Journal of *Imaging*

MDPI

Article

Open Datasets and Tools for Arabic Text Detection and Recognition in News Video Frames

Oussama Zayene [1,2,*], Sameh Masmoudi Touj [1], Jean Hennebert [3], Rolf Ingold [2] and Najoua Essoukri Ben Amara [1]

[1] LATIS Lab, National Engineering School of Sousse (Eniso), University of Sousse, Sousse 4054, Tunisia; samehmasmouditouj@yahoo.fr (S.M.T.); najoua.benamara@eniso.rnu.tn (N.E.B.A.)
[2] DIVA Group, Department of Informatics, University of Fribourg (Unifr), Fribourg 1700, Switzerland; rolf.ingold@unifr.ch
[3] ICoSys Institute, HES-SO, University of Applied Sciences, Fribourg 1705, Switzerland; jean.hennebert@hefr.ch
* Correspondence: oussama.zayene@unifr.ch

Received: 26 November 2017; Accepted: 26 January 2018; Published: 31 January 2018

Abstract: Recognizing texts in video is more complex than in other environments such as scanned documents. Video texts appear in various colors, unknown fonts and sizes, often affected by compression artifacts and low quality. In contrast to Latin texts, there are no publicly available datasets which cover all aspects of the Arabic Video OCR domain. This paper describes a new well-defined and annotated Arabic-Text-in-Video dataset called AcTiV 2.0. The dataset is dedicated especially to building and evaluating Arabic video text detection and recognition systems. AcTiV 2.0 contains 189 video clips serving as a raw material for creating 4063 key frames for the detection task and 10,415 cropped text images for the recognition task. AcTiV 2.0 is also distributed with its annotation and evaluation tools that are made open-source for standardization and validation purposes. This paper also reports on the evaluation of several systems tested under the proposed detection and recognition protocols.

Keywords: video text detection; video text recognition; AcTiV dataset; Arabic Video OCR

1. Introduction

Broadcast news and public-affairs programs are a prominent source of information that provides daily updates on national and world news. Nowadays, TV newscasters archive a tremendous number of news video clips thanks to the rapid progress in mass storage technology. As the archive size grows rapidly, the manual annotation of all video clips becomes impractical.

Since the 80s, research in OCR techniques has been an attractive field in the document analysis and recognition community. Prior work has addressed specific research problems that have bordered on printed and handwritten texts in scanned documents. Recently, embedded text in videos has received increasing attention as it often gives crucial information about the media content [1–3]. News videos generally contain two types of texts [2]: scene text and artificial text (Figure 1). The first type is naturally recorded as part of scene during video capturing, such as traffic and shop signs. The second type of text is artificially superimposed on the video during the editing process. Compared with scene text, the artificial one usually provides brief and direct description of video content, which is important for automatic broadcast annotation. Typically, artificial text in news video indicates speaker's name, location, event information, scores of a match, etc. Therefore, in this context, we particularly focus on this category of text.

Recognizing text in videos, often called Video OCR [4], is an essential task in many applications such as news indexing and retrieval [5], video categorization, large archive managing and speaker identification [6]. A Video OCR system is generally composed of four stages: detection, tracking, extraction and recognition. The two first steps consist in locating text regions in video frames and generating the bounding boxes of text lines as an output. Text extraction aims at extracting text pixels and removing background ones. The recognition task converts image regions into text strings. In this work, we focus especially on the detection and recognition steps.

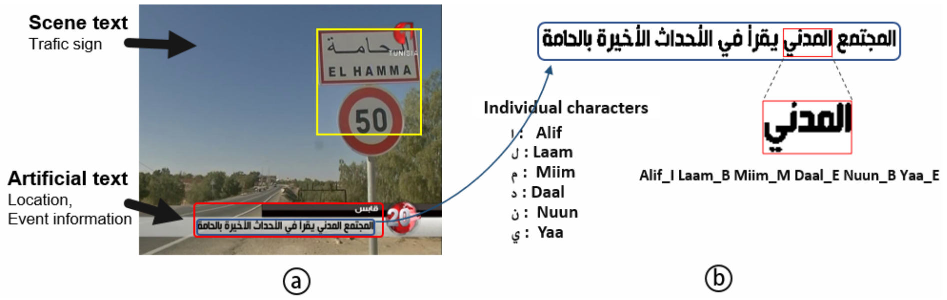


Figure 1. Example of an Arabic video frame including scene and artificial texts (**a**). Decomposition of an Arabic word into characters (**b**).

Compared to scanned documents, text detection and recognition in video frames is more challenging. The major challenges are:

- Text patterns variability: unknown font-size and font-family, different colors and alignment (even in the same TV channel).
- Background complexity: text-like objects in video frames, such as fences, bricks and signs, can be confused with text characters.
- Video quality: acquisition conditions, compression artifacts and low resolution.

All these challenges may give rise to failures in video text detection. The present study focuses on the Arabic video OCR problem. This introduces many additional challenges related to Arabic script [7]. Compared to Latin, the Arabic text has special characteristics such as presence of diacritics, non-uniform inter/intra-word distance and cursiveness of the script, i.e., characters may have up to four shapes depending on their position in the word (for examples, see Figure 1b).

Several techniques have been proposed in the conventional field of Arabic OCR in scanned documents [7–10]. However, few attempts have been made on the development of detection and recognition systems for overlaid text in Arabic news video [11–13]. These systems were tested on private datasets with different evaluation protocols and metrics that make direct comparison and objective benchmarking rather impractical. For instance, in [11], the proposed text detector was evaluated on a private set of 150 video images. In [13], Yousfi et al. evaluated their text detection system on two private test sets of 164 and 201 video frames. Therefore, the availability of an annotated and public dataset is of key importance for the Arabic video text analysis community.

In this paper, we present AcTiV 2.0 as an open Arabic-Text-in-Video dataset dedicated to benchmarking and comparison of systems for Arabic text detection, tracking and recognition. AcTiV 2.0 is an important extension of the one published in ICDAR 2015 [14]. It includes 189 video clips with an average length of 10 min per sequence for a global duration of about 31 h. These video sequences have been collected from four different Arabic news channels during the period between October 2013 and March 2016. In the present work, three video resolutions were chosen: HD (High Definition, 1920 × 1080), SD (Standard Definition, 720 × 576) and SD (480 × 360). The latter resolution concerns video clips that have been downloaded from the official YouTube channel of TunisiaNat1 TV.

The paper is organized as follows: In Section 2, we present related work on datasets for text detection/recognition problems. Then, we present in Section 3 the AcTiV 2.0 dataset in terms of features, statistics and annotations. We detail the evaluation protocols in Section 4 and present the experimental results in Section 5. In Section 6, we draw the conclusions and discuss future work.

2. Literature Review

Recently, several approaches have been proposed to detect and recognize texts in videos and natural scene images [1,2,15,16].

All mentioned work so far are dedicated to Latin or Chinese text detection and recognition methods. Much of the progress that has been made in this field of research is attributed to the availability of standard datasets. The most popular of these is the dataset of ICDAR 2003 Robust Reading Competitions (RRC) [17], prepared for scene text localization, character segmentation (removing background pixels) and word recognition. This dataset includes 509 text images in real environments captured with hand-held devices. 258 images from the database are used for training and the remaining 251 images constitute the test set. Some examples are depicted in Figure 2a. This dataset was also used in the ICDAR 2005 Text Locating Competition [18]. Figure 3 shows the evolution of the Latin text detection research between 2003 and 2013 [18–20] taking as a benchmark the ICDAR 2003 dataset. As can be observed, the method of Huang et al. [19] outperforms other approaches by a large margin. This method enhances the Stroke Width Transform (SWT) algorithm using color information and introduces Text Covariance Descriptors (TCDs). For the word-recognition task, the best accuracy of 93.1%, was achieved by Jaderberg et al. [21] using their proposed Convolutional Neural Networks (CNN) model. The dataset in ICDAR 2011 RRC [22] was inherited from the benchmark used in the previous ICDAR competitions (i.e., 2003 and 2005) but have undergone extension and modification, since there are some missing ground truth information and imprecise word bounding boxes. The final datasets consisted of 485 full images and 1564 cropped word images for localization and word-recognition tasks, respectively. On this dataset, the text detection method of Liao et al. [23] obtains state-of-the-art performance with an F-score of 82%. This algorithm is based on a fully convolutional network (FCN) followed by a standard non-maximum suppression process.

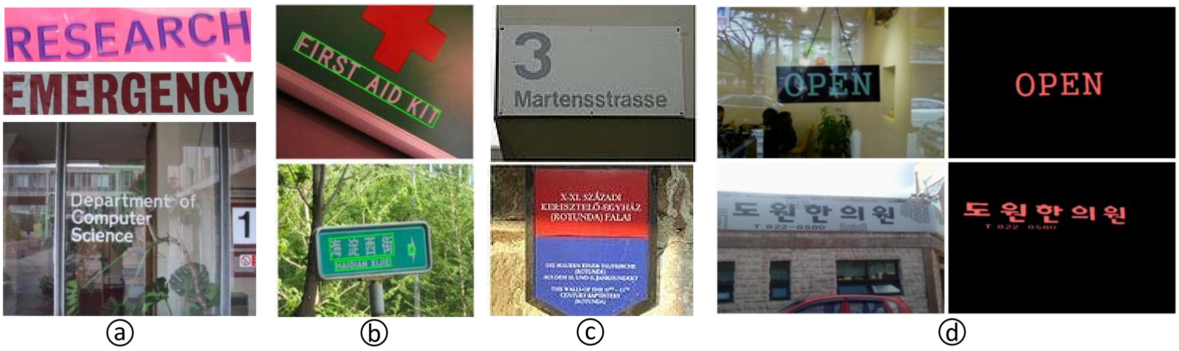


Figure 2. Typical samples from ICDAR2003 (**a**), MSRA-TD500 (**b**), NEOCR (**c**) and KAIST (**d**) datasets.

In the 2013 edition of ICDAR RRC [24], a new database was proposed for video text detection, tracking and recognition. It contains 28 short video sequences. An updated version of this dataset was provided in ICDAR 2015 [25] including a training set of 25 videos and a test set of 24 videos.

The MSRA-TD500 dataset [26] works on multi-oriented scene texts detection. This dataset includes 500 images (300 for training and 200 for testing) with horizontal and slant/skewed texts in complex natural scenes (see Figure 2b for examples). The method of Liu et al. [27] achieves state-of-the-art performance on this database with an F-score of 75%. This method makes use of the Maximally Stable Extremal Regions (MSER) technique as text candidates extractor as well as a set of heuristic rules and an AdaBoost classifier as a two-stages filtering process.

The Street View Text (SVT) dataset [28] is used for scene text detection, segmentation and recognition in outdoor images. It includes 350 full images with 904 word-level annotated bounding boxes. The method of Shi et al. [29] shows superiority over existing techniques with 80.8% as a recognition accuracy. This method is based on Convolutional Recurrent Neural Network (CRNN), which integrates the advantages of both CNN and Recurrent Neural Networks (RNN). For the segmentation task, the best F-score, 90%, was obtained by Mishra et al. [30]. The algorithm is mainly based on two steps: a GMM refinement using stroke and color features and a graph cut procedure.

The KAIST dataset [31] consists of 3000 images taken in indoor and outdoor scenes (see Figure 2d for examples). This is a multilingual dataset, which includes English and Korean texts. KAIST can be used for both detection and segmentation tasks, as it provides binary masks for each character in the image. The text segmentation algorithm of Zhu and Zhang [32] outperforms existing methods on this dataset with an F-score of 88%. The method is based on superpixel clustering. First, an adaptive SLIC text superpixel generation procedure is performed. Next, a DBSCAN-based superpixel clustering is used to fuse stroke superpixels. Finally, a stroke superpixel verification process is applied.

The NEOCR dataset [33] contains 659 natural scene images with multi-oriented texts of high variability (see Figure 2c for examples). This database is intended for scene text recognition and provided multilingual evaluation environments, as it includes texts in eight European languages.

In 2016, Veit et al. [34] proposed a dataset for English scene text detection and recognition called COCO-Text. The dataset is based on the Microsoft COCO dataset, which contains images of complex everyday scenes. The best result on this dataset (67.16%) was obtained by the winner of the COCO-Text ICDAR2017 competition [35]. Note that the participating methods on this competition were ranked based on their Average precision (AP) with an Intersection over Union (IoU) of 0.5.

Recently, Chng and Chan [36] introduced a new dataset, namely Total-text, for curved scene text detection and recognition problems. It contains 1555 scene images and 9330 annotated words with three different text orientations.

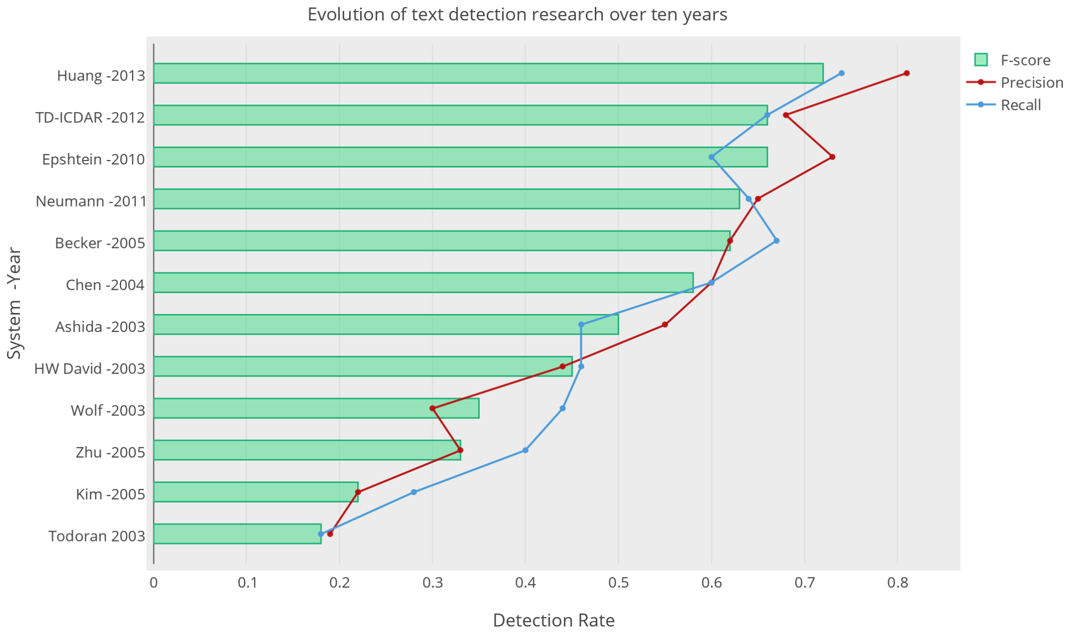


Figure 3. Some examples of text detection systems [18–20] showing the evolution of this area of research over ten years.

As for Arabic language, major contributions have already been made in the conventional field of printed and handwritten OCR systems [7,10]. Much progress of such systems has been triggered thanks to the availability of public datasets. Examples include the IFN/ENIT [37] and KHATT [38] datasets for offline handwriting recognition and writer identification; the APTI database [39] for printed word recognition; and the ADAB dataset [40] that works on online handwriting recognition.

However, handling Arabic text detection and recognition for multimedia documents is limited to very few studies [41–43].

Table 1 presents commonly used datasets for text processing in images and videos, and summarizes their features in terms of text categories, sources, tasks, script, information of training/test samples and best achieved result. As depicted by this table, publicly available datasets for Arabic Video OCR systems are limited to one work for the recognition task and are even non-existent for detection and tracking problems. Yousfi et al. [44] put forward a dataset for superimposed text recognition, called Alif. The dataset was composed of 6532 static cropped text images extracted from diverse Arabic TV channels and with about 12% extracted from web sources. This dataset offered only one image resolution.

Table 1. Most important existing datasets for text processing in videos and scene images. "D", "S" and "R" respectively denote "Detection", "Segmentation" and "Recognition".

Dataset (Year)	**Category**	**Source**	**Task**	**# of Images (Train/Test)**	**# of Text (Train/Test)**	**Script**	**Best Scores**
ICDAR'03 [18] (2003)	Scene text	Camera	D/R	509 (258/251)	2276 (1110/1156)	English	93.1% (R)
KAIST [31] (2010)	Scene text	Camera, mobile phone	D/S	3000	>5000	English, Korean	88% (S)
SVT [28] (2010)	Scene text	Google Street View	D/S/R	350 (100/250)	904 (257/647)	English	80.8% (R) 90% (S)
NEOCR [33] (2011)	Scene text	Camera	D/R	659	5238	Eight languages	
ICDAR'11 [22] (2011)	Scene text	Camera	D/R	485	1564	English	82% (D)
MSRA-TD500 [26] (2012)	scene text	Camera	D	500 (300/200)	–	English, Chinese	75%
ICDAR'13 [24] (2013)	Scene text Artificial text Video scene	Camera Web Camera	D/S/R D/S/R D/T/R	229/233 410/141 28 videos	848/1095 3564/1439 _	Spanish, French, English	
ALIF [44] (2015)	Artificial text	Video frames	R		6532 (4152/2199)	Arabic	55.03%
COCO-Text [34] (2016)	Scene text	MS COCO dataset	D/R	63,686 (43.6k/10k)	173,000	English	67.16% (D)
Total-Text [36] (2017)	Curved scene text	web	D/R	1555 (1255/300)	9330 (words)	English	

3. Proposed Datasets

In this section, we describe the AcTiV 2.0 dataset in terms of characteristics, statistics and annotation guidelines.

3.1. Data Characteristics and Statistics

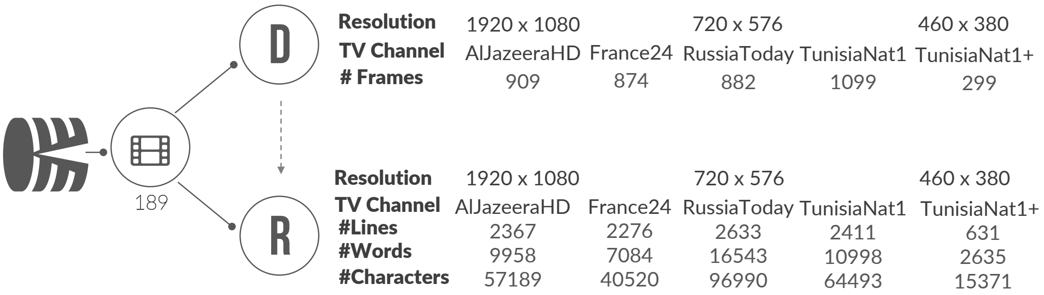
As mentioned in the introduction, AcTiV 1.0 (http://tc11.cvc.uab.es/datasets/AcTiV_1) was presented in the ICDAR'15 conference [14] as the first publicly accessible annotated dataset designed to assess the performance of different Arabic Video OCR systems. This database is currently used by several research groups around the world. It was partially used as a benchmark in the first edition of the "AcTiVComp" contest in conjunction with the ICPR'16 conference [45]. The two main challenges addressed by this dataset are text pattern variability and presence of complex backgrounds with various text-like objects. AcTiV 1.0 consists of 80 video clips recorded from four different Arabic news channels: TunisiaNat1, France24, Russia Today and AljazeeraHD. AcTiV 1.0 is composed of video clips and their corresponding XML files (detailed in Section 3.2). We selected from these video clips 1843 frames dedicated to the detection task. In [14,46], the first results using AcTiV 1.0 were presented.

Based on the obtained results under different evaluation protocols and considering the AcTiV 1.0 users' feed-backs, it was necessary to extend the content in terms of video clips and resolutions offering more training samples, especially for deep learning-based methods.

The new dataset AcTiV 2.0 includes 189 video sequences, 4063 key frames, 10,415 text images and three video-stream resolutions, i.e., the new one is SD (480 × 360). A brief comparison in terms of content between the initial and new version of the proposed dataset is presented in Table 2. The architecture of the new dataset is completely different from the old one. In addition to the videos and their annotation XML files, AcTiV 2.0 includes two appropriate datasets for detection and recognition tasks, (see Figure 4).

Table 2. Statistics of AcTiV 1.0 and AcTiV 2.0.

	#Resolution	#Videos	#Frames	#Cropped Images
AcTiV 1.0	2	80	1843	-
AcTiV 2.0	3	189	4063	10,415



Resolution	1920 x 1080		720 x 576		460 x 380
TV Channel	AlJazeeraHD	France24	RussiaToday	TunisiaNat1	TunisiaNat1+
# Frames	909	874	882	1099	299

Resolution	1920 x 1080		720 x 576		460 x 380
TV Channel	AlJazeeraHD	France24	RussiaToday	TunisiaNat1	TunisiaNat1+
#Lines	2367	2276	2633	2411	631
#Words	9958	7084	16543	10998	2635
#Characters	57189	40520	96990	64493	15371

Figure 4. Architecture of AcTiV 2.0 and statistics of the detection (D) and recognition (R) datasets.

- *AcTiV-D* represents a dataset of non-redundant frames used to build and evaluate methods for detecting text regions in HD/SD frames. A total of 4063 frames have been hand-selected with a particular attention to achieve a high diversity in depicted text regions. Figure 5 provides examples from AcTiV-D for typical problems in video text detection. To test the systems' ability to locate texts under different situations, the proposed dataset includes some frames which contain the same text region but with different backgrounds and some others without any text component.
- *AcTiV-R* is a dataset of textline images that can be utilized to build and evaluate Arabic text recognition systems. Different fonts (more than 6), sizes, backgrounds, colors, contrasts and occlusions are represented in the dataset. Figure 6 illustrates typical examples from AcTiV-R. The collected text images cover a broad range of characteristics that distinguish video frames from scanned documents. AcTiV-R consists of 10,415 textline images, 44,583 words and 259,192 characters. To have an easily accessible representation of Arabic text, it is transformed into a set of Latin labels with a suffix that refers to the letter's position in the word, _B: Begin, _M: Middle; _E: End; and _I: Isolate. An example is shown in Figure 1. During the annotation process, we have considered 164 Arabic character forms:

  - 125 letters, i.e., taking into account this "positioning" variability;
  - 15 additional characters, i.e., combined with the diacritic sign "Chadda";
  - 10 digits; and
  - 14 punctuation marks including the *white space*.

  The different character labels can be observed in Table 3. The same table gives for each character its frequency in the dataset.

  More details about the statistics of the detection and recognition datasets are in Figure 4.

Table 3. Distribution of letters in the AcTiV-R dataset.

Character Label	# of Occurrence	In Arabic I	B	M	E	Character Label	# of Occurrence	In Arabic I	B	M	E
Alif	28,433	ا	-	-	ـا	**HamzaAboveAlif**	1653	أ	-	-	ـأ
Baa	7417	ب	بـ	ـبـ	ـب	**HamzaUnderAlif**	1049	إ	-	-	ـإ
Taaa	8948	ت	تـ	ـتـ	ـت	**TildAboveAlif**	87	آ	-	-	ـآ
Thaa	851	ث	ثـ	ـثـ	ـث	**HamzaAboveAlifBroken**	1022	ىء	-	-	ـىء
Jiim	3270	ج	جـ	ـجـ	ـج	**HamzaAboveWaaw**	268	ؤ	-	-	ـؤ
Haaa	3976	ح	حـ	ـحـ	ـح	**LaamHamzaAboveAlif**	925	لأ	-	-	ـلأ
Xaa	1345	خ	خـ	ـخـ	ـخ	**LaamHamzaUnderAlif**	563	لإ	-	-	ـلإ
Daal	6656	د	-	-	ـد	**LaamTildAboveAlif**	63	لآ	-	-	ـلآ
Thaal	459	ذ	-	-	ـذ	**Space**	31,458				
Raa	11,460	ر	-	-	ـر	**Digit_0**	606		**0**		
Zaay	1478	ز	-	-	ـز	**Digit_1**	655		**1**		
Siin	6348	س	سـ	ـسـ	ـس	**Digit_2**	475		**2**		
Shiin	2353	ش	شـ	ـشـ	ـش	**Digit_3**	276		**3**		
Saad	2123	ص	صـ	ـصـ	ـص	**Digit_4**	306		**4**		
Daad	1085	ض	ضـ	ـضـ	ـض	**Digit_5**	248		**5**		
Thaaa	2184	ط	طـ	ـطـ	ـط	**Digit_6**	203		**6**		
Taa	481	ظ	ظـ	ـظـ	ـظ	**Digit_7**	138		**7**		
Ayn	5989	ع	عـ	ـعـ	ـع	**Digit_8**	148		**8**		
Ghayn	890	غ	غـ	ـغـ	ـغ	**Digit_9**	116		**9**		
Faa	4942	ف	فـ	ـفـ	ـف	**Point**	313		**.**		
Gaaf	4443	ق	قـ	ـقـ	ـق	**Colon**	424		**:**		
Kaaf	2999	ك	كـ	ـكـ	ـك	**Comma**	118		،		
Laam	18,868	ل	لـ	ـلـ	ـل	**Slash**	76		/		
Miim	11,907	م	مـ	ـمـ	ـم	**Percent**	101		%		
Nuun	10,027	ن	نـ	ـنـ	ـن	**QuestionMark**	8		**?**		
Haa	2608	ه	هـ	ـهـ	ـه	**ExclamationMark**	12		**!**		
Waaw	10,614	و	-	-	ـو	**Quote**	445		“”		
Yaa	18,153	ي	يـ	ـيـ	ـي	**Hyphen**	457		–		
AlifBroken	1211	ى	-	-	ـى	**ParenthesisO**	30		)		
Hamza	696		ء			**ParenthesisC**	29		(		
TaaaClosed	7239	ة	-	-	ـة	**Bar**	13		|		
LaamAlif	1916	لَا	-	-	ـلَا	**Overall**	259,192				

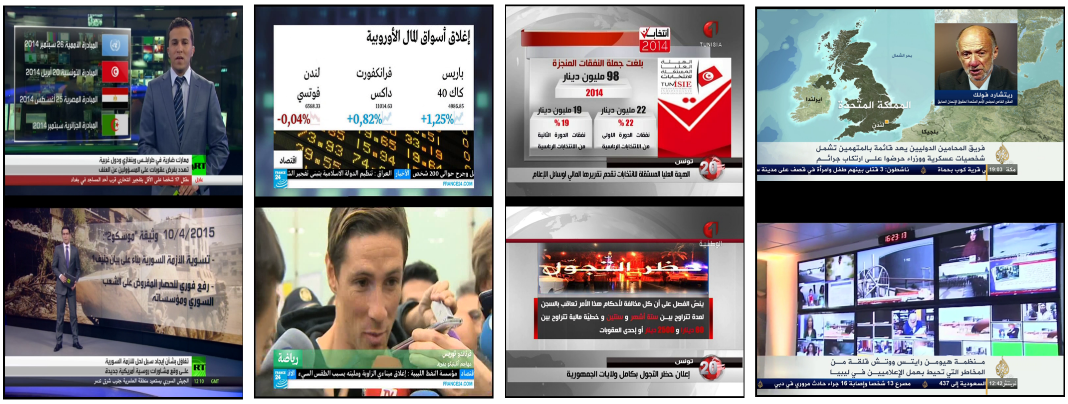


Figure 5. Typical video frames from AcTiV-D dataset. From left to right: Examples of RussiaToday Arabic, France24 Arabe, TunisiaNat1 (El Wataniya 1) and AljazeeraHD frames.

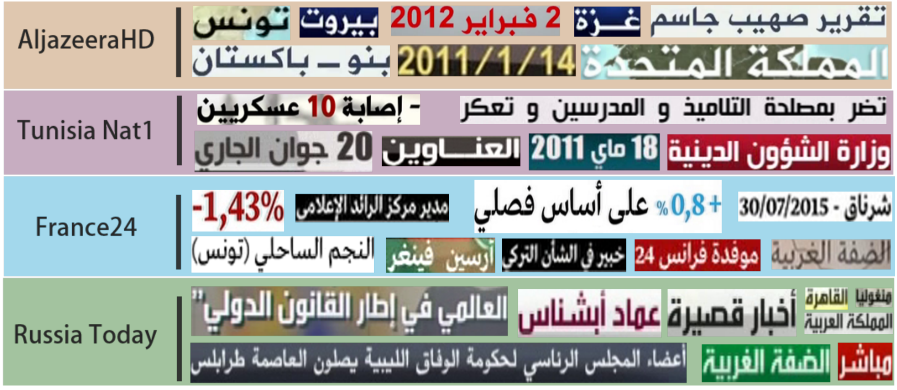


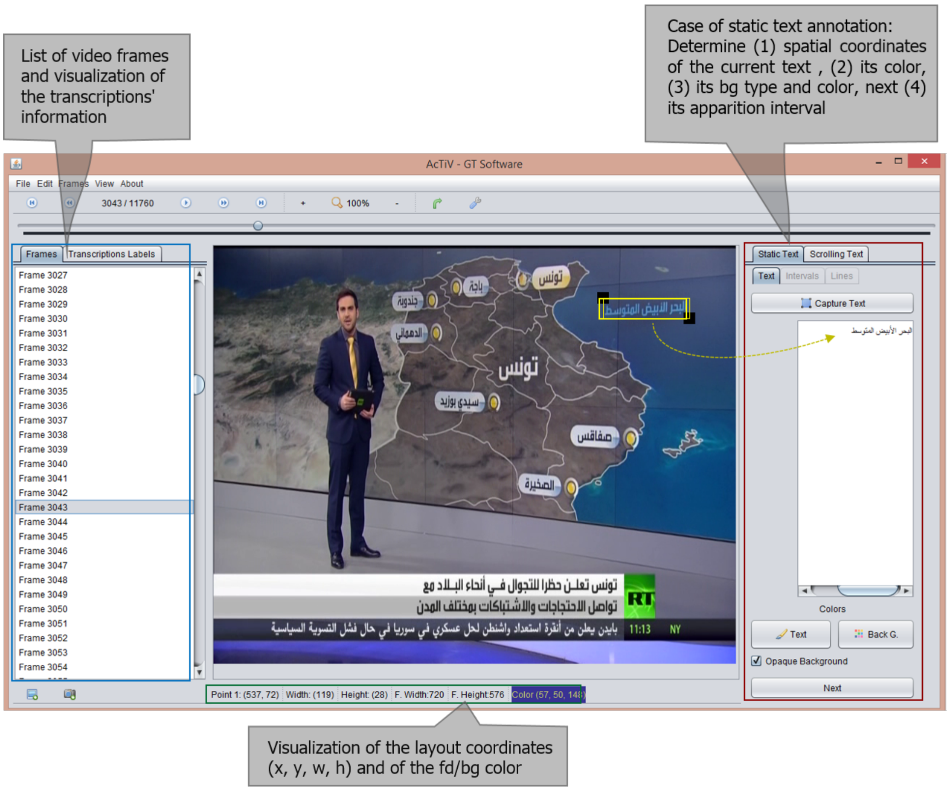
Figure 6. Example of text images from AcTiV-R depicting typical characteristics of video text images.

3.2. Annotation Guidelines

We utilized the AcTiV-GT tool [47] to annotate our collection of data. Figure 7 illustrates the user interface of this tool. In the annotation process, we collect the following information for each text rectangle.

- *position*: x, y, width and height.
- *content*: text strings, text color, background color, background type (transparent, opaque).
- *Interval*: apparition interval of the textline (Frame_S (Start), Frame_E (End)).

Note that a text rectangle can include multiple lines if they share the same font, color and size, and if they are not far from each other.



List of video frames and visualization of the transcriptions' information

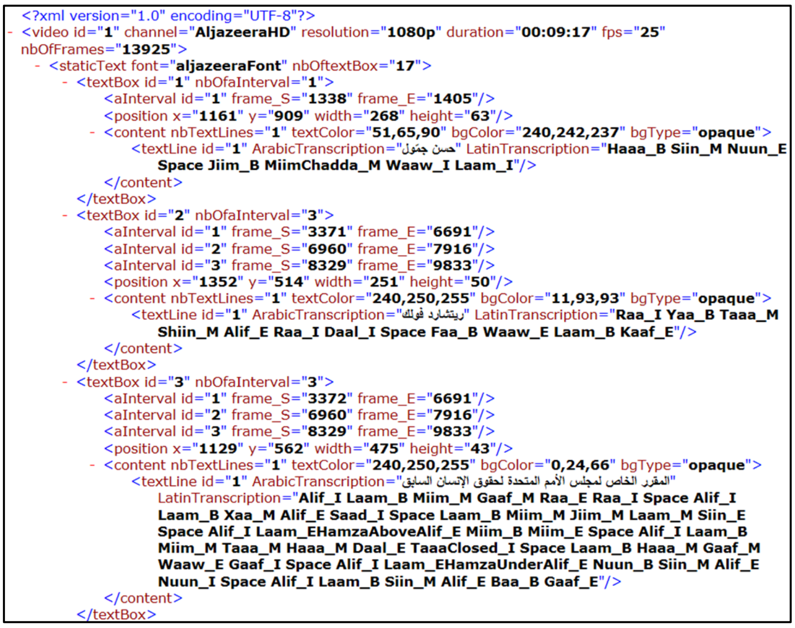
Case of static text annotation: Determine (1) spatial coordinates of the current text , (2) its color, (3) its bg type and color, next (4) its apparition interval

Visualization of the layout coordinates (x, y, w, h) and of the fd/bg color

Point 1: (537, 72) Width: (119) Height (28) F. Width:720 F. Height:576 Color (57, 50, 148)

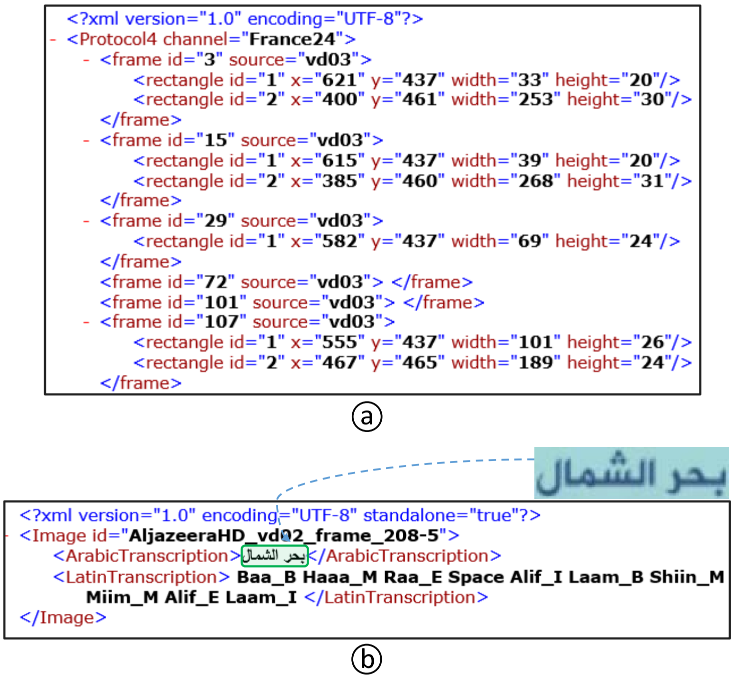
Figure 7. AcTiV-GT open-source tool displaying a labeled frame.

This set of information is saved in a meta file called global XML file (an extract is illustrated in Figure 8). This file can be used for tracking and end-to-end tasks. In AcTiV 2.0, two additional types of XML files have been generated, based on the information contained in the global XML file, one for the detection dataset and the other for the recognition dataset. The detection XML file is provided at the line level for each frame. Figure 9a depicts a part of the detection XML file of France24 TV channel. One bounding box is described by the element *Rectangle* which contains the rectangle attributes: (x, y) coordinates, width and height. The recognition ground-truth files are provided at the line level for each text image. The XML file is composed of two markup sections: *ArabicTranscription* and *LatinTranscription*. Figure 9b depicts an example of a ground-truth XML file and its textline image.



```
  <?xml version="1.0" encoding="UTF-8"?>
- <video id="1" channel="AljazeeraHD" resolution="1080p" duration="00:09:17" fps="25"
  nbOfFrames="13925">
  - <staticText font="aljazeeraFont" nbOftextBox="17">
    - <textBox id="1" nbOfaInterval="1">
        <aInterval id="1" frame_S="1338" frame_E="1405"/>
        <position x="1161" y="909" width="268" height="63"/>
      - <content nbTextLines="1" textColor="51,65,90" bgColor="240,242,237" bgType="opaque">
          <textLine id="1" ArabicTranscription="حسن جمّول" LatinTranscription="Haaa_B Siin_M Nuun_E
            Space Jiim_B MiimChadda_M Waaw_I Laam_I"/>
        </content>
      </textBox>
    - <textBox id="2" nbOfaInterval="3">
        <aInterval id="1" frame_S="3371" frame_E="6691"/>
        <aInterval id="2" frame_S="6960" frame_E="7916"/>
        <aInterval id="3" frame_S="8329" frame_E="9833"/>
        <position x="1352" y="514" width="251" height="50"/>
      - <content nbTextLines="1" textColor="240,250,255" bgColor="11,93,93" bgType="opaque">
          <textLine id="1" ArabicTranscription="ريتشارد فولك" LatinTranscription="Raa_I Yaa_B Taaa_M
            Shiin_M Alif_E Raa_I Daal_I Space Faa_B Waaw_E Laam_B Kaaf_E"/>
        </content>
      </textBox>
    - <textBox id="3" nbOfaInterval="3">
        <aInterval id="1" frame_S="3372" frame_E="6691"/>
        <aInterval id="2" frame_S="6960" frame_E="7916"/>
        <aInterval id="3" frame_S="8329" frame_E="9833"/>
        <position x="1129" y="562" width="475" height="43"/>
      - <content nbTextLines="1" textColor="240,250,255" bgColor="0,24,66" bgType="opaque">
          <textLine id="1" ArabicTranscription="المقرر الخاص لمجلس الأمم المتحدة لحقوق الإنسان السابق"
            LatinTranscription="Alif_I Laam_B Miim_M Gaaf_M Raa_E Raa_I Space Alif_I
            Laam_B Xaa_M Alif_E Saad_I Space Laam_B Miim_M Jiim_M Laam_M Siin_E
            Space Alif_I Laam_EHamzaAboveAlif_E Miim_B Miim_E Space Alif_I Laam_B
            Miim_M Taaa_M Haaa_M Daal_E TaaaClosed_I Space Laam_B Haaa_M Gaaf_M
            Waaw_E Gaaf_I Space Alif_I Laam_EHamzaUnderAlif_E Nuun_B Siin_M Alif_E
            Nuun_I Space Alif_I Laam_B Siin_M Alif_E Baa_B Gaaf_E"/>
        </content>
      </textBox>
```

Figure 8. A part of a global XML annotating a video sequence of Aljazeera TV. This figure contains ground-truth information about three text-boxes from a total of 17.



```
  <?xml version="1.0" encoding="UTF-8"?>
- <Protocol4 channel="France24">
  - <frame id="3" source="vd03">
      <rectangle id="1" x="621" y="437" width="33" height="20"/>
      <rectangle id="2" x="400" y="461" width="253" height="30"/>
    </frame>
  - <frame id="15" source="vd03">
      <rectangle id="1" x="615" y="437" width="39" height="20"/>
      <rectangle id="2" x="385" y="460" width="268" height="31"/>
    </frame>
  - <frame id="29" source="vd03">
      <rectangle id="1" x="582" y="437" width="69" height="24"/>
    </frame>
    <frame id="72" source="vd03"> </frame>
    <frame id="101" source="vd03"> </frame>
  - <frame id="107" source="vd03">
      <rectangle id="1" x="555" y="437" width="101" height="26"/>
      <rectangle id="2" x="467" y="465" width="189" height="24"/>
    </frame>
```

```
  <?xml version="1.0" encoding="UTF-8" standalone="true"?>
- <Image id="AljazeeraHD_vd02_frame_208-5">
    <ArabicTranscription>بحر الشمال</ArabicTranscription>
    <LatinTranscription> Baa_B Haaa_M Raa_E Space Alif_I Laam_B Shiin_M
      Miim_M Alif_E Laam_I </LatinTranscription>
  </Image>
```

Figure 9. Example of AcTiV 2.0 specific XML files: (**a**) a part of the detection XML file of France24 TV; and (**b**) a recognition ground-truth file and its corresponding textline image.

4. Evaluation Protocols and Metrics

As mentioned before, the proposed AcTiV datasets are mainly dedicated to train and evaluate the existing systems for Arabic text detection and recognition in news video. To objectively compare and measure the performance of these systems, we proposed to partition each of the AcTiV-D and AcTiV-R datasets into train, test and closed test subsets taking advantage of the variability in data content. It is to note that the latter subset contains private data (quite similar to the test set) that are used in the context of competitions only. In addition, we suggested a set of evaluation protocols such that different techniques could be directly compared. In other words, the proposed protocols allow us to closely analyze the system behavior towards a given resolution (HD/SD) and/or quality (DBS/Web).

4.1. Detection Protocols and Metrics

Table 4 depicts the detection protocols.

- **Protocol 1** aims to measure the performance of single-frame based methods to detect texts in HD frames.
- **Protocol** 4 is similar to Protocol 1, differing only by the channel resolution. All SD (720 × 576) channels in our database can be targeted by this protocol which is split in four sub-protocols: three channel-dependent (Protocols 4.1, 4.2 and 4.3) and one channel-free (Protocol 4.4).
- **Protocol 4bis** is dedicated to the new added resolution (480 × 360) for the Tunisia Nat1 TV channel. The main idea of this protocol is to train a given system with SD (720 × 576) data i.e., Protocol 4.3 and test it with different data resolution and quality.
- **Protocol 7** is the generic version of the previous protocols where text detection is evaluated regardless of data quality.

Table 4. Detection Evaluation Protocols.

Protocol	TV Channel	Training-Set 1 # Frames	Training-Set 2 # Frames	Test-Set 1 # Frames	Test-Set 2 # Frames	Closed-Set # Frames
1	AlJazeeraHD	337	610	87	196	103
4	France24	331	600	80	170	104
	Russia Today	323	611	79	171	100
	TunisiaNat1	492	788	116	205	106
	All SD	1146	1999	275	546	310
4bis	TunisiaNat1+	-	-	-	149	150
7	All	1483	2609	362	891	563

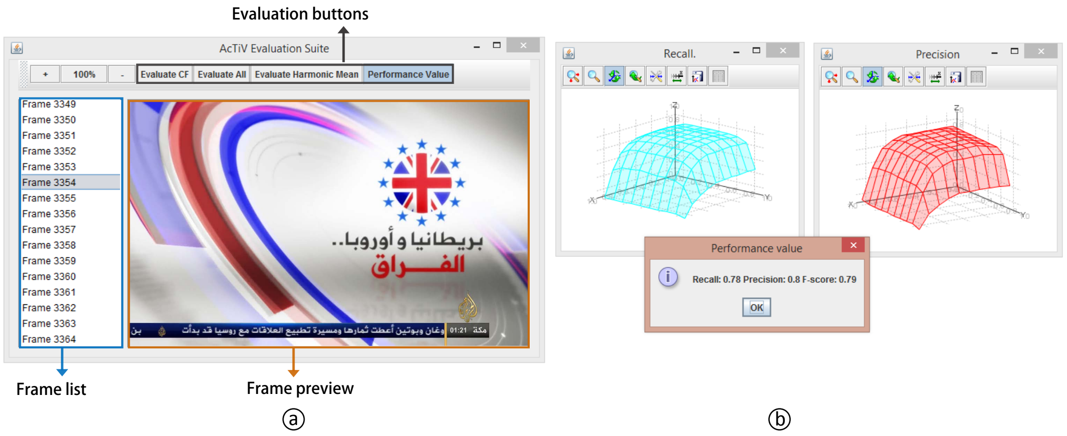
Metrics: The performance of a text detector is evaluated based on precision, recall and F-measure metrics that are defined as:

$$Precision = \frac{\sum_{i=1}^{|D|} matchD(D_i)}{|D|} \tag{1}$$

$$Recall = \frac{\sum_{i=1}^{|G|} matchG(G_i)}{|G|} \tag{2}$$

$$Fmeasure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

where D is the list of detected rectangles, G is the list of ground-truth rectangles and $matchD/matchG$ are the matching functions, respectively. These measures are calculated using our evaluation tool [48] which takes into account all types of matching cases between G bounding boxes and D ones, i.e., one-to-one, one-to-many and many-to-one matching. In the matching procedure, two quality constraints, namely, t_p and t_r are utilized. $t_p \in [0,1]$ is the constraint on area precision and $t_r \in [0,1]$ is the recall constraint. Figure 10 depicts the user interface of our evaluation tool as well as the precision and recall curves, where x-axis denotes t_r values and y-axis denotes t_p ones. The proposed performance metrics and their underlying constraints are similar to those used in ICDAR 2013 [24] and ICDAR 2015 [25] RRCs. It is worth noting that our annotation and evaluation tools are fully implemented in Java and are made open-source for standardization and validation purposes.



Recall: 0.78 Precision: 0.8 F-score: 0.79

Figure 10. AcTiV-D evaluation tool. The user can apply the evaluation procedure to the current frame "Evaluate CF" button or to all video frames "Evaluate All" button (**a**). The "Performance Value" button displays precision, recall and F-score values (**b**).

4.2. Recognition Protocols and Metrics

Table 5 depicts the recognition protocols.

- **Protocol 3** aims to evaluate the performance of OCR systems to recognize texts in HD frames.
- **Protocol 6** is similar to Protocol 3, differing only by the channel resolution. All SD (720 × 576) channels in our dataset can be targeted by this protocol which is split in four sub-protocols: three channel-dependent (Protocols 6.1, 6.2 and 6.3) and one channel-free (Protocol 6.4).
- **Protocol 6bis** is dedicated to the new added resolution (480 × 360) for the Tunisia Nat1 TV channel. The main idea of this protocol is to train a given system with SD (720 × 576) data i.e., Protocol 6.3 and test it with different data resolution and quality.
- **Protocol 9** is the generic version of Protocols 3 and 6 where text recognition is assessed without considering data quality.

Table 5. Recognition Evaluation Protocols. "Lns" and "Wds" respectively denote "Lines" and "Words".

Protocol	TV Channel	Training-Set			Test-Set			Closed Test-Set		
		#Lns	#Wds	#Chars	#Lns	#Wds	#Chars	#Lns	#Wds	#Chars
3	AlJazeeraHD	1909	8110	46,563	196	766	4343	262	1082	6283
6	France24	1906	5683	32,085	179	667	3835	191	734	4600
	Russia Today	2127	13,462	78,936	250	1483	8749	256	1598	9305
	TunisiaNat1	2001	9338	54,809	189	706	4087	221	954	5597
	All SD	6034	28,483	165,830	618	2856	16,671	668	3286	19,502
6bis	TunisiaNat1+	-	-	-	320	1487	8726	311	1148	6645
9	All	7943	36,593	212,393	814	3622	21,014	930	4368	25,785

Metrics: The performance measure for the recognition task is based on the Line Recognition Rate (*LRR*), Word Recognition Rate (*WRR*) at the line and words levels, respectively, and on the computation of insertion (*I*), deletion (*D*) and substitution (*S*) errors at the character level (*CRR*) that are defined as:

$$CRR = \frac{\#characters - I - S - D}{\#characters} \tag{4}$$

$$WRR = \frac{\#words\_correctly\_recognized}{\#words} \tag{5}$$

$$LRR = \frac{\#lines\_correctly\_recognized}{\#lines} \tag{6}$$

Figure 11 shows an example explaining the impact on CRR and WRR metrics resulting from substitution and deletion errors.

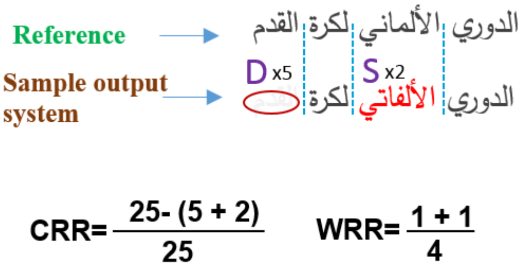


Figure 11. Example of CRR and WRR computation based on output errors.

It is worth noting that the proposed protocols help us understanding how generic is the system, i.e., if a system performs well for Protocols 7 and 9 (independently of the TV channel). For instance, in the AcTiVComp contest, we observed that some participating systems perform well in HD resolution only, some others are quite generic (i.e., good in both SD and HD resolutions). Other systems are incompatible with a specific resolution. Various examples of using these evaluation protocols will be presented in the next section.

5. Application of AcTiV Datasets

The proposed datasets have been used to build and evaluate two systems for Arabic video text detection and recognition. The text detector is based on a hybrid approach composed of CC-based heuristic phase and a machine learning verification procedure. The recognizer system consists of a Multi-Dimensional RNNs (MDRNNs) [49] coupled with a Connectionist Temporal Classification (CTC) layer [50].

5.1. LADI Detector

The LADI text detection system is based in our previous work [14,46], with new added enhancements considering the color consistency of near text regions. Our text detector represents a hybrid approach consisting of two stages: a CC-based heuristic algorithm and a machine learning classification. The main idea of this system is to combine two techniques: an adapted version of the SWT algorithm and a convolutional auto-encoder (CAE). As shown in Figure 12, the first stage starts with a preprocessing step to decrease noise and fine detail. It then computes the edge map and X&Y gradients from the processed frame using Canny and Sobel operators, respectively. After that, the SWT operator is performed as follow.

- Gradient direction $\mathrm{d_p}$ is calculated, at each edge pixel p, which is roughly perpendicular to the stroke orientation.
- A search ray $r = p + n * d_{\mathrm{p}}$ $(n > 0)$ starting from an edge pixel p along the gradient direction $\mathrm{d_p}$ is shot until we find another edge pixel q. If these two edge pixels have nearly opposite gradient orientations, the ray is considered valid. All pixels inside this ray are labeled by the length $|p - q|$.

The next step is to group adjacent pixels in the resulting SWT image into CCs. This is done by applying a flood-fill algorithm based on consistency in stroke width and color. The CCs are then filtered using a set of simple heuristic rules concerning the CC size, position, aspect-ratio and color uniformity. The remaining CCs are iteratively merged into words and textlines based on a proposed textline formation method (see [46] for more details). The second stage uses CAE to automatically produce features, instead of hard-coding them. These features have been learned in an unsupervised way from the textline candidates obtained in the first stage. Then, to discriminate text objects from non-text ones, an SVM classifier with RBF kernel is trained on the patches extracted from the textline candidates by using the generated CAE features.

Note that the whole algorithm is performed twice (for each image) to handle both dark-on-light and light-on-dark texts, once along the gradient direction and once along the inverse direction. The results of two passes are combined to make final decisions.

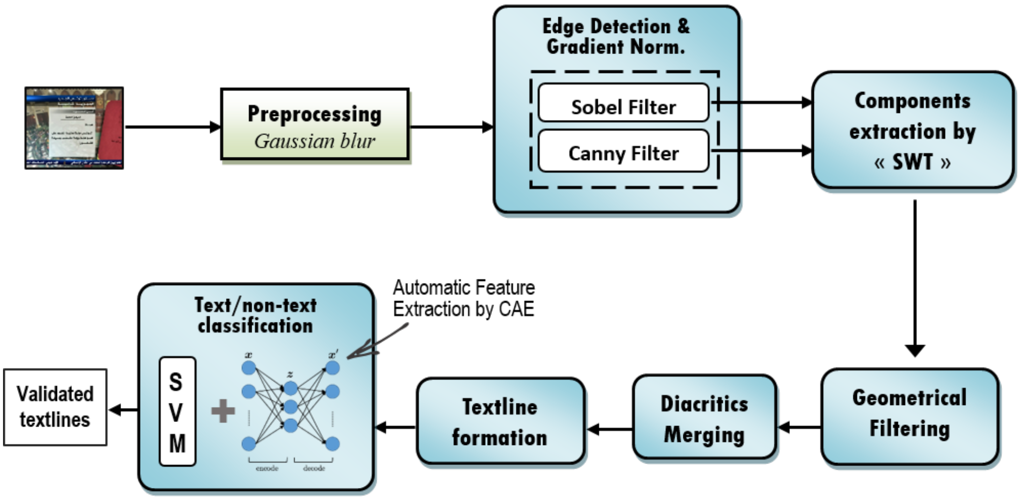


Figure 12. Pipeline of the text detection algorithm. Two passes are performed, one for each text polarity (Dark text on Light background or Light text on Dark background).

5.2. SID OCR

The SID OCR system [51] relies specifically on a Multi-Dimensional Long Short Term Memory (MDLSTM) with a CTC output layer. The proposed network is composed of three levels: an input layer, five hidden layers and an output layer. The hidden layers are MDLSTM that respectively have 2, 10, and 50 cells and separated by feedforward layers with 6 and 20 cells. In fact, we have created a hierarchical structure by repeatedly composing MDLSTM layers with feedforward layers. Firstly, the image is divided into small patches using a pixel window called the "input block", each of which is presented to the first MDLSTM layer as a feature vector of pixel intensities. These vectors are then scanned by four MDLSTM layers in different directions (i.e., up, down, left an right).

After that, the cells activation of the MDLSTM layers are sequentially fed to the first and second feed-forward layers through sub-sample windows, namely "hidden block". This can be seen as a subsampling step with trainable weights, in which the activation are summed and squashed by the hyperbolic tangent (tanh) function. This step aims to extremely reduce the number of weight connections between hidden layers.

The final level is the CTC output layer which labels the sequences of textlines. This layer has n cells, where n is the number of classes, in our case 165 (164 characters and one cell for the 'blank' output). The output activations are normalized at each time step with the softmax activation function. The use of such layer allows working on unsegmented input sequence, which is not the case for standard RNN objective functions. A separate network has been trained for each TV channel of the reference protocol. All input images have been scaled to common heights (70 pixels) and converted to gray-scale. The training is carried out with back-propagation through time (BPTT) algorithm and steepset optimizer has been used with a learning rate of 10^{-4} and with a momentum value of 0.9. We performed several experiments to find the optimal sizes of the MDLSTM layers, feedforward layers, input block and hidden block. Table 6 summarizes the best obtained values of the network parameters.

Note that the size of the input block is set to 1 × 4 for Protocols 6.1 and 3 (not 2 × 4), respectively. To fine-tune these parameters we just pick out a set of 2000 labeled images from AcTiV-R, in which 190 are used as a validation set.

Table 6. Best parameters for training the network.

Parameters	Values
MDLSTM Size	2, 10 and 50
Feed-forward Size	6 and 20
InputBlock Size	2 × 4
HiddenBlock Sizes	1 × 4 and 1 × 4
Learn rate	10^{-4}
Momentum	0.9

5.3. Experimental Results

Several experiments have been conducted using the AcTiV-D and AcTiV-R subsets. These experiments can be divided into two categories: The first one concerns the comparison of our systems with two recent methods. The second category aims at analyzing the effect of increasing the training data on the accuracy of the LADI text detector.

5.3.1. Comparison with Other Methods

As proof of concept of the proposed benchmark, we compare our systems with two recent methods. The first one was proposed by Gaddour et al. [52] to basically detect Arabic texts in natural scene images. The main steps involved are:

- Pixel-color clustering using k-means to form pairs of thresholds for each RGB channel.
- Creation of binary map for each pair of thresholds.
- Extraction of CCs.
- Preliminary filtering according to "area stability" criterion.
- Second filtering based on a set of statistical and geometric rules.
- Horizontal merging of the remaining components to form textlines.

The second method was put forward by Iwata et al. [53] to recognize artificial Arabic text in video frames. It operates as follows:

- Textline segmentation into words by thresholding gaps between CCs.
- Over-segmentation of characters into primitive segments.
- Character recognition using 64-dimensional feature vector of chain code histogram and the modified quadratic discriminant function.
- Word recognition by dynamic programming using total likelihood of characters as objective function.
- False word reduction by measuring the average of the character likelihoods in a word and comparing it to a predefined threshold.

The detection systems have been trained on the training-set1 of Table 4. The evaluation has been done on the test set for the detection and recognition tasks. Table 7 presents evaluation results of the detection protocols in terms of precision, recall and F-measure. The best results are marked in bold. The LADI system scores best for all protocols with an F-measure between 0.73 and 0.85 for AllSD protocol (p4.4) and AljazeeraHD protocol (p1) respectively. In contrast to the SysA that represents a fully heuristic-based method, the LADI system increased the F-measure by 11% for Protocol 1. For Protocols 4.1, 4.2, 4.3 and 4.4 (SD channels), the results are higher, with a gain of, respectively, 11%, 17%, 14% and 24%. This reflects the effectiveness of using a machine-learning solution to filter the results given by the SWT algorithm. The Gaddo system has strong fragmentation and miss detection tendency as depicted by its obtained numerical results. Table 8 presents evaluation results of the recognition protocols in terms of CRR, WRR and LRR metrics. The SID-OCR system has shown superiority in all protocols. The best accuracies are achieved on the TunisiaNat1 channel subset (p6.3) with 0.94 as a CRR and 0.62 as a LRR. The IWATA system performs well for all SD protocols especially for the CRR/WRR metrics. However its current version is incompatible with HD resolution. The result shows that our system has low recognition rate when facing different text patterns and resolutions, i.e., global Protocol 9. Based on our knowledge about the shapes of Arabic characters, we divide the causes of errors into two classes: character similarity and insufficient samples of punctuation, digits and symbols. Several measures can be taken to minimize the character error rate, for instance by integrating language models or dropout mechanism.

Table 7. Performance of text detection systems evaluated on the test set of AcTiV-D.

Protocol	System	Precision	Recall	Fmeasure
1	LADI [46]	**0.86**	**0.84**	**0.85**
	SysA [14]	0.77	0.76	0.76
	Gaddo [52]	0.52	0.49	0.51
4.1	LADI [46]	**0.74**	**0.76**	**0.75**
	SysA [14]	0.69	0.6	0.64
	Gaddo [52]	0.47	0.61	0.54
4.2	LADI [46]	**0.8**	**0.75**	**0.77**
	SysA [14]	0.66	0.55	0.6
	Gaddo [52]	0.41	0.5	0.45
4.3	LADI [46]	**0.85**	**0.82**	**0.83**
	SysA [14]	0.68	0.71	0.69
	Gaddo [52]	0.34	0.49	0.41
4.4	LADI [46]	**0.71**	**0.76**	**0.73**
	SysA [14]	0.5	0.49	0.49
	Gaddo [52]	-	-	-

Table 8. Performance of the recognition systems evaluated on the test set of AcTiV-R.

Protocol	System	CRR	WRR	LRR
3	SIDOCR [51]	0.90	0.71	0.51
	IWATA [53]	-	-	-
6.1	SIDOCR [51]	**0.89**	**0.70**	**0.51**
	IWATA [53]	0.88	0.67	0.46
6.2	SIDOCR [51]	**0.94**	**0.68**	**0.41**
	IWATA [53]	0.9	**0.68**	0.39
6.3	SIDOCR [51]	0.94	0.81	**0.62**
	IWATA [53]	**0.94**	0.77	0.56
6.4	SIDOCR [51]	**0.93**	**0.73**	**0.52**
	IWATA [53]	0.9	**0.73**	0.48
9	SIDOCR [51]	0.73	0.58	0.32
	IWATA [53]	-	-	-

5.3.2. Training with AcTiV 2.0

To examine the effect of increasing the number of training samples on the accuracy of our text detector, we conduct the same experiment of Protocol 6.1, in Table 7, using training-set2, which includes roughly the double of samples (600 frames) than training-set1 (see Table 4). We observed that the detection rates of our text detector have been increased as expected. Specifically, the recall increases by 2% and the precision increases by 5%. This can be explained by the increase in the number of samples provided to the CAE, which leads to more robust feature representations and subsequently better classification results.

6. Conclusions

In this paper, we have presented a new version of the AcTiV dataset for the development and evaluation of text detection and recognition systems targeting Arabic news video. This dataset is freely available to research institutions. We have provided details about the characteristics and statistics of the database. We have also reported about our ground-truthing software used to semi-automatically annotate the video clips and our open text detection evaluation tool. We have evaluated five text detection and recognition algorithms as proof-of-concept of the new dataset. Additionally, a set of evaluation protocols has been made to measure the systems performance under different situations. The experimental results have shown that there is still room for improvement in both detection and recognition of Arabic video text. We look forward to more researchers joining the challenging research topic of Arabic video texts detection and recognition.

Acknowledgments: The researchers would like to thank all the TV channels for providing us with data and multimedia files, especially the archive staff of ElWataniya1 (TunisiaNat1) TV.

Author Contributions: Oussama Zayene conceived the dataset, developed the tools and realized the experiments under the supervision and help of professors Najoua Essoukri Ben Amara, Jean Hennebert and Rolf Ingold. Sameh Masmoudi Touj contributed to the collect, design and annotation of the dataset, and verified the annotated data. All the co-authors have substantially revised the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lu, T.; Palaiahnakote, S.; Tan, C.L.; Liu, W. *Video Text Detection*; Springer Publishing Company, Incorporated: London, UK, 2014.
2. Ye, Q.; Doermann, D. Text detection and recognition in imagery: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1480–1500.
3. Yin, X.C.; Zuo, Z.Y.; Tian, S.; Liu, C.L. Text Detection, Tracking and Recognition in Video: A Comprehensive Survey. *IEEE Trans. Image Process.* **2016**, *25*, 2752–2773.
4. Lienhart, R. Video OCR: A survey and practitioner's guide. In *Video Mining*; Springer: Boston, MA, USA, 2003; pp. 155–183.
5. Yang, H.; Quehl, B.; Sack, H. A framework for improved video text detection and recognition. *Multimed. Tools Appl.* **2014**, *69*, 217–245.
6. Poignant, J.; Bredin, H.; Le, V.B.; Besacier, L.; Barras, C.; Quénot, G. Unsupervised speaker identification using overlaid texts in TV broadcast. In Proceedings of the Interspeech 2012—Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012; p. 4.
7. Märgner, V.; El Abed, H. *Guide to OCR for Arabic Scripts*; Springer: Berlin, Germany, 2012.
8. Touj, S.M.; Amara, N.E.B.; Amiri, H. Arabic Handwritten Words Recognition Based on a Planar Hidden Markov Model. *Int. Arab J. Inf. Technol.* **2005**, *2*, 318–325.
9. Lorigo, L.M.; Govindaraju, V. Offline Arabic handwriting recognition: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 712–724.
10. Chammas, E.; Mokbel, C.; Likforman-Sulem, L. Arabic handwritten document preprocessing and recognition. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 451–455.
11. Jamil, A.; Siddiqi, I.; Arif, F.; Raza, A. Edge-based features for localization of artificial Urdu text in video images. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 1120–1124.
12. Halima, M.B.; Karray, H.; Alimi, A.M. Arabic text recognition in video sequences. *arXiv* **2013**, arXiv:preprint/1308.3243

13. Yousfi, S.; Berrani, S.A.; Garcia, C. Arabic text detection in videos using neural and boosting-based approaches: Application to video indexing. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 3028–3032.
14. Zayene, O.; Hennebert, J.; Touj, S.M.; Ingold, R.; Amara, N.E.B. A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 996–1000.
15. Elagouni, K.; Garcia, C.; Mamalet, F.; Sébillot, P. Text recognition in videos using a recurrent connectionist approach. In Proceedings of the International Conference on Artificial Neural Networks, Lausanne, Switzerland, 11–14 September 2012; pp. 172–179.
16. Khare, V.; Shivakumara, P.; Raveendran, P. A new Histogram Oriented Moments descriptor for multi-oriented moving text detection in video. *Expert Syst. Appl.* **2015**, *42*, 7627–7640.
17. Lucas, S.M.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; Young, R.; Ashida, K.; Nagai, H.; Okamoto, M.; Yamamoto, H.; et al. ICDAR 2003 robust reading competitions: Entries, results, and future directions. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2005**, *7*, 105–122.
18. Lucas, S.M. ICDAR 2005 text locating competition results. In Proceedings of the Eighth International Conference on Document Analysis and Recognition, Seoul, Korea, 31 August–1 September 2005; pp. 80–84.
19. Huang, W.; Lin, Z.; Yang, J.; Wang, J. Text localization in natural images using stroke feature transform and text covariance descriptors. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1241–1248.
20. Zhu, Y.; Yao, C.; Bai, X. Scene text detection and recognition: Recent advances and future trends. *Front. Comput. Sci.* **2016**, *10*, 19–36.
21. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.* **2016**, *116*, 1–20.
22. Shahab, A.; Shafait, F.; Dengel, A. ICDAR 2011 robust reading competition challenge 2: Reading text in scene images. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 1491–1496.
23. Liao, M.; Shi, B.; Bai, X.; Wang, X.; Liu, W. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. In Proceedings of the 2017 AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4161–4167.
24. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L.G.; Mestre, S.R.; Mas, J.; Mota, D.F.; Almazan, J.A.; de las Heras, L.P. ICDAR 2013 robust reading competition. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1484–1493.
25. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.R.; Lu, S.; et al. ICDAR 2015 competition on robust reading. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1156–1160.
26. Yao, C.; Bai, X.; Liu, W.; Ma, Y.; Tu, Z. Detecting texts of arbitrary orientations in natural images. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 1083–1090.
27. Liu, Z.; Li, Y.; Qi, X.; Yang, Y.; Nian, M.; Zhang, H.; Xiamixiding, R. Method for unconstrained text detection in natural scene image. *IET Comput. Vis.* **2017**, *11*, 596–604.
28. Wang, K.; Belongie, S. Word spotting in the wild. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 591–604.
29. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2298–2304.
30. Mishra, A.; Alahari, K.; Jawahar, C. Unsupervised refinement of color and stroke features for text binarization. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2017**, *20*, 105–121.
31. Lee, S.; Cho, M.S.; Jung, K.; Kim, J.H. Scene text extraction with edge constraint and text collinearity. In Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 23–26 August 2010; pp. 3983–3986.
32. Zhu, Y.; Zhang, K. Text segmentation using superpixel clustering. *IET Image Process.* **2017**, *11*, 455–464.

33. Nagy, R.; Dicker, A.; Meyer-Wegener, K. NEOCR: A configurable dataset for natural image text recognition. In Proceedings of the International Workshop on Camera-Based Document Analysis and Recognition, Beijing, China, 22 September 2011; pp. 150–163.
34. Veit, A.; Matera, T.; Neumann, L.; Matas, J.; Belongie, S. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv* **2016**, arXiv:preprint/1601.07140.
35. Gomez, R.; Shi, B.; Gomez, L.; Numann, L.; Veit, A.; Matas, J.; Belongie, S.; Karatzas, D. ICDAR2017 Robust Reading Challenge on COCO-Text. In Proceedings of the 2017 International Conference on Document Analysis and Recognition, Kyoto, Japan, 10–15 November 2017; pp. 1435–1443.
36. Ch'ng, C.K.; Chan, C.S. Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition. In Proceedings of the 2017 International Conference on Document Analysis and Recognition, Kyoto, Japan, 10–15 November 2017; pp. 935–942.
37. Pechwitz, M.; Maddouri, S.S.; Märgner, V.; Ellouze, N.; Amiri, H. IFN/ENIT-database of handwritten Arabic words. In Proceedings of the Colloque International Francophone sur l'Ecrit et le Document (CIFED), Hammamet, Tunisia, 21–23 October 2002; pp. 127–136.
38. Mahmoud, S.A.; Ahmad, I.; Al-Khatib, W.G.; Alshayeb, M.; Parvez, M.T.; Märgner, V.; Fink, G.A. KHATT: An open Arabic offline handwritten text database. *Pattern Recognit.* **2014**, *47*, 1096–1112.
39. Slimane, F.; Ingold, R.; Kanoun, S.; Alimi, A.M.; Hennebert, J. A new arabic printed text image database and evaluation protocols. In Proceedings of the 2009 International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 946–950.
40. Kherallah, M.; Tagougui, N.; Alimi, A.M.; El Abed, H.; Margner, V. Online Arabic handwriting recognition competition. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 1454–1458.
41. Halima, M.B.; Alimi, A.; Vila, A.F.; Karray, H. Nf-SAVO: Neuro-fuzzy system for arabic video OCR. *arXiv* **2012**, arXiv:preprint/1211.2150.
42. Moradi, M.; Mozaffari, S. Hybrid approach for Farsi/Arabic text detection and localisation in video frames. *IET Image Process.* **2013**, *7*, 154–164.
43. Yousfi, S.; Berrani, S.A.; Garcia, C. Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition, Tunis, Tunisia, 23–26 August 2015; pp. 1026–1030.
44. Yousfi, S.; Berrani, S.A.; Garcia, C. ALIF: A dataset for Arabic embedded text recognition in TV broadcast. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1221–1225.
45. Zayene, O.; Hajjej, N.; Touj, S.M.; Ben Mansour, S.; Hennebert, J.; Ingold, R.; Amara, N.E.B. ICPR2016 Contest on Arabic Text Detection and Recognition in Video Frames AcTiVComp. In Proceedings of the 23th International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 187–191.
46. Zayene, O.; Seuret, M.; Touj, S.M.; Hennebert, J.; Ingold, R.; Amara, N.E.B. Text detection in arabic news video based on SWT operator and convolutional auto-encoders. In Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016; pp. 13–18.
47. Zayene, O.; Touj, S.M.; Hennebert, J.; Ingold, R.; Amara, N.E.B. Semi-automatic news video annotation framework for Arabic text. In Proceedings of the 2014 4th International Conference on Image Processing Theory, Tools and Applications, Paris, France, 14–17 October 2014; pp. 1–6.
48. Zayene, O.; Touj, S.M.; Hennebert, J.; Ingold, R.; Amara, N.E.B. Data, protocol and algorithms for performance evaluation of text detection in Arabic news video. In Proceedings of the 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, Tunisia, 21–23 March 2016; pp. 258–263.
49. Graves, A. Offline arabic handwriting recognition with multidimensional recurrent neural networks. In *Guide to OCR for Arabic Scripts*; Springer: Berlin, Germany, 2012; pp. 297–313.
50. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
51. Zayene, O.; Essefi, S.A.; Amara, N.E.B. Arabic Video Text Recognition Based on Multi-Dimensional Recurrent Neural Networks. In Proceedings of the International Conference on Computer Systems and Applications (AICCSA), Hammamet, Tunisia, 30 October–3 November 2017; pp. 725–729.

52. Gaddour, H.; Kanoun, S.; Vincent, N. A New Method for Arabic Text Detection in Natural Scene Image Based on the Color Homogeneity. In Proceedings of the International Conference on Image and Signal Processing, Trois-Rivières, QC, Canada, 30 May–1 June 2016; pp. 127–136.
53. Iwata, S.; Ohyama, W.; Wakabayashi, T.; Kimura, F. Recognition and transition frame detection of Arabic news captions for video retrieval. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 4005–4010.