Dynamic subtitles: A multimodal video accessibility enhancement dedicated to deaf and hearing impaired users

Ruxandra Tapu[1,2], Bogdan Mocanu[1,2], Titus Zaharia[1]
[1]ARTEMIS, Institut Mines-Telecom/TelecomSudParis, CNRS MAP5 8145, France
[2]Telecommunication, Faculty of ETTI, University Politehnica of Bucharest, Romania
ruxandra_tapu@comm.pub.ro, bcmocanu@comm.pub.ro, titus.zaharia@telecom-sudparis.eu

Abstract

In this paper, we introduce a novel dynamic subtitle positioning system designed to increase the accessibility of the deaf and hearing impaired people to video documents. Our framework places the subtitle in the near vicinity of the active speaker in order to allow the viewer to follow the visual content while regarding the textual information. The proposed system is based on a multimodal fusion of text, audio and visual information in order to detect and recognize the identity of the active speaker. The experimental evaluation, performed on a large dataset of more than 30 videos, validates the methodology with average accuracy and recognition rates superior to 92%. The subjective evaluation demonstrates the effectiveness of our approach outperforming both conventional (static) subtitling and other state of the art techniques in terms of enhancement of the overall viewing experience and eyestrain reduction.

1. Introduction

With technological advances, audio-visual productions are governing today, at a large majority, the way on how we access and consume culture and knowledge. The available live programs transmitted over Internet represent multimodal materials that convey information through verbal and non-verbal signs and codes. However, there are millions of people that are suffering from hearing impairments. The recent statistics published by the World Health Organization [1] show that by the year of 2050, it is expected that more than 900 million people across the world to suffer from hearing loss problems [22]. Spoken language in audiovisual production has been conveyed in two ways for the deaf and hard-of-hearing population: sign language interpretation and dynamic subtitling. The dynamic subtitling has gained importance in the audiovisual translation since it is relatively cheap and fast. In addition, the deaf community tends to move nowadays towards orality [24]. Traditionally, the video subtitles are static, positioned in a fixed location, centered, at the bottom of the screen. Although the hearing impaired audience can get certain information from the script there are still some limitations that need to be overcome: (1) Confusion on the speaking characters when multiple characters are involved into a conversation, the hearing impaired user needs to determine from the script the active speaker. In addition, there are multiple cases when the face of the speaking character is not visible or does not appear at all in the video shot; (2) Tracking the subtitle: when the subtitle is displayed on the screen there is no information regarding the duration of each piece of script. If a character is speaking very rapidly, the text will have to update constantly. In this case, the viewer will simply miss a portion of the phrase. For various speaking characters, the display time can vary over a wide range. Therefore, the existing captioning approaches are still far from satisfactory in assisting the hearing impaired people (HIP) in enhancing perception/comprehension over the multimedia content. In this work we propose a novel framework dedicated to dynamically position the video subtitle and designed to help the HIP match the script with the corresponding character. The proposed methodology, illustrated in Fig. 1, jointly exploits deep convolutional neural networks and computer vision algorithms. The main contributions of the paper can be summarized as follows:

(1) A complete framework for automatic, dynamic positioning of the video subtitles/close captions that can suit the needs of any genre of multimedia documents.

(2) A face detection, tracking and recognition approach that exhibits high robustness to variations of the visual appearance due to changes in scale, pose, illumination, expression.

(3) A temporal segmentation method that partitions the video stream into scenes by exploiting the detected and recognized characters.

(4) A novel algorithm for robust active speaker detection, based on a multimodal fusion of text, audio and visual information. Compared with state of the art tech-

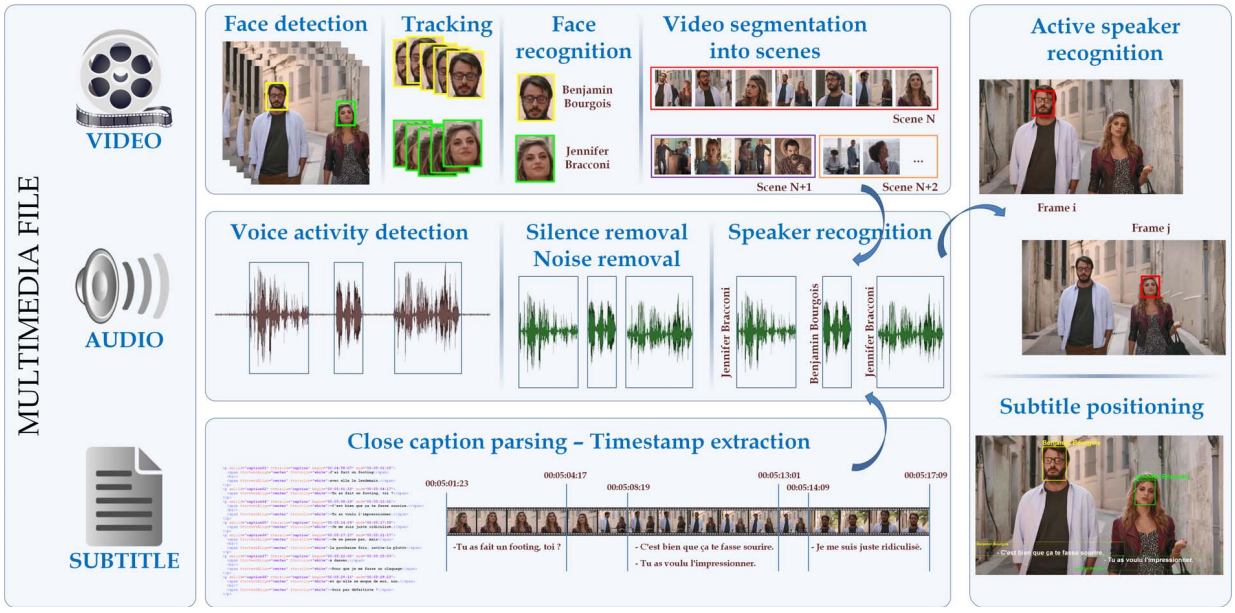

Figure 1. The proposed system architecture with the main steps involved.

niques [9], [10], our approach can efficiently cope with noisy video frames, characters with unfavorable face poses or even speakers that are not visible in the scene. In addition, our method is able to recognize the speaker even if the corresponding speech segments are acquired under unconstrained conditions, with music or background noise.

(5) An efficient optimization algorithm that positions the video subtitle on the screen by taking into account the detected active speaker. Using the proposed strategy, the HIP has an improved experience in watching multimedia documents making it easier to follow both the subtitle and the video document.

To the best of our knowledge, this is the most accurate, robust and complete solution designed to facilitate the access of hearing-impaired people to video content.

The rest of the paper is organized as follows. Section II reviews the state-of-the-art techniques dedicated to the dynamic positioning of video subtitles. Section III introduces the proposed architecture and describes the main steps involved: face detection, tracking, recognition, active speaker recognition, video temporal segmentation and subtitle positioning. Section IV presents the experimental results obtained on a large set of videos. Finally, Section V concludes the paper and opens some perspectives of future work.

2. Related work

Since the birth of subtitle/closed captioning for TV in the early 1970s, little efforts have been dedicated to accommodate the deaf or hearing impaired users in accessing multimedia content. One of the first papers addressing the issue of subtitle positioning is introduced in [14]. The system denoted by Comic Chart is designed to automate several aspects of comics generation such as: balloon construction, selection of character gestures, choice of zoom factor and expressions placement. Similarly, in [5] a method to automatically place the balloon in an image is proposed by keeping the 2D comics text layout rules. The proposed algorithm first positions each word balloon relative to its respective actor while maintaining the reading order of dialogues. The methods in [14] and [5] are focused in handling single frames and cannot be extended to video documents being unable to ensure cross frame coherence.

The earliest study on investigating the issue of dynamic subtitle positioning in multimedia documents is introduced in [9]. The system exploits a rich set of technologies including face detection and recognition, visual saliency analysis and text-to-speech alignment in order to position subtitle segments around the active speaker mouth. The active speaker is identified based on a lip motion analysis while the subtitle is positioned on the region with the lowest saliency score. The framework has been extended in [10], where an improved speaker detection method, exploiting both visual and audio information, is introduced. The system proves to be robust when multiple characters are moving their lips. In addition, the subtitles are positioned using an optimization procedure taking into consideration cross frame coherence and screen layout. However, both approaches suffer from several limitations:

(1) The active speaker needs to face the video camera (i.e., the system works solely on frontal or nearly frontal faces). For profile faces or, more generally, for characters

with unfavorable poses, the system fails. Moreover, the systems are sensitive to low quality videos.

(2) The subtitle positioning strategy can occlude useful information existent in the video scene. In addition, when the active character is moving the subtitle will fluctuate around the speaker mouth, which makes it difficult / tiresome to follow.

(3) Such approaches may be appropriate for relatively short sentences. For phrases covering two lines of text, the subtitle positioned around the speaker mouth may become disturbing.

In [4], authors present a study/analysis of the user experience when watching video with subtitles. Eye-tracking data is here exploited in order to compare the gaze patterns of subtitle users with a baseline corresponding to people that are watching to the videos without subtitles. The main conclusion is that the dynamic subtitle systems create gaze patterns that are closer to the baseline than those induced by regular subtitles. The work has been extended in [15], where a controlled eye-tracking study is developed to compare the impact of regular approach (center bottom subtitles) and context-sensitive, dynamic subtitles. The experimental results show that the static subtitle makes it easier to look around, but more difficult to understand the video content. In addition, speaker-following subtitles lead to higher fixation counts on relevant image regions and reduce saccade length.

A multimodal speaker naming is introduced in [11]. The system is based on deep convolutional neural networks to automatically learn the fusion function of both audio and visual cues. The experimental evaluation performed on three hours videos from two TV series (Friends and The Big Bang Theory) returns an accuracy score around 82%. However, the textual information is not really integrated into the process of speaker naming. In addition, the system performance is obtained for a reduced (11) number of known characters. Moreover, the system is able to recognize the active speaker solely when is visible within the camera field of view.

Another family of approaches concerns the so-called gaze-based subtitling position systems. As representative of such techniques, let us cite the approach recently introduced in [2] and further extended in [3]. In [2], the authors proposed to position the subtitle based on regions of interest that are extracted using the viewers gaze position. The system is sensitive to camera motion and outputs frequent subtitle position changes when the subject is moving. In order to deal with such limitations, in [3], a novel framework is proposed. The required input consists of: the video document, the subtitle file and the multiple viewer gaze data. The system returns as output the video with the subtitles displayed at different positions for each textual segment. However, the authors assume the active speaker eyes and mouth are visible in the scene, which is not always true. Furthermore, the gaze information is difficult to obtain, since it requires the set-up of a complex acquisition protocol that involves multiple users.

The state-of-the-art analysis highlights that the existing subtitle positioning approaches are still far from satisfactory in assisting the deaf and hearing-impaired users for video content consumption and understanding. None of the existent frameworks can address the issue of identifying the active speakers when they are not in a frontal position, with poor quality visual appearances or not present at all in the video scene.

In this paper we introduce a novel dynamic positioning system, designed to solve the above-mentioned limitations and to automatically position the subtitle segment in the vicinity of the active speaker. The following section describes the proposed approach, and details the various modules involved.

3. Proposed approach

Figure 1 illustrates the proposed system architecture with three different parts dedicated to each media channel involved in a video document: visual content, audio and text.

3.1. Video processing

Let us now detail each of the modules starting with the visual face detection, tracking and recognition.

A. Face detection, tracking and recognition

The face detection module is based on the Faster R-CNN architecture introduced in [21], extended with the Region Proposal Networks (RPN) approach [13]. We initialized the CNNs with a model pre-trained on the ImageNet database [7] and we trained the CNN on the WIDER database [27]. The system has been run for 100k iterations, at a learning rate of 0.001.

The face tracking is performed with the help of the ATLAS algorithm introduced in [17] and extended to the case of multiple moving instances. In order to identify the eventual presence of a new character in the scene, the face detection algorithm is applied to each individual face of the video stream. If a new character is detected, a new instance of the tracker is initialized with the novel face.

For each face detected and tracked between successive frames a low level representation is generated from the last layer before the classification layer of a CNN. In our framework, we have adopted the VGG16 [23] network architecture with batch normalization strategy introduced in [19].

The output of the VGG16 is a 4096-dimensional feature representation that is further normalized to the unit vector. Let us denote by $F = \{x_1, x_2, ..., x_L\}$ a face tracked in a video sequence of length L frames, where $x_i, i = 1, ..., L$ is

a face instance in the i^{th} frame of the considered video. For each x_i 4096-dimesional features f_i are extracted from the VGG16 CNN architecture.

Then, our objective is to create a global descriptor g(F) that aggregates all the face instances (tracked between the successive frames of the video stream) into a compact feature representation:

$$g(F) = \sum_i w_i \cdot f_i; \quad (1)$$

where $\{w_i\}_{i=1}^{L}$ is a set of real-valued, positive and unitary-normalized weights, with w_i the coefficient associated to the feature of the i^{th} frame.

The most important parameter in Eq. (1) is the set of weights that determines the relevance of a face pose in the global descriptor. In a nave approach [19] all face instances are treated similarly (equally important). However, such a strategy shows quickly its limitation when confronted to the high diversity of face poses existent in commercial video documents.

In order to deal with such limitations, we have designed a learning-based weight adaptation scheme that estimates the face degree of relevance depending on its viewing angle, pose, noise, compression artifacts. We propose to train a different CNN with only two categories denoted: relevant and trivial that will help us differentiate between various face instances.

In the relevant class we have included high-quality, aligned faces with little variation for the yaw, roll or pitch angles that are appropriate for recognition purposes. While, the trivial class includes low-quality face images: profile face poses, with linear motion / optical blur, with important video compression noise and at a reduce resolution (scale), whose impact on the recognition process should be minimized. The output of the CNN is the probability (sign(i)) of a face instance to be assigned to the relevant category. Higher scores are returned by frontal, unblurred and unoccluded faces.

Finally, the sign(i) coefficients are passed through a softmax operator:

$$w_i = \frac{exp(sign(i))}{\sum_j exp(sign(j))}; \quad (2)$$

in order to obtain normalized to unity weights w_i with $\sum_i w_i = 1$.

By assigning a global descriptor to a face track we are able to construct a fixed size representation for each face regardless on its number of instances. In Figure 2 we present some results of the weight adaptation module. As it can be observed, partially occluded, blurred or profile face instances have a reduce impact on the global descriptor that is further used for classification.

B. Video temporal segmentation

The video temporal segmentation involves two stages: shot boundary detection and scene segmentation.

The video shots s_i are identified using the graph partition strategy presented in [25]. Then, the shots are used in order to create video scenes that satisfy certain homogeneity with respect to a semantic criterion. By definition, a scene needs to respect three continuity rules, corresponding to consistency/homogeneity in space, time and action [20]. The proposed scene segmentation algorithm exploits the detected and recognized characters existent in various shots. Each shot is characterized by a set of face tracks and their associated global descriptors (cf. Section 3.1.A). By exploiting the visual similarity between the low level descriptors we construct a connected graph at the level of the video sequence. The cut edges of the graph correspond to scene boundaries.

Let us denote with $S = \{s_i\}_{i=1}^{N}$ the set of shots extracted from a video sequence, where:

$$s_i = \{g_{i,j}\}_{j=1}^{K_i}; \quad (3)$$

$g_{i,j}$ is the global face feature descriptor extracted from the j^{th} face track in the current shot s_i, N represents the total number of shots extracted for the video sequence and K_i denotes the number of faces tracked within the current shot. By using a temporal sliding window of size T we start evaluating the video shots in order to determine their degree of similarity. Two shots s_i and s_j are considered similar according to the following similarity measure:

$$D(s_i, s_j) = \max_{m,n}(\delta(g_{i,m}, g_{j,n})); \quad (4)$$

if they contain at least one common character. $\delta(.,.)$ denotes the cosine distances computed between the global low level face descriptors. Two video shots are clustered into the scene if the similarity score $D(s_i, s_j)$ is superior to an empirical established threshold.

When the same video shot is assigned to more than one cluster, the two clusters are directly merged.

Based on the generated clusters and the temporal order of shots, a graph is constructed, where the nodes are represented by clusters of shots and a direct edge is drawn from one node to another if the clusters are temporally interleaved. Each disconnected sub-graph will represent a video scene. Singular shots (i.e., not assigned to a scene) will be merged to its adjacent scenes based on the maximum similarity value computed between the shots color histograms. Finally, to each scene a list of characters is associated with.

However, the detected video scenes may contain in their structure multiple active personages. In order to position the subtitle segment to its corresponding character, we introduce an additional active speaker recognition framework, described in the following section.

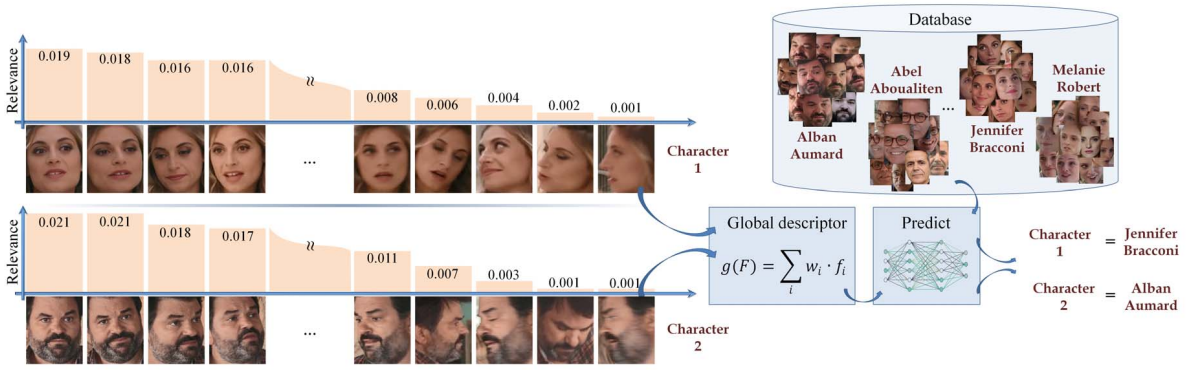

Figure 2. The proposed weight adaptation scheme.

3.2. Audio and text processing

The active speaker detection and recognition system receives as input the list of characters (cf. Section 3.1.B) associated to each video scene, the audio stream and the subtitle file. The system is designed to determine the speaker identity and to assign each textual segment to the corresponding character (Fig. 1).

A. Subtitle and audio stream processing

The video subtitle is a textual document containing the spoken lines and the timing information. The timestamps from a synchronized subtitle file provide the temporal interval during which a specific text segment is displayed on the user screen. Since this interval is always longer than the actual speaking duration, with additional time added at the beginning and end of the phrase, this information cannot be used directly in the process of active speaker identification. In addition, a subtitle segment may contain lines from more than one speaker (e.g., two characters having a conversation).

In order to deal with such limitations we propose to initially split the audio file into smaller audio segments (i.e., audio chucks) using directly the timestamps extracted from the synchronized subtitle file. So, the start and end times of an audio chunk will correspond to the display time associated to its corresponding text segment. Then, based on the observation that a subtitle segment belonging to two speaking personages contains in its structure a dedicated character (usually "-"), we proposed dividing the segment, together with its associated audio chunk, into two subsegments.

The start time (ST_0 and ST_1) and the end time (ET_0 and ET_1) for each audio sub-segment can be determined using the segment timestamps and its total number of letters ($No_{LETTERS}$) using the following equations:

$$T_{Letter} = \frac{ET_{Segment} - ST_{Segment}}{No_{LETTERS}}; \quad (5)$$

$$ST_0 = ST_{Segment}; ET_0 = ST_0 + T_{Letter} \cdot No_{S_0}; \quad (6)$$

$$ST_1 = ET_0; ET_1 = ET_{Segment}; \quad (7)$$

where $ST_{Segment}$ and $ET_{Segment}$ are the start time and the end time respectively, of the considered audio segment, while No_{S_0} represents the total number of letters from the first line of the textual segment.

Finally, in order to extract the actual speaking segments without pauses we applied on all audio chunks and subchunks traditional speech processing techniques, including voice activity detection, silence and unvoiced speech removal as indicated in [28].

The obtained filtered audio chunks will be used for the active speaker recognition purposes as presented in the following section.

B. Active speaker recognition

The speaker recognition framework receives as input the filtered audio chunks (cf. Section 3.2.A) and the list of characters for each video scene (cf. Section 3.1.B) and predicts the active character. In this context, the audio chunks are converted into spectrograms and we treat the speaker identification task as a multicategory classification problem.

The spectrogram is represented as vectors of size (257 x T x 1). We used a 512 point FFT (Fast Fourier Transform), returning 256 spectral frequency components, which together with the DC component give a STFT (Short Time Fourier Transform) of size 257. T corresponds to the temporal length of the audio chunk and 1 is the number of color channels used to represent the spectrogram. Then, the spectrogram is normalized by subtracting the mean and dividing with the standard deviation of all frequency components in one stage.

Since we treat the speaker identification as a multi-class classification problem, we have used a modified version of the residual-network ResNet-34 [8] architecture. As recommended in [6], we modified ResNet-34 in a fully convolutional way to adapt it to 2D spectrogram inputs. In addition as in [26], in order to produce a fixed-length output descriptor we extended the network with a NetVLAD layer for feature aggregation along the temporal axis. In this way, the output descriptor has a reduced dimensionality, is efficient to store and retrieve and requires reduced memory capabilities. Then, we applied the softmax operator in order to determine the distribution over all the considered classes of known characters. The top-3 candidates for each prediction are saved for a further analysis.

Finally, by comparing the current scenes characters list with the top-3 candidates returned by the modified ResNet-34 architecture, the system can make a prediction about the active speaker identity. If a unique match is obtained than the output is considered as correct and the sentence is assigned to the recognized character. When, multiple matches are identified, the category with the highest probability score is considered as correct. If no match is obtained than no prediction is generated.

3.3. Dynamic subtitle positioning

As previously mentioned, the proposed framework has as main objective to present the scripts near the speaking character such that the deaf and hearing impaired audience can easily identify from whom the script came from.

In order to dynamically position the subtitle near the speaker, several aspects need to be considered: (1) The script should be placed in the near vicinity of the active speaker so that no confusion with other personages can arise. (2) The script should not occlude important visual content and notably other faces that are present in the scene. (3) Across consecutive frames, the distance between the subtitle placements should be small enough to reduce eyestrain.

In order to address such requirements, we have defined 15 potential regions that are candidates for placing the subtitle (Fig. 3). The 15 areas that can receive the subtitle segment are determined by subdividing the horizontal axis into considering 3 parts and the vertical one into 5.

Regarding the horizontal locations, the subtitle can admit an offset relative to the left edge of a video frame of 0% (left), 20% (center) or 40% (right). For the vertical direction, the subtitle can admit an offset relative to the top part of the frame of 60%, 65%, 70%, 75% and 80%. In addition, we impose for the subtitle region not to cover more than 20% of the video frame height and 60% of its width.

Different from other subtitle positioning techniques such as those introduced in [9] and [10] that propose to place the subtitle segment near the character mouth, our approach has the advantage to reduce the subtitle position fluctuations for moving characters. Thus, a unique subtitle location is considered for a given face track within the entire video shot.

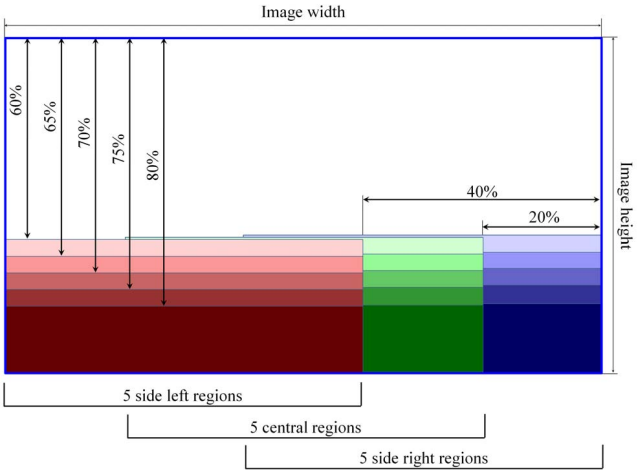

Figure 3. The proposed subtitle placement regions.

In order to determine the horizontal location for a subtitle segment we take as reference the active speaker face centroid and we determine the optimal location based on a majority voting scheme. The vertical position is established so that no overlap with important visual content and other faces occurs. Finally, let us mention that for the subtitle segments not assigned to any character, the default central position (i.e., as for regular subtitles) is retained.

4. Experimental evaluation

The proposed dynamic subtitle positioning framework has been tested on a variety of videos, in which the active speaker is moving within the video shot, the face is partially occluded or is not visible for several seconds. We have compared the proposed methodology with other state of the art techniques ([9] and [10]) and we show that our system allows an increase in the precision and recall scores with more than 8%. In addition, in order to completely evaluate our system, we have conducted a comprehensive usability study and compare it with conventional fixed position subtitling or with other dynamic subtitling methods ([9] and [10]).

4.1. The benchmark

The experiments were performed on a dataset with 30 video elements of 20 minutes, recorded at a resolution of 1024 x 576 pixels, at a frame rate of 25 fps. Specifically, ten movies were selected from the sitcoms Friends and The Big Bang Theory and twenty videos from the France Tlvisions TV series Un si grand soleil. All videos contain multiple speaking characters with significant switching dialogs.

4.2. CNN training

For face recognition purposes, the CNN has been trained on 110 categories of known persons. For each person, a maximum number of 800 face instances have been retained. The training has been performed with 50k iterations, with a batch size of 64 and at a learning rate of 0.0001. In addition, the network weights are initialized with the VGG16 face model [19] that achieves state of the art accuracy for face recognition tasks on static images.

For the weight adaptation module, the CNN architecture (VGG16) uses the same training parameters as for the face recognition module. The training process is quite effective and it takes less than 20 minute for 1M face instances on a GPU card (Nvidia 1080Ti) mounted on a desktop computer.

The active speaker recognition module shares the same list of personages as for the visual face recognition with 110 identities of which approximate half are male and half are female. We used for training the first 100 episodes for the Un si grand soleil series, while for Friends and The Big Bang Theory series we used all episodes from season I.

In this stage, the challenge was to automatically construct the learning database. To this purpose, we have applied the proposed face detection, tracking and recognition algorithm in order to identify the potential active speaker. Then, we determined if the visible face is actually the speaker. This is done by using a multi-view adaptation of SyncNet [12] a two stream CNN that establishes the correlation between the audio stream and the mouth motion. This strategy helps us reject clips with voice-over or where the speaker is not visible. The generated database contains approximate 13000 appearances for all the considered characters. The lowest length for an audio segment is restricted to 1.25 seconds. The retained audio streams are very challenging and include: background chatter, laughter, noise, music and overlapping speech.

4.3. Objective system evaluation

The system evaluation was performed on a set of 30 video elements (cf. Section 4.1) for which the ground truth was established by manually annotation. For the test dataset, 15208 sentences have been identified. Let us underline that more than 3500 unknown individuals are present in the test videos. The objective evaluation has been performed using traditional metrics including accuracy (A), recognition rate (R) and F1 norm [18]. The experimental results are summarized in Table 1. As it can be observed, from the 15208 sentences, a number of 13547 represent known active speakers existent in the training database. We have compared the proposed system with state of the art techniques introduced in [9] and [10]. The following conclusions can be highlighted:

(1) The lowest performance is obtained by Hong et al. [9] method that determines the active speaker based solely on lip motion. This approach fails when other people are in the scene moving their lips, or when the active speaker is not visible in the camera field of view;

(2) The methods of Hu et al based on both visual and audio information returns superior recognition scores than Huangs. However, the system cannot handle scenes where the active speaker mouth is occluded or when the speaker is not visible. In addition, the system is sensitive to the character movement;

(3) Our approach returns gains in precision and recall of more than 8%. This result can be explained by an enhanced robustness of the proposed framework to important face pose variations and relatively important camera/face motion. Some sample outputs of the proposed dynamic subtitle positioning system are presented in Fig. 4.

4.4. Subjective system evaluation

The qualitative system evaluation has been performed with the help of 30 anonymous volunteers (19 male and 11 females). During testing, the audio channel was set to mute when examining the video subtitles. We compared the following three paradigms:

(1) Static Caption (SC) the participants were shown the videos with static caption (here we adopted the regular cinematic captioning);

(2) Speech Bubbles (SB) the subtitle is dynamically positioned near the active speaker mouth using the strategy in [9] and [10];

(3) The proposed Dynamic Caption (DC) the subtitle is positioned in one of the 15 potential regions of interest described in Section 3.3.

We randomly divide all participants into 3 groups in order to avoid repeatedly playing a video to a set of subjects which will cause knowledge accumulation. Therefore, each group of 10 participants merely evaluates one of the three paradigms on the assigned set of video clips. After the evaluation, the following conclusions can be highlighted:

(1) The dynamic caption remarkably outperforms the SC in terms of naturalness and enjoyment;

(2) The SB is somehow disturbing because there are several cases when script is continuously fluctuating around the subjects head. Under these circumstances the users spent too much attention on the subtitle variation and some of them missed the video action;

(3) Systems relying solely on mouth analysis to detect the active speaker usually fail in complex scenes or when the speaker mouth is occluded;

(4) Users expressed interest in using the proposed DC system because our framework can handle shots where the active speaker is not visible. In addition, the users said that the information added to the subtitle regarding the identity of the active speaker is useful to increase the comprehension over the video subject.

Table 1. Comparative experimental results for the proposed system

Method	Ground truth (Total sentences / Known individuals)	True Positive (TP)	False Positive (FP)	False Negative (FN)	Accuracy (%)	Recognition Rate (%)	F1 score (%)
Hong [9]	15208/13547	10567	2478	2980	0.81	0.78	0.79
Hu [10]		10973	1496	2574	0.88	0.81	0.84
Ours		11803	226	1744	0.98	0.87	0.92

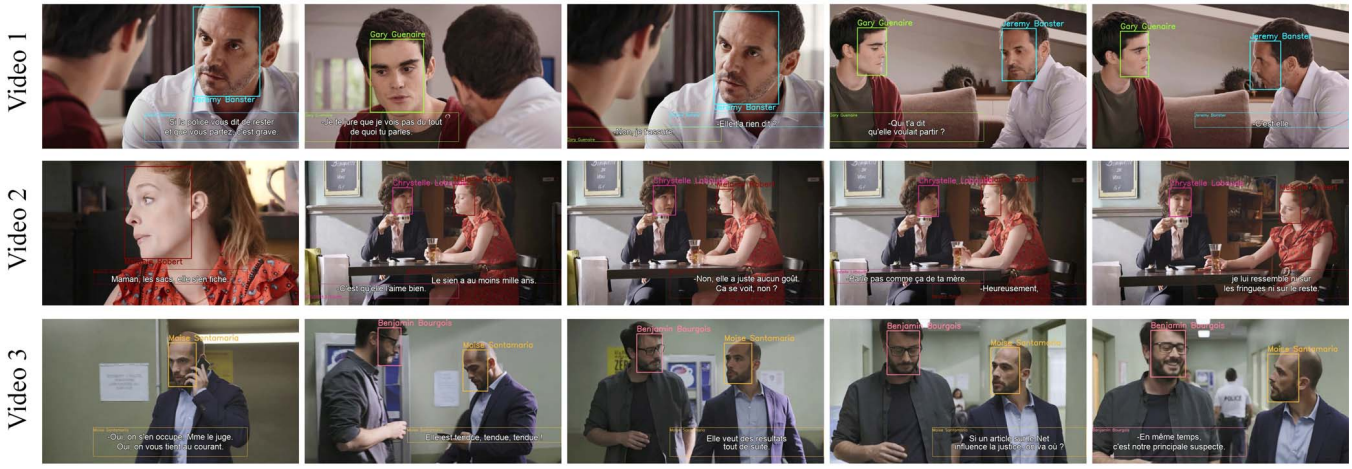

Figure 4. Experimental results of the proposed dynamic captioning system.

5. Conclusions and perspectives

In this paper, we have introduce a novel dynamic subtitle positioning framework designed to enhance the deaf and hearing impaired users perception and viewing experience when watching multimedia documents.

The proposed methodology is based on a multimodal fusion of information from textual subtitles, audio and video streams and proves to be robust to important object/camera motion, face pose variation or audio and visual noise. The experimental evaluation performed on a large dataset validates the proposed method with average F1-scores superior to 92%. When compared with other techniques from the state of the art [9] and [10], our framework shows an increase in performances of more than 8%.

For further work and developments, we aim to extend the system [16] to cope with sound to text libraries and online video streams for which the subtitle file is not available in advance. In addition, we envisage performing a wider study on hearing impaired users.

References

[1] http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss - accessed on the 29 of may 2019.

[2] W. Akahori, T. Hirai, S. Kawamura, and S. Morishima. Region-of-interest-based subtitle placementusing eye-tracking data of multiple viewers. *Proc. ACM International Conference on Interactive Experiences for TV and Online Video,*, pages 123–128, 2016.

[3] W. Akahori, T. Hirai, and S. Morishima. Dynamic subtitle placement considering the region of interest and speaker location. *Proc. of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 102–109, 2017.

[4] A. Brown, R. Jones, M. Crabb, J. Sandford, M. Brooks, M. Armstrong, and C. Jay. Dynamic subtitles: the user experience. *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pages 103–112, 2015.

[5] B.-K. Chun, D. S. Ryu, W. I. Hwang, and H. G. Cho. An automated procedure for word balloon placement in cinema comics. *Proc. Int. Symp. Visual Computing*, 2006.

[6] J.S. Chung, A. Nagrani, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. *Proc. Interspeech*, 2018.

[7] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[9] R. Hong, M. Wang, M. Xu, S. Yan, and T.S. Chua. Dynamic captioning: Video accessibility enhancement for hearing impairment. *Proc. ACM Conf. Multimedia*, pages 421–430, 2010.

[10] Y. Hu, J. Kautz, Y. Yu, and W. Wang. Speaker-following video subtitles. *ACM Trans. Multimedia Comput. Commun. Appl.*, 11(2):1–17, 2014.

[11] Y. Hu, J. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang. Deep multimodal speaker naming. *Proc. 23rd Annu. ACM Int. Conf. Multimedia*, 2015.

[12] A. Zisserman J. S. Chung. Out of time: Automated lip sync in the wild. *Proc. ACCV*, pages 251–263, 2016.

[13] H. Jiang and E. G. Learned-Miller. Face detection with the faster r-cnn. *12th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 650–657, 2017.

[14] D. Kurlander, T. Skelly, and D. Salesin. Comic chat. *ACM Computer Graphics Annual Conf. Series*, pages 225–236, 1996.

[15] K. Kurzhals, E. Cetinkaya, Y. Hu, W. Wang, and D. Weiskopf. Close to the action: Eye-tracking evaluation of speaker-following subtitles. *Proc. of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6559–6568, 2017.

[16] M. Leo, A. Furnari, G. Medioni, M. Trivedi, and G.M. Farinella. Deep learning for assistive computer vision. *In Proceedings of the European Conference on Computer Vision*, page 11134, 2018.

[17] B. Mocanu, R. Tapu, and T. Zaharia. Single object tracking using offline trained deep regression networks. *Seventh International Conference on Image Processing Theory, Tools and Applications*, pages 1–6, 2017.

[18] B. Mocanu, R. Tapu, and T. Zaharia. Deep-see face: A mobile face recognition system dedicated to visually impaired people. *IEEE Access*, 6:51975–51985, 2018.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *In British Machine Vision Conference*, 1:1–6, 2015.

[20] Z. Rasheed and M. Shah. Detection and representation of scenes in videos. *IEEE Transactions on Multimedia*, 7(6):1097–1105, 2005.

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real- time object detection with region proposal networks. *NIPS*, 2015.

[22] B. Safadi, M. Sahuguet, and B. Huet. When textual and visual information join forces for multimedia retrieval. *Proceedings of International Conference on Multimedia Retrieval*, pages 265–272, 2014.

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *In Proc. ICLR*, 2015.

[24] A. Tamayo and F. Chaume. Subtitling for d/deaf and hard-of-hearing children: Current practices and new possibilities to enhance language development. *Brain Sci*, 7(7):75–77, 2017.

[25] R. Tapu, T. Zaharia, and F. Preteux. A scale-space filtering-based shot detection algorithm. *IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pages 919–923, 2010.

[26] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman. Utterance-level aggregation for speaker recognition in the wild. *Proc. ICASSP*, 2019.

[27] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.

[28] R. Yin, H. Bredin, and C. Barras. Speaker change detection in broadcast tv using bidirectional long short-term memory networks. *in Proc. Interspeech 2017*, pages 3827–3831, 2017.