

[38] Miguel Grinberg, Flask Web Development, 2nd Edition, vol. O'Reilly Media, Inc. 2018.