Video Accessibility Enhancement for Hearing-Impaired Users

RICHANG HONG and MENG WANG, Hefei University of Technology
XIAO-TONG YUAN and MENGDI XU, National University of Singapore
JIANGUO JIANG, Hefei University of Technology
SHUICHENG YAN and TAT-SENG CHUA, National University of Singapore

There are more than 66 million people suffering from hearing impairment and this disability brings them difficulty in video content understanding due to the loss of audio information. If the scripts are available, captioning technology can help them in a certain degree by synchronously illustrating the scripts during the playing of videos. However, we show that the existing captioning techniques are far from satisfactory in assisting the hearing-impaired audience to enjoy videos. In this article, we introduce a scheme to enhance video accessibility using a Dynamic Captioning approach, which explores a rich set of technologies including face detection and recognition, visual saliency analysis, text-speech alignment, etc. Different from the existing methods that are categorized as static captioning, dynamic captioning puts scripts at suitable positions to help the hearing-impaired audience better recognize the speaking characters. In addition, it progressively highlights the scripts word-by-word via aligning them with the speech signal and illustrates the variation of voice volume. In this way, the special audience can better track the scripts and perceive the moods that are conveyed by the variation of volume. We implemented the technology on 20 video clips and conducted an in-depth study with 60 real hearing-impaired users. The results demonstrated the effectiveness and usefulness of the video accessibility enhancement scheme.

Categories and Subject Descriptors: C.4 [Performance of Systems]: Design studies; H.1.2 [Models and Principles]: User/Machine Systems—Human factors; H.5.1 [Information Interfaces and Presentation (e.g., HCI)] Multimedia Information Systems

General Terms: Experimentation, Human Factors, Performance

Additional Key Words and Phrases: Accessibility, dynamic captioning, hearing impairment

This work was supported by NRF/IDM Program of Singapore, under Research Grants NRF2007IDM-IDM002-047 and NRF2008IDM-IDM004-029.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.


1. INTRODUCTION

Video is an important information carrier that presents visual and audio content in live form. With rapid advances of capturing and storage devices, networks and compression techniques, videos are growing in an explosive rate and play an increasing important role in people's daily life. However, there are millions of people that are suffering from hearing impairment. They are fully or partially unable to perceive sound. It is estimated that there are more than 66 million people with hearing impairment, of which about 41% cannot hear any speech at all and 59% are able to hear only if words are shouted around their ears.\(^1\) This disability brings them great difficulty in comprehending video content as auditory information is lost or incomplete in their hearing.

There are two typical approaches to helping these special audience better access videos. The first is “direct access”, which provides access as part of the previously developed system [Nielsen 1995]. However, merely providing access is not sufficient for the hearing-impaired audience. Therefore, most attempts have focused on the second approach, that is, the so-called assistive approach. For a large family of videos that have associated scripts,\(^2\) such as movies, television programs and documentary, captioning is the most widely applied assistive technique. By synchronously illustrating the scripts during the playing of videos, hearing impaired audience can obtain the necessary information from texts. Generally, captioning can be categorized into open captions and closed captions in accordance with whether it is able to be activated by users; and it can also be presented in a variety of styles. However, the existing captioning methods mostly resemble each other as the scripts are simply demonstrated in a fixed region and they are illustrated statically. Although the hearing-impaired audience can get certain information from the scripts, they still encounter difficulty in the following aspects.

1. Confusion on the Speaking Characters. When multiple characters are involved in a scene, the hearing-impaired audience needs to judge from which person the scripts come, and this increases their difficulty of content understanding and also degrades their experience of video enjoyment.

2. The Tracking of Captioning. In video playing, there is no hint on the duration of each piece of script. As speaking pace can vary significantly, the duration of the script presentation will also vary over a wide range. This brings the hearing-impaired audience difficulty in the tracking of scripts. For example, they may miss a part of a sentence when the character is speaking rapidly.

3. The Loss of Volume Information. The variation of volume conveys important information about the emotion [Xu et al. 2008; Juslin and Scherer 2009]. For example, the sound of a character will be loud if he/she becomes happy or angry. However, such information is not illustrated in the existing captioning technology.

Therefore, the existing captioning approach is far from satisfactory in assisting the hearing-impaired audience. Recent studies reported that the conventional captioning approach can hardly add significant

\(^1\)http://en.wikipedia.org/wiki/Deaf.

\(^2\)These videos are also called multimedia videos. Although there are also many videos that have no script information, as mentioned in Section 5, our scheme can be extended to deal with general videos by further exploring speech recognition and speaker identification technologies. Actually, this work is just our primary step towards helping hearing-impaired people better access video content.
information for the hearing impaired audience's perception [Gulliver and Ghinea 2003a, 2003b]. One major reason is that the audience can hardly track the scripts and match them with visual content rapidly.

In this work, we propose a novel approach named dynamic captioning to enhance the accessibility of videos for hearing impairment. Compared with the existing captioning methods, which are categorized as static captioning, dynamic captioning is able to help hearing-impaired users match the scripts with the corresponding characters. Dynamic captioning is also able to synchronize the scripts word-by-word with the speech as well as highlight the variation of voice volume. In this way, the aforementioned three problems can be addressed.

The dynamic captioning is accomplished by exploring a diverse set of technologies, including face detection and recognition, lip motion analysis, visual saliency analysis, etc. The scheme contains three main components: script location, script-speech alignment, and voice volume estimation. Script location determines the region in which scripts will be presented. It first performs a script-face matching to identify the talking face for each piece of scripts (i.e., establish the person from whom the scripts are coming) based on face detection and recognition techniques. It then selects a nonintrusive region around the face via visual saliency analysis in order to avoid the occlusion of important visual content. Script-speech alignment temporally matches each piece of script and the corresponding speech segment, and in this way the scripts can be highlighted word-by-word in synchrony with the speech. Voice volume estimation computes the magnitude of audio signal in a small local window, and visually demonstrates its variation within the scripts. The main contributions of this article can be summarized as follows.

(1) We propose a video accessibility enhancement scheme for the hearing-impaired audience. To the best of our knowledge, this is the first integrated solution of this kind to facilitate hearing-impaired users in video access.
(2) Our scheme involves the combination of a variety of technologies as well as novel methodologies. For example, the script-face mapping is an important topic per se and our algorithm can be applied to many other applications.
(3) We conduct an in-depth user study to compare different captioning paradigms with hearing-impaired subjects. Results also shed light on further research in this direction.

The organization of the rest of this article is as follows. In Section 2, we provide a review on related work. Section 3 introduces the system overview of video accessibility enhancement. In Section 4, we describe the components of the scheme in detail. Experimental results and user study are presented in Section 5. Finally, we conclude the article in Section 6.

2. RELATED WORK

Accessibility is defined as the capability of being used or seen in the Merriam-Webster dictionary. However, this term actually has more specific meaning, accessibility is a general term used to describe the degree to which a product is accessible by as many people as possible. It is often used to focus on people with disability and their rights of access to entities, often through the use of assistive technology. Many efforts have been dedicated to accommodate users with disability in accessing multimedia. A typical example is that: with the popularity and widespread use of search engines, the accessibility aspect is receiving increasing attention in order to help people with disabilities more efficiently find and enjoy web content. In this scenario, “image search accessibility” is put forward to provide a scheme to

help people suffering from colorblindness to access image search. Besides the Google Accessible Search and Baidu Elderly Search, Azzopardi et al. [2009] proposed a PuppyIR system that aims to serve children in web search. Arrue and Vigo [2007] proposed a scheme that estimates the accessibility of search content and then generates ranking lists accordingly. Fajardo et al. investigated approaches to improve deaf users’ accessibility in hypertext search with the help of graphical interfaces. Wang et al. [2009c] evaluated the accessibility for each image and proposed an efficient scheme for search improvement.

Deafness refers to conditions in which individuals are fully or partially unable to detect or perceive at least some frequencies of sounds. Efforts on accommodating hearing impaired people in accessing television have been made since 1970s when closed captioning was demonstrated at the First National Conference on Television in Nashville, Tennessee. As previously described, there are two approaches to facilitating hearing impaired individuals’ access to multimedia content, which are direct access and assistive technologies. More recently, assistive technologies stand out as being able to fill any accessibility shortfalls of its original design. We can categorize the assistive technologies into communication systems and captioning techniques. Communication systems interpret between different media by performing translation of speech to text, speech to video sign language and text to computer-generated voice or video sign language [Cox et al. 2002]. On the other hand, captioning techniques is the most common form of assistive technology for hearing-impaired people and is a synchronized textual alternative to audio.

Although similar in appearance, there are slight differences between captioning techniques used in television program and movie due to their different production process and watching settings. The paramount difference attributes to the different standards. For television, captions are encoded into Line 21 of the vertical blanking interval in NTSC programming, while teletext is used in captioning transmit and storage in Phase Alternate Line and Sequential Color with Memory. For movies, probably the best-known closed caption in theatres is the Rear Window Captioning System from the National Center for Accessible Media. Besides that, other captioning technologies for movie include hand-held displays similar to PDA (Personal Digital Assistant), eyeglasses fitted with a prism over one lens and projected bitmap captions. More recently, special effort has been made to build accessibility features into digital cinema. Other difference in captioning techniques between television programs and movies is that television program may contain a combination of transcripts available beforehand and those available lively during the program. Such differences, from another point of view, verified that captioning techniques have been widely used in multimedia video. However, there are evident drawbacks of the existing captioning techniques. On one hand, there is little interactivity and flexibility, that is, audiences have to passively receive caption. On the other hand, some methods require hardware equipment support such as the assistive systems in theater that need to be distributed to audience to enable them to work. Certainly, this will increase the cost in management.

Despite many captioning standards and technologies have been made, the analysis of the impact of captioning in hearing impaired audience is fairly scarce. The earliest study on investigating caption perception of the hearing-impaired audience shows that adjusting captions to suitable linguistic level and reading rate is able to significantly improve the information gain from captions [Boyd and Vade 1972]. Braveman and Hertzog [1980] concluded that the language level but not the rate of captioning affected deaf users’ comprehension. A rate of 60 words per minute was found to improve program comprehension while the reduction of language level in captions frustrated many within the deaf community as information is summarized or ignored. Jelinek and Jackson [2001] assessed both hearing impaired and hearing students’ comprehension of caption with and without video. They declared that comprehension test scores for students who were deaf were consistently below the scores of hearing.

---

students. Furthermore, the captioned video provides deaf students with a significantly better comprehension of the script. Garrison et al. [1997] studied how working memory affected language comprehension for deaf students and found that reading comprehension depends on readers’ background knowledge and lexical abilities. Because of the reduced level of hearing, deaf individuals usually had a reduced or limited variety of learning interactions that were considered as the key factor in increasing language ability. Gulliver and Ghinea [2003a, 2003b] investigated the impact of captions on the perception of video clips for the hearing-impaired audience. They concluded that much information can be gained from caption, but the information from other sources such as visual content and video text would be significantly reduced, that is, the caption had no significant effect on the average level of assimilated information across all sources. This indicates that it is not easy for the special audience to track, perceive and learn from the caption efficiently.

Therefore, the existing captioning technology is still far from satisfactory in assisting the hearing-impaired audience (in Section 1 we have introduced several shortcomings). Recently, the closed captioning on YouTube is a meaningful exploration towards helping hearing impaired users access web videos. There also exists software, such as Captioneer,\(^5\) that is able to support manual editing of captions or even add several attractive effects. However, they still can not fully address the aforementioned problems and manual editing is also not an ideal solution due to the high labor cost. In this work, we investigate an automatic approach to intelligently present caption. It puts scripts in suitable regions, aligns them with speech and also illustrates the variation of voice volume. The user study with hearing-impaired audience has demonstrated the effectiveness of this approach.

3. DYNAMIC CAPTIONING: SYSTEM OVERVIEW

Figure 1 demonstrates the schematic illustration of our video accessibility enhancement process. It contains three main components: script location, script-speech alignment, and voice volume estimation. Given a video along with its script and subtitle file,\(^6\) we first extract video segments with speech

---

\(^5\)http://www.tsstech.org/captioneer.html.

\(^6\)In some cases, “subtitle” may assume the audience can hear but cannot understand the language or accent but “caption” aims to describe to the hearing-impaired all significant audio content. In this study, we restrict subtitle as the text that has time information and dialog. There also exists subtitle that contains richer content but it is not easily to acquire.
in accordance with the time information in subtitle. We then map the character faces to the corresponding scripts with face detection and recognition techniques. After that, a nonintrusive region is detected around the face based on visual saliency analysis, in which the scripts are presented.

In parallel, the scripts are aligned with the audio track based on script-speech technology [Moreno 1998], and the starting and end time of each word are recorded. Based on this information, we synchronously highlight the scripts word-by-word along with the speech so that the hearing-impaired audience can better track them. The voice volume estimation component estimates the local power of the audio signal. We visualize it near the scripts to help audience understand the emotion of the corresponding characters.

Our scheme thus generates a set of metadata in addition to the scripts, including the region information of each piece of script, the starting and ending time of each word and the voice volume information. In our work, we use an XML file to record these metadata. With these metadata, we can easily develop an intelligent player to display videos with dynamic caption.

4. DYNAMIC CAPTIONING: TECHNOLOGIES

In this section, we introduce the three components in the proposed scheme in details.

4.1 Script Location

This subsection describes our script location algorithm that finds suitable regions for presenting the scripts. As shown in Figure 1, it comprises three steps: (1) face detection, tracking and grouping; (2) script-face mapping; and (3) nonintrusive region detection.

4.1.1 Face Detection, Tracking and Grouping. In several cases, script file only contain speech content and speaker identity and there is another subtitle file that records the time information, as illustrated in Figure 2. Therefore, we need to merge the speech content, speaker identity and time information from the subtitle and script. Here we utilize a dynamic time warping method [Everingham et al. 2006] to align subtitle and script. Figure 2 demonstrates an example of such alignment. Of course, this step can be eliminated if there is only a script file encoding all the information.

Next we implement a face detector to extract faces from the frames in the video segments with speech. Here we adopt the face detection algorithm in Viola and Jones [2001]. As a video may contain thousands of or even more detected faces, we group continuously detected faces of a particular
character as a face “track” with a robust foreground correspondence tracker [Yang et al. 2005]. The tracker mainly works as follows. Given a pair of faces in adjacent frames, the size of overlapped area between the two bounding boxes of faces is estimated. If this value is greater than a given threshold, a match is declared. This tracking procedure is also able to deal with the cases that faces are not continuously detected due to pose variation or expression change. In this way, the number of tracks can be significantly reduced (typically, we only need to deal with hundreds of such tracks). As a consequence, face track is adopted as the unit for labeling.

4.1.2 Script-Face Mapping. Now we consider the script-face matching problem. The difficulty mainly lies on the following two facts: (1) In many cases, there are more than one face within a frame and we need to judge who is the speaker; and (2) Even when there is only one face in the frame, he/she may not be the speaker and scripts come from another character is an example and the girl in the frame is actually not speaking). To deal with these problems, first we adopt lip motion analysis [Saenko et al. 2005] to establish whether the character is speaking when the frame contains only one face based on the fact that speaking is associated with distinctive lip movement.

The lip motion analysis is performed as follows. First, we detect a rectangular mouth region within each detected face region using Haar feature based cascade mouth detector. We then compute the mean squared difference of the pixel values within the mouth region between each two consecutive frames. To keep the translation invariance, the difference is calculated over a search region around the mouth region in the current frame and we then take the minimal difference for decision. Two thresholds are set to establish three statuses, namely “speaking”, “nonspeaking”, and “difficult to judge”.

Now we consider the cases that a frame contains more than one face. Our approach is to first label faces with speaker identities and then match them with scripts accordingly (the script file contains the speaker identity information). Note that in cases in which the frame only contains one face, we can easily label the face with speaker identity. We then label the face tracks with identities based on such information. For example, if over half of the faces in a track are detected as in the speaking status and the script shows that merely “EDWARD” is speaking in this period, then we can label this track as “EDWARD” with high confidence. Those labeled tracks with high confidence are treated as training exemplars to predict other tracks that are unlabeled due to not containing enough established identities. Each unlabeled face track is simply represented as a set of history image feature vectors. One simple method for identification, as conducted in Everingham et al. [2006] and Wang et al. [2009b], is to directly calculate the feature distance between a test face track and exemplar face tracks, and then assign test face track to the nearest neighborhood. Another feasible method is to classify each history image independently via certain classification methods such as sparse representation-based classification [Wright et al. 2009; Wang et al. 2009c; Tang et al. 2009], and then assign the face track to the class that achieves the highest frequency.

In this work, by regarding the identification of each history image in a test face track as a task, we formulate the face track identification challenge as a multitask face recognition problem. This motivates us to apply the multitask joint sparse representation model [Obozinski et al. 2009] to accomplish the identification. The key advantage of multitask learning is that it can efficiently make use of the complementary information embedded in different subtasks. We construct the representation of face appearance with a part-based descriptor extracted around local facial features [Everingham et al. 2006]. Here we first use a generative model [Arandjelovic and Zisserman 2005] to locate nine facial key-points in the detected face region, including the left and right corners of two eyes, the two nostrils and the tip of the nose and the left and right corners of the mouth. We then extract the 128-dim Sift descriptor from each key-point and concatenate them to form a 1152-dimensional face descriptor (SiftFD).
The employed face detection, tracking as well as speaker detection are able to offer a number of face tracks where the proposed identity is correct with high probability. For tracks that contain only a single identity, they can be treated as exemplars for labeling other tracks that contain no, or uncertain, proposed identity. Each unlabeled face track is, nevertheless, simply represented as a set of history image vectors. For such history image in the track, the identification can be efficiently done via sparse representation classification [Wright et al. 2009].

Our multitask joint sparse representation model works as follows. Suppose that we have a set of exemplar face tracks with $M$ subjects. Denote $X = [X_1, \ldots, X_M]$ as the feature matrix in which the track $X_m \in \mathbb{R}^{d \times p_m}$ is associated with the $m$th subject consisting of $p_m$ samples. Here $d$ is the dimensionality of feature and $\sum_{m=1}^{M} p_m = p$ is the total number of samples. Given a test face as an ensemble of $L$ history images $y^l \in \mathbb{R}^d$, we consider a supervised $L$-task linear representation problem as follows:

$$ y^l = \sum_{m=1}^{M} X_m w_m^l + \epsilon^l, \quad l = 1, \ldots, L, $$

(1)

where $w_m^l \in \mathbb{R}^{p_m}$ is a reconstruction coefficient vector associated with the $m$th subject, and $\epsilon^l$ is the residual term. Denote $w^l = [(w_1^l)^T, \ldots, (w_M^l)^T]^T$ as the representation coefficients for probe image feature $y^l$, and $w_m = [w_1^m, \ldots, w_M^m]^T$ as the representation coefficients from the $m$th subject across different case images. For simplicity, we denote $W = [w_m^l]_{m,l}$. Our proposed multitask joint sparse representation model is formulated as the solution to the following multitask least square regressions with $\ell_{1,2}$ mixed-norm regularization problem:

$$ \min_W F(W) = \frac{1}{2} \sum_{l=1}^{L} \left\| y^l - \sum_{m=1}^{M} X_m w_m^l \right\|_2^2 + \lambda \sum_{m=1}^{M} \| w_m \|_2, $$

(2)

here we use the Accelerated Proximal Gradient (APG) approach [Tseng 2008] to solve the optimization problem in Eq. (2).

When the optimum $\hat{W} = [\hat{w}_m^l]_{m,l}$ is obtained, a test image $y^l$ can be approximated as $\hat{y}^l = X_m \hat{w}_m^l$. For classification, the decision is ruled in favor of the class with the lowest total reconstruction error accumulated over all the $L$ tasks:

$$ m^* = \arg \min_m \sum_{l=1}^{L} \left\| y^l - X_m \hat{w}_m^l \right\|_2^2. $$

(3)

After labeling each face track with a speaker identity, we can identify the speaking character even when there are more than one face in a frame. Hitherto we have accomplished the mapping between scripts and faces. It is worth mentioning that there also exist scripts that cannot be successfully mapped to faces, and in this work we directly display them at the bottom of frames just like static captioning (off-screen voice is also processed in the same way).

4.1.3 Nonintrusive Region Detection. Up to now, we are able to establish the speaker identity of each piece of scripts. As previously mentioned, our target is to present the scripts near the speaking face such that hearing impaired audience can easily identify the character from whom the scripts come from. However, we also need to select a region that will not occlude important visual content and especially other faces. Therefore, we perform a visual saliency analysis to select the nonsalient regions.
Given an Image $I$, the contrast of each pixel is an accumulated Gaussian distance between it and its neighbors:

$$c_{i,j} = \sum_{q \in \Theta} d(I_{i,j}, q),$$

where $I_{i,j}$ is the pixel position in $I$ and $\Theta$ is the neighborhood of $I_{i,j}$. The contrasts $c_{i,j}$ thus form a saliency map [Ma and Zhang 2003; Wang and Zhang 2009]. In Hong et al. [2010], Figure 5(b) shows an example of the saliency map of the image in Figure 5(a) where the distance is measured in LUV color space. The brighter the pixel in the saliency map, the more important or salient it is.

For the detection of the nonintrusive regions, $I$ is represented by a set of blocks $B = \{b_i\}_{i=1}^{N_b}$ that are obtained by partitioning image $I$ into $M \times M$ grids ($N_b = M^2$). Each grid corresponds to a block $b_i$ and it gives a candidate region of caption insertion. For each block $b_i$, a saliency energy $s_i$ ($0 \leq s_i \leq 1$) is computed by averaging all the normalized energies of the pixels within $b_i$. As previously analyzed, the region should be selected around the speaking face. Therefore, a face weighting map $W = \{w_i\}_{i=1}^{N_b}$ is designed to weight the energy $s_i$, so that the caption will be restricted around the face. The face weighting map is generated by simply assigning the blocks around the speaker’s face block constant weights and all other regions are assigned a weight of 0. More specifically, the weights of the left and right regions around the face region are set to 1, and the weights of the upper-left, bottom-left, upper-right and bottom-right regions are set to 0.8. Figure 5(c) shows an example of the weighting map. Hence, the score for region selection is given by:

$$P(b_i) = w_i \times (1 - s_i).$$

The region with the maximal score is finally established for caption insertion. In our work, the parameter $M$ is empirically set to 5, but it is also found that an adaptive setting of the parameter will be able to improve performance.

It is worth mentioning that although the nonintrusive region detection approach is effective, it cannot fully guarantee that informative visual content will not be occluded by caption. Thus, in dynamic caption, we choose to overlay the scripts with parent background such that audience can still see partial content behind the caption.

### 4.2 Script-Speech Alignment

This section describes our script-speech alignment approach. As previously mentioned, based on this component we can synchronously highlight the scripts word-by-word and help the hearing-impaired audience better track the scripts. Here we adopt a method based on recursive speech recognition with a shrinking dictionary and language model, which is analogous to the approach in Moreno [1998]. Figure 3 demonstrates a schematic illustration of our scheme. We use 39-dimensional MFCC features. The text analysis module processes the text file and we use the CMU pronouncing dictionary to translate each word into a phonetic sequence. For those words that are out of the dictionary, we use an automatic module introduced in Daelons and Bosch [1993] to process them. To reduce the computation cost, we build a simple bigram and trigram word model instead of a complete language model based on $N$-gram. We then use SPHINX II [Huang et al. 1993], a speaker-independent speech recognition engine, to recognize the speech based on the previously generated language model and dictionary. When a complete hypothesis text string is produced for the whole audio stream, we employ dynamic programming to find the globally optimum alignment. The detailed process is as follows. We compare the scripts and the recognition results and the matched parts that contain more than $N$ words are regarded as anchors. In our work, we empirically set $N$ to 3.
We then iterate the algorithm on each unmatched segment. In each iteration, the language model and dictionary are rebuilt to limit the list of active words and word sequence to those found in the script of this segment. This can speed up the recognition as we only search for those words and their word pairs and triples that are available in the segment. These steps are repeated on the unmatched segments until all the texts have been matched. The iteration also terminates if the recognizer is unable to find any additional words in the audio segment. Our test on 20 video clips (the data are described in Section 5) shows that this approach is able to obtain accuracy, that is, the percentage of correctly aligned words, of above 90%.

4.3 Voice Volume Analysis

Existing studies reveal that the variation of voice volume conveys important information about human emotion [Cu et al. 2006; Xu et al. 2008]. However, for hearing-impaired audience, the volume information is fully or partially lost. Therefore, in our dynamic captioning scheme, we symbolize and illustrate the voice volume to help the special audience get more information.

We estimate the sound volume by computing the power of the audio signal in a small local window (the size of the window is set to 30 ms in our scheme). After a normalization process, the estimated volume is visualized near the scripts with an “indicator”. The volume is indicated by the highlighted part of a strip and the size of the part will vary in accordance with the estimated power.

5. EXPERIMENTS

In this section, we conduct experiments to evaluate the effectiveness and usefulness of the proposed scheme.

5.1 Evaluation of Script-Face Mapping

Our experiments involve 20 clips from three movies, namely “Titanic”, “Twilight”, and “Up in the Air”, and one teleplay, namely “Friends”. Table I presents the information about these clips.

For script-face matching, we have proposed a novel algorithm, namely multitask joint sparse representation classification. We thus compare it against two existing methods: (1) nearest-neighbor (NN) classifier; and (2) the sparse representation (SR) classifier [Wright et al. 2009]. For each clip, we use the labeled exemplar faces with high confidence as the training set, and all the detected face tracks are regarded as the test set. The parameter $\lambda$ in Eq. (2) is set to 0.1 throughout the experiment. In Hong et al. [2010], Figure 7 shows several exemplar training faces and test face tracks. The accuracies...
Table I. The Information About the Video Clips and the Script-Face Mapping Accuracy

<table>
<thead>
<tr>
<th>Movie Name</th>
<th>Clips</th>
<th>Frames</th>
<th>Face Tracks</th>
<th>Accuracy (%)</th>
<th>Accuracy (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NN</td>
<td>SR</td>
<td>Script-Face Mapping</td>
</tr>
<tr>
<td>&quot;Titanic&quot;</td>
<td>C_1</td>
<td>2,864 (1.99min)</td>
<td>22</td>
<td>73.33</td>
<td>72.26</td>
<td>76.19</td>
</tr>
<tr>
<td></td>
<td>C_2</td>
<td>7,449 (5.17min)</td>
<td>31</td>
<td>86.90</td>
<td>90.90</td>
<td>93.24</td>
</tr>
<tr>
<td></td>
<td>C_3</td>
<td>2,868 (1.99min)</td>
<td>18</td>
<td>80.89</td>
<td>87.47</td>
<td>93.75</td>
</tr>
<tr>
<td></td>
<td>C_4</td>
<td>7,022 (4.88min)</td>
<td>22</td>
<td>83.74</td>
<td>91.51</td>
<td>88.89</td>
</tr>
<tr>
<td></td>
<td>C_5</td>
<td>9,801 (6.80min)</td>
<td>43</td>
<td>95.00</td>
<td>95.00</td>
<td>87.50</td>
</tr>
<tr>
<td>&quot;Twilight&quot;</td>
<td>C_6</td>
<td>4,543 (3.15min)</td>
<td>42</td>
<td>88.21</td>
<td>88.21</td>
<td>89.29</td>
</tr>
<tr>
<td></td>
<td>C_7</td>
<td>8,193 (5.69min)</td>
<td>47</td>
<td>81.89</td>
<td>80.60</td>
<td>81.45</td>
</tr>
<tr>
<td></td>
<td>C_8</td>
<td>5,788 (4.02min)</td>
<td>51</td>
<td>93.60</td>
<td>93.10</td>
<td>95.89</td>
</tr>
<tr>
<td></td>
<td>C_9</td>
<td>6,317 (1.95min)</td>
<td>47</td>
<td>75.30</td>
<td>81.90</td>
<td>86.50</td>
</tr>
<tr>
<td></td>
<td>C_10</td>
<td>4,745 (3.30min)</td>
<td>24</td>
<td>96.15</td>
<td>96.15</td>
<td>96.15</td>
</tr>
<tr>
<td>&quot;Up in the Air&quot;</td>
<td>C_11</td>
<td>9,707 (6.74min)</td>
<td>22</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>C_12</td>
<td>7,955 (5.52min)</td>
<td>87</td>
<td>89.01</td>
<td>90.20</td>
<td>92.52</td>
</tr>
<tr>
<td></td>
<td>C_13</td>
<td>3,852 (2.67min)</td>
<td>31</td>
<td>91.7</td>
<td>93.55</td>
<td>92.98</td>
</tr>
<tr>
<td></td>
<td>C_14</td>
<td>6,285 (4.36min)</td>
<td>50</td>
<td>91.94</td>
<td>90.45</td>
<td>92.20</td>
</tr>
<tr>
<td></td>
<td>C_15</td>
<td>6,533 (4.54min)</td>
<td>44</td>
<td>92.75</td>
<td>93.33</td>
<td>94.40</td>
</tr>
<tr>
<td>&quot;Friends&quot;</td>
<td>C_16</td>
<td>4,549 (3.16min)</td>
<td>38</td>
<td>74.26</td>
<td>74.07</td>
<td>77.78</td>
</tr>
<tr>
<td></td>
<td>C_17</td>
<td>5,779 (4.01min)</td>
<td>26</td>
<td>90.56</td>
<td>92.41</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td>C_18</td>
<td>3,621 (2.51min)</td>
<td>39</td>
<td>78.08</td>
<td>78.61</td>
<td>80.66</td>
</tr>
<tr>
<td></td>
<td>C_19</td>
<td>5,036 (3.56min)</td>
<td>44</td>
<td>61.52</td>
<td>68.03</td>
<td>71.90</td>
</tr>
<tr>
<td></td>
<td>C_20</td>
<td>4,748 (3.30min)</td>
<td>31</td>
<td>74.07</td>
<td>86.41</td>
<td>89.10</td>
</tr>
</tbody>
</table>

of script-face mapping achieved by our proposed algorithm and two existing methods are given in Table I. We can see that our proposed algorithm outperforms the other two methods on 15 out of the 20 clips. We can also see that for most clips the recognition accuracy is above 80%. This is important for our scheme, as putting scripts around an incorrect face will be misleading for hearing-impaired participants.

5.2 User Study I: Performance Evaluation

There are 60 anonymous hearing-impaired users participating in the study (21 male and 39 female). These participants come from Anhui Special Education School, Huangshan Branch, China. Their ages vary from 11 to 22. Most of them have prelingual deafness, which means that they sustained hearing impairment prior to the acquisition of language and the impairment occurred as a result of a congenital condition or through hearing loss in early infancy. Sign language is their first or preferred language. A small part of participants are postlingual hearing-impaired that occurs as a result of disease, trauma or as a side-effect of medicine after the acquisition of language. In our study, two teachers from a deaf-mutes school helped us to communicate with the participants. Before the study, all the participants were required to carefully read the investigation questionnaire and made sure that they understood their roles in the experiment.

We compare the following three paradigms.

(1) No Caption (NC), that is, the hearing-impaired participants were shown videos without caption.

(2) Static Caption (SC), that is, the hearing-impaired participants were shown videos with static caption (here we adopt the cinematic captioning).

(3) Dynamic Caption (DC), that is, the hearing-impaired participants were shown videos with dynamic caption.
We randomly divide all the participants into three groups (each group has 20 participants) to avoid the repeated playing of a video that will cause knowledge accumulation. Therefore, each group merely evaluates one of the three paradigms for each video clip.

During the video playing process, participants were informed to stop and answer a number of questions that is related to the content of the movie clips after each showing. To sufficiently investigate the effectiveness of dynamic captioning, we first measure how much advantage our proposed scheme is able to gain on content comprehension and user impression, and then we further evaluate the components. Here content comprehension indicates the extent of understanding from the hearing-impaired participants and user impression reflects whether the presentation of such dynamic caption is enjoyable and natural.

5.2.1 Content Comprehension. As we know, some questions such as “how many characters are there in this movie clip” have a single definite answer. Thus, it is possible to estimate the percentage of correctly answered questions for a predefined question set. In our study, we have designed 50 questions for each movie clip. These questions are carefully designed to broadly cover the content in the video clip.

The questions can also be categorized according to the information source of their answers. For example, the question “Who wore the sports clothes numbered 23?” can only be answered based on the video text information in the video, while “What’s the name of hero” can merely be answered based on caption information. Therefore, we can also estimate the percentage of correctly answered questions that are related to different information sources. Thus, we categorize the questions as follows.

1. **Caption Related.** Information from the captions only (34 questions in total).
2. **Video Text Related.** Textual information contained in video but not in the caption (6 questions in total).
3. **Visual Content Related.** Visual information contained in movie (10).

We can see that most questions (34 among 50) are related to caption. This is because caption is paramount to understanding the story of video. Hearing-impaired participants were asked to answer the questions independently. For performance evaluation, we take the metric of Quality of Perception (QoP), which is defined as the ratio of the correctly answered questions in the full question set.

Figure 4 gives average QoP scores of each video clip (the scores are averaged across participants) with different captioning paradigms. From the figure, we can see that both the static (SC) and the dynamic (DC) caption can greatly improve the comprehension level in comparison with the NC paradigm. It is worth noting that this does not contradict with the study of Gulliver and Ghinea [2003a, 2003b], which reports that SC can hardly improve the information gain of hearing-impaired participants, as in our study most questions are related to caption and thus participants cannot answer these questions without watching the captions. Next, we can see that for most clips the DC paradigm outperforms SC. Only for clips C-1 and C-7 the DC paradigm performs slightly worse. This is mainly due to the relatively low script-face mapping accuracies (76.19% for clip C-1 and 81.45% for clip C-7). We also perform a one-way ANOVA test [Fisher 1970], and the results are illustrated in Table II and Table III. From the results, we can see that the superiority of DC is statistically significant and the difference among users is statistically insignificant.

We then estimate the QoP scores for different question sets. Figure 5 illustrates the results. We can see that for questions that are related to caption (Figure 5(a)), the performance of NC is very poor as...
Video Accessibility Enhancement for Hearing-Impaired Users

Fig. 4. The QoP scores of: (1) no caption; (2) static caption; and (3) dynamic caption. We can clearly see the superiority of dynamic caption.

Table II. The ANOVA Test Results on Comparing DC and NC

<table>
<thead>
<tr>
<th>The factor of schemes</th>
<th>The factor of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-statistic</td>
<td>p-value</td>
</tr>
<tr>
<td>86.75</td>
<td>2.47 × 10^{-11}</td>
</tr>
</tbody>
</table>

The conclusion is that the difference of the two schemes is significant, and the difference of users is insignificant.

Table III. The ANOVA Test Results on Comparing DC and SC

<table>
<thead>
<tr>
<th>The factor of schemes</th>
<th>The factor of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-statistic</td>
<td>p-value</td>
</tr>
<tr>
<td>32.27</td>
<td>1.93 × 10^{-11}</td>
</tr>
</tbody>
</table>

The conclusion is that the difference of the two schemes is significant, and the difference of users is insignificant.

expected, and SC and DC are very close. This indicates that the conversion from static to dynamic caption doesn’t add much information. Figure 5(b) shows that the QoP scores of DC are remarkably higher than SC for the questions that are related to video text or visual content. The QoP scores of SC are even much worse than NC for the question related to video text. This indicates that the conventional captioning styles are rather distracting, and the hearing-impaired participants manage to view both the visual content and the caption at the bottom of frames. Our dynamic captions scheme can be more easily glimpsed as the scripts are presented around the character faces.

Overall, it is clear that captioning is important for understanding the story in videos, but the conventional static captioning approach will degrade the information assimilation of hearing-impaired subjects from other sources, and our dynamic captioning scheme can help them better perceive the content.

5.2.2 Component Evaluation. Now we further evaluate the components in the dynamic captioning scheme. We compare the following paradigms.

(1) **Dynamic Captioning (DC).** That is, Hearing-impaired participants were shown videos with dynamic caption.

(2) **DC without Volume Demonstration (DC-VD).** That is, We removed the voice volume demonstration from the dynamic captioning.

Fig. 5. The QoP scores of: (a) caption related; (b) video text related; and (c) visual content related. We can again clearly see the superiority of dynamic caption.

(3) **DC without Volume Demonstration and Synchronous Highlight (DC-VD-SH).** That is, We removed both voice volume demonstration and script synchronous highlight from the dynamic captioning.

(4) **Static Captioning (SC).** That is, Hearing-impaired participants were shown videos with static caption.

Analogous to the previous study, we test the content comprehension of the participants with the above paradigms. We conduct this study with the remaining 5 video clips (C-16 to C-20) and, for each video, we also design 50 questions. We randomly divide the 60 participants into four groups and then implement the question-answering test, and the process is the same with the previously introduced study. Figure 6(a) illustrates the average QoP scores for different clips. We can see that removing volume demonstration and synchronous highlight will reduce QoP scores, but DC-VD-SH is still able to outperform SC. This demonstrates the effectiveness of each component.

5.3 User Study II: Preference

We focus on user impression and preference between static and dynamic caption in this section. For user impression, we compare static captioning and dynamic captioning with the following two criteria: *enjoyment* and *naturalness*.
Video Accessibility Enhancement for Hearing-Impaired Users

Fig. 6. (a) The comparison of QoP scores of dynamic captioning; dynamic captioning without volume demonstration; dynamic captioning without volume demonstration and script highlight; and static captioning; (b) Study results of Enjoyment and Naturalness. We have demonstrated the scores averaged over users and video clips.

—Enjoyment. It measures the extent to which users feel that the video is enjoyable.
—Naturalness. It measures whether the users feel the visual appearance of caption is natural.

In this test, we do not need to divide the users into groups, and thus each user was asked to assign a score of 1 to 10 (higher score indicates better experience) to these two criteria. Figure 6(b) shows the results that are averaged over video clips and users. We can see that the dynamic caption remarkably outperforms static captioning in terms of enjoyment. However, the naturalness scores of the two captioning schemes are close. By communicating with the participants, it is found that this is due to the fact that, in several cases, the regions of script presentation vary abruptly. One possible solution to address this problem is to smooth the variation of the regions for presenting the scripts.

For the preference between static and dynamic caption, we ask each user to choose between the static captioning and dynamic captioning that he/she prefers and wishes to use in the future considering all these factors. The results show that 53 among the 60 users chose dynamic captioning. The remaining 7 users chose static mainly because they were already familiar with static captioning. This clearly demonstrates the usefulness of our scheme.

5.4 User Study III: Semiautomatic Scheme

From Table I, we see that although the average accuracy of mapping between face and script is above 80%, there are still some cases where scripts are not accurately positioned. Furthermore, we merely combine the detected face location and saliency analysis to obtain the script position. In this scenario, we can conjecture that there may be some scripts that are placed inside or far from the face region. Here, we employ a semiautomatic scheme, that is, manually align the script location for smoothing its presentation. An interactive user interface is designed to help users to make adjustment and “remember” the adjusted positions.

We conducted experiments on twenty-six anonymous users (15 female and 11 male, all are not hearing-impaired users). These participants’ ages range from 21 to 27 years. They were required to carefully read the instructions and make sure that they understood their roles in the experiment. We compared two schemes as follows.

(1) Automatic Dynamic Captioning (A-DC).\textsuperscript{8} That is, participants were shown videos with dynamic caption.

\textsuperscript{8}We denote the dynamic caption as an automatic dynamic caption that indicates no interactivity is involved in the process of video generation.
Fig. 7. The QoP scores of: (1) automatic dynamic caption (A-DC) and (2) semiautomatic dynamic caption (S-DC). We can clearly see the superiority of semiautomatic dynamic caption.

Table IV. The ANOVA Test Results on Comparing A-DC and S-DC

<table>
<thead>
<tr>
<th>The factor of schemes</th>
<th>The factor of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F )-statistic</td>
<td>( p )-value</td>
</tr>
<tr>
<td>54.31</td>
<td>( 3.55 \times 10^{-9} )</td>
</tr>
<tr>
<td>0.332</td>
<td>0.918</td>
</tr>
</tbody>
</table>

The conclusion is that the difference of the two schemes is significant, and the difference of users is insignificant.

(2) **Semi-automatic Dynamic Caption (S-DC).** That is, participants were shown videos with dynamic caption where the script positions have been manually smoothed.

We randomly divided all the participants into 2 groups (each group has 13 participants) to avoid the previously mentioned knowledge accumulation. Each group merely evaluates one of the two paradigms for each video clip. In experiments, the audio track was removed and participants were informed to stop and answer a number of questions that were related to the content of the movie clips after each showing. We measure how much advantage of semiautomatic scheme is able to gain on content comprehension and user impression compared to automatic dynamic caption. For content comprehension, these 50 questions for each movie clip are again employed as the questionnaire.

Figure 7 is the average QoP scores of each video clip. We see that S-DC outperforms A-DC over all the video clips. The clips C-1 and C-7 are improved much more under the S-DC scheme. This demonstrated the effectiveness of manual adjustment. We also perform a one-way ANOVA test [Fisher 1970], and illustrate the results in Table IV. From the results we can see that the superiority of S-DC is statistically significant and the difference among users is statistically insignificant. Note that the QoP scores of A-DC in Figure 7 are not equivalent to the scores in Figure 4 because of the differences between the two participants groups (e.g., hearing status, ages, etc.).

We again compare the two criteria of user impression: enjoyment and naturalness. Each user was asked to assign a score of 1 to 10 to the two criteria. Figure 8(b) illustrates the results that are averaged over video clips and users. We can see that the S-DC is remarkably superior to the A-DC in terms of enjoyment and naturalness. It agrees to our intuition that the involvement of interactivity can smooth the presentation. Here we detail the time-cost introduced by the interactivity. Figure 8(a) illustrates the percentage of scripts that need to be smoothed. We emphasize that there are two scenarios: one is inaccurate script-face mapping, and another is that more than half of participants prefer the adjusted caption positions. An average of 33.31% of the script entries need to be smoothed. Therefore, more time-cost is needed for processing (We estimate 5 seconds for each operation of the position adjustment and each video clip has thirty script entries. The added time-cost for each movie clip is about 1 minute.)
5.5 Discussion

In our experiments, it takes less than 4 minutes to process a video clip on average on a PC with Pentium 4 3.0G CPU and 2G memory. The average duration of the 20 video clips is 3.96 minutes, and this means that the processing time is roughly equivalent to the video duration (of course, the processing time also depends on many factors such as the number of characters and the appearing frequency of dialogues). Semiautomatic dynamic caption scheme is able to improve the performance and smooth the presentation while human labor is required. The total time cost can be significantly reduced, such as by speeding up the solution process of Eq. (2) and visual saliency analysis.

We would like to mention that in this work we mainly focus on the technical part of dynamic captioning and care less about user interface, such as the visualization of volume variation (currently, we just use a very simple stripe with a highlighted part) and the style of script highlight. However, even with a simple interface, our scheme has shown clear advantages through the study of user impression. User interface design is beyond the scope of this article although it is crucial for real-world application. We will leave it to our future work. Finally, we want to emphasize that although the focuses of our scheme are on videos along with scripts, it can be extended to process general videos without scripts. Actually what we need is to employ a speech-recognition engine to convert speech to scripts, and use speaker clustering [Aimer and Wooters 2003; Stadelmann and Freisleben 2009] and identification [Reynolds et al. 2000; Wan and Campbell 2000] to replace the face grouping and recognition techniques in our current scheme. Of course, this task will be much more challenging, but it will be an important topic along this research direction.

6. CONCLUSION

This article describes a dynamic captioning scheme to enhance the accessibility of videos towards helping the hearing-impaired audience better enjoy videos. Different from the existing static captioning methods, dynamic captioning put scripts at suitable positions to help the hearing-impaired audience better recognize the speakers. It also synchronously highlights the scripts by aligning them with the speech signal and illustrates the variation of voice volume to help the hearing-impaired audience better track and perceive scripts. Comprehensive user study with 60 hearing-impaired participants has demonstrated the effectiveness of our scheme.

As this is the first work to our knowledge to help hearing-impaired individuals better access videos with dynamic captioning, there is a lot of future work along this research direction. We will further improve the script-face mapping component to further boost the mapping accuracy and we will also...
investigate the extension of the scheme to deal with videos without script. We also plan to conduct a more comprehensive user study on a larger dataset.

REFERENCES


Video Accessibility Enhancement for Hearing-Impaired Users


Received February 2011; revised May 2011; accepted July 2011