Journal of Computer Science

Original Research Paper

Classification Program and Story Boundaries Segmentation in TV News Broadcast Videos via Deep Convolutional Neural Network

Mounira Hmayda, Ridha Ejbali and Mourad Zaied

RTIM: Research Team in Intelligent Machines, University of Gabes, National Engineering School of Gabes (ENIG), Tunisia

Article history
Received: 24-01-2020
Revised: 13-04-2020
Accepted: 08-05-2020

Corresponding Author:
Mounira Hmayda
RTIM: Research Team in Intelligent Machines, University of Gabes, National Engineering School of Gabes (ENIG), Tunisia
Email: hmayda.mounira@gmail.com

Abstract: Given the amount of video information on the net, the user has had difficulty finding the information in a reasonable amount of time. Thus, all video content must be segmented and annotated so that he/she can access the information directly. The goal of the proposed approach is to allow a better exploitation of video by multimedia services (TV-On-Demand, catch-up TV), social community and video-sharing platforms (Youtube, Facebook…). In this work, an approach to classify TV programs and story boundaries segmentation in TV news broadcast video using Deep Convolutional Neural Network (DCNN) is presented. The first step is to extract features from video. This characteristics will modeled as video corpus governings the organization of TV stream content. This organization is carried out on two levels. The first consists in the identification of anchorperson by Single-Linkage Clustering through CNN faces and the second level aims to identify the story of news program due to the large audience because of the pertinent information they contain. In addition, we implement a 360-h broadcast video dataset obtained from five French news channels with ground-truth marked semantic shot categories, program genres and story boundaries. Experiments on this dataset prove the relevance of our approach for news broadcast video segmentation.

Keywords: Anchorperson, Clustering, Deep Learning, News Program, AlexNet CNN, Convolutional Neural Network

Introduction

The amount of audiovisual material broadcast daily has increased enormously in the last two decades. For example, nowadays in France, Digital Terrestrial Television (DTT) offers 32 television channels (excluding local programs) broadcasting content 24 h a day totaling 768 h of broadcast per day. Some packages of satellite TV channels broadcast more than 1000 channels or more than 24,000 h of content each day. Finally, YouTube site declares that more than 24 h of videos are put online each minute, i.e., more than 34,560 h of daily additional content.

A television stream is an audiovisual stream, i.e., an uninterrupted series of images and sounds produced by a television channel. It may also be accompanied by some description information, called metadata, provided by the chain. This metadata is either broadcast with the feed or available on the internet. Our problem is to automatically and accurately identify, in such a TV stream, the beginning, end and title or category of each game, newspaper, magazine, film, documentary, advertising, trailer, etc. The present work belongs to the general domain of structuring TV streams.

We aim to delimit the boundaries of the diffused elements. However, different levels of granularity in flow structuring exist. Indeed, some broadcast elements may themselves belong to a larger element, such as a weather report inserted in a magazine. Disseminated items can also be grouped by theme. Some elements may have a clean structure; For example, newscasts consist of anchorpersons and reports that can also be grouped by theme or by news story. In addition, some elements such as sponsorship or trailer for a next broadcast are linked (Fig. 1). In this context, television on demand is targeted, therefore the elements diffused should considered according to two notions: the inter-programs and the programs.

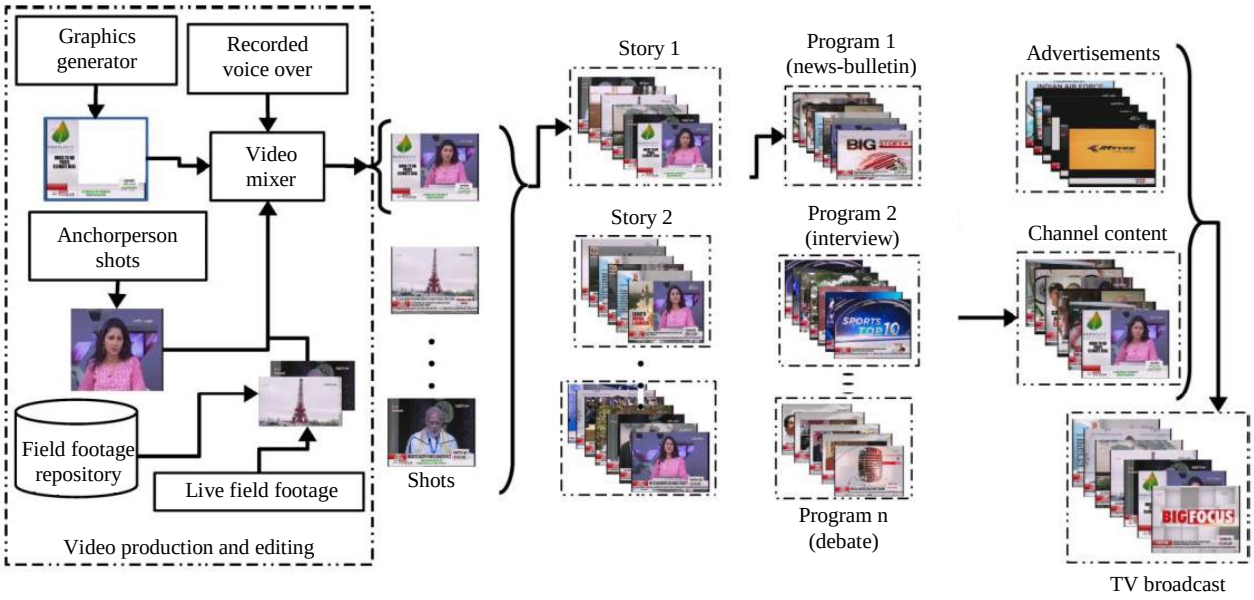


Fig. 1: broadcast video production pipeline

Due to the presence of large quantities of content videos, effective and rapid access to multimedia information has become a difficult endeavor. This difficulty has created a growing demand for appropriate methods to permit quick access to the content of unstructured video contents that require automatic video segmentation and indexing methods. The inter-segmentation of video streams is an essential step for a multimedia indexing system; However, for it to be robust, it must deal with the internal structuring of the programs to promote its audiovisual content.

Hence the choice to focus particularly on the analysis and automatic identification of television news structure which is justified by many reasons: First, the emergence of mass-communication appliances and amenities is crucial because of the ubiquitousness of huge amounts of numerical videos. In fact, TV streams represent the key source of enormous amounts of video-based content. As a result, the usage of TV streams is becoming a staple in everyday life which has led to a growing substantial financial burden. Furthermore, many applications aim at the proper structuring of TV stream. Indeed, the identification of television programs could be used in establishing mechanisms for the control of broadcasting of television channels.

Finally, each story can be segmented into one or several shots. Video shot segmentation is largely discussed in the literature. In addition, there are well developed approaches for detecting shot boundaries of edited videos (e.g., news broadcast) (Smeaton *et al*., 2010). This work focuses on the classification of channel content into programs using auto-encoder technique and segmentation of story boundaries in TV news broadcast video. In what follows, we give a brief overview of TV news broadcast structure.

Structure of TV News Broadcast

Two main reasons may justify the choice of news programs treatment. First, thanks to the important content they have, news programs are produced for and followed by a large number of TV viewers who get easy access to them as they are frequently published on the web. Thus, they have become a means of communication.

In fact, news topics retrieval from large databases is the focus of some systems. For example, (Wu *et al*., 2010) suggested a system for dealing with the different stories of the news assessed on large-scale broadcast video database which includes pre-segmented topics.

However, this kind of approaches must go through automatic segmentation as a first step. Therefore, their performance is strongly dependent on the quality of the results of the segmentation step. This represents one of the many reasons behind proposing an efficient method of news topic segmentation. Moreover, in spite of the varieties of styles of different TV channels, they do adopt the same production rules, which help to provide significant cues to analyze their content automatically (Goyal *et al*., 2009), (Misra *et al*., 2010). At this stage, the present work opted for standardized features of TV news programs since some approaches are based on predetermined features of these programs. For instance, the technique proposed in (Poulisse *et al*., 2010) is only applied to the BBC channel news. Therefore, it uses

keywords such as labeled entities in English language only. Similarly, the work of (Xie *et al*., 2011) faces the same problem as they use features in the form of sub words of the Chinese language. Thus, their investigation can only be applied to restricted categories of news, yet it cannot be efficient for others. Hence the choice to apply the principles and visual features found in any kind of TV news regardless of its language.

News has an operating structure defined as a syntactic rule that governs the organization of the content. This rule is founded on the studio/Topic notion (Fig. 2), where the anchorperson is the main actor who announces the changes in topics. The Topic is the demonstration of what the anchorperson says in the studio. Each subject usually has an average duration between 1 and 3 min (Misra *et al*., 2010) and is preceded by the appearance of an anchorperson.

Proposed Approach

Video dissection is considered to be an indispensable phase in all applications of video exploitation. Nevertheless, despite their satisfactory outcomes, their performance is based on the sort of video (news, action films, documentaries, sports program, etc.,). This can be attributed mainly to the recent techniques that are based on specific video production rules. For instance, the method introduced in (Berrani *et al*., 2008) consists in detecting recurrent sequences in a video. This technique ensures the identification of Inter-Programs (IP), such as advertisements, jingles, credits, etc., which allows the TV broadcast segmentation and useful program extraction. It is based on a technique of micro-clustering method which puts comparable audio/video characterestic vectors together.

Zlitni and Mahdi (2010) developed a technique for the identification of TV programs through two stages: the first consists of video grammars used as a benchmark catalog. The second consists in identifying TV programs by investigating the resemblance of video signals to benchmark video grammars (Zlitni and Mahdi, 2010; Zlitni *et al*., 2015).

However, in the present work, we propose an automatic approach to determine the TV stream. The novelty of the technique lies in using the sparse auto-encoder as an extractor of characteristics corresponding to visual jingles for training the classifier of a video category. Feature extraction directs the generics, which mark the beginning of TV programs, structuring and determination.

Moreover, the second proposal consists in an approach of structuring news content based on segmentation into stories. This approach is based on two major steps: First; the identification of the anchorperson by Single-Linkage Clustering through the CNN faces; Then, the segmentation of news into stories using AlexNet CNN. The first step consists in modeling the indices by image processing techniques, whereas the second relies on the direct exploitation of the features resulting from the first step to segment the content of news into different stories.

The proposed technique is illustrated in Fig. 3. We have also created a TV news broadcast dataset to validate the proposed technique. This dataset has been acquired from five French news channels and consists of 360 h of broadcast. After removing advertisements, the non-commercial contents of these videos are marked for ten semantic shot categories, program and story boundaries. The proposed technique benchmarked on this dataset and compared with a few baseline algorithms.

The remainder of the paper is organized as follows: In section 2, briefly reviews the existing methods for program and story segmentation. The classification of a program in a TV stream by means of deep learning is described in section 3. Section 4 describes the second proposed technique for story boundary detection using AlexNet CNN. The dataset, the design of our experiments and performance analysis are presented in Section 5.

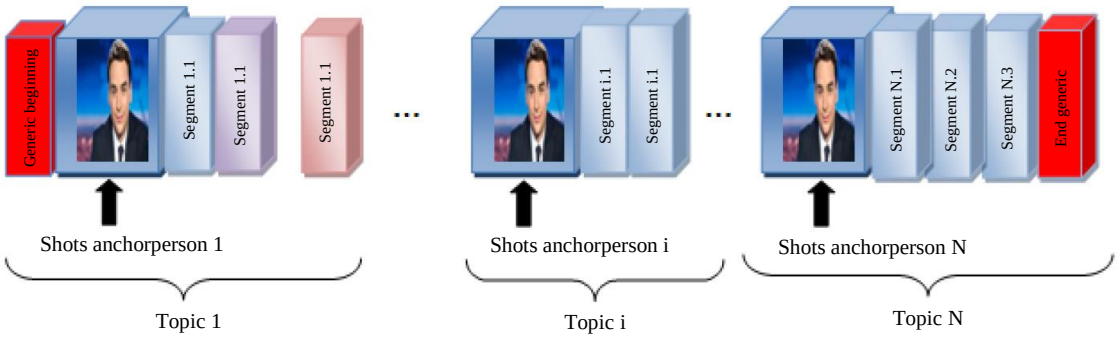


Fig. 2: Traditional structure of TV news

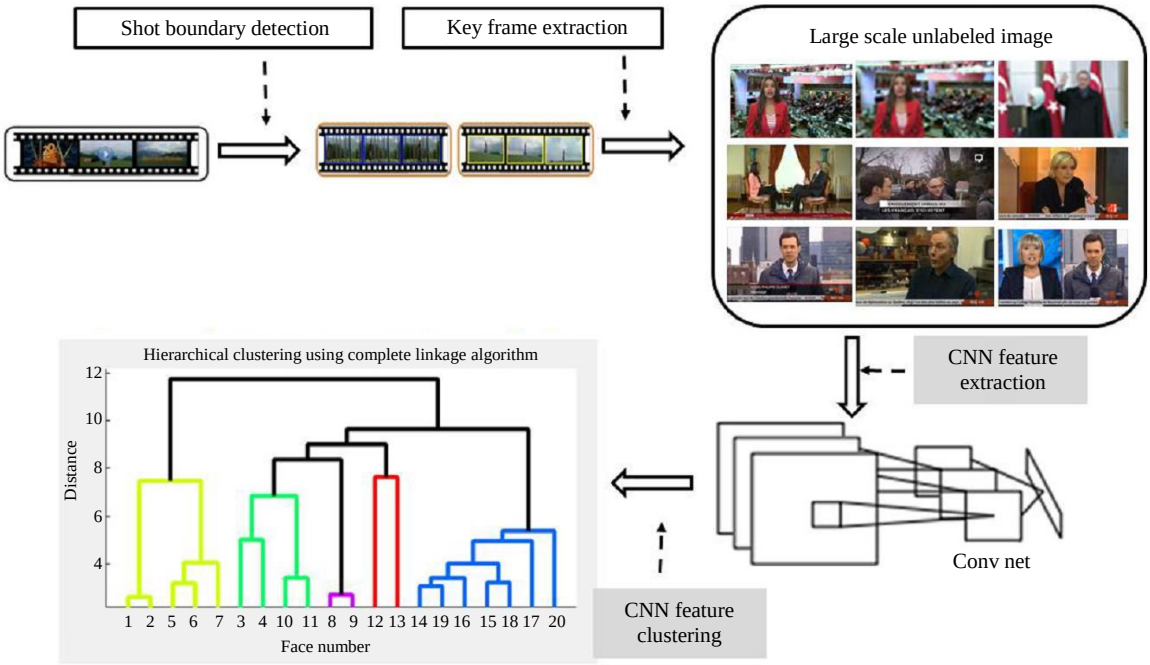


Fig. 3: Identification of anchorperson by Single-Linkage Clustering through CNN Features extraction

Related Works

Temporal segmentation of videos has been an extensively studied problem in the multimedia community. The definition of temporal video segmentation depends on the task and the source of the video. For example, in the case of videos obtained using surveillance or body worn cameras, the segmentation task consists in identifying a portion of the video in which a particular object is present (Shaban *et al*., 2017). The Segmentation task is defined as the identification of various scenes for movie videos (Iwan and Thom, 2017) Sport videos are segmented according to various events occurring during the game (Ullah *et al*., 2018). Lecture videos segmentation problem is formulated in terms of the topic of discussion (Shah and Zimmermann, 2017). However, in the case of television news broadcast, the problem is that videos have to be segmented into shots, stories or programs (Xu *et al*., 2013). Detailed review on these different problems is out of the scope of this work which only deals with works concerning the segmentation of TV news videos.

In previous works on TV news broadcast segmentation, for example in (Ibrahim and Gros, 2011) the detection of the repetitions can be considered as a key tool for stream structuring. After the detection stage, a classification method is applied to separate the repetitions of programs and breaks. Breaks repetitions, in their turn, are then used to classify the segments which appear only once in the stream. Finally, the stream is aligned with an Electronic Program Guide (EPG) in order to annotate the programs.

Besides, the approach introduced by Manson *et al.* (2009) is based on the detection of Inter-Programs (IPs), which have several shared characteristics, particularly, when shown repeatedly in the stream as it is much simpler to spot short audiovisuals than longer ones which are varied (series, documentary films, shows) and thus have no characteristics in common.

Furthermore, Weiming (2011) dealt with indexing and retrieval of content-based videos, emphasizing techniques for the analysis of video structure, including the detection of shot boundary, the extraction of key frame, the segmentation of scenes, extraction of features including static key frame features, object features and motion features, video data mining, video annotation, video retrieval including query interfaces, similarity measure and relevance feedback as well as video browsing.

The technique introduced in (Bingqing, 2015) tackled the issue of structuring unsupervised program with minimum prior information about the programs. This method aimed to identify multiple structures and deduce structural grammars for recurring TV programs of various sorts. It has three sub-problems: (i) the structural elements contained in programs are determined with minimal information about the type of elements it can present, (ii) multiple structures for the programs are identified and the programs structures are modeled and (iii) the structural grammar is generated for each respective structure.

The second group of program and story segmentation approaches rely on presentation styles used in the production of TV news. Presentation styles are characterized using a variety of features, such as anchor shots (Feng *et al*., 2014; Zlitni *et al*., 2015) video jingles (Zlitni *et al*., 2015; Feng *et al*., 2012) occurrence of the face in a special area of screen (Qu *et al*., 2004; Poulisse *el al*.,

2010), audio events (Browne *et al*., 2002; Meinedo *el al*., 2003), presence or absence of text at particular locations (Jindal *et al*., 2011; Ghosh *et al*., 2010), etc.

Video segmentation is the first important step in video content analysis. It aims at dividing the video stream into a set of meaningful and manageable segments (shots) that are used as basic elements for indexing (Weiming *et al*., 2011). In fact, video shots segmentation is the initial procedure of anchorperson detection that plays an important role in further video processing.

O'Hare *et al*. (2004) adopted a framework in which a news broadcast can be segmented into individual stories based on the location of the anchorperson shots within the program. The program is first segmented into individual shots and then a number of analysis tools are run on the program to extract the features representing each shot. The results of these feature extraction tools are then combined using a Support Vector Machine (SVM) trained to detect anchorperson shots.

A system based on partitioning of a news video into stories and the classification of the detected stories within a certain set of categories (world news, national news, sports, political news, weather, advertising, etc.,) was presented in (Colace *et al*., 2005). The system uses Markov chains and Bayesian networks for segmentation and topics classification. The whole analysis is carried out by exploiting information extracted from video and audio tracks using techniques of superimposed text recognition, speaker identification, speech transcription and anchorperson detection. The segmentation of news is based on feature recognition, such as anchorperson shots or interviews. The combination of that knowledge by a descriptor should be applied to recognize a change in topic.

The technique presented in (Misra *et al*., 2010) shows that the segmentation of the video stream into stories is achieved through the detection of anchorperson shots, then the text stream is divided into stories applying the approach of Latent Dirichlet Allocation (LDA). Goyal *et al.* (2009), however, suggested a frame-work for the segmentation of semantic stories on the basis of anchorperson detection. They opted for a mechanism of split-and-merge in order to detect topic boundaries. The approach is based on visual features and text transcripts.

Various techniques have been proposed in the literature; For example, Dumont and Quénot (2012) developed a sensor using Multiple Modalities for Systematic segmentation of stories for News Videos. This system is based on classification techniques and machine learning methods. It combines both audio descriptors (silence segments and parole) with visual features such as anchors or logos. Poulisse *et al*. (2010) proposed an approach based on multiple multimedia features to segmenting news video. which was inspired from the approach based on text segmentation. The authors opted for different methods to achieve topic segmentation with different approaches: Text, video, audio and layout features. At an advanced step of the analysis, those features were used to detect story breaks after training a maximum entropy classifier. To structure TV streams, Hmayda *et al*. (2017) presented an approach for TV stream programs identification by means of deep learning.

As for textual approaches, we can cite the work of Wang *et al*. (2006) which focuses on a multimodal fusion between visual and textual information for Story Segmentation and Concept Association and in which video subtitle is used to identify the most relevant concepts/topics addressed in each independent segment.

Program Shot Classification Using Deep Learning Approach

Auto-encoder for TV Program Classification

The efficiency of deep learning inspired us to use this learning principle in the recognition and classification of TV stream programs by means of Stacked Sparse Auto-Encoders (SSAE).

Auto-encoder is an feature learning algorithm which is unsupervised whose objective is to develop better feature representation of high-dimensional data input by identifying the correlation between the data. It is merely a multi-layer feedforward neural network that is trained to denote the input using back-propagation. The auto-encoder attempts to reduce the discrepancy between input and reconstruction by applying back- propagation as much as possible by learning an encoder and a decoder (Fig. 4).

The Basic Sparse Auto-Encoder

Given $X = (x^{(1)}, x^{(2)}. \ldots, x^{(N)})$ where $x^{(i)} \epsilon R^{dx}$, N is the number of training images and d_x is the number of pixels in each of the images.

$h^{(1)}(k) = \left(h_1^{(1)}(k), h_2^{(1)}(k), ..., h_{d_n}^{(1)}(k)\right)^T$ signifies the high-level feature learned on layer l for the k-th image and the number of hidden units in layer is presented with d_h. In this work, to introduce the hidden layer, the superscript is used; however, to present the unit of this layer, the subscript is opted for. For example, in Fig. 1 the i-th unit in the first hidden layer is presented by $h_1^{(1)}$ claiming that x is the input frame and $h_i^{(1)}$ is its representation in hidden layer l.

Figure 1 shows the construction of basic SAE. Generally, the input layer of the auto-encoder is an encoder that transforms input x into its corresponding representation h, while the hidden layer h, can be considered as a new feature representation of input data. The output layer is an effective decoder that is trained to reconstruct an approximative features $\hat{x}$ of the input from the hidden representation h. Principally, training an auto-encoder aims to obtain optimal parameters through minimizing the divergence between input x and its reconstruction $\hat{x}$.

The divergence is described using a cost function. The cost function of an SAE consists of the following three terms (Hassairi *et al*., 2015; Ejbali *et al*., 2012):

$$\delta_{SAE} = \frac{1}{N}\sum_{k=1}^{N}\left(L\left(x(k), d_{\hat{\theta}}\left(e_{\hat{\theta}}\left(x(k)\right)\right)\right)\right) + \alpha\sum_{j=1}^{n} KL\left(\rho \,\|\, \hat{\rho}_j\right) + \beta\|W\|_2^2 \quad (1)$$

Stacked Sparse Auto-Encoder (SSAE)

The SAE is a neural net that involves basic SAE 's multiple layers where the each layer's output is linked to the input of the consecutive layer. The SAE used in our work is composed of two auto-encoder layers and a softmax layer. Various auto-encoder layers are stacked together form an unsupervised pretraining stage from Layer one to Layer three. The latent representation obtained by an auto-encoder is used as the input to the last auto-encoder layer. After this phase of pretraining, a fine tuning using back propagation is used to improve the results. In this work, we construct two SSAE layers that involve two basic SAE. Figure 5 shows the architecture of SSAE. However, for simplicity reasons, the decoder parts of each basic SAE are ignored (Fig. 3).

Function *f*, which modifies an input raw pixel of a patch to a novel characteristic depiction $h^{(2)}$, is the result of the SSAE. In the input layer, which is actually the first layer, the input represents the raw pixel intensity of an image which is denoted as a column vector of 64*64 pixel intensity. The number of input units can be presented as 64*64 = 4096 in the input layer. There are 150 Hidden units in the first hidden layer compared to 100 hidden units in the second one.

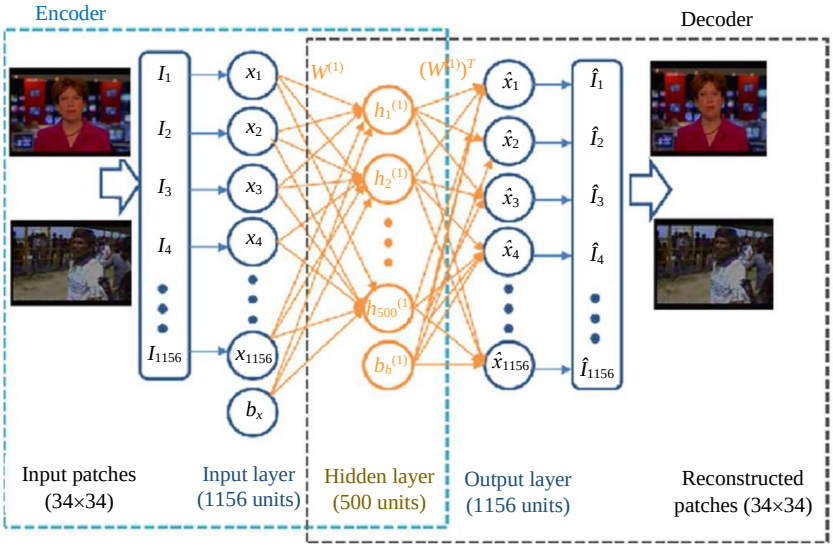


Fig. 4: Basic SAE architecture for TV program classification

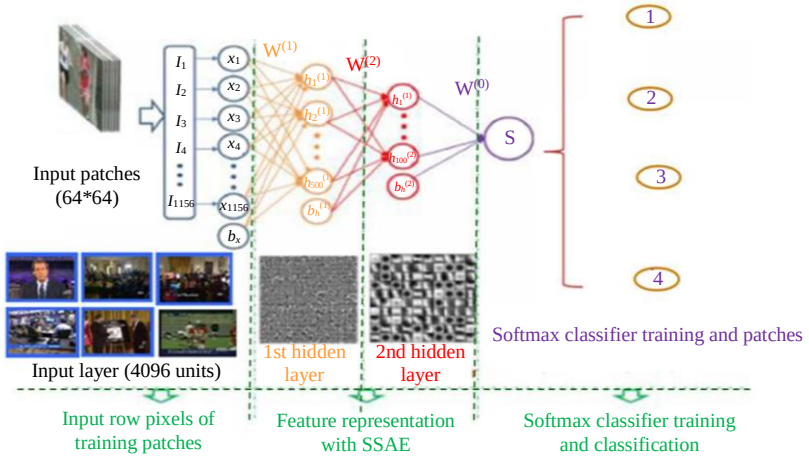


Fig. 5: The architecture of SSAE

Automatic Story Segmentation for News TV

At this stage, no prior information about the structural features of News Programs (NP) is provided. It is, therefore, essential to implement a fully automatic process for modeling this structure using image processing features. This modeling is realized in three steps: Step one is dedicated to shot boundary detection. Step two involves identifying shots where the face of the anchorperson appears using optical flow to extract the keyframe from these shots. The last step is based on unsupervised classification to identify the anchorperson by Single-Linkage Clustering through CNN Features extraction.

Shot Boundary Detection

Segmentation into shots is considered to be among the first contributions to the analysis and structuring of video by content as it is the starting point of any process of macro-segmentation of video content. Indeed, the units (shots) resulting from this segmentation constitute the inputs of any phase of content structuring and are also the basis for identifying other units with a semantic level of higher granularity, such as topics for the case of News.

The variety of segmentation into shots techniques proposed in the literature is strongly linked to the diversity of the types of changes (abrupt transition, fade, etc.,). Yuan *et al*. (2007) conducted an analytical study of the main approaches proposed in the literature. Based on this study, we opted in the present paper for the Edge Change Ratio (ECR) (Jacobs *et al*., 2004) technique based on its performances in shots detection. This method consists in changing the edges of objects (Fig. 6) in the frames across a border. In other words, structural discontinuity is accompanied by temporal visual discontinuity.

On the basis of this assumption, the technique starts with the calculation of the percentage of incoming and outgoing contours between two images.

Thus, the value of ECR (*n*, *k*) between the images *n*-*k* and *n* is calculated as in (Equation 2):

$$ECR(n,k) = \max\left(\frac{x_n^{in}}{\sigma_n}, \frac{x_{n-k}^{out}}{\sigma_{n-k}}\right) \quad (2)$$

where, σ_n is the number of edge pixels in the frame *n* and x_n^{in} and x_{n-k}^{out} are the entering and exiting edge pixels in frames *n* and *n* - *k*, respectively.

During the shots' detection, the ECR technique is used with $k = 10$ for a temporal distance of 10 images. For the detection of hard cuts, two values are calculated: The first is the near-far ratio which describes the ratio between the ECR values of two successive images (near ECR) and the ECR value between the current image and the 10th image (far ECR). The second one is the far last-far ratio which is the ratio between the current value of far ECR and the previous value of far ECR.

This phase consists in the assembly of all the shots composing the news. In order to delimit the topics, these shots then undergo two levels of filtering: The first aiming at extracting the keyframe from these shots and the second at identifying the anchorperson.

Key Frame Extraction Using Optical Flow

After segmentation of the news video sequence into shots, the next step is to retrieve the keyframes from these shots. The extraction of keyframes (representative frame) has an important influence on the performance of content multimedia, like the extraction of the keyframes in the video and representative view selection for objects (Gao *et al*., 2011). An anchorperson shot must last more than 2 s, thus those shots with a lifetime less than 2 s are rejected for anchorperson shot detection. In addition, the frames in one anchorperson shot should be highly similar, whereas there may be a bigger difference in news report shots for camera and object movement. Therefore, if big enough change occurs in one shot, it cannot be an anchorperson shot candidate. In an anchorperson shot, all frames are very similar (Fig. 7), whereas in the news report shot (Fig. 8), with camera movement, the change between different frames in the shot may be greater and can be detected by the difference between the first and the fourth frame based on optical flow to analyze the change in one shot.

Fig. 6: The edges in two consecutive video frames

Fig. 7: Anchorperson shot frames

Fig. 8: News report shot frames with object movement

The extraction of keyframes algorithm is described as follows:

Step 1: Test the length of each shot and cast those shots with a lifetime lower than 2s

Step 2: Calculate the moving area value between the first and the fourth frame positions for each shot using the differential method of Lucas and Kanade (Gnouma *et al*., 2016)

Optical flow algorithms estimate the deformations between two images. The basic assumption for the optical flow calculation is that pixel intensity is conserved. It is assumed that the intensity, or color, of

the objects has not changed significantly between the two images. Based on this idea, we have the following assumption:

$$I(x,y,t+1) = I(x+V_x, y+V_y, t) \tag{3}$$

where, $\vec{V}$ = (V_x,V_y) is the vector of velocity. Then, by derivation, we obtain the well known optical flow constraint equation:

$$\vec{\nabla} I.\vec{\Lambda} + \frac{\partial I}{\partial t} = 0 \tag{4}$$

where, $\vec{\nabla} I = \left(\frac{\partial I}{\partial x}, \frac{dI}{dy}\right)$ is the gradient of the image. The gradient constraint equation given in (2) is in two unknowns and cannot be solved. This is known as the aperture problem of the optical flow. To solve this aperture problem, Lucas and Kanade (Dhara and Saurabh, 2015) compute the optical flow on a point (u,v) considering that the motion is constant in a fixed neighborhood of this point. Let us note θ = (a,b) the vector parameters of motion, the Lucas-Kanade's method consists to search the velocity vector $\vec{V}$ the point (u,v) as the solution of:

$$\vec{V}(u,v) = \arg\max_{\theta} \sum_{x,y \in V_{u,v}} \left[\frac{\partial I}{\partial x}.a + \frac{\partial I}{\partial y}.b + \frac{\partial I}{\partial t}\right]^2 \tag{5}$$

To explain the representation of the route of the optical flow field, the vectors in the original image are superimposed speeds. We can also use the color map to represent the direction of flow as well as its intensity (Fig. 9).

In the color map, the velocity vectors are represented by the colors contained within the circle. Each vector is encoded by the color that indicates its origin at the center of the circle. The intensity varies from black to full color until a maximum speed to display white.

If the intensity of light between the first and fourth frames position (Fig. 10) is less than 30, then it can be concluded that this is an anchorperson shot, if not, i.e., if the value of the intensity is very important and higher then 50, it can be considered as a report shot.

Fig. 9: Real-time estimation of Optical Flow for key frame extraction. Intensity of white is very low to designate an anchorperson shot(top). The intensity of white is very important to indicate a report shot (bottom)

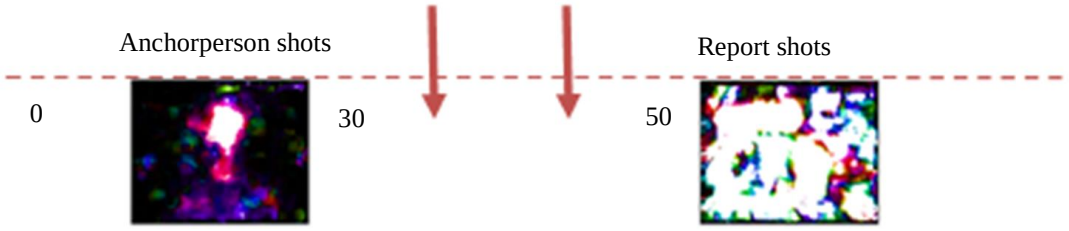


Fig. 10: Light intensity range

Identification of Anchorperson Shots

The main idea of our approach is to classify all the faces detected according to an unsupervised classification. Thus, the anchorperson face belongs to the cluster having the largest size.

To achieve this classification, we go through the step of extracting the features of all the faces, as inputs to the classification process, using the CNN technique.

Features Extraction Using Convolutional Neural Network

Deep learning refers to a set of automatic learning methods that are based on the artificial neural network. This new type of learning is used to model the data with high level of abstraction. Indeed, this technique provides a significant and rapid progress in fields of signal analysis, object recognition and computer vision. It is also based on the use of a set of non-linear processing layers for extracting and transforming features. Thus, each layer takes as input the output of the previous one. Deep learning is characterized by a multi-level learning of details or data representations, called levels of data abstraction. We find several architectures of deep learning, namely the CNN and the convolutive pre-trained neurons (the transfer learning). The latter regroups two types of classification that are fine-tuning and the automatic extraction of features. In the present work, we use the transfer learning technique for anchorperson classification and especially the fine tuning method. In what follows, we will detail the previously cited architectures of deep learning.

Classification Using Hierarchical Clustering Algorithm

CNN features vectors are given as input to Hierarchical Clustering Algorithm. That partitions the related dataset of key frame by building a hierarchy of clusters.

It uses the distance matrix requirements for clustering the key frame and constructs clusters step by step. At each step of the hierarchical clustering, the data are not partitioned into a particular cluster. It takes a sequence of partitions, which may run from a single cluster containing all objects to '*n*" clusters, each containing a single object.

In the present work, we have a set of *N* key frames (faces) to be clustered, thus these steps should be followed:

Step1: Start with assigning each face to a cluster so that for N faces, we now have N clusters each containing only one face. Let the distances between the clusters be the same as the distances between the faces they contain

Step2: Find the nearest pair of clusters and merge them into one cluster so that there is one cluster less

Step3: Calculate the distances (similarities) between the new cluster and each of the ancient clusters

Step4: Repeat steps 2 and 3 until all faces are clustered into one cluster of size *N*

Single-linkage clustering is used in step 3: We consider that the distance between one cluster and another one is equal to the shortest distance from any face in one cluster to any face in the other.

Algorithm of Single-Linkage Clustering:

```
L: Level of clustering
m: Sequence number
n: Number of clusters
Ci: a cluster, i belongs to {1,..,n}
r, s, j belong to {1, .., n}
D: The proximity matrix, D(i,j)=d(Ci,Cj)
Start
L(0) = 0, m = 0, min = 10**20 (very big number ~ infinity)
while (n not equal to 1):
    for i in range(n-1):
        for j in range(i+1,n):
                if d(Ci,Cj)<min:
                        min = d(Ci,Cj)
                        (r,s) = (i,j)
    m = m+1
    L(m) = d(Cr,Cs) k = r
    Ck = merge (Cr,Cs)
    for i in {1,..,n}:
        remove D(i,s)
        remove D(s,i)
  for i in {1,..,n-1}:
        D(k,i) = min(d(Cr,Ci),d(Cs,Ci))
        D(i,k) = D(k,i)
  n = n-1
```

After the hierarchy of clusters is provided, the optimal number of clusters is presented based on a data representation tree. Therefore, the cluster which contains the largest number of faces is considered an anchorperson. The data representation tree in blue (Fig. 11) is considered an anchorperson. Thus, here our base is labeled as an anchorperson and as not anchorperson.

News Story Segmentation Using AlexNet Convolutional Neural Network

According to the proposed approach, the structuring of News TV content by segmentation into topics is carried out according to two modes: Stand-alone mode or Rewire mode. The former is operated in off-line and is used when the NP we want to structure is presented for the first time as input to the system. In this case, all the steps of the learning phase of the structure are performed. Segmentation into topics is, therefore, an immediate use of the results provided at the output of the anchorperson identification step using the AlexNet CNN technique.

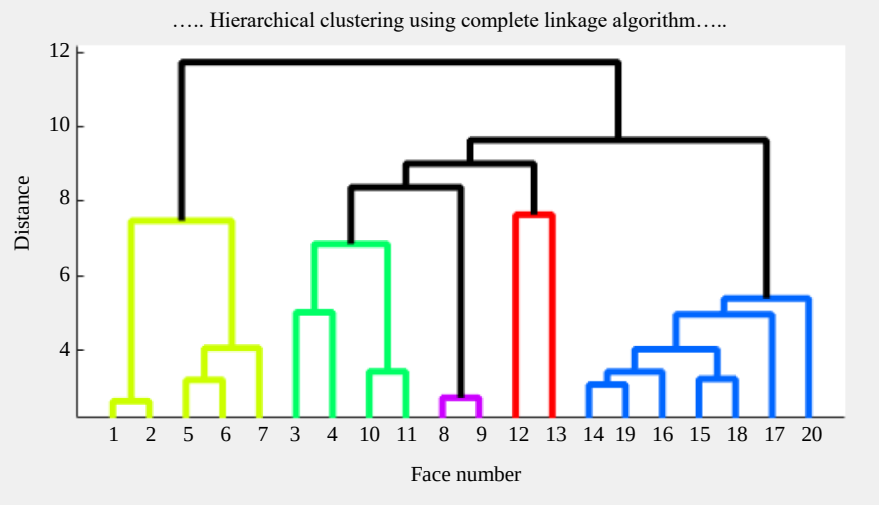


Fig. 11: Hierarchical clustering of images

This can be operated on-line. Indeed, in the case where the NP has already been processed by the system once, it is obvious that it is stored in a database. This database is composed of a set of anchorpersons. Thus, when the NP is presented a second time as input to the system, the frames of the anchorperson are deployed from the database and the problem of structuring is reduced to identify the anchorperson in the video stream. Our database is trained using AlexNet CNN to categorize two classes of our database. If the anchorperson is detected, a change of subject is reported.

AlexNet Architecture

Here, the AlexNet CNN deep learning architecture (Krizhevsky *et al*., 2012) is used for images classification in news videos. The network is profounder than the standard five convolution layer CNN with pursued by three maximum pooling layers. A decrease of 0.5% is carried out on the fully connected layers to circumvent data overfitting. The architecture encompasses the following components:

- 1 Convolution with 11×11 kernel size (1CONV)
- Rectified Linear Unit Layer Activation (RELU)
- Response Normalization Layer
- 1 Maximum Pooling (4×4 kernel)
- 2 Convolution with 5×5 kernel size (2CONV)
- Rectified Linear Unit Layer (RELU)
- Response Normalization Layer
- 2 Maximum Pooling (3×3)
- 3 Convolution with 3×3 kernel size (3CONV)
- Rectified Linear Unit Layer Activation (RELU)
- 4 Convolution with 3×3 kernel size
- Rectified Linear Unit Layer Activation (RELU)
- 3 Maximum Pooling (3×3)
- Fully Connected Layer (4096 nodes)
- Rectified Linear Unit Layer Activation (RELU)
- Fully Connected Layer (4096 nodes)
- Rectified Linear Unit Layer (RELU)
- Soft-max out

The proposed AlexNet CNN architecture is illustrated in Fig. 12. In the proposed method, image input layer is a pre-processing layer in which the input frames are down-sampled from 640×480 to 227×227 in terms of spatial resolution to minimize the calculation cost of the deep learning framework. The proposed system utilised five Convolutional (CONV) layers pursued by three Pooling Layers (POOL) and Rectified Linear Unit (RELU). A total of 96 kernels of relatively large size 11×11×3 are used for the first convolutional layer and 256 kernels of size 5×5 for the second convolutional layer. For the third, fourth and fifth layers, 384 kernels with size 3×3 have be employed. Each convolutional layer produces a feature map. The feature maps of the first, second and fifth convolutional layers are used in combination with pooling layers of 3×3 and stride of 2×2. The framework is composed of eight layered architectures with 4096 nodes. This is generates the trainable feature maps. These feature maps are exposed to Fully Connected (FC) layers then Soft-max activation is effected to identify the classification probabilities used by the final output classification layer. These classification probabilities in the Soft-max layer can generate categories of up to 1000 different classes, but in our dataset, we have only two classes.

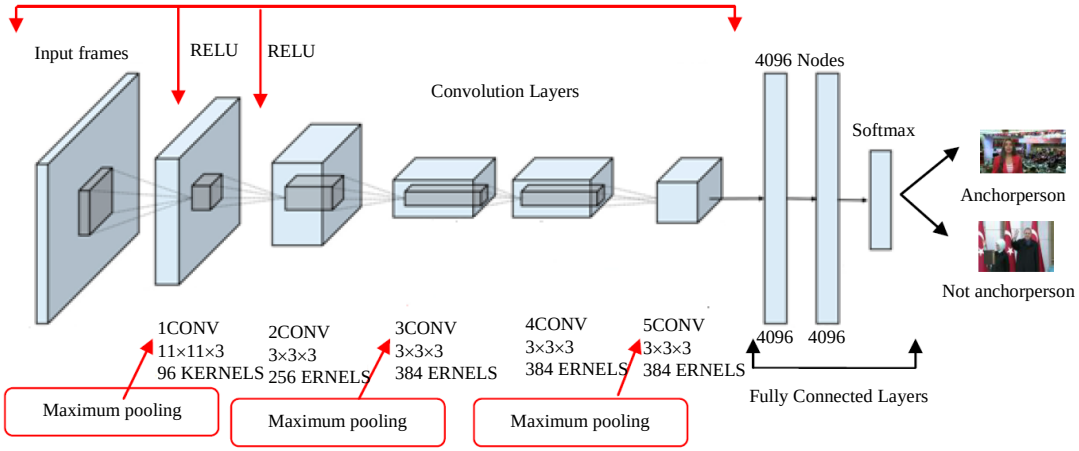


Fig. 12: AlexNet CNN architecture of the proposed framework

Convolution Network Layer

This is the most significant layer in deep learning phenomena of neural networks that generates the feature maps which are subjected to classification layers. It consists of a kernel that slides over the input frame, which generates the output known as feature map. At every location on the input, we performed matrix multiplication followed by integrating the result. The output feature map is defined as:

$$N_x^r = \frac{N_x^{r-1} - L_x^r}{S_x^r} + 1; N_y^{r-1} = \frac{N_y^{r-1} - L_y^r}{S_y^r} + 1 \quad (6)$$

where, (N_x, N_y) is the width and height of the output feature map of the last layer and (L_x, L_y) is the kernel size, (S_x, S_y) that defines the number of pixels skipped by the kernel in horizontal and vertical directions and index r indicates the layer i.e., r = 1. Convolution is applied on the input feature map and a kernel to get the output feature map that is defined as:

$$X_1(m,n) = (J * N)(m,n) \quad (7)$$

where, X_1 (m, n) is a two-dimensional output feature map obtained by convolving the two-dimensional kernel R of size (L_x, L_y) and input feature map J. The sign * is used to represent the convolution between J and R. The convolution operation is expressed as:

$$X_1(m,n) = \sum_{p=\frac{Ly}{2}}^{p=+\frac{L_x}{2}} \sum_{q=\frac{Ly}{2}}^{q=+\frac{L_x}{2}} J(m-p, n-q) R(p,q) \quad (8)$$

In order to train the dataset with maximum accuracy, we applied five CONV layers with RELU layer and response normalization layer for extracting the maximum feature maps from the input frames.

Experimentation and Results

Experiments of Programs Identification in TV Streams

For the evaluation of programs identification in TV stream, we used a corpus from several TV streams from five digital TV channels: LCI, Itele, M6, France 24 and RTV. These streams were either captured using a satellite map or downloaded from live streaming addresses on the web.

The dataset is composed of Standard-Definition Television (SDTV) resolution videos with a frame rate equal to 25. Videos resolutions are varied (512×288, 320×240 and 640×360). Table 1 recapitulates the dataset description and the results of programs identification in TV stream.

SSAE Training

As can be seen in Figure 13, the layered gourmet approach is used in the pre-training of SSAE through training each layer at a time. Then, we use the trained SSAE to classify TV programs.

Firstly, we use the SAE figures on the raw inputs x to learn principal characteristics $h^{(1)}(x)$ on the raw input x by adjusting the weight $w^{(1)}$.

Then, the raw input is introduced into this trained sparse auto-encoder to obtain the primary feature activations $h^{(1)}(x)$ for each of the input images x, which are used as the "raw input" to another SAE to learn their subordinate characteristics $h^{(2)}(x)$.

Subsequently, the principal features are fed into the second SAE to activate the subordinate feature $h^{(2)}(x)$ for each of the principal characteristics $h^{(1)}(x)$ (those are linked to the principal characteristics of the respective input images x).

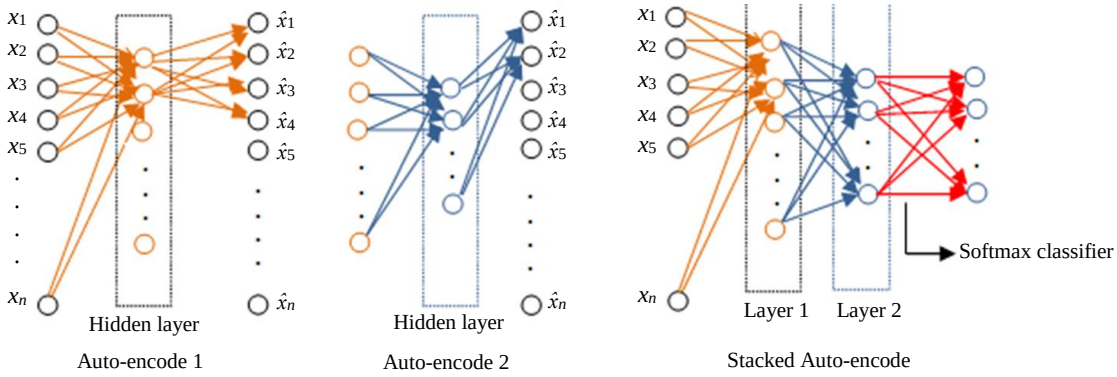


Fig. 13: Stacked sparse auto-encoders

Table 1: Results of programs identification

TV channel	Number of programs	Number of program identification
M6	48	46
RTV	52	44
LCI	112	114
Itele	78	75
France 24	66	70
All channels	356	349

Table 2: The results of program classification on the TRECVID database for different techniques

Methods	Rate of classification
Our Approach	96%
IP using PL	93.51%
POI and CCV	89.08%

Those subordinate characteristics are used as "raw input" of a softmax or SVM classifier, to categorize the TV programs input images.

Afterwards, the three layers are grouped to form a 2 hidden-layer SSAE. As can be seen in (Fig. 13), a Softmax classifier is introduced in the final layer in order to evaluate the proposed technique given that the Softmax is an activation function used in classification. Finally, we apply the fine tuning with the backpropagation algorithm to all the hidden layers to enhance the stacked auto-encoder performance.

In what follows, we outline the experimentations and we detail the results. To experimentally assess the proposed technique, a vast and varied corpus of video streams of programs jingles (Fig. 14).

A set of stacked auto-encoders is applied to our features. The obtained SSAE has two hidden layers with a Softmax classifier in the final layer (Fig. 15).

It is clear that the results obtained are quite encouraging. In fact, we get a classification rate of 96% (Fig. 16). These results validate our approach (Table 2) and further confirm the value of auto-encoder compared to the (Zlitni *et al*., 2015) technique, which uses the signatures of the spatial and temporal descriptors (POI and CCV) and (Manson *et al*., 2009), which uses IP segments (IP using PL) without using reference databases or metadata to segment the stream into program segments.

Experiments of News Story Segmentation

At this level, two experiments are conducted. We build a dataset of news program from different channels to achieve comparative results with similar works and we carry out a second experiment on the TRECVID dataset.

The performance of the proposed algorithm is assessed in terms of precision rate, recall rate and *F1* score. The terms are defined as follows:

$$\text{Precision rate}=\frac{Correct}{Corrent+False}\times 100\%$$

$$\text{Recall rate}=\frac{Correct}{Corrent+Missed}\ 100\%$$

$$\text{F1}=2\times\frac{Recall\times Precision}{Recall+Precision}$$

Experiments on News Dataset from TV Streams

In the first experiments, we use ten recordings of news programs from different channels. Among which, two news program recordings are used from each channel: The first one for the segmentation in stand- alone mode and the other for the rewire mode. We also create a ground truth which contains the number of shots, face shots and anchorperson shots for each news program (Table 3). First, we conduct the detection of all shots and the face identification. Good rates for shots and faces shot (Fig. 17) improve anchorperson shot detection and therefore topic detection.

We train our dataset using AlexNet CNN to classify two different classes presented in our dataset. Transfer learning of a network is presented in Fig. 18.

Table 3: Ground truth of TV news segmentation

	TF1		LCI		France24		Itele		M6	
	NP1	NP2	NP1	NP2	NP1	NP2	NP1	NP2	NP1	NP2
Shot	353	290	350	310	414	370	360	270	446	442
Face shot	250	160	186	120	231	198	210	135	280	275
Anchorperson shot	50	46	58	35	117	106	55	30	75	68

Fig. 14: Samples of TV programs jingles

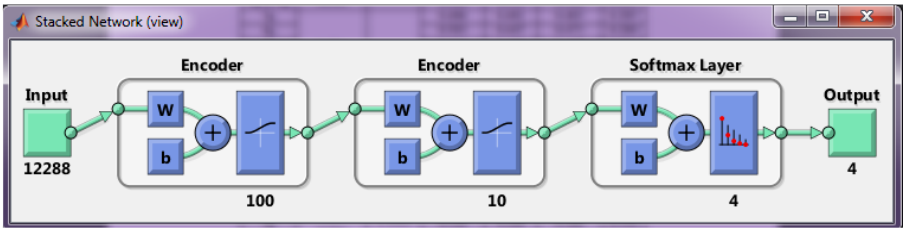


Fig. 15: An illustration of SSAE composed of 2 hidden layers and a Softmax classifier

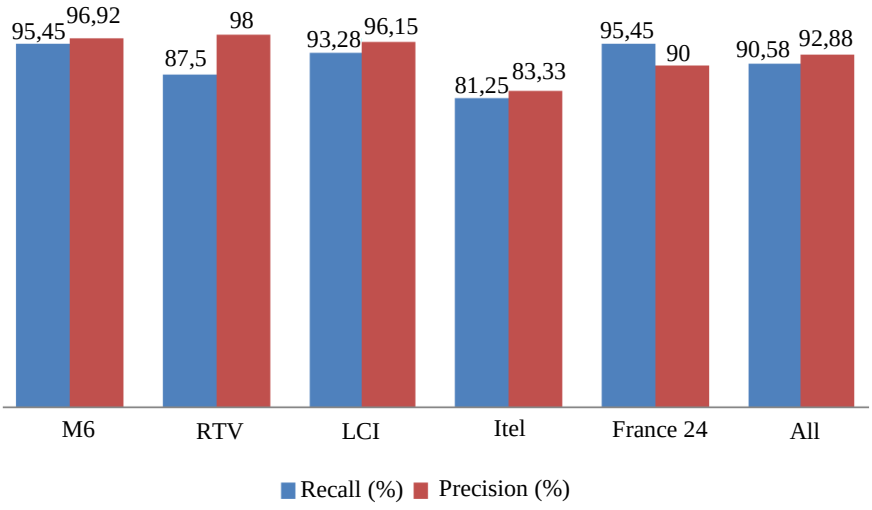


Fig. 16: Recall and precision rates of programs identification

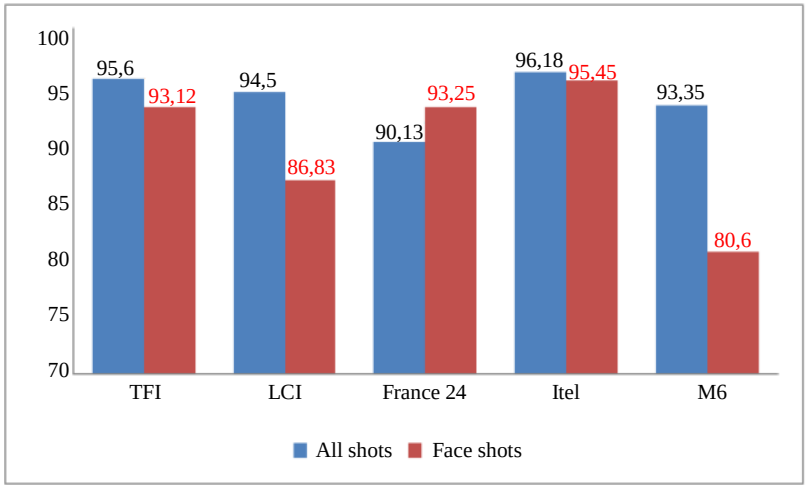


Fig. 17: Face and shot detection

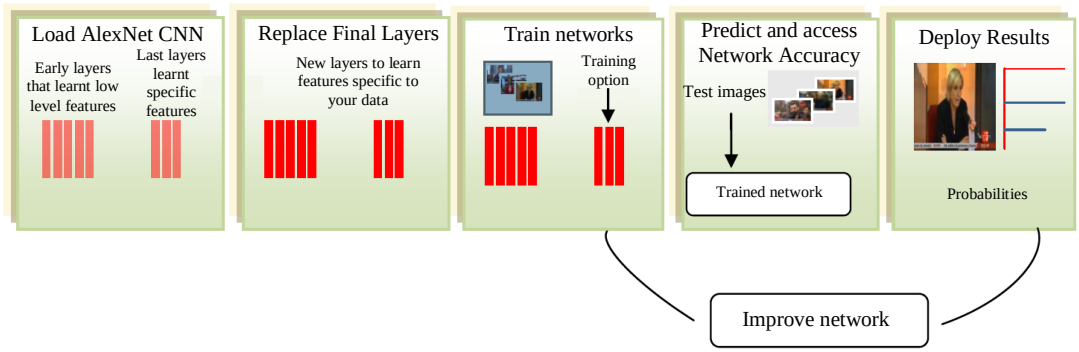


Fig. 18: Transfer learning of AlexNet CNN network

Training

The network takes four epochs in four to five days to train on two GTX 580 Graphic Processing Units (GPU). An epoch is the number of times training vectors are used to update the weights. In the proposed system, each epoch has 500 iterations of the dataset. A stochastic approximation of gradient descent is used to perform training iterations on the dataset. The Stochastic Gradient Descent (SGD) is applied with a learning rate of 0.0001, momentum of 0.9 and weight decay of 0.0005.

We then carried out the evaluation of our approach based on the two modes: "stand-alone" and "Rewire" (Fig. 19).

These results show very satisfactory performance in both modes with an average detection rate of about 90%: Only 10 out of 100 detections are missed. This rate of missed detections can be explained, first, by the missed detections of the face shots and then by the low detection rate obtained in the case of the France 24 NP with the "Rewire" mode. In fact, in this case the anchorperson in NP2 is different from the anchorperson in NP1. As a result, the face tracking method stored in the dataset and the video stream of the NP2 is undetermined. One of the immediate improvements to overcome this problem is to take into account several anchorpersons' faces per news program in the dataset.

Experiments on TRECVID Dataset

The second experiment is carried out on TRECVID 2003 benchmark. We select this dataset, which is the only available benchmark and the most adopted by scholars, to compare our results to those of recent works. The TRECVID 2003 collection include more than 2900 story boundaries (Smeaton *et al*., 2003). TV news programs of this dataset are selected from various channels (CNN, ABC, etc.,). In these experiments, we compare the results obtained by our approach of structuring the news program with other recent works of the state of the art (Zlitni *et al*., 2015; Dumont and Quénot, 2012; Kannao and Guha, 2019). These works for the structuring of the news program are based on the exploitation of the anchorperson as anchor points of reference for the identification of topics. In addition, all these works used the TRECVID 2003 benchmark during the evaluation process.

To obtain an accurate comparison with other works, we use the same metrics as those defined in (Zlitni *et al*., 2015). We therefore evaluate the performance of news segmentation of stand-alone and rewire modes using the precision, recall and F1 metrics. The comparative results (Fig. 20) demonstrate the performance of the proposed. This performance is better when the stand-alone mode is used.

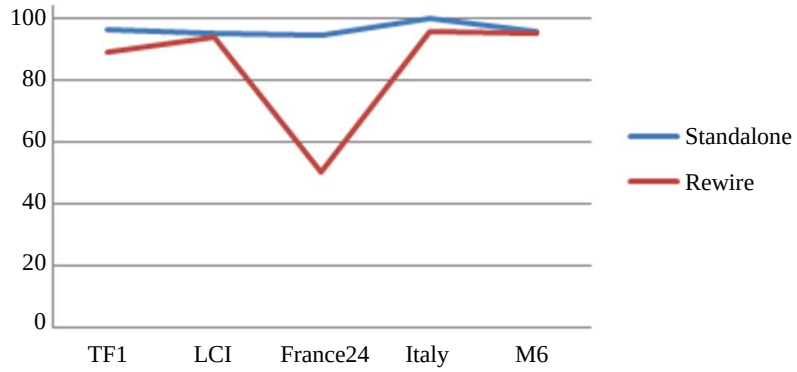
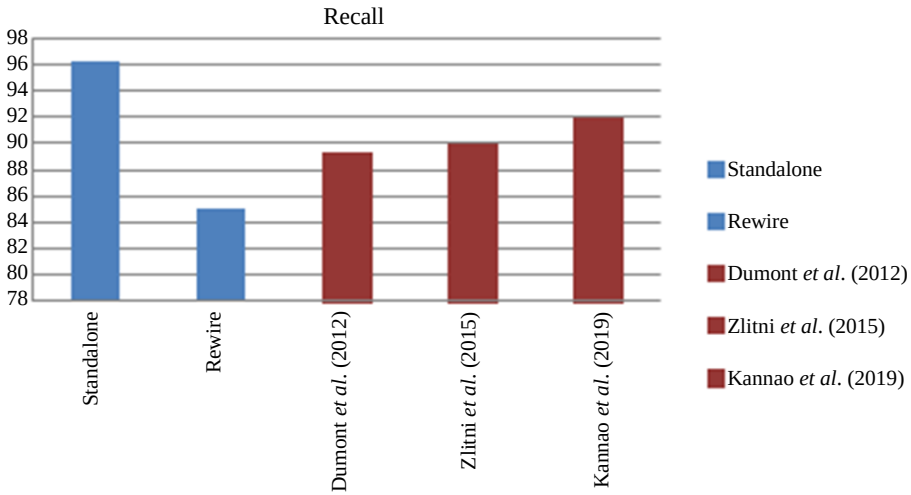
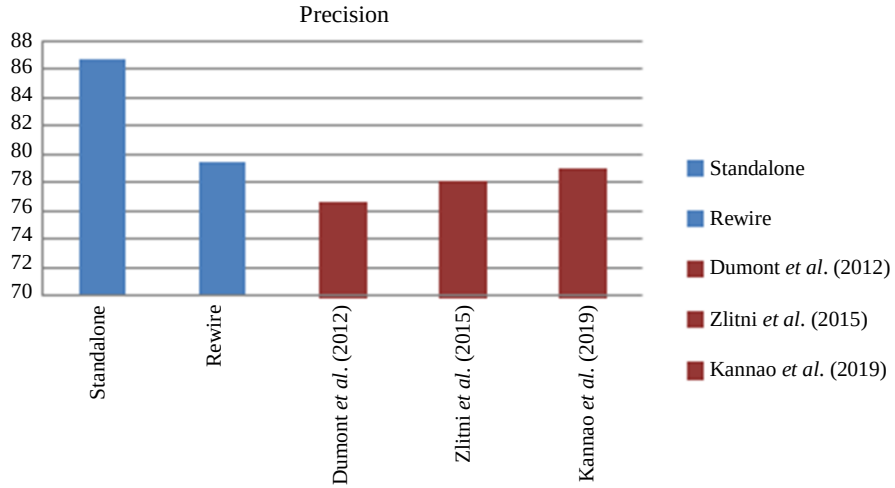


Fig. 19: Rate of topics detection using stand-alone and Rewire modes





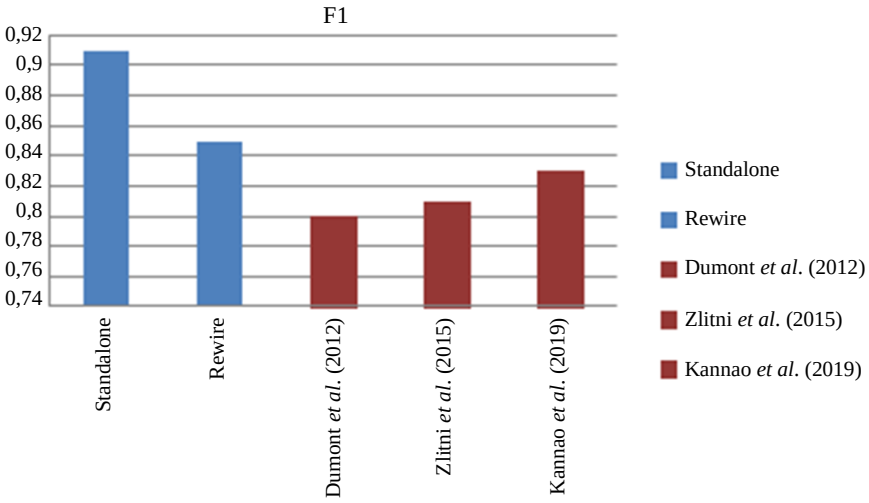


Fig. 20: Comparison of experimental results

Conclusion and Future Works

In this study, we presented a new approach to structure news programs integrated in TV streams, based on production rules of TV channels. A first method was proposed for the segmentation of video streams. This technique consisted of detecting the generics of a program belonging to a TV stream by means of SSAE. In the second method, we presented an approach for automatic segmentation of TV streams into topic using AlexNet CNN. We also described the necessary steps for this structuring. This contribution draws its originality from the use of both the contextual and operational characteristics that regulate the organization of the contents of a news program. The modeling of these characteristics was realized by image processing techniques and statistical models. From this structuring, we also showed that we could extract the topics of a news program according to two modes: Stand-alone mode and Rewire mode. Currently, the proposed approach is dedicated to the analysis of television news only. In further research works, we will try to adapt this approach to deal with a continuous stream of programs. One of the interesting perspectives in this context is the identification of Talk Show in the topics of news program.

Acknowledgement

The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUBprogram.

Author's Contributions

Mounira Hmayda: Designed the research plan, organized the study, participated in all experiments, coordinated the data-analysis and participated in the manuscript writing.

Ridha Ejbali: Advise research project, proof reading of the paper and contributed to the writing of the manuscript.

Mourad Zaied: Advise research project and designed the research plan and contributed to the paper writing.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and there are no ethical issues involved.

References

Berrani, S.A., G. Manson and P. Lechat, 2008. A Non-supervised approach for repeated sequence detection in TV broadcast streams. Signal Proc. Image Commun., 23: 525-537. DOI: 10.1016/j.image.2008.04.018

Bingqing, Q.U., 2015. Content-based discovery of multiple structures from episodes of recurrent TV programs based on grammatical inference. Proceedings of the International Conference on Multimedia Modelling, (CMM' 15). DOI: 10.1007/978-3-319-14445-0_13

Browne, P., C. Czirjek, C. Gurrin, R. Jarina and H. Lee *et al*., 2002. Dublin city university video track experiments for trec.

Colace, F., P. Foggia and G. Percannella, 2005. A probabilistic framework for TV-news stories detection and classification. Proceedings of the International Conference Multimedia Expo, (CME' 05), pp: 1350-1353. DOI: 10.1109/ICME.2005.1521680

Dhara, M.P. and U. Saurabh, 2013. Optical flow measurement using Lucas Kanade method. Int. J. Comput. Applic., 61: 6-10. DOI:10.5120/9962-4611

Dumont, E. and G. Quénot, 2012. Automatic story segmentation for TV news video using multiple modalities. Int. J. Digital Multimedia Broadcast.

Ejbali, R., M. Zaied and C. Ben Amar, 2012. Multi-input Multi-output beta wavelet network: Modeling of acoustic units for speech recognition. Int. J. Adv. Comput. Sci. Applic.

Feng, B., P. Ding, J. Chen, J. Bai and S. Xu *et al*., 2012. Multi-modal information fusion for news story segmentation in broadcast video. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (SSP' 12), pp: 1417-1420. DOI: 10.1109/ICPR.2016.7900085

Feng, B., Z. Chen, R. Zheng and B. Xu, 2014. Multiple style exploration for story unit segmentation of broadcast news video. Mult. Syst., 20: 347-361. DOI: 10.1007/s00530-013-0350-0

Gao, Y., M. Wang, Z. Zha, Q. Tian and Q. Dai *et al*., 2011. Lessismore: Efficient 3D object retrieval with query view selection. IEEE Trans. Multimedia, 11: 1007-1018. DOI: 10.1109/TMM.2011.2160619

Ghosh, H., S.K. Kopparapu, T. Chattopadhyay, A. Khare and S.S. Wattamwar *et al*., 2010. Multimodal indexing of multilingual news video. Int. J. Digital Multimedia Broadcasting. DOI: 10.1155/2010/486487

Gnouma, M., R. Ejbali and M. Zaied, 2016. Detection of abnormal movements of a crowd in a video scene. Int. J. Comput. Theory Eng. DOI: 10.7763/IJCTE.2016.V8.1078

Goyal, A., P. Punitha, F. Hopfgartner and J.M. Jose, 2009. Split and merge based story segmentation in news videos. Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Apr. 6-0, ACM, Toulouse, France, pp: 766-766. DOI: 10.1007/978-3-642-00958-7_82

Hassairi, S., R. Ejbali and M. Zaied, 2015. Supervised image classification using deep convolutional wavelets network. Proceedings of the 27th International Conference on Tools with Artificial Intelligence, Nov. 9-11, IEEE Xplore Press, Vietri sul Mare, Italy. DOI: 10.1109/ICTAI.2015.49

Hmayda, M., R. Ejbali and M. Zaied, 2017. Program classification in a stream TV using deep learning. Proceedings of the 18th International Conference on Parallel and Distributed Computing, Applications and Technologies, Dec. 18-20, Taipei, Taiwan. DOI: 10.1109/PDCAT.2017.00029

Ibrahim, Z.A.A. and P. Gros, 2011. TV stream structuring. ISRN Signal Proc.

Iwan, L.H. and J.A. Thom, 2017. Temporal video segmentation: Detecting the end-of-act in circus performance videos. Multimedia Tools Applic., 76: 1379-1401. DOI: 10.1007/s11042-015-3130-3

Jacobs, A., A. Miene, G.T. Ioannidis and O. Herzog, 2004. Automatic shot boundary detection combining color, edge and motion features of adjacent frames. Proceedings of the TRECVID Workshop Notebook Papers, (WNP' 04), pp: 197-206.

Jindal, A., A. Tiwari and H. Ghosh, 2011. Efficient and language independent news story segmentation for telecast news videos. Proceedings of the International Symposium on Multimedia, Dec. 5-7, IEEE Xplore Press, Dana Point CA, USA, pp: 458-463. DOI: 10.1109/ISM.2011.81

Kannao, R. and P. Guha, 2019. Segmenting with style: Detecting program and story boundaries in TV news broadcast videos. Multimedia Tools Applic., 78: 31925-31957. DOI: 10.1007/s11042-019-7699-9

Krizhevsky, A., I. Sutskever and G.E. Hinton, 2012. Imagenet Classification with Deep Convolutional Neural Networks. In: Neural Information Processing System Foundations, Krizhevsky, A., I. Sutskever and G.E. Hinton (Eds.), San Diego, CA, USA, pp: 1097-1105.

Manson, G., X. Naturel and S.A. Berrani, 2009. Automatic program extraction from TV streams. Proceedings of the European Interactive TV Conference EuroITV'09, (ICE' 19), Leuven, Belgium. DOI: 10.1155/2010/153160

Meinedo, H. and J. Neto, 2003. Audio segmentation, class cation and clustering in a broadcast news task. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Apr. 6-10, IEEE Xplore Press, Hong Kong, China. DOI: 10.1109/ICASSP.2003.1202280

Misra, H., F. Hopfgartner, A. Goyal, P. Punitha and J. Jose, 2010. TV news story-based segmentation one semantic coherence and content similarity. Proceedings of the 16th International Conference Multimedia Model, Jan. 6-8, Springer, Chongqing, China, pp: 347-357. DOI: 10.1007/978-3-642-11301-7_36

O'Hare, N., A.F. Smeaton, C. Czirjek, N. O'Connor and N. Murphy, 2004. A generic news story segmentation system and its evaluation. Proceedings of the International Conference Acoust Speech Signal Process, May 17-21, IEEE Xplore Press, Montreal, Que., Canada, pp: 1028-1031. DOI: 10.1109/ICASSP.2004.1326723

Poulisse, G.J., M.F. Moens, T. Dekens and K. Deschacht, 2010. News story segmentation in multiple modalities. Multimedia Tools Applied, 48: 3-22. DOI: 10.1007/s11042-009-0358-9

Qu, G., D. Mararu, S. Ayache, M. Charhad and L. Besacier *et al*., 2004. Clips-lis-lsr-labri experiments.

Shaban, A., A. Firl, A. Humayun, J. Yuan and X. Wang *et al*., 2017. Multiple-instance video segmentation with sequence-specie object proposals. Proceedings of the CVPR Workshops, (CVW' 17).

Shah, R. and R. Zimmermann, 2017 Lecture Video Segmentation. In: Multimodal Analysis of User-Generated Multimedia Content, Shah, R. and R. Zimmermann (Eds.), Springer,
ISBN-13: 978-3-319-61806-7, pp: 173-203.

Smeaton, A.F., P. Over and A.R. Doherty, 2010. Video shot boundary detection: Seven years. Comput. Vision. Image Understanding, 114: 411-418.
DOI: 10.1016/j.cviu.2009.03.011

Smeaton, A.F., W. Kraaij and P. Over, 2003. TRECVID 2003 an overview. Proceedings of the TRECVID 2003-Text REtrieval Conference TRECVID Workshop, (CTW' 03).

Ullah, J., A Khan and M.A. Jaffar, 2018. Motion cues and saliency based unconstrained video segmentation. Multimedia Tools Applic., 77: 7429-7446.
DOI: 10.1007/s11042-017-4655-4

Wang, J., L. Duan, H. Lu and S. Jin, 2006. A semantic image category for structuring TV broadcast video streams. Proceedings of the Conference Pac Rim Multimedia, Nov. 2-4, Hangzhou, China, pp: 279-286.
DOI: 10.1007/11922162_33

Weiming, H.U., 2011. A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybernet., 41: 797-819.
DOI: 10.1109/TSMCC.2011.2109710

Weiming, H.U., N. Xie, L. Li, X. Zeng and S.J. Maybank, 2011. A Survey on Visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybernet., 41: 797-819.
DOI: 10.1109/TSMCC.2011.2109710

Wu, X., I. Ide and S. Satoh, 2010. PageRank with text similarity and video near-duplicate constraints for news story re-ranking. Proceedings of the 16th International Conference on MultiMedia Modeling, Lecture Notes in Computer Science, Jan. 6-8, Chongqing, China, pp: 533-544.
DOI: 10.1007/978-3-642-11301-7_53

Xie, L., Y.L. Yang and Z.Q. Liu, 2011. On the effectiveness of subwords for lexical cohesion based story segmentation of Chinese broadcast news. Int. J. Infect. Sci., 181: 2873-2891. DOI: 10.1016/j.ins.2011.02.013

Xu, S., B. Feng, Z. Chen and B. Xu, 2013. A general framework of video segmentation to logical unit based on conditional random fields. Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, Apr. 16-20, ACM, Texas, Dallas, USA, pp: 247-254.
DOI: 10.1145/2461466.2461506

Yuan, J., H. Wang and L. Xiao, 2007. A formal study of shot boundary detection. EEE Trans. Circuits Syst. Video Technol.

Zlitni, T. and W. Mahdi, 2010. A visual grammar approach for TV program identification. Int. J. Comput. Network Security.

Zlitni, T., B. Bouaziz and W. Mahdi, 2015 Automatic topics segmentation for TV news video using prior knowledge. Multimedia Tools Applic., 75: 5645-5672. DOI: 10.1007/s11042-015-2531-7