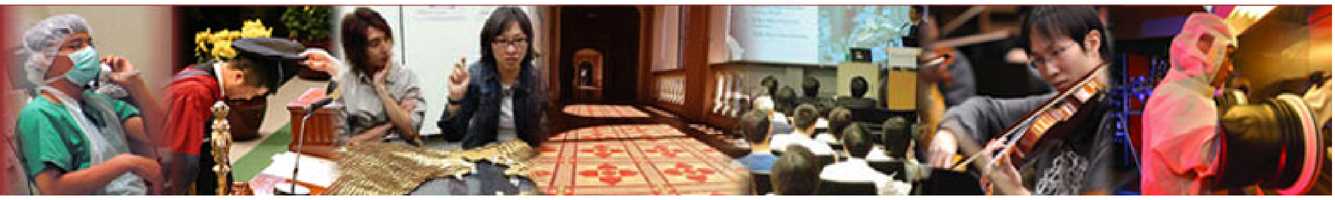


Thesis Title:

Multimodal Speaker Localization and Identification for Video Processing

Author:

HU Yongtao / 胡永涛

Degree:

Doctor of Philosophy

Multimodal Speaker Localization and Identification for Video Processing

by

Yongtao Hu

Department of Computer Science

The University of Hong Kong

Supervised by

Prof. Wenping Wang

A thesis submitted in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy
at The University of Hong Kong

December, 2014

Declaration of Authorship

I declare that this thesis represents my own work, except where due acknowledgement is made, and that it has not been previously included in a thesis, dissertation or report submitted to this University or to any other institution for a degree, diploma or other qualifications.

Signed: ____________________________________

Yongtao Hu

December, 2014

Abstract of thesis entitled

"Multimodal Speaker Localization and Identification for Video Processing"

Submitted by

Yongtao Hu

for the degree of Doctor of Philosophy
at The University of Hong Kong
in December, 2014

With the rapid growth of the multimedia data, especially for videos, the ability to better and time-efficiently understand them is becoming increasingly important. For videos, speakers, which are normally what our eyes are focused on, have played a key role to understand the content. With the detailed information of the speakers like their positions and identities, many high-level video processing/-analysis tasks, such as semantic indexing, retrieval summarization. Recently, some multimedia content providers, such as Amazon/IMDb and Google Play, had the ability to provide additional cast and characters information for movies and TV series during playback, which can be achieved via a combination of face tracking, automatic identification and crowd sourcing. The main topics includes speaker localization, speaker identification, speech recognition, etc.

This thesis first investigates the problem of speaker localization. A new algorithm for effectively detecting and localizing speakers based on multimodal visual and audio information is presented. We introduce four new features for speaker detection and localization, including *lip motion*, *center contribution*, *length consistency* and *audio-visual synchrony*, and combine them in a cascade model. Experiments on several movies and TV series indicate that, all together, they improve the speaker detection and localization accuracy by 7.5%–20.5%. Based on the locations of speakers, an efficient optimization algorithm for determining appropriate locations to place subtitles is proposed. This further enables us to develop an automatic end-to-end system for subtitle placement for TV series and movies.

The second part of this thesis studies the speaker identification problem in videos. We propose a novel convolutional neural networks (CNN) based learning framework to automatically learn the fusion function of both faces and audio cues. A systematic multimodal dataset with face and audio samples collected from the real-life videos is created. The high variation of the samples in the dataset, including pose, illumination, facial expression, accessory, occlusion, image quality, scene and aging, wonderfully approximates the realistic scenarios and allows us to fully explore the potential of our method in practical applications. Extensive experiments on our new multi-modal dataset show that our method achieves state-of-the-art performance (over 90%) in speaker naming task without using face/person tracking, facial landmark localization or subtitle/transcript, thus making it suitable for real-life applications.

The speaker-oriented techniques presented in this thesis have lots of applications for video processing. Through extensive experimental results on multiple real-life videos including TV series, movies and online video clips, we demonstrate the ability to extend our previous multimodal speaker localization and speaker identification algorithms in video processing tasks. Particularly, three main categories of applications are introduced, including (1) combine applying our speaker-following video subtitles and speaker naming work to enhance video viewing experience, where a comprehensive usability study with 219 users verifies that our subtitle placement method outperformed both conventional fixed-position subtitling and another previous dynamic subtitling method in terms of enhancing the overall viewing experience and reducing eyestrain; (2) automatically convert a video sequence into comics based on our speaker localization algorithms; and (3) extend our speaker naming work to handle real-life video summarization tasks.

[**499 Words**]

Acknowledgements

First I would like to thank my PhD supervisor Prof. Wenping Wang for his great guidance and continuous support on my research. I am really grateful that he has given me lots of freedom on determining the direction of my research topics that I am interested in. He appreciated the ideas I came up with and encouraged me to think independently and push myself to figure out own solutions for the challenging problems that I encountered. This thesis would not be possible without his sound advice, encouragement and guidance during my PhD study.

Part of my work is collaborated with Prof. Yizhou Yu from HKU, Prof. Jan Kautz from UCL and Prof. Yanwen Guo from NJU. I wish to thank them for their excellent guidance. I would also like to thank Dr. Jingwen Dai, Dr. Sijie Ren and Dr. Chang Yuan for their patient help and discussion during my internship at IVC Lab of Lenovo R&T (HK) and Dr. Xin Tong during the internship at MSRA. My thanks also goes to Prof. Changhe Tu and Prof. Haodi Feng from SDU for their referrals during my PhD application.

I am very grateful to thank all my friends in HK. Without them, I could not enjoy the joyful life during these years. My thanks go to "HKU CG Group" including Bin Chan, Yi-King Choi, Pengbo Bo, Lin Lu, Feng Sun, Yufei Li, Ruotian Ling, Yuanfeng Zhou, Li Cao, Zhan Yuan, Hao Pan, Wenni Zheng, Xinghua Zhu, Yanshu Zhu, Zhengzheng Kuang, Shuiqing He, Chuan Wang, Wenchao Hu, Ruobing Wu, Guangmei Jing, Rui Wang, Yujing Sun, Weikai Chen, Xiaolong Zhang, Yating Yue, Lingjie Liu and Changjian Li, "SDU2HK Group" including Fei Deng, Guodong Han, Xiaoyang Li and Mengmeng Wang, as well as "HW335A fellas" including Li Ning, Jing Li, Hao Wang, Weida Zhang, Sirui Li, Xiangzhong Xiang, Yupeng Li, Guanbin Li, Li Zhang, Xiaoguang Han, Wei Zhang, Yatong An; and "Pokfulam Basketball Group" including Qinliang Su, Jingrong Zhou, Hongfei Zeng, Bin Luo and Shuai Zhang.

Finally I would like to dedicate this thesis to my parents, my elder sister and brother and my girl friend Lily for their love, never-ending support and long-time waiting. Without them, I would not have come so far in my education.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

AFR	**A**utomatic **F**ace **R**ecognition
AV	**A**udio-**V**isual Synchrony
CC	**C**enter **C**ontribution
CNN	**C**onvolutional **N**eural **N**etworks
DOA	**D**irection **o**f **A**rrival
FPS	**F**rames **P**er **S**econd
GPU	**G**raphics **P**rocessing **U**nits
ITSP	**I**nnovation and **T**echnology **S**upport **P**rogramme
LBP	**L**ocal **B**inary **P**atterns
LC	**L**ength **C**onsistency
MCMC	**M**arkov **c**hain **M**onte **C**arlo
MFCC	**M**el-**F**requency **C**epstral **C**oefficient
MRF	**M**arkov **R**andom **F**ield
MSD	**M**ean **S**quared **D**ifference
RBF	**R**adial **B**asis **F**unction
SGD	**S**tochastic **G**radient **D**escent
SVM	**S**upporting **V**ector **M**achine
TDOA	**T**ime **D**ifference **O**f **A**rrival
TV	**T**ele**v**ision

Dedicated to my beloved family.

Chapter 1

Introduction

1.1 Motivation

With the rapid growth of the multimedia data, especially for videos, the ability to better and time-efficiently understand them is becoming increasingly important. For videos, speakers, which are normally where our eyes are focused, have played a key role to understand the content. With the detailed speakers' information like their positions and identities, many high-level video processing/analysis tasks, such as semantic indexing, retrieval [116] and summarization [117]. The main topics includes speaker localization, speaker identification, speech recognition, etc. See Figure 1.1 for an example.

1.1.1 Speaker Localization

In literature, audio information from microphone arrays is frequently used to localize the speakers (audio sources), i.e. estimate the Direction of Arrival (DOA) [21], which can be achieved based on the time difference of the sound received at different microphone pairs, i.e. Time Difference Of Arrival (TDOA) [15]. Such methods, however, are not suitable for video processing due to their special setup requirement of multiple microphones. For videos, more commonly used

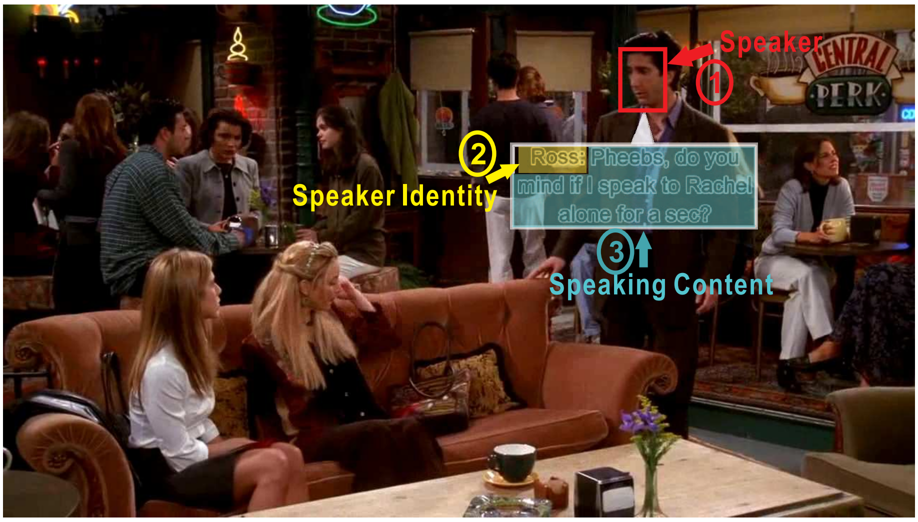

Speaker
1
2
Speaker Identity
Ross: Pheebs, do you mind if I speak to Rachel alone for a sec?
3
Speaking Content
CENTRAL PERK

FIGURE 1.1: Novel multimedia representation based on our proposed techniques [47, 48]. We highlight three key problems in real-life multimedia data processing: (1) ***speaker localization*** to detect and localize the speaker; (2) ***speaker identification*** to identify the speaker and (3) ***speech recognition*** to translate of spoken words into text. Copyright info: Screenshots from "*Friends*" TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

techniques are based on the visual cue [30, 46]. Several works have been done to jointly apply audio and visual cues together to better localize the speakers in videos [35, 63, 73]. However, performances of speaker localization in real-life videos with complex scenes are still limited.

In this thesis, we propose two novel speaker localization algorithms based on audio and visual cues. In the first speaker localization algorithm, see details in Chapter 2, we take several visual and audio features into consideration in a cascaded model to achieve the robust speaker localization accuracy of over 90% for real-life videos. For the second one described in Chapter 3, we propose a learning based framework that automatically fuse audio and visual information together, at the same time achieving the state-of-the-art performance.

1.1.2 Speaker Identification

The task of speaker identification is to identify who that individual is based on the human speech information [97]. Identifying speakers in movies, TV series and live shows is a significant problem, since speaker identity is an important cue in many high-level video analysis tasks. Recently, some multimedia content providers, such as Amazon/IMDb [1] (see Figure 1.2 for an example) and Google Play [2], have had the ability to provide additional cast and characters information for movies and TV series during playback, which can be achieved via a combination of face tracking, automatic identification and crowd sourcing. Normally, only audio information are used for speaker identification [40, 56, 113] in literature. Besides audio, several works have tried to use multimodal information like face, lip texture and motion [29, 91] to identify the speaker. Recently, more researchers have tried to make use of video context to increase the accuracy of character or speaker naming. Most of these works focused on *identifying/naming face tracks* [11, 30, 86, 95, 96, 112].

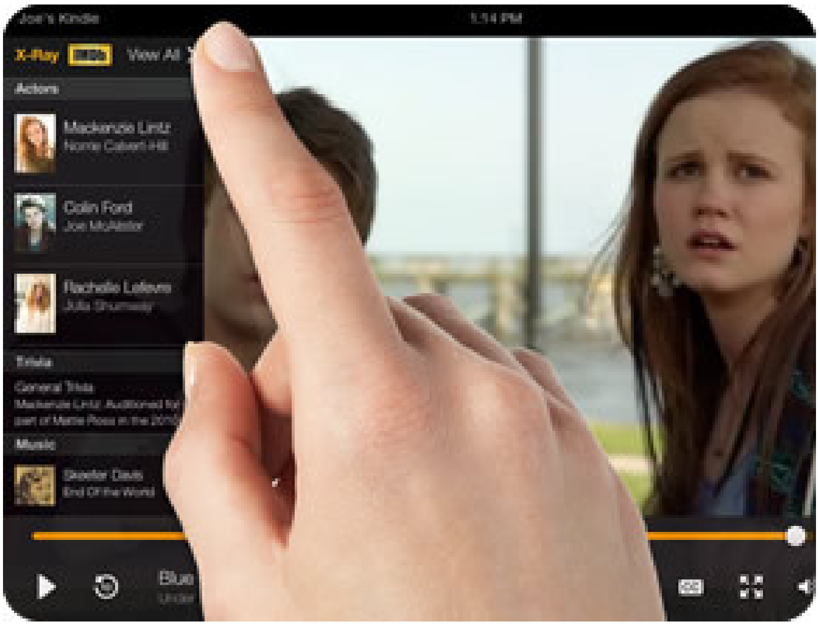


FIGURE 1.2: X-Ray service available on Kindle and Wii U. X-Ray has the ability to display the cast members shown in the current scene. Copyright info: Source: http://www.imdb.com/x-ray/.

Unlike all these previous works, we propose a learning based speaker identification algorithm in Chapter 3 that does not rely on any face/person track and does not need motion detection, facial landmark localization or subtitle/aligned transcript. With only the input of cropped face regions and corresponding audio segment, our approach recognizes the speaker in each frame in real-time.

1.1.3 Multimodal Data Fusion

In literature, multimodal data fusion has been widely used in different research areas, including biometrics [33], medical imaging [9], person authentication [20], person identity verification [13], visual tracking [82], segmentation [79] and human-computer interface [80], to further boost performance. However, most of previous works only conduct the multimodal data fusion step staying in the score-level [101, 104].

In this thesis, we apply multimodal data fusion (audio + visual) for both speaker localization (Chapter 2) and speaker identification (Chapter 3). In contrast, in both works, we conduct the multimodal data fusion in the feature extraction level, whose effectness will be verified in real-life video processing applications, including speaker-following video subtitles, automatically converting video to comics and speaker naming tasks (Chapter 4).

1.2 Outline

This thesis addresses several problems of real-life video processing and it is organized as follows:

Chapter 2 proposes a novel method for speaker detection and localization in videos using multimodal data fusion. It is further applied for improving the presentation of subtitles in video (e.g., TV and movies). Our method places on-screen subtitles next to the respective speakers to allow the viewer to follow the visual content while simultaneously reading the subtitles. The placement of the subtitles is determined using global optimization. We outperform previous work both in speaker detection and in subtitle placement. (The presented material has been published in [47].)

Chapter 3 introduces a novel convolutional neural networks (CNN) based learning framework to automatically learn the fusion function of both faces and audio cues. Extensive experiments on our new multi-modal dataset indicate that our method can achieve state-of-the-art performance (over 90%) in speaker naming task without using face/person tracking, facial landmark localization or subtitle/transcript, thus making it suitable for real-life applications. (The presented material has been published in [48, 49].)

Chapter 4 evaluates our proposed techniques in real-life video processing tasks. Particularly, we highlight three main applications: (1) combine applying our speaker-following video subtitles work and speaker naming work to enhance video viewing experience; (2) automatically convert a video sequence into comics based on our speaker localization algorithms in the speaker-following video subtitles work and speaker naming work and (3) extend our speaker naming work to handle real-life video summarization tasks. (The presented material has been published in [47, 48, 54].)

Chapter 5 summarizes the thesis and concludes possible future research.

Chapter 2

Speaker-following Video Subtitles

We propose a new method for improving the presentation of subtitles in video (e.g., TV and movies). With conventional subtitles, the viewer has to constantly look away from the main viewing area to read the subtitles at the bottom of the screen, which disrupts the viewing experience and causes unnecessary eyestrain. Our method places on-screen subtitles next to the respective speakers to allow the viewer to follow the visual content while simultaneously reading the subtitles. We use novel identification algorithms to detect the speakers based on audio and visual information. Then the placement of the subtitles is determined using global optimization. We outperform previous work both in speaker detection and in subtitle placement. A comprehensive usability study (see detailed information in Section 4.1) indicated that our subtitle placement method outperformed both conventional fixed-position subtitling and another previous dynamic subtitling method in terms of enhancing the overall viewing experience and reducing eyestrain.

2.1 Introduction

Subtitles are necessary for television programs, movies and other visual media for people with hearing impairments to help them understand and follow the dialog. Translated subtitles are necessary for media in a foreign language that is not dubbed to help viewers understand the spoken dialog. Subtitling is also useful for learning foreign languages.

Conventionally, the position of the subtitles is in a fixed location such as at the bottom of the screen. Humans can see objects within their field of vision but can only read text clearly in a narrow vision span. This vision span is an angular span (vertically and horizontally) of approximately 6 degrees of arc, which yields a region with a diameter of 5.23cm when viewed from 50cm away (Figure 2.1) [55, 70, 87, 109]. The conventional location of subtitles at the bottom of the screen means that to follow both the facial expression of speakers and the subtitles, the viewer has to constantly move their eye gaze between the main viewing area and the bottom of the screen, leading to a high level of eyestrain.

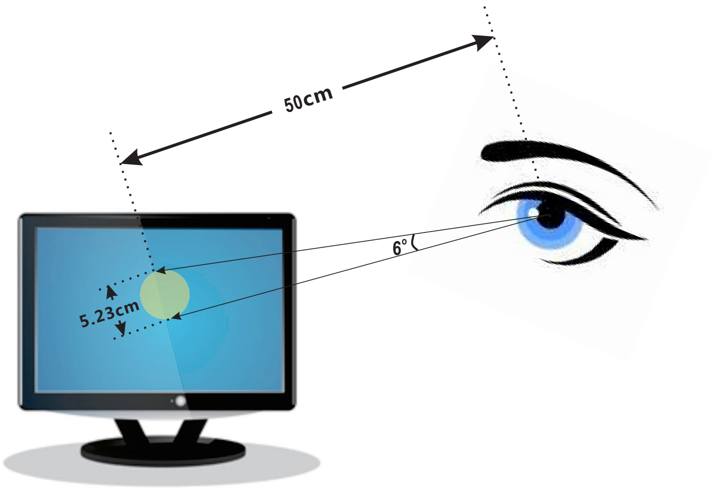


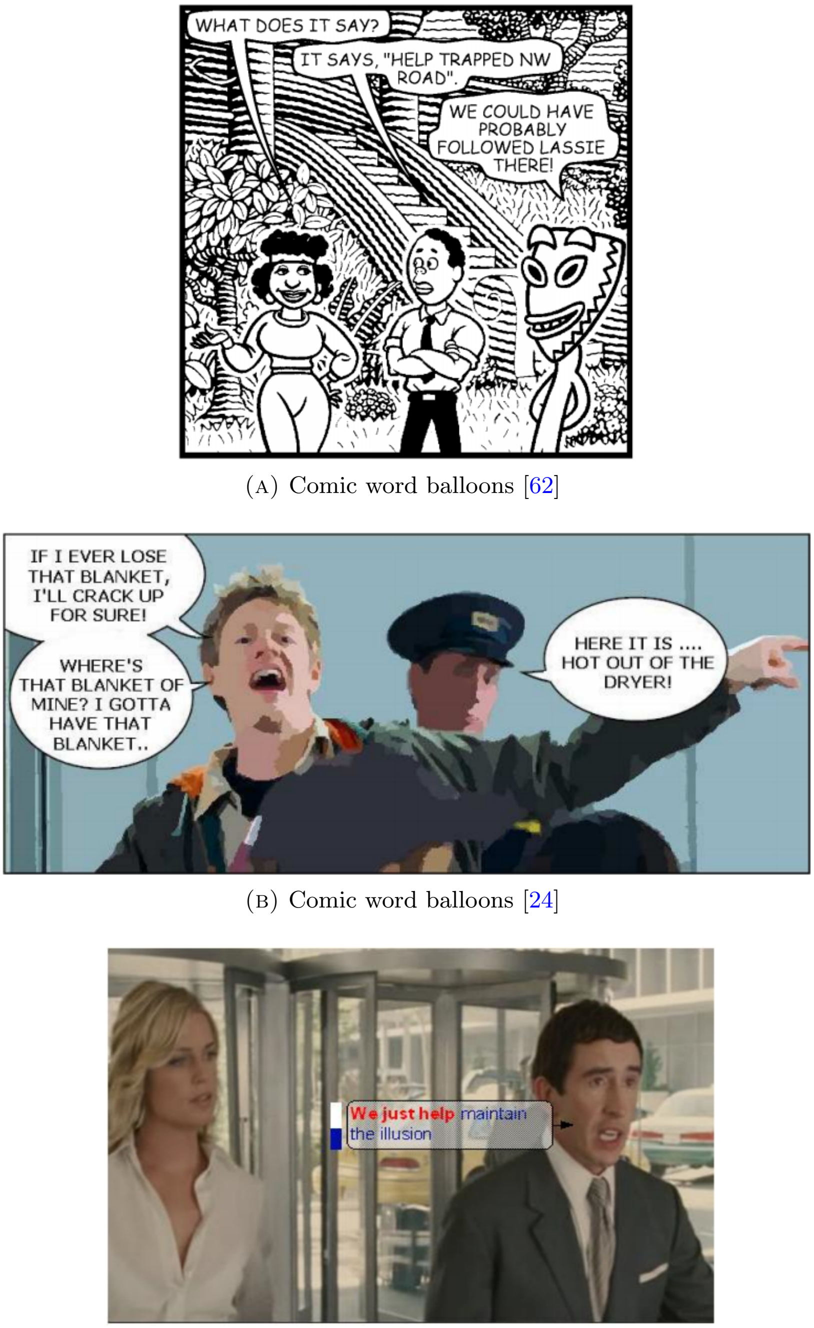
Figure 2.1: Limited vision span of human eyes. The human vision span is an angular span of 6 degrees of arc both vertically and horizontally, which yields a region with a diameter of 5.23cm when viewed from 50cm away.

Considerable effort has been made to enable the viewer to understand conversations through better presentation of the spoken dialog. In comics, the word balloon construction and layout of the graphics are techniques used to achieve this goal [24, 62]. Word balloon layouts in comics are used to help readers better visualize the written dialog, and the layout is carefully arranged to ensure that the flow of the story is clear (Figure 2.2 (A) and (B)). The previous work by Hong et al. [46] is the first and only research that extends these concepts to enhance the accessibility of videos by placing on screen subtitles next to the respective speaker (Figure 2.2 (C)).

In this chapter, we shall analyze the placement of subtitles so the viewer can comfortably follow the video content and subtitles simultaneously, which enhances the viewing experience. We shall improve on the work by Hong et al. [46] in two main aspects. We present a new speaker detection algorithm to accurately detect the speakers based on visual and audio information. An efficient optimization algorithm is then used to place the subtitle based on a number of factors. The framework of our approach is shown in Figure 2.3.

2.2 Previous Work

The work by Hong et al. [46] is the first and only previous research to study how dynamically placed subtitles can enhance video accessibility for the hearing impaired. The speaker is determined by examining lip motion features and then subtitles are placed in a non-salient region (based on a saliency map) around the speaker. Their usability study showed that their method effectively improved video accessibility for viewers with hearing impairments. We have extended this work to improve video subtitling and enhance viewing experience. First, we developed a new algorithm for improving speaker detection accuracy. The method is based on both visual and audio information rather than using only lip motion because other people in the scene may also be moving their lips. Second, the subtitles are positioned using an optimization procedure that takes several factors into consideration, including absence of speakers, cross-frame coherence, and screen layout. The combined use of these factors has proven to be more robust than using image saliency alone.


WHAT DOES IT SAY?
IT SAYS, "HELP TRAPPED NW ROAD".
WE COULD HAVE PROBABLY FOLLOWED LASSIE THERE!
IF I EVER LOSE THAT BLANKET, I'LL CRACK UP FOR SURE!
WHERE'S THAT BLANKET OF MINE? I GOTTA HAVE THAT BLANKET..
HERE IT IS .... HOT OUT OF THE DRYER!
We just help maintain the illusion

(A) Comic word balloons [62]

(B) Comic word balloons [24]

(C) Video subtitle placement [46]

FIGURE 2.2: Previous work to improve accessibility and understanding of conversations through word balloon layout in comics (A, B), and subtitle placement in videos (C). Copyright info: All images with permission of the authors of corresponding papers.

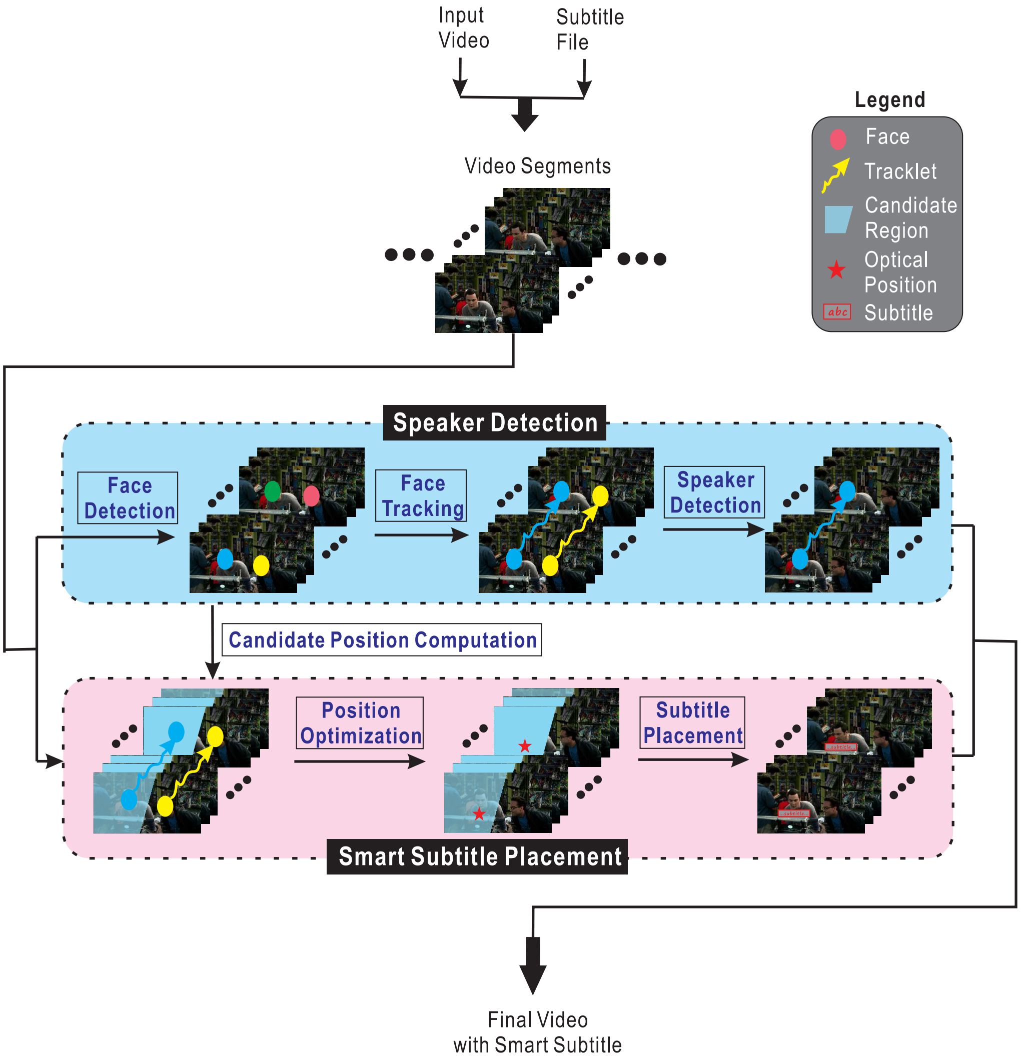


FIGURE 2.3: Our framework consists of two main components: (1) speaker detection; and (2) smart subtitle placement.

Speaker detection in video is an active research topic. In the task of speaker diarization, "who spoke when" can be resolved based on audio signals [6]. The problem of speech recognition based on visual information has been explored by Gordan et al. [42] and Saenko et al. [88, 89]. Visual and audio data are often used together in speaker localization ([75, 84]). Generally speaking, methods based on training data are more accurate but are more time-consuming. Out of efficiency consideration, we developed a method without involving a training process but still achieved robust detection results because of our use of several novel features and cues.

Subtitle placement has been previously investigated by Kurlander et al. [62] and Chun et al. [24] in their work on word balloon layouts in comics. Their methods are focused on handling single frames and so cannot easily be extended to video subtitles because they cannot ensure cross-frame coherence. In contrast, our approach employs a comprehensive optimization procedure to ensure cross-frame coherence to improve video subtitle presentation.

2.2.1 Contributions and Overview

In summarise, our work in this chapter has the following contributions:

- Development of an automatic end-to-end system for subtitle placement for TV/movies.
- A new algorithm for effectively detecting speakers based on visual and audio information.
- An efficient optimization algorithm for determining appropriate locations to place subtitles.

The rest of this chapter is organized as follows. The problem formulation and the whole framework are introduced in Section 2.3. Experiments results are given and discussed in Section 2.6, with conclusions followed in Section 2.7.

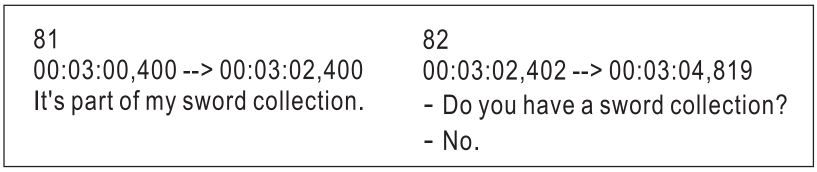
2.3 Problem Formulation and Preliminaries

2.3.1 Problem Formulation

Given a video (TV, movie, etc.) and its subtitle (caption) file as input, our goal is to generate the same video but with the subtitles positioned next to their corresponding speakers to provide better viewing experience. Our method is consisted of two major components: (1) speaker detection; and (2) optimization of subtitle positions. The workflow of the method is shown in Figure 2.3.

2.3.2 Preliminaries

The subtitle file is a text file in SubRip text file format (with the extension .srt) consisting of subtitle segments that contain the spoken lines and timing information. Time information provides only an approximate time interval during which the subtitle are shown on the video screen. Since this interval is always longer than the speaking time with extra time added at the beginning and the end. The input video consists of video segments corresponding the timing information in the subtitle file. Those video segments that have corresponding subtitle segments are called "*speaking video segments*", while the others called "*non-speaking video segments*". Figure 2.4 shows an example of two consecutive subtitle segments. A subtitle segment may contain lines from more than one speaker, as in the second example. The first example in Figure 2.4 corresponds to one speaking video segment, while the second example to two speaking video segments for two speakers.



```
81
00:03:00,400 --> 00:03:02,400
It's part of my sword collection.
```

```
82
00:03:02,402 --> 00:03:04,819
- Do you have a sword collection?
- No.
```

FIGURE 2.4: Two examples of subtitle segments. A subtitle segment has three parts: the segment index (the first line), the timing information of the segment (the second line), and the words spoken in this segment (the third line). The first subtitle segment is for one speaker and the second subtitle segment for two speakers.

We shall explain next how to process all the speaking video segments for speaker detection and optimization of subtitle placement.

2.4 Speaker Detection

For each speaking video segment, we perform the speaker detection procedure to detect the speakers. The workflow for speaker detection is shown in light blue in Figure 2.3. Our strategy is to initially detect the faces and combine them to form face tracklets. We examine all the face tracklets to determine the speaker for a given subtitle segment. The details of each step are discussed below.

2.4.1 Face Tracking

We first detect faces using the OpenCV frontal + profile Viola-Jones face detector [103]. Then these detected faces are linked using a low-level association approach to obtain face tracklets, following Kuo et al. [61]. Regions with similar positions, sizes and appearances are linked. These low level tracklets are further linked into final face tracklets if their size, appearance and coherent direction of movement are very similar to each other.

Face detections are not directly used for face tracking because matching the appearances of faces is extremely challenging due to similar colors, expressions, poses of different faces, in addition to lighting changes and motion blur [115]. Therefore, the clothing appearance has been used as additional cues for matching the same face [52]. Previous works show that face tracking in certain conditions, where matching is not possible if only face appearance is used, can be overcome by matching clothing appearances [30]. This improvement is mainly due to the richer texture variety of clothing compared to faces. Hence, we incorporated clothing appearance in our face tracking algorithm and found that this improved the performance significantly. With the detected faces including the localization and scale information, the bounding boxes of clothing can be obtained. We estimate the clothing appearance in an area under the face at a distance of $0.2\times$ face height. The scale of the rectangular clothing area can be approximated by the face area with a width $2d$ and height $1.5d$, where d is the width of the face. Examples are given in Figure 2.5. We chose these coefficients based on the ground-truth box fitting of our learning images.

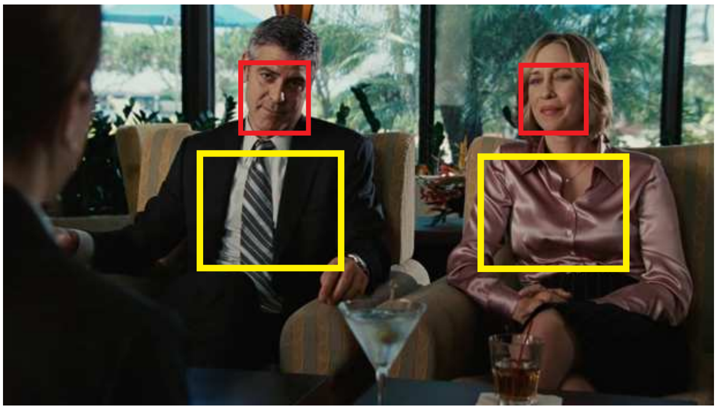


FIGURE 2.5: Clothing appearance used in face tracking. Red rectangles show the detected faces and yellow rectangles show the corresponding clothing appearance. Copyright info: Screenshot from movie "*Up in the Air*" ©DW Studios, The Montecito Picture Company and Rickshaw Productions.

2.4.2 Speaker Detection

Lip motion is the most widely used feature for speaker detection [30, 46]. However, the lip motion feature alone is not robust enough for speaker detection because the detection of a mouth region is often not accurate. We have observed that in TV and movies the positions and duration of the speakers are highly distinguishable from non-speakers. In addition to lip motion, we introduce a center contribution feature and length consistency feature for better speaker detection. We further consider the audio-visual synchrony feature to take into account the close relationship between audio and visual information to improve the robustness of our method. Based on these observation, in the presence of multiple speaker candidates, we detect the true speaker by examining the following four features: (1) *lip motion*, (2) *center contribution*, (3) *length consistency* and (4) *audio-visual synchrony*. The detailed procedures are explained below.

2.4.2.1 Lip Motion

A speaker often has more significant lip motion than non-speakers. Similar to Everingham et al. [30], we can detect the speaker from speaker candidates by detecting those faces with significant lip motion. For each face tracklet, we first use a facial landmark detector [100] to detect mouth corners. The mouth region can then easily be predicted based on the mouth corners. Based on the difference of pixel values between current and previous frames, the mean squared difference (MSD) in the mouth area can be then obtained. We compute the average of MSDs as the lip motion feature as

$$\mathcal{MSD} = \frac{1}{N-1}\left[\sum_{i=1}^{N-1}\mathcal{MSD}_{(M_i,\ M_{i+1})}\right], \tag{2.1}$$

where N is the length (i.e., # of frames) of the speaker tracklets, M_i the mouth region of the speaker in the i-th frame.

2.4.2.2 Center Contribution

It is observed that, in most TV and movies, a speaker is more likely than non-speakers to be located towards the center of the screen. To leverage this observation, we introduce the *center contribution* (CC) feature as

$$\mathcal{CC} = \frac{1}{N}\left[\sum_{i=1}^{N}\mathcal{CC}_i\right], \text{ with } \mathcal{CC}_i = 100 \times \left[1 - \frac{d\left(P_i,\ P_c\right)}{d\left(\mathcal{O},\ P_c\right)}\right], \tag{2.2}$$

where P_i is the center of the speaker's face in i-th frame, and $\mathcal{O}$ and P_c are the origin and center of the image plane, respectively. Here $d(\cdot,\ \cdot)$ denotes the distance function of two given points in Euclidean space.

2.4.2.3 Length Consistency

The accuracy of speaker detection can be further improved by considering the consistence between the length of a candidate face tracklet and the length of speaking time of a speaking subtitle segment. We call this *length consistency* (LC). Similar to MSD and CC, the candidate face tracklet with a higher LC is more likely to be the true speaker. We compute the length consistency as follows

$$\mathcal{LC} = \frac{1}{|L - L_{std}|}, \quad \text{with} \quad \begin{cases} L_{std} = \dfrac{L_{words} \cdot F_{video}}{\overline{V_{speaking}}}, \\ \overline{V_{speaking}} = \dfrac{L_{total\_words}}{T}, \end{cases} \tag{2.3}$$

where L is the length of the candidate face tracklet, L_{words} for # of the spoken words within the current subtitle segment, and L_{total_words} the total # of words in the input video, T the total speaking time in the input video, F_{video} the frame rate of input video, and $\overline{V_{speaking}}$ the average speaking speed in the input video (i.e., the number of words spoken in unit time).

2.4.2.4 Audio-Visual Synchrony

Previous work [28, 105] has demonstrated that audio cues and video cues can be combined to enhance the understanding of an environment. For example, sounds appear to correspond to motion synchronous with acoustic stimuli. Audio-visual (AV) synchrony has been used to resolve speaker localization [73] by computing the synchronization score of audio with the lower half of the faces, with audio-visual co-occurrence measured by the *synchronization score* $\mathcal{AV}$ of time slot $\Delta = [T_1, T_2]$ which is computed as

$$\mathcal{AV} = \langle y_a(t),\ y_v(t) \rangle\,, \ \text{with } t \in \Delta, \tag{2.4}$$

where $y_a,\ y_v$ are the audio and visual feature vectors and $\langle \cdot, \cdot \rangle$ indicates the scalar product between the vectors.

Speaker detection that is solely based on lip motion [46] relies too much on the accuracy of the mouth region detection and therefore may fail for complex scenes. Our method includes *center contribution* and *length consistency* features, and we further improve speaker detection by examining *audio-visual synchrony.* Moreover, unlike Monaci [73], we use extended facial landmarks [100] for better prediction of the motion region instead of using only the lower half of the face when examining audio-visual synchrony. This will help to filter out any motion disturbance in the background, especially when speakers or cameras are moving. Our method also compensates for the errors due to poor face detection and facial landmarks (see Figure 2.6).

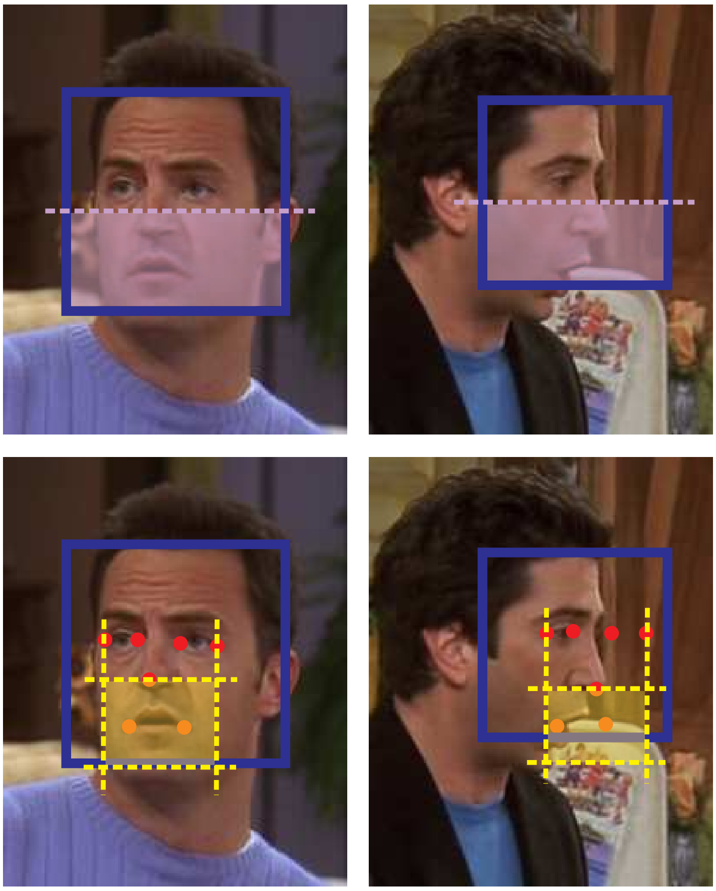


Figure 2.6: Improving audio-visual synchrony through better motion region prediction. First row: motion region (light pink area) used by Monaci [73]. Second row: better motion region (light yellow area) prediction based on facial landmarks (red points) from our method. Copyright info: Screenshots from "*Friends*" TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

For better accuracy and efficiency, we chain the four features above in a cascade in the following order: *lip motion*, *center contribution*, *length consistency* and *audio-visual synchrony* (see Figure 2.7). The process can be seen as a degenerate decision tree [103]. The design of this cascade structure is based on our observation that the majority of the candidate speakers are non-speakers and only candidate speakers that satisfy the current constraint pass to the next level. We put *lip motion* first in the cascade because it can effectively reject most non-speakers. *Center contribution* and *length consistency* are put second and third in the cascade as they can be evaluated very quickly. *Audio-visual synchrony* is placed last in the chain because this evaluation is more time-consuming. Experiments have also shown that a cascade in this order gives the top three best results among all combinations at the same time with relatively fast speed. Detailed implementation is shown in Algorithm 1[1].

Figure 2.14 shows the performance improvement in speaker detection accuracy[2]. The left figure shows the effect of adding the *center contribution*, *length consistency* and *audio-visual synchrony* (with better motion prediction) besides *lip motion*, and the right figure shows our speaker detection accuracy compared with previous methods [46, 73][3]. Overall, our speaker detection method outperformed these two methods.

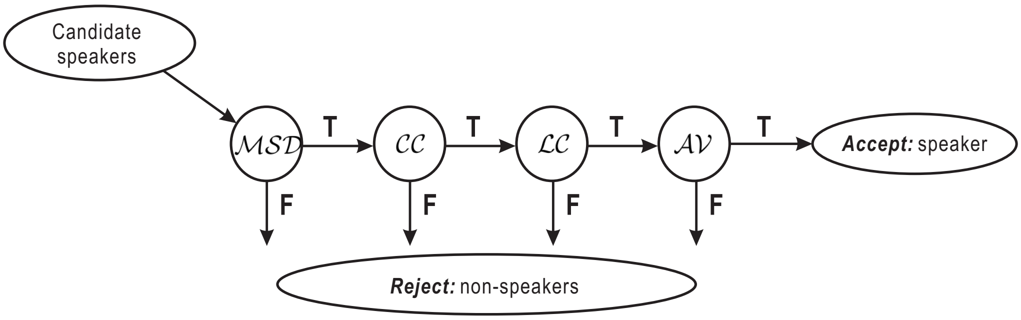


Figure 2.7: Schematic depiction of our cascade speaker detection method. Only candidate speakers that pass the current test will go through to the next level.

[1] We apply $\theta_1 = 20$, $\theta_2 = 2.5$, $\theta_3 = 2$, $\theta_4 = 0.1$, $\theta_5 = 2$ throughout.

[2] Speaker detection accuracy is computed in terms of video segments. It is considered to be correct for those video segments where the subtitle is put in the default position (i.e., the bottom of the screen) when no speaker is detected.

[3] Hong et al. [46] used only the *lip motion* feature (i.e., MSD) for speaker detection.

Algorithm 1: Speaker detection algorithm.

```
Input  : $N$ face tracklets, each of which stands for a speaker candidate.
Output: The speaker (tracklet) or no speaker.

Compute $\mathcal{MSD}$ feature for each face tracklet.
Delete all tracklets whose $\mathcal{MSD} < \theta_1$ (assume remaining $N_1$ tracklets
after this).
if $N_1 >= 2$ then
    Delete all tracklets whose $\mathcal{MSD} *\theta_2 < \max_{\mathcal{MSD}}$ (assume remaining
    $N_2$ tracklets after this).
    if $N_2 >= 2$ then
        Compute $\mathcal{CC}$ feature for each face tracklet.
        Delete all tracklets whose $\mathcal{CC} *\theta_3 < \max_{\mathcal{CC}}$ (assume remaining $N_3$
        tracklets after this).
        if $N_3 >= 2$ then
            Compute $\mathcal{LC}$ feature for each face tracklet.
            if $\max_{\mathcal{LC}} - second \max_{\mathcal{LC}} > \theta_4$ then
                return the tracklet with $\max_{\mathcal{LC}}$.
            else
                Compute motion region based on facial landmarks for
                each tracklet.
                Compute $\mathcal{AV}$ feature for each face tracklet
                if $\max_{\mathcal{AV}} > \theta_5$ then return the tracklet with $\max_{\mathcal{AV}}$.
                else return no speaker.
        else if $N_3 == 1$ then return current tracklet.
        else return no speaker.
    else if $N_2 == 1$ then return current tracklet.
    else return no speaker.
else if $N_1 == 1$ then return current tracklet.
else return no speaker.
```

2.5 Subtitle Placement

Generally speaking, our goal is to place subtitles close to their corresponding speakers. There are several considerations in attaining this goal:

- In each frame, the subtitle should be close enough to its speaker so not to cause any confusion that it is spoken by another person;

- Across consecutive frames, the overall distance between all subtitle placements should be small to reduce eyestrain;
- Subtitles should not be placed near screen boundaries so as not to detract the viewer from the central viewing area; and
- Subtitles should not occlude any important visual contents (e.g., faces).

In this section, we shall present an optimization framework (detailed in in Algorithm 2) for subtitle placement. First, we shall explain some preprocessing steps before performing subtitle placement. The following operations are needed to process the input video file and the subtitle file.

1. Split a speaking video segment when the speaker moves significantly during a subtitle segment;
2. Split a subtitle segment if there is a major shot change in the corresponding video segment;
3. Split a subtitle segment when it corresponds to multiple speakers; and
4. Refine the display time of a subtitle from the provided timing information.

Algorithm 2: Subtitle placement algorithm.

```
Input : Speaker detection result (i.e., face tracklets for each subtitle
        segment).
Output: Positions for subtitle placement at each subtitle segment.
foreach subtitle segment do
    Determine the proper rectangle to hold the subtitle[4].
    Compute candidate positions for placing the subtitle.
Compute the optimal positions of subtitle based on candidate positions.
return the position of each subtitle segment.
```

[4] Similar to Hong et al. [46], we use a rectangle to bound the subtitle and an arrow to the speaker to avoid confusion. The dimensions of the bounding rectangle of the subtitle can easily be determined by processing the subtitle text based on its length and font.

2.5.1 Computing Candidate Subtitle Positions

Any position close to the speaker may be a good subtitle position. In order to reduce the search space for better efficiency, we only consider the following eight candidate positions around the speaker's face: above left, above, above right, below left, below, below right, left and right, which are indicated by the yellow dots in Figure 2.8). Note that the exact locations of these eight candidate positions are computed based on the speaker's location and size of the speaker's face and the length and font size of the subtitle.

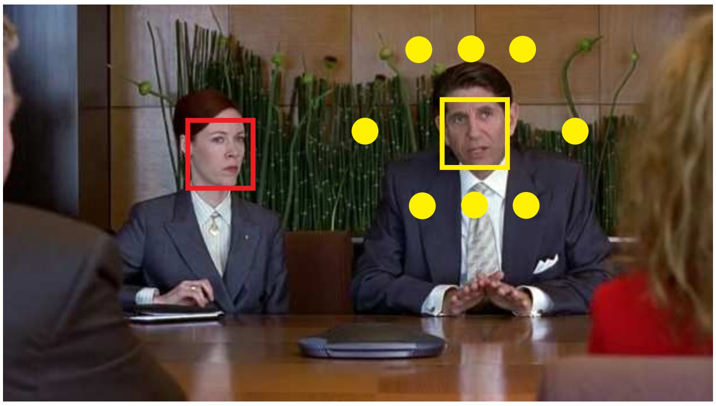


FIGURE 2.8: Candidate positions for the placement of the subtitle around the speaker's face. Speaker (yellow rectangle), non-speaker (red rectangle), candidate subtitle positions (yellow circles, 8 positions in our experiment). Copyright info: Screenshot from movie "*Erin Brockovich*" ©Universal Pictures (USA) and Columbia Pictures (International).

2.5.2 Subtitle Position Optimization

Subtitle position optimization is the key step in Algorithm 2. We take into account layout information to obtain the best position to place the subtitle. Here, we assume that each subtitle will be assigned a fixed position.

2.5.2.1 Local Optimization

Locally, we require the subtitle to be close to the speaker and far away from non-speakers. Let P denote the optimal subtitle position of the current subtitle segment, P_S the speaker's position, and $P_{NS_k}(k = 1, 2, ..., K)$ the positions of the K non-speakers in the scene. The energy function of P due to local optimization is defined as

$$E_{local} = d_{(P,\, P_S)} - \sum_{k=1}^{K} d_{\left(P,\, P_{NS_k}\right)} \tag{2.5}$$

2.5.2.2 Position Consistency Over Time

We require that the positions of consecutive subtitles do not change their positions too much across consecutive frames in order to alleviate eye-strain. To this end, we constrain subtitles of subsequent speaking video segments to follow the positions of the preceding subtitles. Let P_{pre_opt} denote the optimaized position of the preceding subtitle. Then the energy function due to this consideration is defined as

$$E_{global} = d_{(P,\, P_{pre\_opt})} \tag{2.6}$$

2.5.2.3 Layout with Respect to Screen Boundary

Subtitles should not be placed at the screen boundaries in order to allow the viewer to focus more on the central viewing part of the screen. This preference is reflected in following energy function:

$$E_{layout} = d_{(P,\, boundary)} \tag{2.7}$$

Now, combining these three energy terms, we defined the following total energy function.

$$\min \quad E = w_1 \cdot E_{local} + w_2 \cdot E_{global} + w_3 \cdot E_{layout}. \tag{2.8}$$

We use 1.0 for w_1 and w_2, -1.0 for w_3 for all our experiments. General speaking, higher w_1 will generate subtitles better following the speakers, higher w_2 for more position-consistent subtitles among different subtitle segments to help reduce eye-strain and higher w_3 for subtitles further away from screen boundaries. We minimize this function to compute the optimized position of the current subtitle.

2.5.3 Some Further Considerations

2.5.3.1 Advanced Video Segment Splitting

When the speaker moves around during a subtitle segment, we would still like to have the subtitle tagged to the speaker. However, rather than letting the subtitle float with the speaker, we split both the subtitle segment and its corresponding speaking video segment into shorter segments. Such subtitle segments are first identified, based on the speaker tracklet, and are split into shorter subtitle segments. Then the positions of these shorter subtitle segments are determined using the same procedure (Algorithm 2) as for the normal subtitles, so that the viewer can better follow the moving speaker together with the subtitles. Note that, the position of each short subtitle is fixed for each small video segment. Two such examples are shown in Figure 2.10.

2.5.3.2 Shot Change Handling

A significant shot change in TV/movies may make a speaker appear or disappear. In this case, subtitle positions need to change accordingly. For example, it should be placed at a default position (such as the bottom of the screen) if the speaker disappears after a shot change. We propose a simple shot change detector for video segments solely based on color histogram (see Algorithm 3[5]) instead of using shot change detectors based on complex models [27, 32, 92]. Our shot change detector proves to be accurate enough even for some very challenging TV

[5] We computed the correlation coefficient as the similarity of their RGB color histogram and applied a threshold of 0.99.

series such as "*Friends*" and "*The Big Bang Theory*", as shown in Table 2.1. Some results of detected shot changes are shown in Figure 2.9.

```
Algorithm 3: Shot change detector for video segments.
Input   : A video segment.
Output: All shot change positions.
foreach pair of adjacent two frames do
    Compute the similarity of their color histogram as $\delta$.
    if $\delta < threshold$ then
        Label the second frame as a shot change position.
return all these labeled positions.
```

Table 2.1: Performance of our shot change detection procedure. The ground truth of shot changes are manually labeled in our experiments.

Video Input	True	Detected	False	Missed	Recall	Precision
Friends S01E05	115	105	3	13	88.70%	97.14%
Friends S01E13	103	96	2	9	91.26%	97.92%
Friends S10E15	106	108	3	1	99.06%	97.22%
The Big Bang Theory S03E05	31	32	1	0	100%	96.88%

Based on the shot change detection, video segments are further split into shorter video segments, each of which has no big significant shot changes and the subtitle position in it can be kept relatively constant. After splitting, the video segment containing the scene that has the longest time overlap with the speaking video subtitle will be assigned the subtitle. The subtitle for all the other video segments will be set at a default position (e.g., the bottom of the screen). Some sample outputs are shown in Figure 2.11.

2.5.3.3 Splitting Multiple Speakers

There are often two or more speakers in a single subtitle segment, such as in the second example in Figure 2.4, where different speakers are denoted by a '-' in these segments. In this case, we split such a subtitle segment into different parts

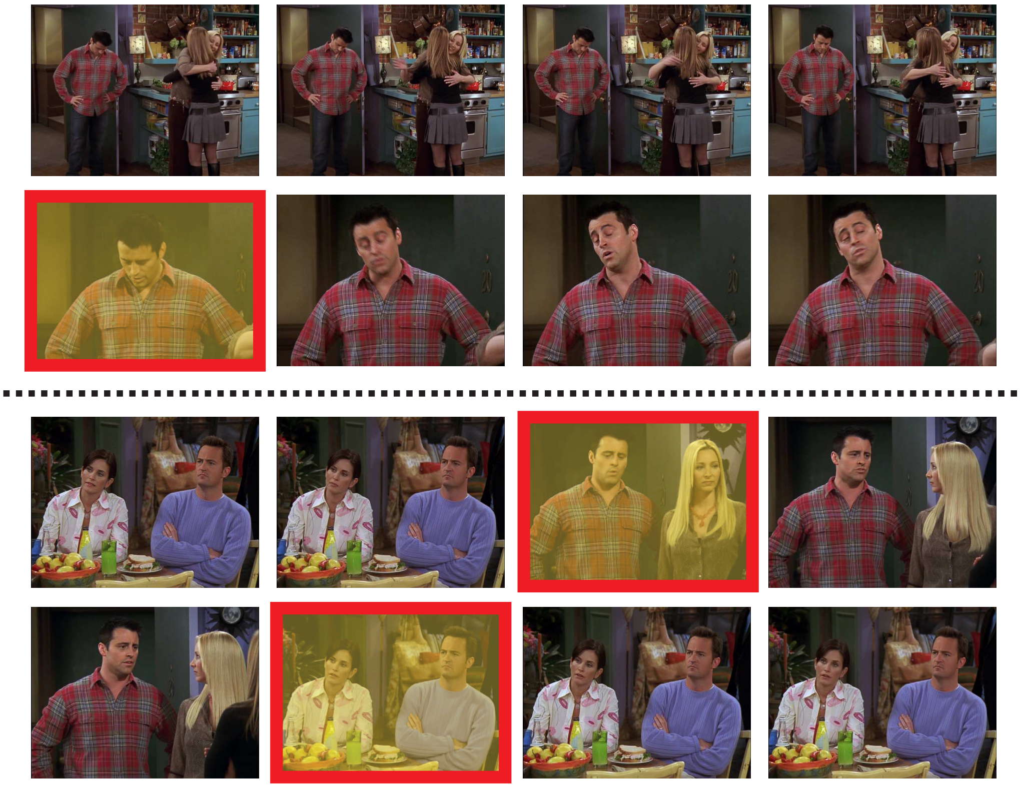


FIGURE 2.9: Results of shot change detections. Each two rows show a video segment. The first frame of the shot change is marked by a red rectangle with a yellow shade. There is one shot change in the first video segment and two shot changes in the second segment. Copyright info: Screenshots from "*Friends*" TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

corresponding to different speakers, using the separating character '-'. Accurate active speaking time for each speaker is determined based on all the face tracklets of the whole segment. Then, within each small time range, Algorithm 1 is used to find the specific speaker. One example is shown in Figure 2.12.

2.5.3.4 Refining the Speaking Time

The time information in the subtitle file is usually not the actual speaking time but only provides an approximate interval to show the subtitle on screen. The time interval is manually specified and is normally longer than the actual speaking time, with extra time padded at the beginning and the end. We refine the speaking time based on the span of the speaker tracklet, which is more reliable and coherent with the video content. Some examples are shown in Figures 2.10, 2.11 and 2.12.

2.6 Experiment Results and Discussion

We tested our system on a variety of professional videos including those in which people speak very fast and move quickly. The videos were from the following TV series and movies: "*Friends*", "*The Big Bang Theory*", "*Kramer vs. Kramer*", "*Erin Brockovich*", "*Up in the Air*", "*The Man From Earth*", "*Scent of a Woman*" and "*Lions for Lambs*". All these video clips can be accessed at the website[6]. Some sample outputs (TV/movie screenshots) are shown in Figure 2.10, 2.11, 2.12 and 2.13.

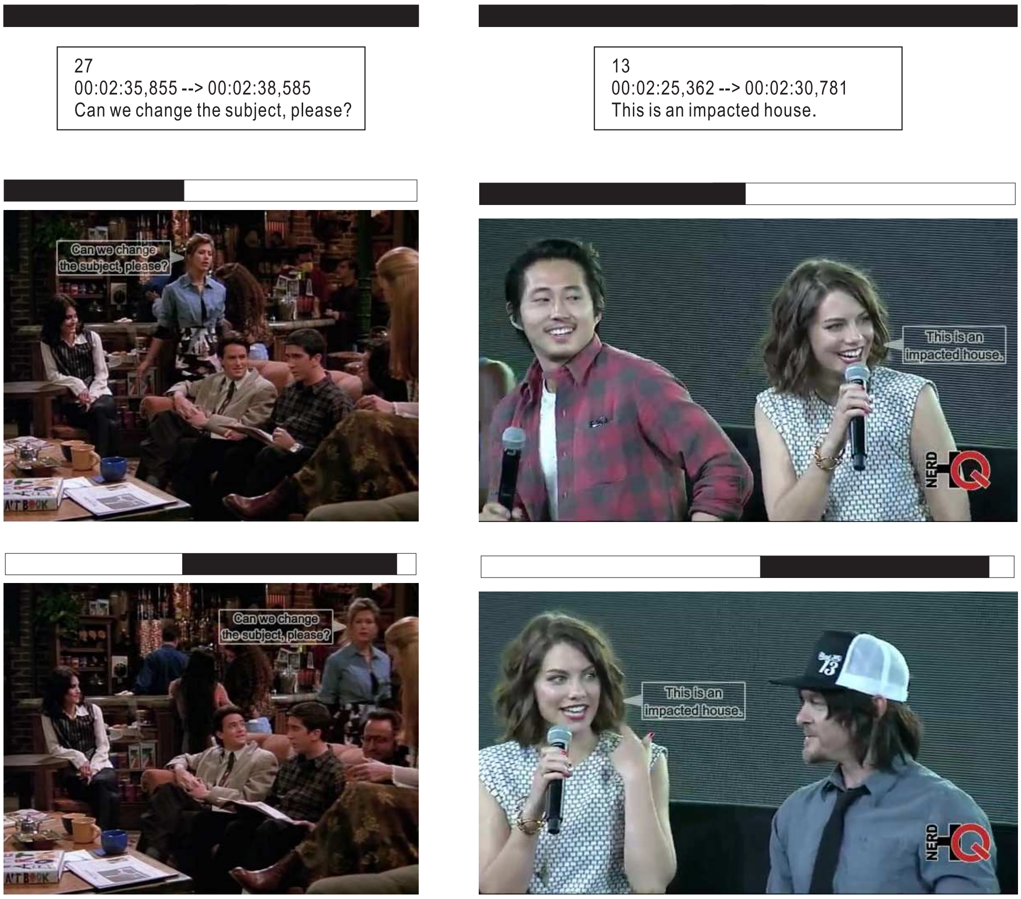
2.6.1 Speaker Detection Accuracy

We have introduced four new features for speaker detection. Their contributions to improving detection accuracy are shown in Figure 2.14. The *center contribution* improved the speaker detection accuracy by 0.5%–2.5%, *length consistency* improved accuracy by 1.5%–4%, and *audio-video synchrony* with better motion prediction improved accuracy by 3.5%–6.5%. Taken together, our speaker detection method outperformed Monaci [73] by 8.5%–20.5% and Hong et al. [46] by 7.5%–17%.

2.6.2 Subtitle Placement

In the previous method [46], subtitles are positioned in the least salient region around the speaker based solely on a saliency map. This performs well only with a relatively simple background. For videos with a complex background such as the example shown in Figure 2.15, the subtitles generated that way

[6] https://sites.google.com/site/smartsubtitles/

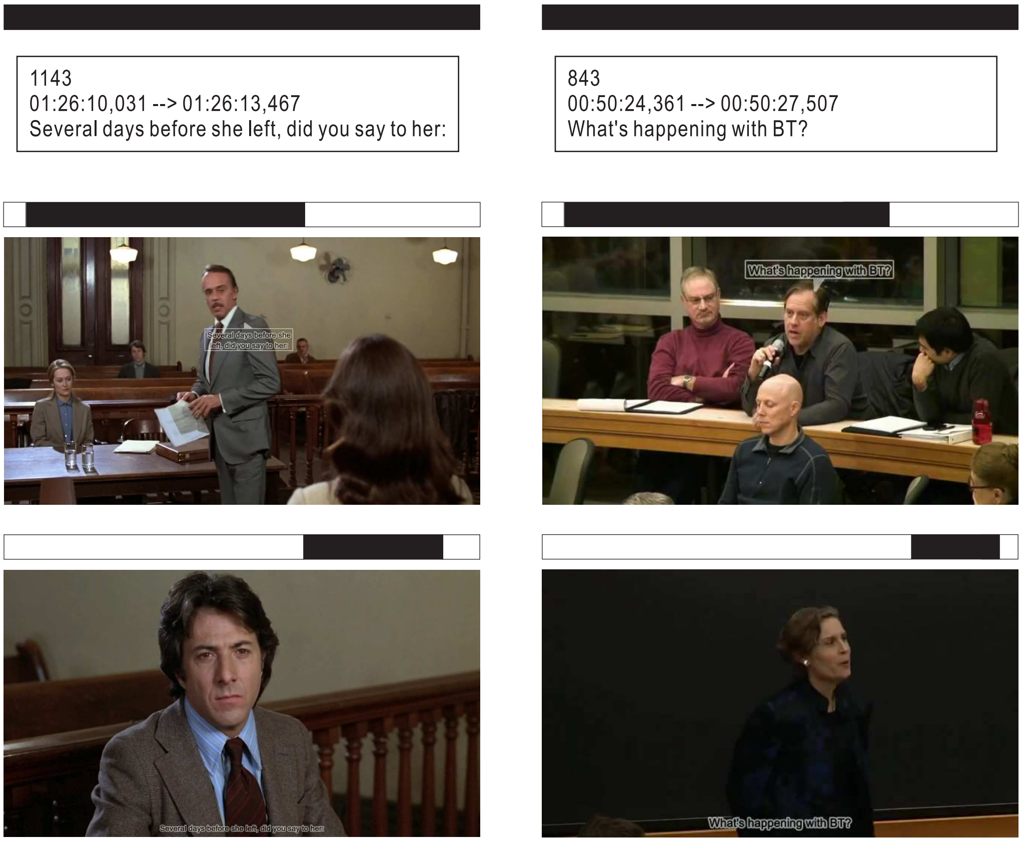


27
00:02:35,855 --> 00:02:38,585
Can we change the subject, please?

13
00:02:25,362 --> 00:02:30,781
This is an impacted house.

Can we change the subject, please?
This is an impacted house.
Can we change the subject, please?
This is an impacted house.
NERD HQ
NERD HQ

Figure 2.10: Sample results of our system. We highlight two real-life examples in each column to show our ability to split a video segment when the speaker moves significantly during a subtitle segment. Black/white bars are the active time (active in black, non-active in white). Copyright info: Screenshots of the first column from "*Friends*" TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television and second column from YouTube video https://www.youtube.com/watch?v=v0y1rZXthmI.



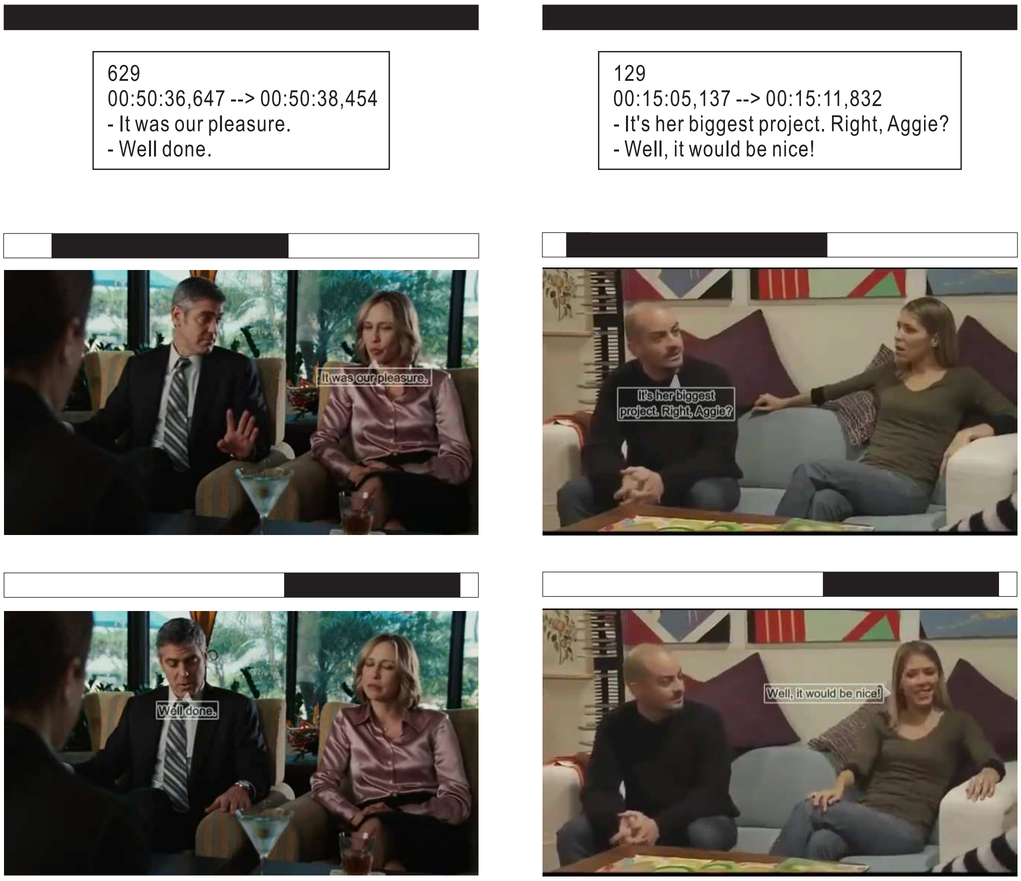
```
1143
01:26:10,031 --> 01:26:13,467
Several days before she left, did you say to her:
```

```
843
00:50:24,361 --> 00:50:27,507
What's happening with BT?
```

Several days before she left, did you say to her:
What's happening with BT?
Several days before she left, did you say to her:
What's happening with BT?

FIGURE 2.11: Sample results of our system. We highlight two real-life examples in each column to show our ability to split a subtitle segment where there is a major scene change in the video segment. Black/white bars are the active time (active in black, non-active in white). Copyright info: Screenshots of the first column from movie "*Kramer vs. Kramer*" ©Columbia Pictures and second column from YouTube video https://www.youtube.com/watch?v=7R4xhwy-1oI.

may block some faces. In contrast, our subtitle placement procedure is more robust for complex backgrounds. The local energy term in Equation 2.5 tends to position the subtitle close to the speaker and at the same time far away from non-speakers, thus lessening confusion and reducing eyestrain. Furthermore, our subtitle placement algorithm is consistent across frames thanks to the global energy term in Equation 2.6.



```
629
00:50:36,647 --> 00:50:38,454
- It was our pleasure.
- Well done.
```

```
129
00:15:05,137 --> 00:15:11,832
- It's her biggest project. Right, Aggie?
- Well, it would be nice!
```

FIGURE 2.12: Sample results of our system. We highlight two real-life examples in each column to show our ability to split a subtitle when it corresponds to multiple speakers. Black/white bars are the active time (active in black, non-active in white). Copyright info: Screenshots of the first column from movie "*Up in the Air*" ©DW Studios, The Montecito Picture Company and Rickshaw Productions and second column from YouTube video https://www.youtube.com/watch?v=CXP6KU4urgE.

2.6.3 Limitations

Our system cannot always correctly handle scenes where there are no speakers such as subtitles from a gramophone or a telephone message. If the speaker detection algorithm does not detect a speaker, then our system can still correctly put the subtitle into the default bottom screen position. However, if people in the scene move in such a way that confuses the speaker detection algorithm, our system may assign the subtitle to a wrong speaker, causing misunderstanding.

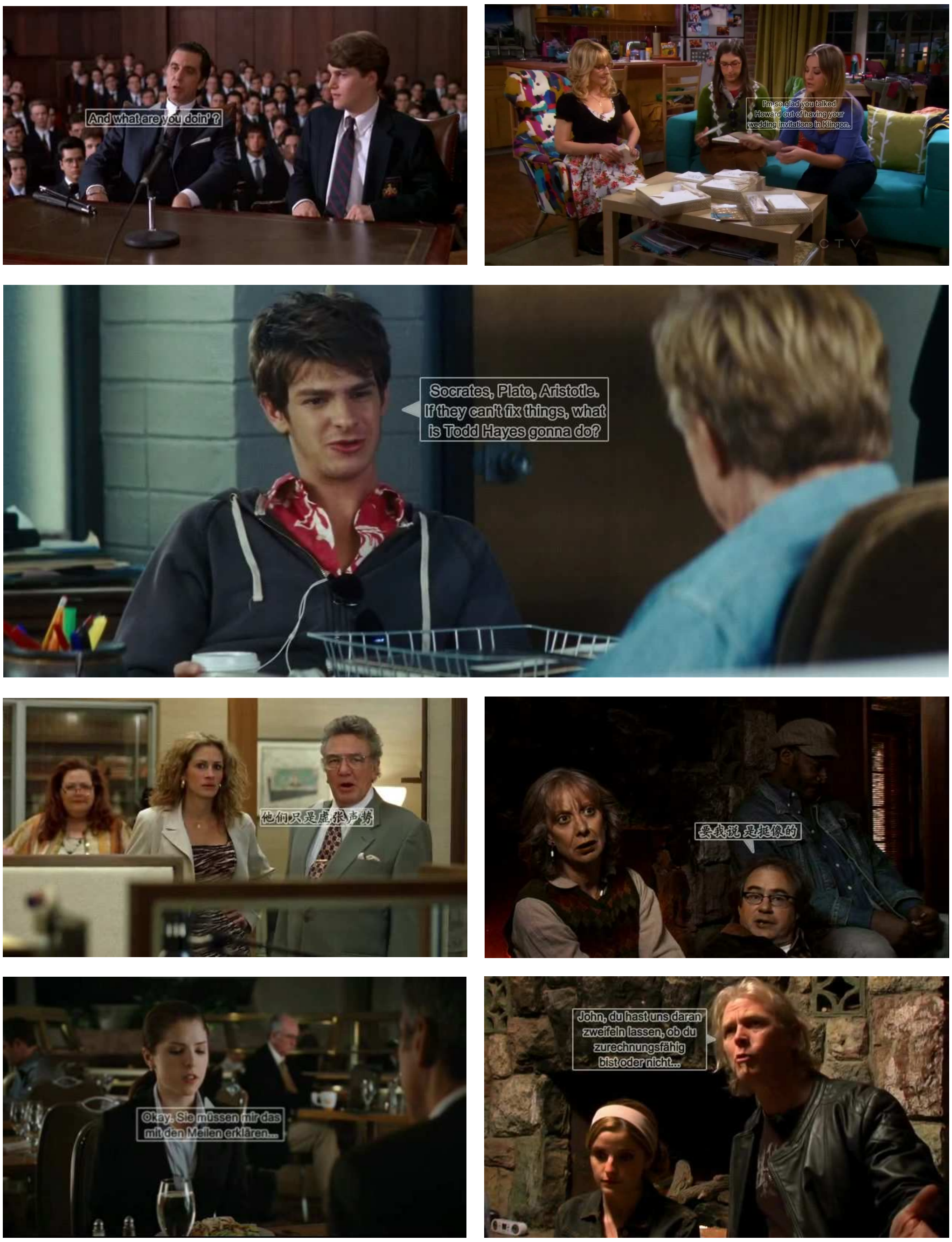

And what are you doin' ?
I'm so glad you talked Howard out of having your wedding invitations in Klingon.
Socrates, Plato, Aristotle. If they can't fix things, what is Todd Hayes gonna do?
他們只是虛張声势
要我说 是挺像的
Okay. Sie müssen mir das mit den Meilen erklären...
John, du hast uns daran zweifeln lassen, ob du zurechnungsfähig bist oder nicht...
CTV

Figure 2.13: Sample results of our system. We highlight several real-life examples to show our ability to handle subtitles of different languages, including English in first two rows, Chinese in third row and German in last row. Copyright info: Screenshot of the left one in the first row from movie "*Scent of a Woman*" ©City Light Films, the right one in the first row from "*The Big Bang Theory*" TV series ©Chuck Lorre Productions and Warner Bros. Television, the second row from movie "*Lions for Lambs*" ©United Artists, Cruise/Wagner Productions, Andell Entertainment, Brat Na Pont Productions and Wildwood Enterprises, the left one in the third row from movie "*Erin Brockovich*" ©Universal Pictures (USA) and Columbia Pictures (International), the left one in the fourth row from movie "*Up in the Air*" ©DW Studios, The Montecito Picture Company and Rickshaw Productions and the other two right ones of last two rows from movie "*The Man From Earth*" ©Anchor Bay Entertainment and Shoreline Entertainment.

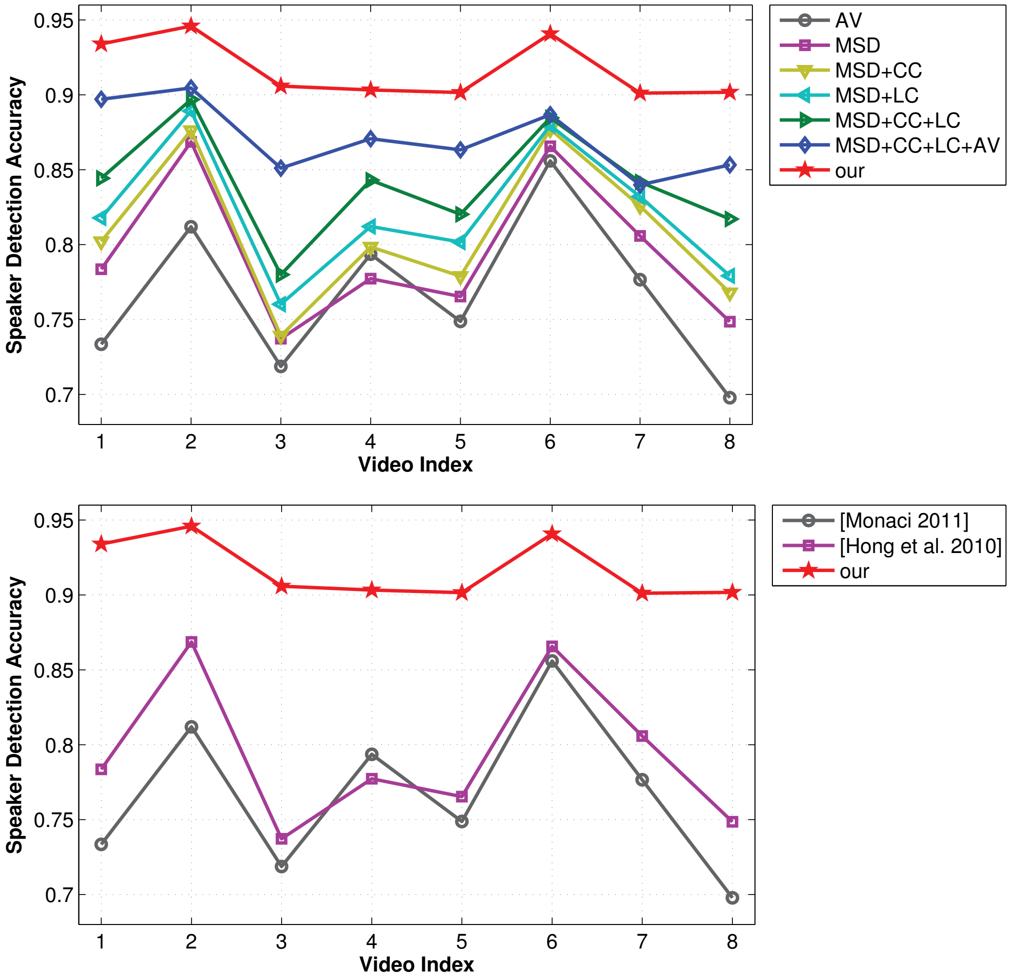


FIGURE 2.14: Performance comparison in terms of speaker detection accuracy. Up: performance improvement of different features. Bottom: performance improvement compared with the previous method. Video indexes 1 to 8 refer to the TV clips from "*Friends S10E15*" and "*The Big Bang Theory S05E16*", and the movie clips from "*Kramer vs. Kramer*", "*Erin Brockovich*", "*Up In The Air*", "*The Man From Earth*", "*Scent of a Woman*" and "*Lions for Lambs*" respectively. Ground truth speakers are manually labeled in our experiments.

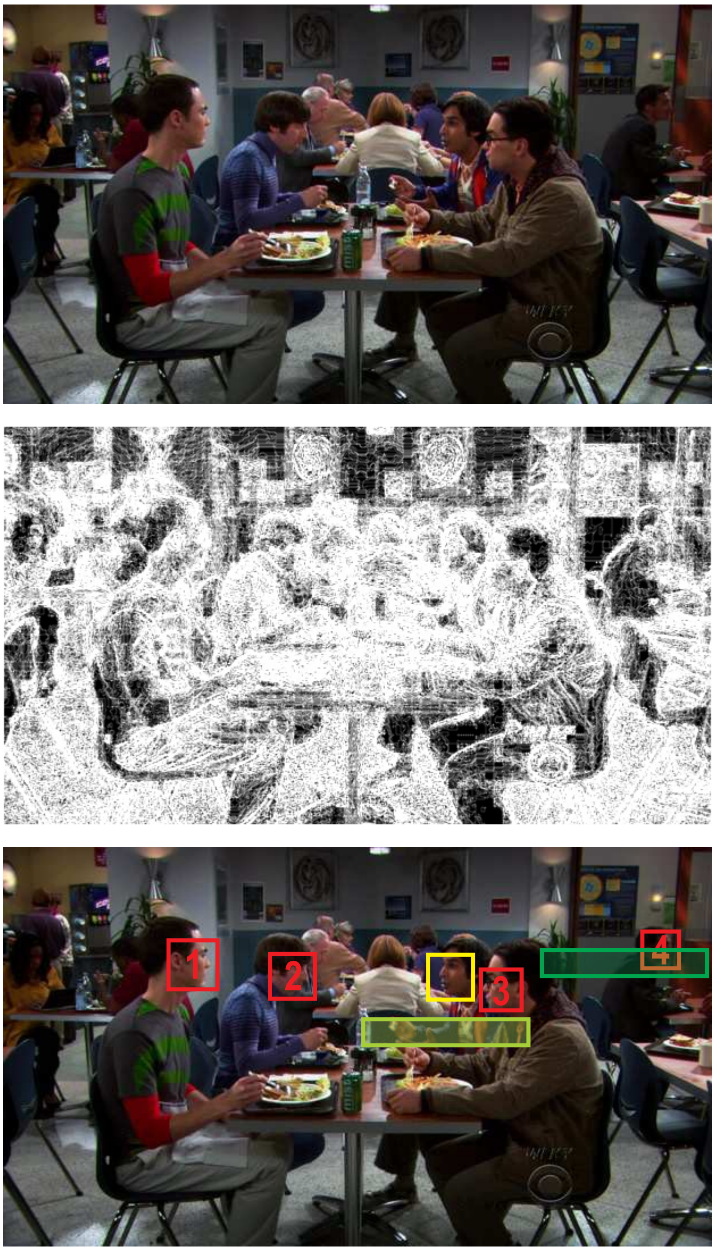


FIGURE 2.15: Use of saliency map in a complex background. Left: a screen shot of the TV clip "*The Big Bang Theory S05E16*"; Middle: a saliency map used in the method by Hong et al. [46]; Right: subtitle placement by our method (yellow shadowed rectangle) vs. that by Hong et al. [46] (green shadowed rectangle) for the speaker (yellow rectangle), red rectangles are non-speakers. The subtitle placement by the method of Hong et al. [46] mistakes the non-speaker #3 to be speaking and blocks non-speaker #4, thus confusing the viewer. Copyright info: Screenshots from "*The Big Bang Theory*" TV series ©Chuck Lorre Productions and Warner Bros. Television.

2.7 Conclusion

To enhance video viewing experience, we have developed an automatic system to place video subtitles next to their corresponding speaker. The proposed system detects and localizes speakers, and then computes optimal positions to place the subtitles. Improving the performance of both the speaker detection step and subtitle placement step enables a better presentation of video subtitles in TV and movies. A comprehensive usability study (see detailed information in Section 4.1) with 219 users validated the effectiveness of our system.

Claim: the presented material in this chapter has been published in [47].

Chapter 3

Multimodal Speaker Naming using Convolutional Neural Networks

Automatic speaker naming is the problem of identifying the name of each speaker present in a TV/movie/live show video. It is challenging because there can be big appearance variations of each character, especially in multi-season TV series. Although previous approaches demonstrated promising performance in some special videos with simple scenes, their performance is degraded in complex scenes due to their high dependency on face tracking and facial landmark detection.

We propose a novel convolutional neural networks (CNN) based learning framework to automatically learn the fusion function of both faces and audio cues. Extensive experiments on our new multi-modal dataset indicate that our proposed method can achieve state-of-the-art performance (over 90%) in speaker naming without using face/person tracking, facial landmark localization or subtitle/transcript, thus making it suitable for real-life applications.

3.1 Introduction

Identifying speakers, or speaker naming, in movies, TV series and live shows is a significant problem, since speaker identity is an important cue in lots of high-level video processing/analysis tasks, including semantic indexing, retrieval [116] and summarization [117]. Recently, some multimedia content providers, such as Amazon/IMDb [1] and Google Play [2], have had the ability to provide additional cast and characters information for movies and TV series during playback, which can be achieved by the combination of face tracking, identification and crowd sourcing.

As noted by previous authors [30], automatic speaker naming is extremely challenging because there are normally large changing of visual appearance including variation attributes in illumination, pose, scale, expression, dress, hair style and accessaries over the characters. In addition, problems with video acquisition, such as motion blur and poor image quality, make the matter even worse. Previous studies using only a single visual cue, such as face features, have failed to generate satisfactory results.

Movies, TV series and live shows are all multimedia data consisting of multiple sources of information, such as image frame, video sequence, audio, subtitle and transcript. It is therefore natural to combine these multiple sources of information to solve the speaker naming problem. In particular, audio provides reliable supplementary information for speaker naming task because it is closely associated with the video.

We propose a new method based on convolutional neutral networks (CNN) for addressing the *speaker naming* problem, that is, identifying each speaker appearing in a video/movie or a live show with a name from a fixed list. Unlike other previous methods making use of multimedia data, our method automatically learns the fusion function of both face and audio cues and achieves the state-of-the-art performance without using face/person tracking, facial landmark localization or subtitle/transcript. Our system is also trained end to end, providing an effective way to generate high quality intermediate unified features to distinguish outliers.

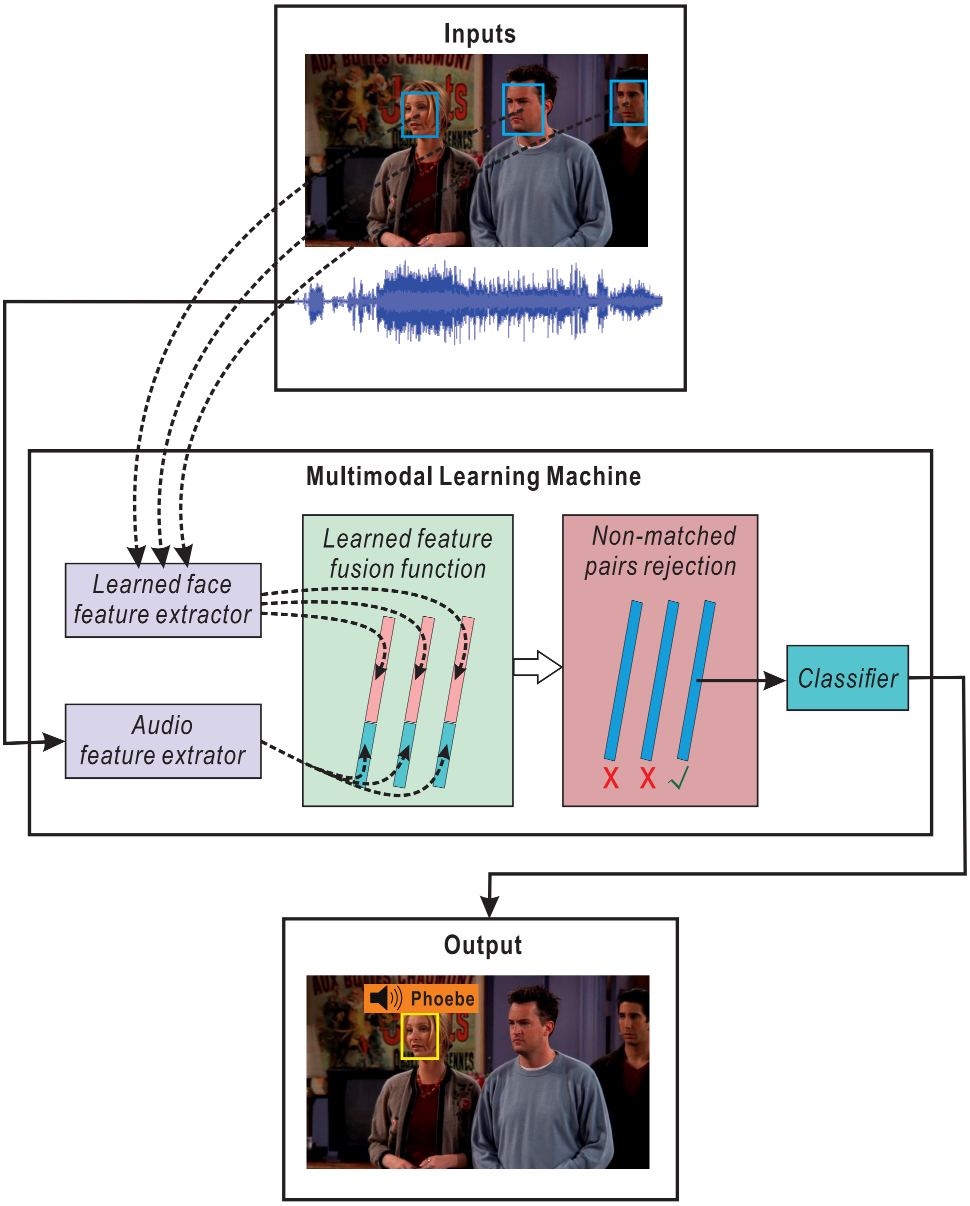

Inputs
Multimodal Learning Machine
Learned face feature extractor
Audio feature extrator
Learned feature fusion function
Non-matched pairs rejection
X X ✓
Classifier
Output
Phoebe

Figure 3.1: Multimodal learning framework for speaker naming. Copyright info: Screenshots from "*Friends*" TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

3.2 Previous Work

Automatic character/speaker naming in TV series, movies and live shows has received increasing attention in the past decade. In previous works [7, 14, 31, 34], character/speaker naming was considered as an automatic face recognition (AFR) problem. These works followed the typical AFR pipeline to detect and segment faces, localize facial landmarks, extract certain features and finally classify the faces. Due to large appearance variations of faces in movies and TV series, all these methods have difficulty matching the performance of AFR.

Recently, more researchers have tried to make use of video context to increase the accuracy of character or speaker naming. Most of these works focused on *naming face tracks* [11, 30, 86, 95, 96, 112]. Bauml et al. [11, 16] extended face tracks to person tracks to increase coverage. Besides face tracks and person tracks, other sources of information embraced in multimedia data, such as subtitle and transcript [11, 16, 25, 30], voice [96] or action [16] of actor/actress, have also been used to improve performance.

In [30], cast members in video are automatically labelled by first detecting the speakers and then aligning them with transcripts and subtitles to resolve their identities. Several work has been done by following and further refine this strategy [25, 95]. In [95], they trained a character specific multiple kernel classifier to tackle this problem. Bauml et al. [11] use a similar method to automatically identify the obtained face tracks, and determine which one is speaking. However, the ability of identifying such face tracks are very limited, where normally less than 30% of the total face tracks can be assigned [11]. That is because speaker detection relies heavily on lip movement detection, which is not reliable for videos of low quality or videos with large face pose variation.

In [96], Markov Random Field (MRF) is used to model each episode of the TV series in a probabilistic style, which combines face recognition, clothing appearance, speaker recognition and contextual constraints together. An minimization problem is then applied to formulate the identification problem. In [11], they propose a unified learning framework to resolve multiclass classification problem. It's achieved by integrates labeled and unlabeled data input, together with constraints between feature pairs in the training stage, which is further applied to train multinomial logistic regression classifiers to handle multi-class face recognition task. Bojanowski et al. [16] utilize scripts as weak supervision to learn a joint model of actors and actions in movies for character naming. It demonstrated significant improvements over previous methods by extracting features from tracked faces together with corresponding motion. Although these methods try to solve character or speaker naming problem in new machine learning frameworks, they still heavily rely on accurate face/person tracking, motion detection, landmark detection and aligned transcripts.

Unlike all these previous works, our approach does not rely on face/person track and does not need motion detection, facial landmark localization or subtitle/aligned transcript. With only the input of cropped face regions and corresponding audio segment, our approach recognizes speaker in each frame in real-time.

3.2.1 Contributions and Overview

Our main contributions are:

1. a novel CNN based framework which automatically learns high quality multimodal feature fusion functions;
2. a systematic approach to reject outliers for multimodal classification tasks typified by speaker naming;
3. a state-of-the-art system for practical speaker naming applications with accuracy over 90%, and
4. a large multi-modal set of faces and voice data with large variability in pose, illumination, facial expression, accessory, occlusion, image quality, scene and age.

The remainder work in this chapter is structured as follows. Our multimodal CNN learning framework is introduced in Section 3.3. In Section 3.4, a new large-scale multimodal face+audio dataset derived from a 10-season TV series is introduced. We show experimental results Section 3.5, followed by the conclusion in Section 3.6.

3.3 Multimodal CNN Learning Framework

The goal of this study is to build a system which is able to accurately and efficiently recognize speakers among a number of subjects in real world video streams with soundtracks. It is a challenging problem because it is not reasonable to directly formulate it as a standard classification or a feature matching problem. A single face classifier only tells the identity of one subject, however, there might be multiple subjects in the scene among which only one of them is the speaker.

Our approach is a learning based system in which we fuse the face and audio cues in the feature extraction level. The face feature extractor is learned from data rather than handcrafted. Then our learning framework is able to leverage both face and audio feature extractors and learns a unified multimodal feature extractor. This enables a larger learning machine to learn a unified multimodal classifier which takes both face image and speaker's sound track as inputs. In practical applications, it is not uncommon to encounter multimodal samples with mismatched identity (e.g., person A's face with person B's audio), therefore it's important for the framework to incorporate the function of identifying those type of samples to ensure high classification accuracy. The overview of our learning framework is illustrated in Figure 3.1.

The following subsections elaborate our detailed design of the components in this framework and the algorithm to train it. In the experiment section, we will report the superior properties of this framework and the effectiveness of the system in carrying out speaker naming.

3.3.1 Multimodal CNN Architecture

We adopted convolutional neural networks (CNN) [59, 65] as the baseline model in our learning machine. While various learning algorithms may fit in our multimodal learning machine, several unique characters of CNN make it a superior candidate including

1. It is an end to end learning machine tailored for vision tasks. It not only efficiently processes the imagery data by sharing weights, it also allows one to train image feature extractors and classifiers in a holistic fashion.
2. CNN is vectorizable. Fully vectorized CNN usually brings 10× to 40× speed up by offloading the computation to modern graphics processing units (GPU) [59]. This is very important for video processing applications.
3. As we will see shortly, CNN's architecture is inherently extensible.

This makes our extension to multimodal learning concise, efficient but powerful.

The role of CNN in our framework is two-fold. Firstly, it learns a face feature extractor from face imagery data so that we have a solid face recognition baseline. Secondly, it combines both face feature extractor as well as the audio feature extractor and learns a unified multimodal classifier.

Figure 3.2 illustrates CNN's architecture and our extension for multimodal processing. In the trainable face feature extractor part, each layer of the network can be expressed as

$$N_c(\mathcal{I}) = \sigma(\mathcal{P}(\sigma(\mathcal{I} * K^l + b^l))),\ l = 1, 2, ..., n, \tag{3.1}$$

where $\mathcal{I}$ denotes the input for each layer. $\mathcal{I}$ is usually a 3D image volume, namely 3-channel input images when $l = 1$, multi-channel feature maps when $1 < l \leq n$. K^l and b^l are the trainable convolution kernels and trainable bias term in layer l respectively. σ represents the nonlinearity in the network, which is modeled by a rectifier expressed as $f(x) = \max(0, x)$. $\mathcal{P}$ is a pooling function which subsamples the inputs by a factor of 2. Same nonlinearity is applied after the

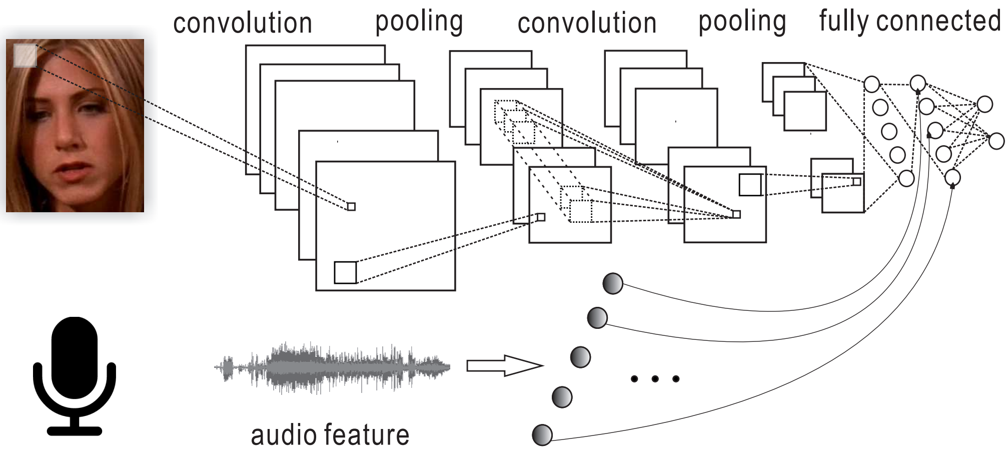


FIGURE 3.2: Multimodal CNN architecture.

pooling function. When $l = n$ the output of $N_c(\mathcal{I})$ is an one dimensional high level feature vector.

For extracting the audio features, mel-frequency cepstral coefficients (MFCCs) [26] are applied. The MFCCs of one audio frame is also a one dimensional feature vector. This allows us to ensemble a unified multimodal feature by stacking $N_c(\mathcal{I})$ and MFCCs together.

It is worth noting that the stacking of the face feature with MFCCs in this stage is non-trivial in terms of the classification. The reason is the ensuing trainable classifier essentially learns a higher dimensional nonlinear features representation of the previous layer by mapping the stacked multimodal feature to a feature space with higher dimensions. It can be expressed as

$$N_f(\mathcal{F}) = \sigma(\mathcal{F} \cdot \mathcal{W}^l + b^l),\ l = n+1, n+2, ..., m, \tag{3.2}$$

where $\mathcal{F}$ is the stack of face feature and MFCCs with the layer $l = n + 1$, and m the number of fully connected layers. When $n + 1 < l < m$ we impose the following constraint, $Dim(\mathcal{F}^{l-1}) < Dim(\mathcal{F}^l)$ where $Dim()$ denote the dimension of the intermediate feature vector, which promotes the learning of higher dimensional feature mapping during training. Such feature mapping is realized by the trainable weights $\mathcal{W}$ and b as well as the nonlineary σ. The system outputs the decision values of each class label by going through a softmax layer when $l = m$. The cross-entropy error function $\sum_{i=0}^{n} \ln(o_i) \cdot t_i$ is used as the error function during training, where o_i is the ith element in $N_f(\mathcal{F})$, t is the ground truth class label.

3.3.2 Learning Multimodal CNN

Though the conciseness of the model, one key insight of this approach is the whole system is trained end to end such that the influence of face feature extractor and MFCCs to the whole network is interwinding through learning.

3.3.2.1 Learning Procedure

During the feed-forward pass, the computation of the error function reflects the influence of both face feature extractor and MFCCs by going through the procedure described in Equation 3.2. During training, such collectively accumulated error will not only be propagated back to the classification layers of the network it will also influence the face feature extractor. Concretely, during back-propagation, when computing the error derivative

$$\delta^l = \frac{\partial E}{\partial \sigma^l} \cdot \frac{\partial \sigma^l}{\partial a^{l-1}}, \ l = 1, 2, ..., m, \tag{3.3}$$

where E is the error function, a is the activation of the layer $l-1$, δ^{n+2} to δ^m will directly influence the updates of the weights in the classification portion of the network. On the other hand, when further propagating the error back to the face feature extraction portion of the network, though the computation of δ^n directly depends on the connected fraction of δ^{n+1}, we should notice that δ^n as a whole is significantly influenced by the automatically learned multimodal feature in which MFCCs is a major player. As a result, the updates of the weights in the face extractor portion in the network will be driven by both face and audio cues. During training stochastic gradient descent (SGD) is used for optimization and a large number faces and audio samples are exposed to the network.

3.3.2.2 Multimodal Feature Extraction

One important character of the CNN based classifier is its intermediate layers are essentially high level features extractors. A number of previous studies [76, 114] showed that such high level features is very expressive and can be applied in tasks such as recognition and content retrieval. It was not clear if such high level feature extraction mechanism works well in the context of multimodal learning.

We will show in our experiments that our method is able to generate high quality multimodal features which is highly expressive in distinguishing outlier samples. This discovery forms one of the most important building blocks of making our system superior for real life speaker naming applications.

3.4 Multimodal Face+Audio Dataset (MUD4SN)

3.4.1 Properties

MUD4SN is a large-scale multimodal face+audio database with a large range of variations built upon two well-known TV series namely “Friends” and “The Big Bang Theory”.

3.4.1.1 Multimodal

For each subject in our dataset, we not only provide the face images, but also his/her speaking audio. Multimodal information has shown improved performance in many applications like image classification [44], human tracking [118], speaker detection [47], etc.

3.4.1.2 Scale

In total, our dataset contains $427,068$ face images and corresponding speaking audio segment extracted from over four hours videos of eleven episodes from two TV series ("Friends" and "The Big Bang Theory"). We extract the face images and audio for the eleven leading actors/actresses (six for "Friends" and five for "The Big Bang Theory"). There are over 38K face images for each subject in average. All images are in PNG format and audio in WAV.

3.4.1.3 Diversity

Faces built from archival videos have "natural" variability in pose, illumination, facial expression, accessory, occlusion, quality and scene [86]. Especially for "Friends", the whole TV series of 10 seasons is taken over a large time range of 10 years. To leverage such a long time span, we intentionally selected five episodes that spans the whole range, namely *S01E03* (Season 01, Episode 03), *S04E04*, *S05E05*, *S07E07* and *S10E15*. This further makes our dataset cover large aging variability. Image samples of our dataset can be seen in Figure 3.3, in which we categorize our face images in terms of different variation types.

3.4.1.4 MUD4SN and Related Datasets

Lots of face datasets have been available for research purpose. However, most of them are either captured in controlled setting [17, 37, 71, 83, 94] or too narrow in variability [8, 53] or both [72]. LFW [50], PersonID [11], PubFig [60] and YouTube Faces [111] are good candidates with large variations, but none of them are multimodal. Also, none of them has the large variation in terms of aging of subjects as MUD4SN that provides ten years of aging for the "Friends" part, though REPERE Corpus [39] can be viewed as a good candidate for person recognition. In addition, most of them have limited number of images per subject though they provide much broader in terms of subjects. Detailed comparison of our dataset with previous ones can be seen in Table 3.1.

Table 3.1: Detailed comparison of our dataset with available ones. Among the variability attributes, Po is short for pose (view points), Il for illumination, Ex for facial expression, Ac for accessory (including hat, jewelry and hair style), Oc for occlusion, Sa for image scale, Qu for image quality, Se for image scene (background), Ag for aging, MM for multimodal and N_{PS} for number of images per subject. For all attributes, we only consider them on the same subject, e.g., aging varying across different subjects will not be counted as holding aging variability. For Agi attribute, the time is also given with abbreviation of w for weeks, m for months and y for years. For N_{PS}, average values are used and $\star$ means it cannot be directly obtained from the dataset as it provides videos instead of images.

Dataset	Po	Il	Ex	Ac	Oc	Sa	Qu	Se	Ag	MM	N_{PS}
AR Face [69]		✓	✓		✓				✓ (2w)		26
BANCA [8]							✓	✓	✓ (3m)	✓	$\star$
CAS-PEAL Face [37]	✓	✓	✓	✓							30
CMU PIE [94]	✓	✓	✓								60
CVL Face [81]	✓		✓								7
FERET [83]	✓	✓	✓	✓		✓	✓		✓ (2y)		5 – 11
BIOID [53]		✓				✓		✓			66
IAM Faces [3]	✓										10
JAFFE [66]			✓								7
KFDB [17]	✓	✓	✓								52
LFW [50]	✓	✓	✓	✓	✓	✓	✓	✓			2
MUCT [72]	✓	✓									10 – 15
ORL (AT&T) [90]	✓	✓	✓	✓							10
Physics-Based Face [68]		✓									19
PubFig [60]	✓	✓	✓			✓	✓	✓			300
Richard's MIT [4]	✓			✓							6
UMIST Face [43]	✓										19 – 36
XM2VTS Face [71]	✓			✓					✓ (5m)	✓	$\star$
Yale Face B [38]	✓	✓									576
YouTube Faces [111]	✓	✓	✓	✓	✓	✓	✓	✓			390
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓ (10y)	✓	40059

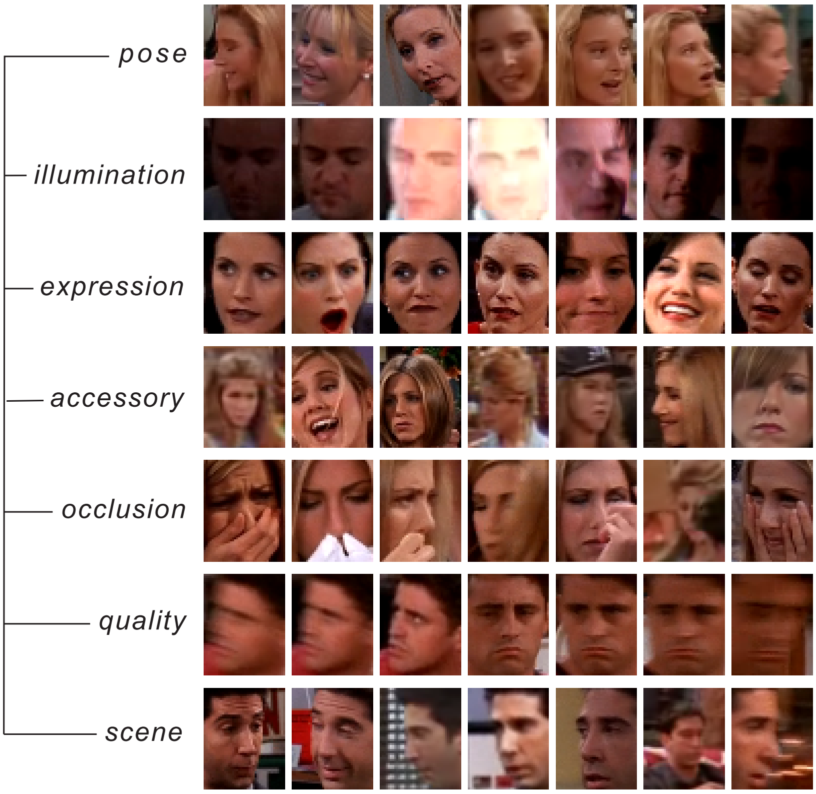


Figure 3.3: Image samples of our dataset.

3.4.2 Constructing MUD4SN

Like LFW [50], PersonID [11] and YouTube Faces [111], our faces and their corresponding audio are obtained from unconstrained archival videos, from over four hours videos of eleven episodes of two TV series, naming “Friends” and “The Big Bang Theory”. The eleven episodes that we selected are *S01E03* (Season 01, Episode 03), *S04E04*, *S05E05*, *S07E07*, *S10E15* from “Friends” and *S01E01*, *S01E02*, *S01E03*, *S01E04*, *S01E05*, *S01E06* from “The Big Bang Theory”.

3.4.2.1 Collecting Face Images

For “Friends”, we first ran a multi-view face detector [102] and then manually assigned their identifies. For “The Big Bang Theory”, we adopt the face annotations provided by [11]. To be able to span long time of aging and keep a high quantity per subject, we only keep the face images of the six leading actors/actrees for “Friends”, i.e. *Rachel*, *Monica*, *Phoebe*, *Joey*, *Chandler* and *Ross*, and five leading actors/actrees for “The Big Bang Theory”, i.e. *Sheldon*, *Leonard*, *Howard*, *Raj* and *Penny*.

3.4.2.2 Collecting Speaking Audio

We extract the audio segment based on the timing information from the associated video subtitles. Minor timing shifts are done in order to filter out noise (e.g. laughter) on both ends to make the speaker’s voice clearly stand out. One audio segment is further split into multiple parts if it contains several speakers in the single subtitle segment. Their identities are then manually assigned. All the audio data are further re-sampled to 16 kHz.

3.5 Experiments

3.5.1 Experimental Setup

For all the experiments in this chapter, faces and audio from the four videos including “*Friends S01E03”*, “*Friends S04E04”*, “*Friends S07E07”* and “*Friends S10E15”* serve as the training set and those from video “*Friends S05E05”* as the evaluation set.

We conduct three experiments to evaluate our system in terms of evaluating the performance of (1) face recognition; (2) identifying non-matched face-audio pairs and (3) real world speaker naming respectively. For face recognition using both face and audio information, we only identify matched face-audio pairs. We further show how our model be able to classify matched face-audio pairs from non-matched ones. It is worth noting that the first two experiments provide solid foundations towards achieving promising performance in our third real world experiment. It also justifies the effectiveness of the building blocks in our resulting system. And finally, in experiment 3 we illustrate the performance of our framework applied in real-life speaker naming problem.

Our CNN's detailed setting is described as follows. The network has 2 alternating convolutional and pooling layers in which the sizes of the convolution filters used are 15 by 15 and 5 by 4 respectively. The connection between the last pooling layer and the fully connected layer uses filters of size 7 by 5. The number of feature maps generated by the convolutional layers are 48 and 256 respectively. For fully connected layers, the number of hidden units are $1,024$ and $2,048$ respectively. Such architecture requires more than 11 million trainable parameters. All the bias terms are initialized to 0.01 to prevent the dead unit caused by rectifier units during training. All the other parameters are firstly initialized within the range of -1 to 1 draw from a Gaussian distribution and then scaled by the number of fan-ins of hidden unit they connect to. Average pooling of factor 2 is used throughout the network.

3.5.2 Face Model

We first evaluate our face model for face recognition purpose. To start, we tested on the small sized face images of our dataset. Specifically, by randomly selecting, $95,000$ images are used for training and $26,719$ ones for testing. To better evaluate our method, we also tested four previous algorithms under the same setting, including Eigenface [99], Fisherface [12], LBP [5] and OpenBR[1] [58].

Results are shown in the second column in Table 3.2 and their corresponding confusion matrices in Figure 3.4. We can see that all previous four algorithms fail to work well (all $< 70\%$), on the other hand, our method works better for every subject and achieves an overall accuracy of 86.695%. The results are

[1] We use the default face recognition algorithm in OpenBR, which is based on the Spectrally Sampled Structural Subspaces Features (4SF) algorithm [57].

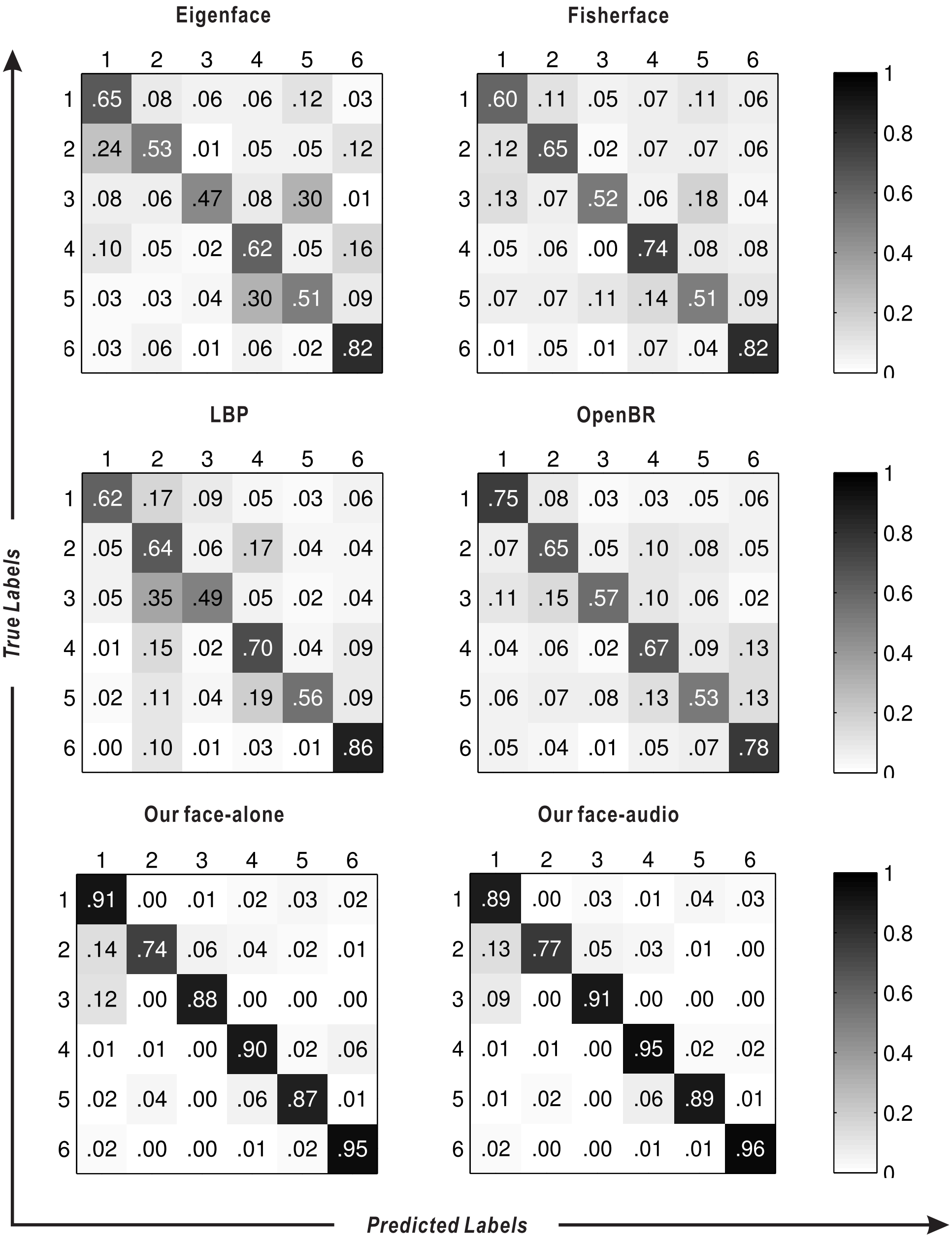
Table 3.2: Face recognition accuracy of different algorithms. Accuracy (small) is evaluated on the small fix-sized images and accuracy (large) on the original sized ones.

Algorithm	Accuracy (small)	Accuracy (large)
Eigenface [99]	60.702%	61.720%
Fisherface [12]	64.467%	67.254%
LBP [5]	65.627%	71.095%
OpenBR [58]	66.054%	73.807%
Our face-alone	86.695%	-
Our face-audio	88.525%	-

expected as previous algorithms either require alignment of the face images or detecting facial feature points or both. This makes them not able to work well in the small sized face images that are extracted from unconstrained videos, which has no guarantee of alignment of the images, challenging large pose variation, small scales, different illumination conditions, large facial expression variation and subjects spanning ten years aging, etc.

We also evaluate all the previous algorithms on the original sized face images. In specifically, 75,444 images are for training and 25,110 ones for testing. Larger images normally will result in better face recognition performance, as this will improve the opportunity of detecting facial feature points and thus further aligning them. Results are shown in the third column in Table 3.2. As expected, all algorithms improves their performances by $1-8\%$ when using larger face images. We didn't test our face model on the original sized face images because of memory limitation. But still, none of them works well (all $< 75\%$), compared with 86.695% accuracy of our algorithm even on the smaller face images.

We further used audio information to fine-tune our face model. The weights in this extended network is initialized by the parameters in the face-alone network. For the newly introduced parameters introduced by the new audio inputs, they are initialized in the same way as we presented before. Concerning the audio features, we choose 20ms as the window size and 10ms as the frame shift range. We have then selected mean and standard deviation of 25 MFCCs, and standard deviation of 2-ΔMFCCs, resulting in a total of 75 features per audio sample. In



Eigenface

	1	2	3	4	5	6
1	.65	.08	.06	.06	.12	.03
2	.24	.53	.01	.05	.05	.12
3	.08	.06	.47	.08	.30	.01
4	.10	.05	.02	.62	.05	.16
5	.03	.03	.04	.30	.51	.09
6	.03	.06	.01	.06	.02	.82

Fisherface

	1	2	3	4	5	6
1	.60	.11	.05	.07	.11	.06
2	.12	.65	.02	.07	.07	.06
3	.13	.07	.52	.06	.18	.04
4	.05	.06	.00	.74	.08	.08
5	.07	.07	.11	.14	.51	.09
6	.01	.05	.01	.07	.04	.82

LBP

	1	2	3	4	5	6
1	.62	.17	.09	.05	.03	.06
2	.05	.64	.06	.17	.04	.04
3	.05	.35	.49	.05	.02	.04
4	.01	.15	.02	.70	.04	.09
5	.02	.11	.04	.19	.56	.09
6	.00	.10	.01	.03	.01	.86

OpenBR

	1	2	3	4	5	6
1	.75	.08	.03	.03	.05	.06
2	.07	.65	.05	.10	.08	.05
3	.11	.15	.57	.10	.06	.02
4	.04	.06	.02	.67	.09	.13
5	.06	.07	.08	.13	.53	.13
6	.05	.04	.01	.05	.07	.78

Our face-alone

	1	2	3	4	5	6
1	.91	.00	.01	.02	.03	.02
2	.14	.74	.06	.04	.02	.01
3	.12	.00	.88	.00	.00	.00
4	.01	.01	.00	.90	.02	.06
5	.02	.04	.00	.06	.87	.01
6	.02	.00	.00	.01	.02	.95

Our face-audio

	1	2	3	4	5	6
1	.89	.00	.03	.01	.04	.03
2	.13	.77	.05	.03	.01	.00
3	.09	.00	.91	.00	.00	.00
4	.01	.01	.00	.95	.02	.02
5	.01	.02	.00	.06	.89	.01
6	.02	.00	.00	.01	.01	.96

FIGURE 3.4: Confusion matrices of different algorithms for face recognition on the small sized face images. Labels 1-6 stand for the six subjects accordingly, i.e., *Rachel*, *Monica*, *Phoebe*, *Joey*, *Chandler* and *Ross*.

total, we have $73,974$ audio samples for training, which we used to fine-tune the previous face model. For each face in the evaluation set, we catenate it with five audio samples of the same subject that are randomly selected from the training audio data, which results in a total of $25,110 \times 5 = 125,550$ face-audio samples for evaluation.

Compared with our previous face-alone model (accuracy: 86.695%), our face-audio model further improved this to 88.525% with corresponding confusion matrix shown in the last sub-figure in Figure 3.4. We can clearly see that, by adding audio information to the model, the accuracies of identifying all the subjects improved by $1\sim5\%$ except a slight drop for *Rachel*.

3.5.2.1 Discussion

Both experiments on the small and original sized face images verify that our dataset is a very challenging dataset for face recognition with large pose variation, different scales, different illumination conditions, large facial expression variation and subjects spanning ten years aging, etc. This also indicates that, compared with previous methods, our face model can achieve a much better performance on such a challenging dataset. Further adding audio data to fine-tune the model improve the face recognition to another level.

3.5.3 Identifying Non-matched Pairs

In all the above experiments of face models, all the training and evaluation face-audio samples are matched pairs, i.e., belong to the same person. However, this condition cannot be fulfilled in practice. Consider a speaking frame in video, there are N detected faces, one of which is speaking, see Figure 3.1 as an example where $N = 3$. In order to find the correct speaker, we need to examine all the pairs by concatenating each of these N face features with the associated audio feature. All the pairs are non-matched ones except the one of the real speaker. And, it is almost impossible to train all possible non-matched pairs because new faces are unpredictable.

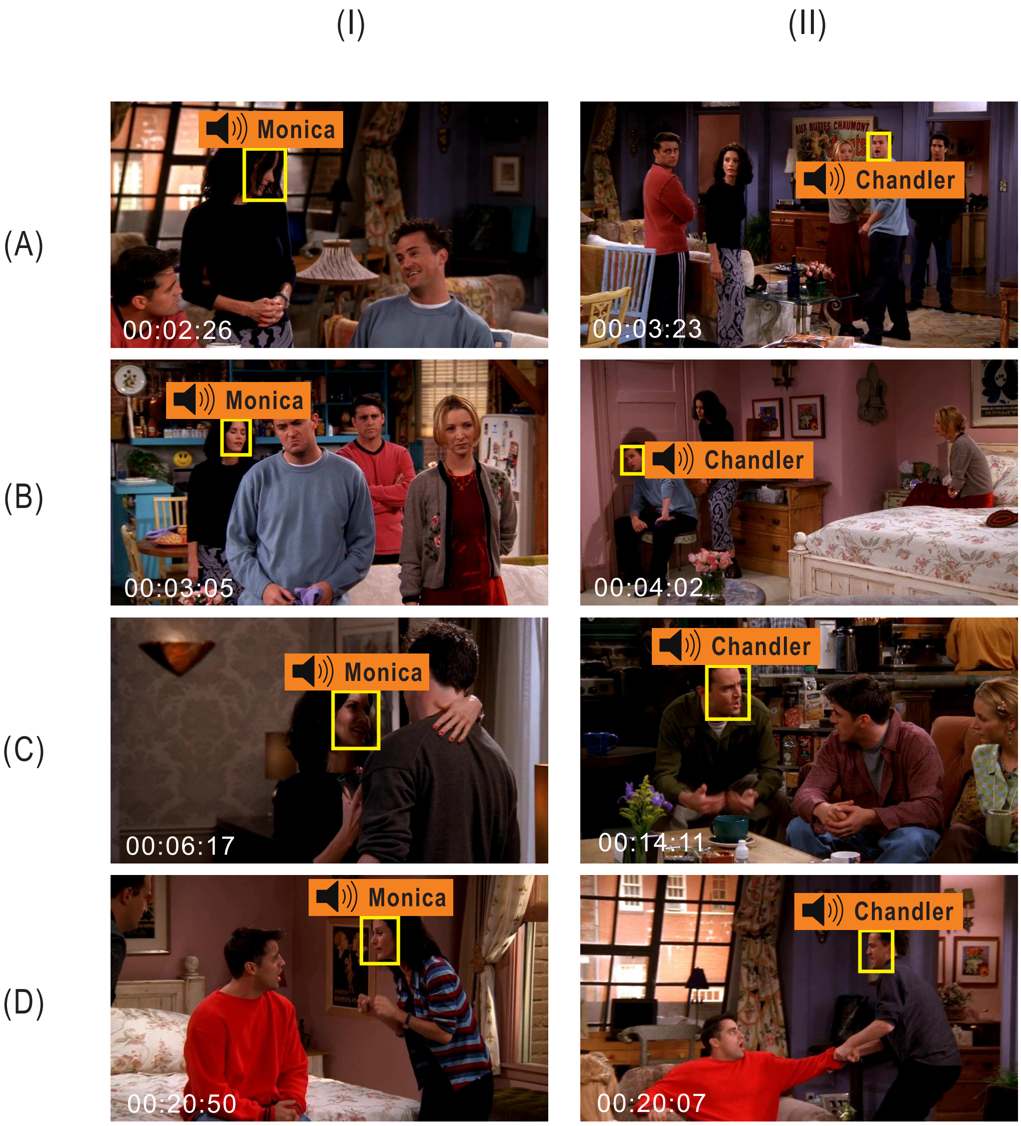


FIGURE 3.5: Some experimental results of speaker naming. Each column shows a particular subject cross the whole video, with time stamp shown at the bottom left of each sub-figure. We highlight the result under various conditions, including pose (I(A), I(D) and II(D)), illumination (I(C) and II(B)), small scale (I(B) and II(B)), occlusion (I(A)) and clustered scene (I(A) and II(A), etc.). Copyright info: Screenshots from "*Friends*" TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

On the other hand, we also find that adding some non-matched pairs during training process cannot result in good face models. Such issues appear in many real-life applications like the speaker naming problem which we will address in details in next section, where we can only achieve 72.669% if we directly use our previous model.

Thus, to identify non-matched pairs, a better way is to develop new strategies at the same time guarantee the quality of the face model. Instead of using the final output label of our face models, we explore the effectiveness of the features returned from the model in the last CNN layer. As baseline, we train two binary supporting vector machines (SVM) [19]. One is trained on the 1024-dimensional fused feature that returned from our face-audio model and the other trained on the 1024-dimensional face feature returned from our face-alone model and then concatenate it with 75-dimensional audio feature (MFCC). We then train another SVM model using the same setting with the second SVM expect that we replace the 1024-dimensional face feature by the same dimensional fused feature from our face-audio model. We randomly select $40,000$ samples for training, including $20,000$ positive and $20,000$ negative ones. In all the SVMs, radial basis function (RBF) kernel are applied.

We evaluate these three models on the evaluation video, in which there are in total $17,131$ speaking frames. It will count as correct if the most confident (has the largest distance to the supporting plane of SVM) face-audio pair match, i.e., face and audio are both from the same person. Detailed results are listed in Table 3.3. The two baseline SVM models can obtain comparable accuracy of 82.154% and 82.896% respectively, whilst the third SVM using *fused feature+MFCC* can achieve 84.069%. The results clearly justify that fused feature is more discriminative than original face feature. On the other hand, we believe the results also show that the fused feature and the MFCCs capture different but complimentary dimensions of the required information in distinguishing non-matched pairs.

Table 3.3: Accuracy of identifying non-matched pairs using different features.

Feature	Accuracy
fused feature	82.154%
face feature+MFCC	82.896%
fused feature+MFCC	84.069%

3.5.4 Speaker Naming

To show the feasibility of real-life applications of our framework, we apply it to solve the problem of *Speaker Naming*, the goal of which is to identify the speaker in each frame. In detail, for a speaking frame (with audio information), if N faces were detected, the input to our framework would be N face-audio pairs by concatenating each face feature with the audio feature, as shown in Figure 3.1. The goal here is to find out the matched face-audio pair and also identify this pair to recognize the speaker.

It's worth noting that such a problem can be viewed as an extension of previous experiment of identifying non-matched pairs, the goal of which is to find the matched pair and reject all non-matched ones. The goal of speaker naming, however, is not only to find the matched pair but also be able to obtain its identification.

For the in total $17,131$ speaking frames in the evaluation video "*Friends S05E05*" that we manually labelled, we applied the SVM trained on *fused feature+MFCC* to reject all non-matched pairs. The remaining pair will be assigned with the label returned of our face-audio model of it. We further examined this identification process in sliding window of 60 frames with a voting strategy, i.e., voting a final label during 60 frames range. The application of such a range means, for a video with 30 frames per second (FPS), we only need to cache the video content of two seconds before providing the real-time speaker naming service. Under such setting, we can achieve the speaker naming accuracy of 90.032%. Sample output of our result can be seen from Figure 3.5.

We also tested different sliding windows in the voting strategy. Detailed results can be found in Table 3.4. As expected, a larger window of 70 frames for voting

Table 3.4: Accuracy of using different sizes of the voting window.

Voting window size (N)	Accuracy
10	79.568%
30	84.413%
50	88.218%
60	90.032%
70	90.036%

results in a slightly higher accuracy of 90.036%, and vice versa. However, there should be a compromise between high accuracy and low latency as larger window means the higher latency when dealing with online problems like in real-time video streaming.

3.5.4.1 Discussion

Previous work [11, 96] have addressed similar problem of speaker naming. They were both evaluated on $1-6$ episodes of the first season of "*The Big Bang Theory*" TV series. Although they released the dataset, there is no associated audio data for the subjects, making us very hard to objectively compare with their methods in the same setting. Experiments [11, 96] showed that they can achieve person identification accuracy of 82.6% and 83.7% respectively in their setting.

In our dataset (from "*Friends*"), which we believe the setting is more challenging because the data spans a much larger time range of ten seasons across ten years, our framework can achieve person identification accuracy of over 90% in terms of the speakers. Despite the difference in settings, we argue that the figures well justified the effectiveness of our approach.

On the other hand, previous methods rely on face tracks within the time range specified by the transcript segments. Thus, their methods can be viewed as voting on the range of transcript segments. Take our evaluation video "*Friends S05E05*" as an example, the average frame of a transcript segment is 63.2. This means that they identified a person in a window of 63.2 frames, whilst our framework works better in a narrower window (90.032% accuracy over 60 frames). As mentioned above, larger windows will result in higher latency, which gives our framework higher feasibility in applications with live-streaming videos where timing requirement is crucial.

3.6 Conclusion

In this chapter, we propose a novel CNN based multimodal learning framework to tackle the task of *speaker naming*. A systematic multimodal dataset with face and audio samples collected from the real-life videos is created. The high variation of the samples in the dataset wonderfully approximates the realistic speaker naming scenarios and enables us to fully explore the potential of our method for practical applications. Our approach is able to automatically learn the fusion function of both face and audio cues. We show that our multimodal learning framework not only obtains high face recognition accuracy but also extracts representative multimodal features which is the key to distinguish sample outliers. By combining the aforementioned capabilities, our system achieved state-of-the-art performance without introducing any face/person tracking, facial landmark localization or subtitle/transcript.

Claim: the presented material in this chapter has been published in [48, 49].

Chapter 4

Applications for Video Processing

In Chapter 2, we demonstrated how to localize the speaker given video input by jointly using audio and visual information, i.e. multimodal speaker localization, and the ability to generate video subtitles that can follow the speakers. In Chapter 3, we introduced how to localize and identify the speaker in a unified framework by our novel CNN based learning framework that automatically learns the fusion function of both faces and audio cues, i.e. multimodal speaker (localization and) identification.

In this chapter, we will show the ability to extend our previous multimodal speaker localization and speaker identification algorithms (refer to Chapter 2 and Chapter 3 for details) in real-life video processing tasks. Particularly, we highlight the following applications:

1. combine applying our speaker-following video subtitles work and speaker naming work to enhance video viewing experience, described in Section 4.1;
2. automatically convert a video sequence into comics based on our speaker localization algorithms in the speaker-following video subtitles work and speaker naming work, see Section 4.2;

3. extend our speaker naming work to handle real-life video summarization tasks, see Section 4.3.

4.1 Video Viewing Experience Enhancement

The goal of our work on generating speaker-following video subtitles in Chapter 2 is to allow the viewer to follow the visual content while simultaneously reading the subtitles. Due to the limited vision span of human, video viewing experience will be greatly enhanced by applying our such system in real-life videos.

To further evaluate our system, we conducted a comprehensive usability study[1] to compare it with conventional fixed position subtitles and a previous dynamic subtitling method.

4.1.1 Video Clips Selection

For the usability study, we selected 11 video clips from the above eight TV/movie videos. Specifically, two clips were selected from the movies "*Erin Brockovich*", "*Up in the Air*" and "*The Man from Earth*" and each one clip from the other five videos. We select those video clips from video segments that contain several speakers with significant switching of dialogs. The average length of the 11 video clips is 2.3 minutes.

For a fair comparison, the video clips were chosen such that the speaker detection accuracy of them is consistent with that of the entire TV/movie videos from which they were selected (±1.84% in our user study). The consistency here is only considered for the video clips selected for our results (*Dynamic_III*) instead of to all versions (*Static*, *Dynamic_I* and *Dynamic_II*) which was probably not possible. Video clips of other versions were selected accordingly.

[1] User study of speaker-following video subtitles: https://sites.google.com/site/smartsubtitles/user_study

4.1.2 Video Group Setup

For each of the 11 video clips, we produced a group of four different versions, namely *Static*, *Dynamic_I*, *Dynamic_II* and *Dynamic_III*. Details of the version for each video group are shown in Table 4.1.

Table 4.1: Details of the four versions for each video group.

Video Versions	**Description**
Static	Traditional fixed version where subtitles are always put at the bottom of the screen
Dynamic_I	[46]
Dynamic_II	Combine our speaker detection with [46]'s subtitle placement
Dynamic_III	Ours

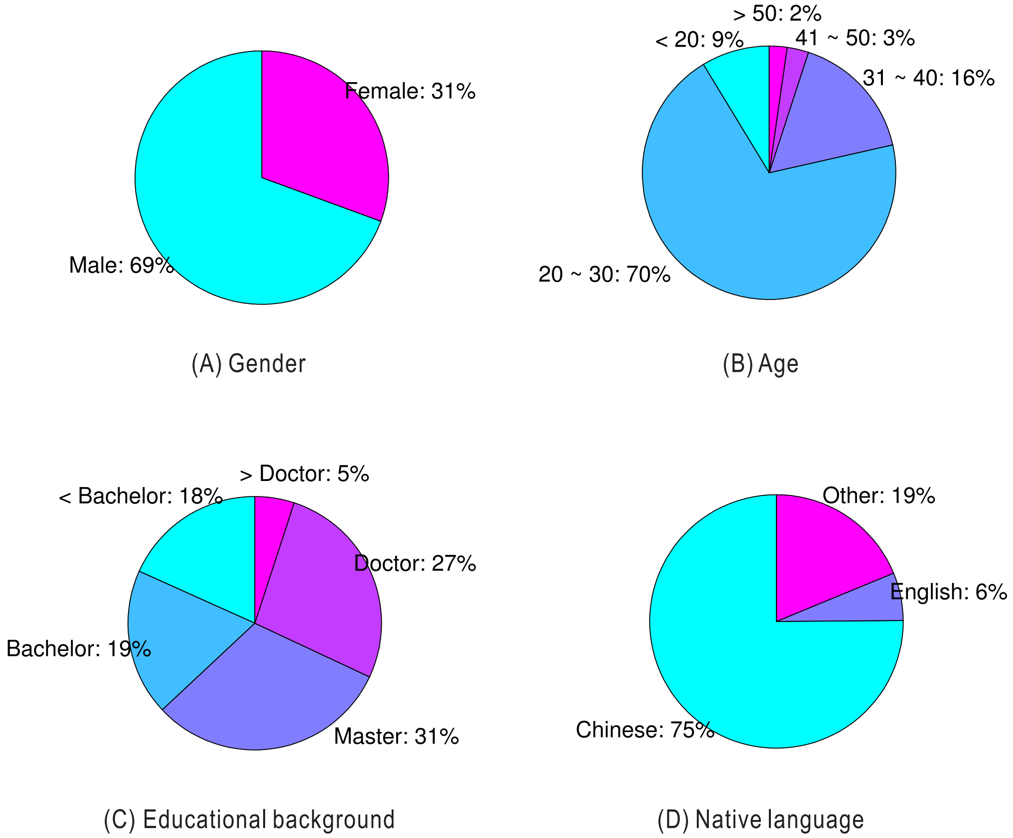
Note that volume demonstration and subtitle highlighting are included in the method by Hong et al. [46] because their method was mainly developed for people with hearing impairment. However, their usability study showed that the effects of volume demonstration and subtitle highlighting were minor for people with hearing impairments. Therefore, these cues were not used in the *Dynamic_I* version because it may be a distraction for the users in our study group. Users were asked to score (on a scale of 1 to 10) each of the four versions for overall viewing experience and for eyestrain level. The scores are uniformly defined with a higher score corresponding to a better overall viewing experience or a less eyestrain level.

4.1.3 Audio and Subtitle Language Setup

Based on the assumption that subtitles help the viewer understand the videos better, we chose specific videos with such audio so that subtitling would be useful rather than redundant. We chose clips with audio in French, Spanish and Hindi depending on the participant's language background. Subtitles, where possible, were in the native language of the reviewer, that is, Chinese for native Chinese speakers and English for native English speakers (or others).

4.1.4 Results

Each user was asked to view at least one video group of the four versions in entirety. In total, there were 219 participants in our usability study, amounting to about 20 for each video group. The users were of various ages and genders, from many different countries, and had different language and educational backgrounds (Figure 4.1).



(A) Gender

(B) Age

(C) Educational background

(D) Native language

FIGURE 4.1: Information about the participants in our user study (*gender*, *age*, *educational background* and *native language*).

The results of the user study are shown in Figure 4.2. We can see that all our results outperformed the *Dynamic_I* version [46] in terms of the overall viewing experience and degree of eyestrain level. In terms of overall viewing experience, the results showed that the *Dynamic_I* version [46] was worse or equal to the

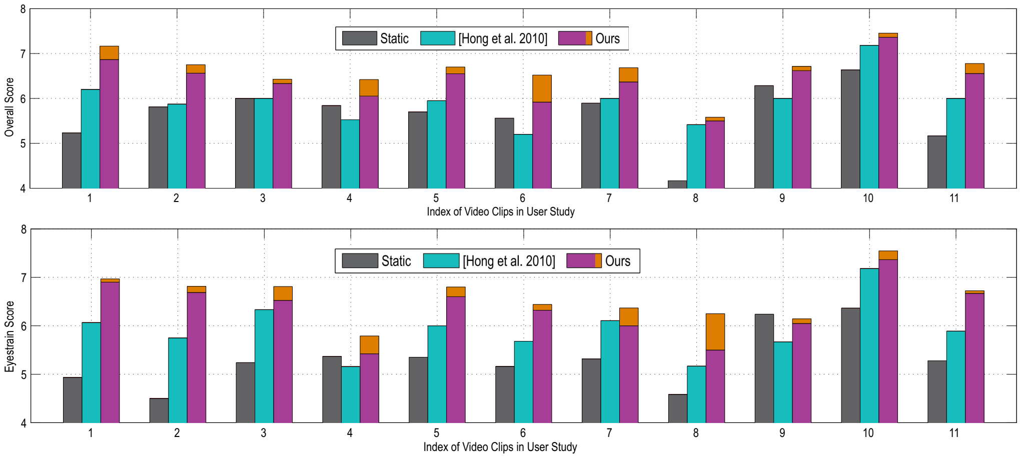


Figure 4.2: Result of the user study. Overall viewing experience (above) and eyestrain level (below). Higher scores indicate better outcome. *Dynamic_I* version is the method by Hong et al. [46], the purple colored part of the third bar is for *Dynamic_II* version and the whole third bar is our final version.

experience of Static version in 4 of the 11 groups. In comparison, the results from the *Dynamic_III* version (our method) outperformed the Static version in all groups. In terms of eyestrain, the method of Hong et al. [46] was worse than or equal to the Static version in 2 of the 11 groups, whereas our method was better than the Static version in all but one group. Nevertheless, the results for that one group (video #9) were still better than the *Dynamic_I* version [46].

The *Dynamic_II* version outperformed *Dynamic_I* version in all groups for overall viewing experience and in 10 of the 11 groups for eyestrain. Our results indicate that our speaker detection is more accurate than the method by Hong et al. [46]. Moreover, our results outperformed the *Dynamic_II* versions in terms of both overall viewing experience and eyestrain level, further validating our improved subtitle presentation method.

It should be noted that, in video #9 the level of eyestrain in all three dynamic subtitle versions were worse than in the Static version, although our method was the best. The corresponding scene in the clip showed seven people arguing with the camera shot changing rapidly, which probably caused viewers to follow the rapid switching of the subtitle position in the changing conversation.

4.1.5 Speaker Identity for Further Video Viewing Enhancement

Our speaking naming work, introduced in Chapter 3, not only resolves “where is the speaker?”, but also determines “who is the speaker?”. The main requirement of our previous work on generating speaker-following video subtitles, discussed in Chapter 2, is the position of the speaker. Thus, we can generate speaker-following video subtitles directly based on the output of our speaker naming framework.

In addition, the identity of the speaker can be also displayed along with the speaking content to enhance video viewing experience, e.g., with style “ID: spoken-words”, see Figure 4.3. More results of such merged rendering are demonstrated in Figure 4.4.

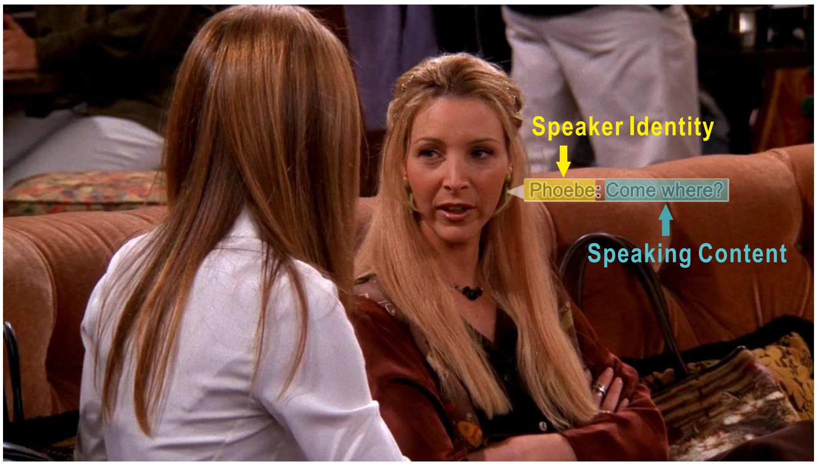


FIGURE 4.3: Add speaker identity to speaker-following video subtitles to further enhance video viewing experience. The identity of the speaker is also displayed along with the speaking content to enhance video viewing experience with style “ID: spoken-words”. Copyright info: Screenshots from “*Friends*” TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

4.1.6 Other Video Enhancing Applications

Our speaker-following video subtitles system can also be used in other video enhancing applications. In the following, we will highlight some of them.

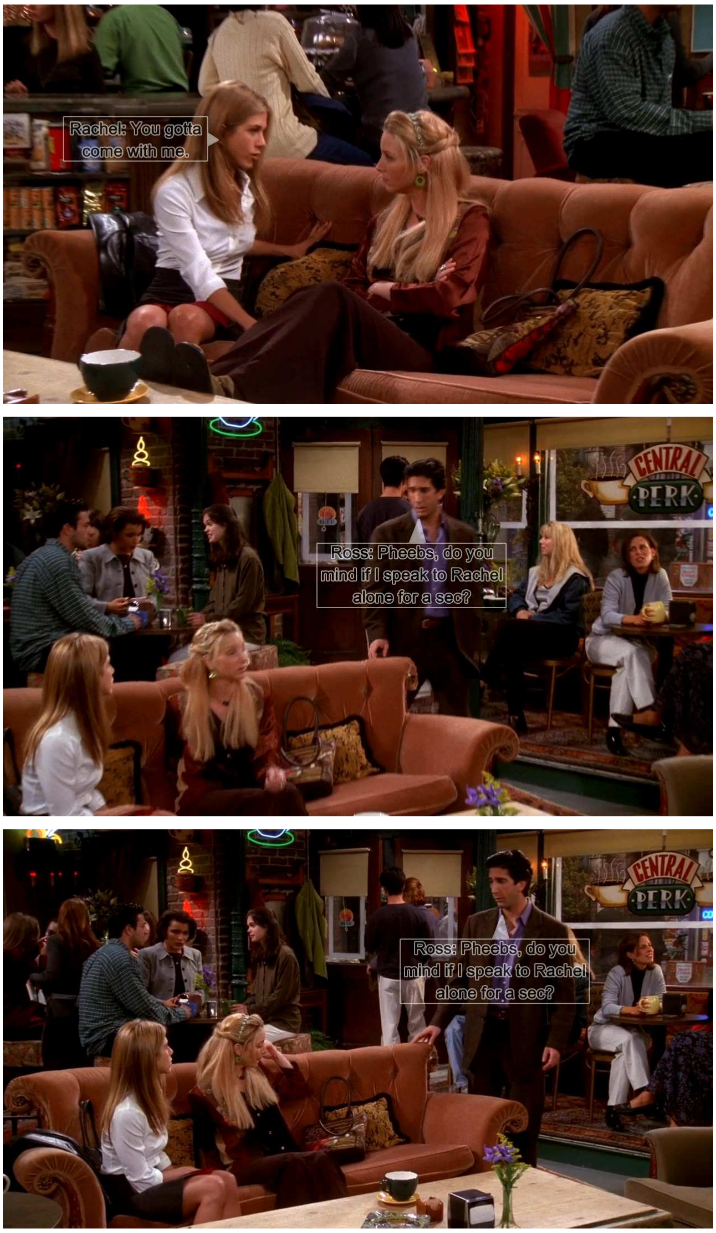

Rachel: You gotta come with me.
Ross: Pheebs, do you mind if I speak to Rachel alone for a sec?
CENTRAL PERK
Ross: Pheebs, do you mind if I speak to Rachel alone for a sec?
CENTRAL PERK

Figure 4.4: More results of adding speaker identity to speaker-following video subtitles to further enhance video viewing experience. The identity of the speaker is also displayed along with the speaking content to enhance video viewing experience with style "ID: spoken-words". The last two rows further show the ability to handle the case where the speaker moves too much in single subtitle segment, see detailed discussion in Section 2.5.3.1. Copyright info: Screenshots from "*Friends*" TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

4.1.6.1 As a Tool to Assist the TV/Movie Industry

The usability study has demonstrated that our system can assist the TV/movie industry as a speaker detection tool for subtitling to provide a better viewing experience. Although not 100% accurate (>90% accuracy), our speaker detection tool can still replace more than 90% of the manual work in generating the speaker-following subtitles for TV/movies.

4.1.6.2 Better Presentation of Multimedia in Noisy Environments and for Users with Hearing Impairment

In noisy environments (e.g., subways and squares) or in a quiet public area, audio from the video may be hard to hear clearly or is muted. The viewer may then need speaker-following subtitles to know "who is speaking or what is being said" in order to better understand and enjoy the multimedia such as news, music videos, and television interviews.

As shown in the user study conducted by Hong et al. [46], videos with dynamic subtitles can help users with hearing impairment understand the video contents. More than 66 million people in the world are suffering from hearing impairment, who can potentially benefit from the technology presented here.

Beyond this, we are now working with The Hong Kong Society for the Deaf[2] in a Innovation and Technology Support Programme (ITSP) project called *New Methodology and Software for Improving Subtitle Presentation in Movies and Videos*[3], the outcome of which holds the promise of benefitting all the people viewing a subtitled video and in particular those with hearing difficulty.

[2] The Hong Kong Society for the Deaf: http://www.deaf.org.hk/eng/index.php.

[3] Project reference: ITS/226/13. More info: https://www.itf.gov.hk/l-eng/prj_Profile.asp?code=F0DE0C0B56CE6EF16A6A20A79794F6A4A340B951F03CA1B275119611005CFA7F.

4.1.6.3 Automatic Speech Balloon Layout for 2D/3D Game Avatars

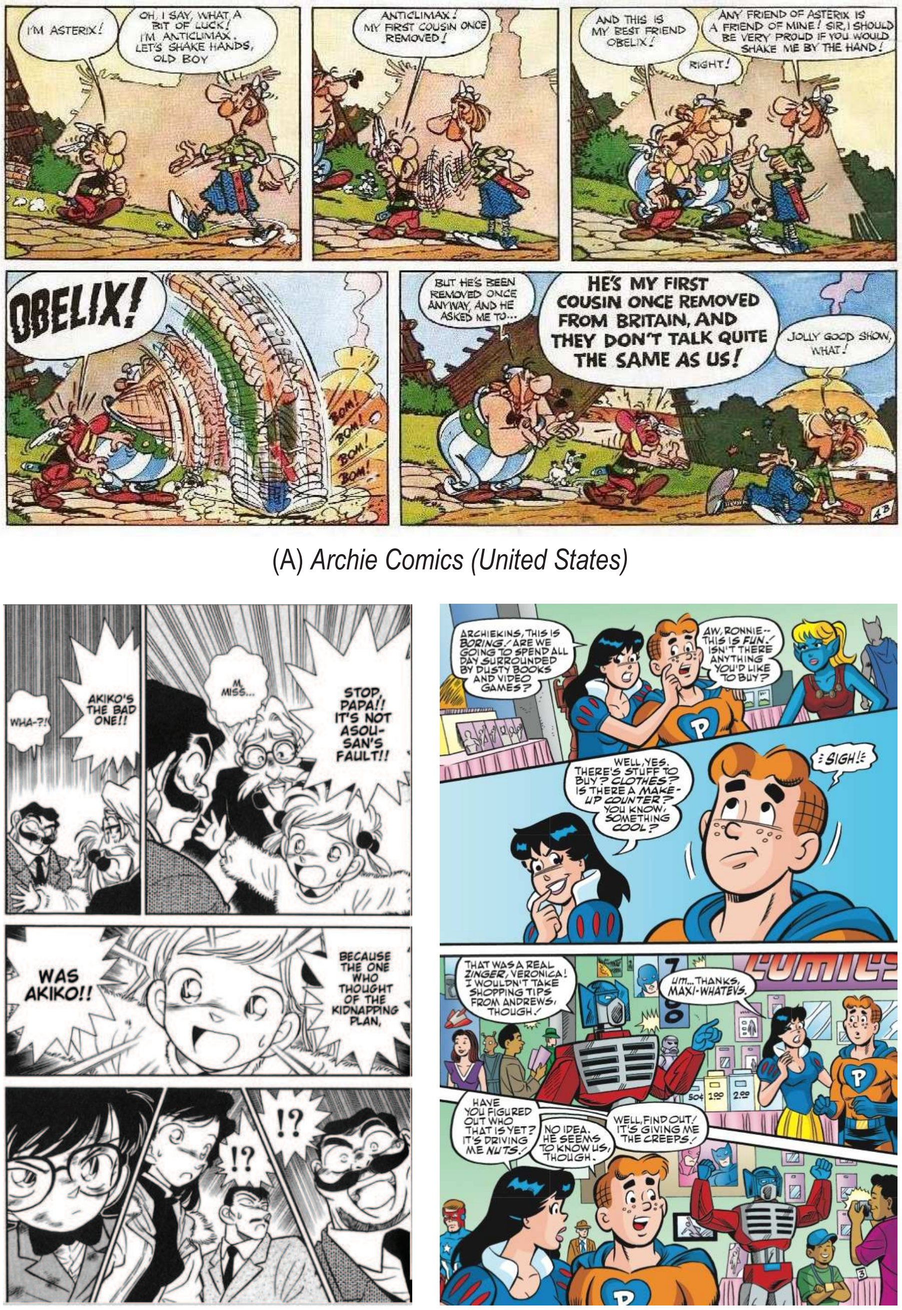
Previous methods [77, 78] simply put speech balloons above game avatars' heads. Our system can be used for automatic word balloon positioning for avatars in 2D/3D games to provide better presentation of speech for game avatars.

4.2 Converting Video to Comics

Nowadays, comics is becoming more and more popular all over the world as a prevalent artwork. Nevertheless, it is still a labor-intensive and high time-consuming process for comics creation despite the availability of auxiliary tools and softwares [106]. In this section, we will show how to automatically convert a video sequence into comics based on the output of speaker localization in Chapter 2 and 3 and the word balloon placement technique introduced in Chapter 2. With the help of Comics, people can enjoy the video content in a totally different representation, i.e. in the form of image sequence with juxtaposed panels [108]. Unlike previous methods that can only generate comics with regular rectangular layout like the one shown in Figure 4.5(A), our method can achieve comics with more flexible layouts, varying panel sizes and irregular panel shapes like those shown in Figure 4.5(B) and 4.5(C).

4.2.1 Motivation

As a type of artwork, comics describes a story in the form of image sequence via a graphical medium. It can date back to the early 20th century when it started to gain popularity with comic strip in newspapers like William Hogarth and Trajan's Column [106, 108]. Recently, comics can be widely seen in other mediums like graphic novels and magazines. Figure 4.5 shows several representative comic pages. As can be seen, comics is usually presented with regular rectangular panels, with some more flexible layout like different panel sizes, and irregular panel shapes [18] to make them more attractive and appealing [54].


I'M ASTERIX!
OH, I SAY, WHAT A BIT OF LUCK! I'M ANTICLIMAX. LET'S SHAKE HANDS, OLD BOY
ANTICLIMAX! MY FIRST COUSIN ONCE REMOVED!
AND THIS IS MY BEST FRIEND OBELIX!
ANY FRIEND OF ASTERIX IS A FRIEND OF MINE! SIR, I SHOULD BE VERY PROUD IF YOU WOULD SHAKE ME BY THE HAND!
RIGHT!
OBELIX!
BOM! BOM! BOM! BOM!
BUT HE'S BEEN REMOVED ONCE ANYWAY, AND HE ASKED ME TO...
HE'S MY FIRST COUSIN ONCE REMOVED FROM BRITAIN, AND THEY DON'T TALK QUITE THE SAME AS US!
JOLLY GOOD SHOW, WHAT!
STOP, PAPA!! IT'S NOT ASOU-SAN'S FAULT!!
M, MISS....
AKIKO'S THE BAD ONE!!
WHA-?!
BECAUSE THE ONE WHO THOUGHT OF THE KIDNAPPING PLAN,
WAS AKIKO!!
!?
!?
ARCHIEKINS, THIS IS BORING! ARE WE GOING TO SPEND ALL DAY SURROUNDED BY DUSTY BOOKS AND VIDEO GAMES?
AW, RONNIE-- THIS IS FUN! ISN'T THERE ANYTHING YOU'D LIKE TO BUY?
WELL, YES. THERE'S STUFF TO BUY? CLOTHES? IS THERE A MAKE-UP COUNTER? YOU KNOW, SOMETHING COOL?
SIGH!
THAT WAS A REAL ZINGER, VERONICA! I WOULDN'T TAKE SHOPPING TIPS FROM ANDREWS, THOUGH!
UM... THANKS, MAXI-WHATEVS.
HAVE YOU FIGURED OUT WHO THAT IS YET? IT'S DRIVING ME NUTS!
NO IDEA. HE SEEMS TO KNOW US, THOUGH.
WELL, FIND OUT! IT'S GIVING ME THE CREEPS!

(A) *Archie Comics (United States)*

(B) *Detective Conon (Japan)*

(C) *Asterix (France)*

FIGURE 4.5: Sample comics pages. Copyright info: "Archie Comics" ©Archie Comic Publications, Inc, "Detective Conan" ©AYOYAMA Gosho and Shogakukan Inc and "Asterix" ©Dargaud (France).

In literature, converting video sequences to comics has started to receive lots of attention in recent years [23, 24, 51, 85, 93, 106]. Nevertheless, it's still not an easy task to automatically convert videos to comics. Most existing algorithms and frameworks mentioned above are not automatically. As a matter of fact, lots of existing approaches that try to convey video content in comics involve lots of labor-intensive human interactions [106]. For example, the video frames selected to later placed in the final comics pages are manually selected in [51]. In [85], comics layout and the word balloon placement (contains the speaking content) are designed manually.

On the other hand, although [106] proposes a possible automatic solution for this task, it can only generate comics with regular rectangular layouts like the one shown Figure 4.5(A). In contrast, our method can achieve comics with more flexible layouts, varying panel sizes and irregular panel shapes like those shown in Figure 4.5(B) and 4.5(C). Moveover, our novel work on speaker localization and work balloon placement introduced in Chapter 2 and 3 further ensure more accurate representation of the video sequence.

4.2.2 Our Approach

Our system consists of three main components:

1. **Informative frame extraction.** Informative frames are those ones that are selected to place in the final comics pages. In this step, we will utilize our previous speaker localization algorithms introduced in Chapter 2 and 3 and try to extract one representative frame for each subtitle segment, particularly within the speaking video segments.

2. **Initial layout determination.** We will initialize the comics layout following the layout templates that we learn from real-life comics. It's worth noting that, to avoid the visual content in a panel being too small to be seen clearly on an e-book reader like Kindle or a standard A4 paper, we set 3 rows per comic page.

3. **Layout optimization.** We formulate it as an optimization problem by concerning layout geometry, visual content in each panel, and word balloon placement, relating to the display of a comic page all together. An efficient Markov chain Monte Carlo (MCMC) sampling algorithm is carefully designed for the optimization process.

Note that, the detailed workflow of this work can be found in [54], which is a cooperative work with others. In the following, I will highlight several results of our system, at the same time skip the technical details.

4.2.3 Experimental Results

We conduct experiments on several video clips that are extracted from four movies: "*Up in the Air*", "*Les Choristes*", "*The Man from Earth*" and "*Harry Potter*", and two TV series: "*Friends*" and "*The Big Bang Theory*". All videos are associated with subtitle files. Sample outputs for the six videos can be viewed from Figure 4.6, 4.7, 4.8, 4.9, 4.10 and 4.11.

4.2.3.1 Further Enhancement by Stylization

Stylization of photographs has become a tool for effective visual communication. Our system provides two ways to stylize the comics we produce. Specifically, we produce the abstraction results with simplified color illustrations by [64] and black-white, pencil-shading effects by [110]. Sample stylization results can be seen in Figure 4.11. Please refer to our paper [54] for more sample outputs. Note that other stylization methods can also apply in our system.

4.3 Video Summarization

The goal of video summarization is to provide a compact representation of given video input. It has shown its powerfulness in terms of video retrieval, indexing and condensing the original videos' storage. In literature, lots of work has been

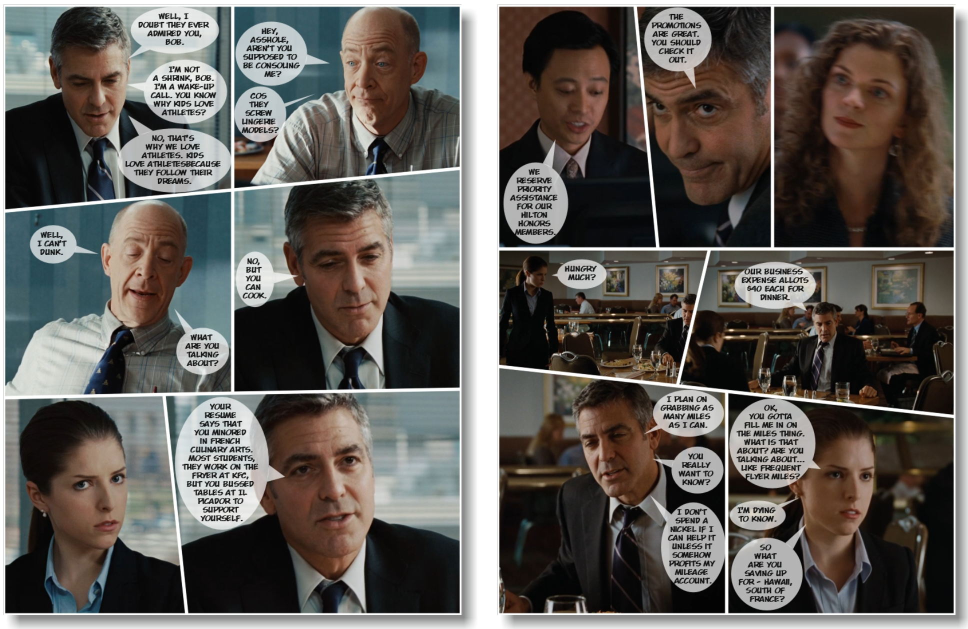

WELL, I DOUBT THEY EVER ADMIRED YOU, BOB.
I'M NOT A SHRINK, BOB. I'M A WAKE-UP CALL. YOU KNOW WHY KIDS LOVE ATHLETES?
NO, THAT'S WHY WE LOVE ATHLETES. KIDS LOVE ATHLETESBECAUSE THEY FOLLOW THEIR DREAMS.
HEY, ASSHOLE, AREN'T YOU SUPPOSED TO BE CONSOLING ME?
COS THEY SCREW LINGERIE MODELS?
WELL, I CAN'T DUNK.
WHAT ARE YOU TALKING ABOUT?
NO, BUT YOU CAN COOK.
YOUR RESUME SAYS THAT YOU MINORED IN FRENCH CULINARY ARTS. MOST STUDENTS, THEY WORK ON THE FRYER AT KFC, BUT YOU BUSSED TABLES AT IL PICADOR TO SUPPORT YOURSELF.
WE RESERVE PRIORITY ASSISTANCE FOR OUR HILTON HONORS MEMBERS.
THE PROMOTIONS ARE GREAT. YOU SHOULD CHECK IT OUT.
HUNGRY MUCH?
OUR BUSINESS EXPENSE ALLOTS $40 EACH FOR DINNER.
I PLAN ON GRABBING AS MANY MILES AS I CAN.
YOU REALLY WANT TO KNOW?
I DON'T SPEND A NICKEL IF I CAN HELP IT UNLESS IT SOMEHOW PROFITS MY MILEAGE ACCOUNT.
OK, YOU GOTTA FILL ME IN ON THE MILES THING. WHAT IS THAT ABOUT? ARE YOU TALKING ABOUT... LIKE FREQUENT FLYER MILES?
I'M DYING TO KNOW.
SO WHAT ARE YOU SAVING UP FOR - HAWAII, SOUTH OF FRANCE?

FIGURE 4.6: Comic pages generated from movie "*Up in the Air*".
Copyright info: ©DW Studios, The Montecito Picture Company and Rickshaw Productions.

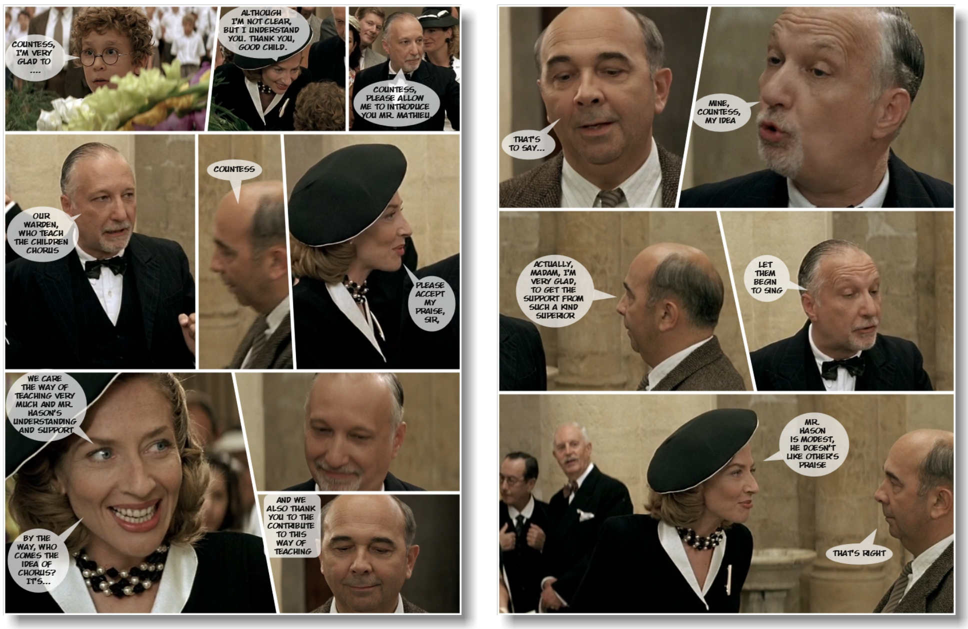

COUNTESS, I'M VERY GLAD TO ....
ALTHOUGH I'M NOT CLEAR, BUT I UNDERSTAND YOU. THANK YOU, GOOD CHILD.
COUNTESS, PLEASE ALLOW ME TO INTRODUCE YOU MR. MATHIEU.
OUR WARDEN, WHO TEACH THE CHILDREN CHORUS
COUNTESS
PLEASE ACCEPT MY PRAISE, SIR,
WE CARE THE WAY OF TEACHING VERY MUCH AND MR. HASON'S UNDERSTANDING AND SUPPORT
BY THE WAY, WHO COMES THE IDEA OF CHORUS? IT'S...
AND WE ALSO THANK YOU TO THE CONTRIBUTE TO THIS WAY OF TEACHING
THAT'S TO SAY...
MINE, COUNTESS, MY IDEA
ACTUALLY, MADAM, I'M VERY GLAD, TO GET THE SUPPORT FROM SUCH A KIND SUPERIOR
LET THEM BEGIN TO SING
MR. HASON IS MODEST, HE DOESN'T LIKE OTHER'S PRAISE
THAT'S RIGHT

FIGURE 4.7: Comic pages generated from movie "*Les Choristes*".
Copyright info: ©Pathé (UK/France) and Miramax Films (USA).

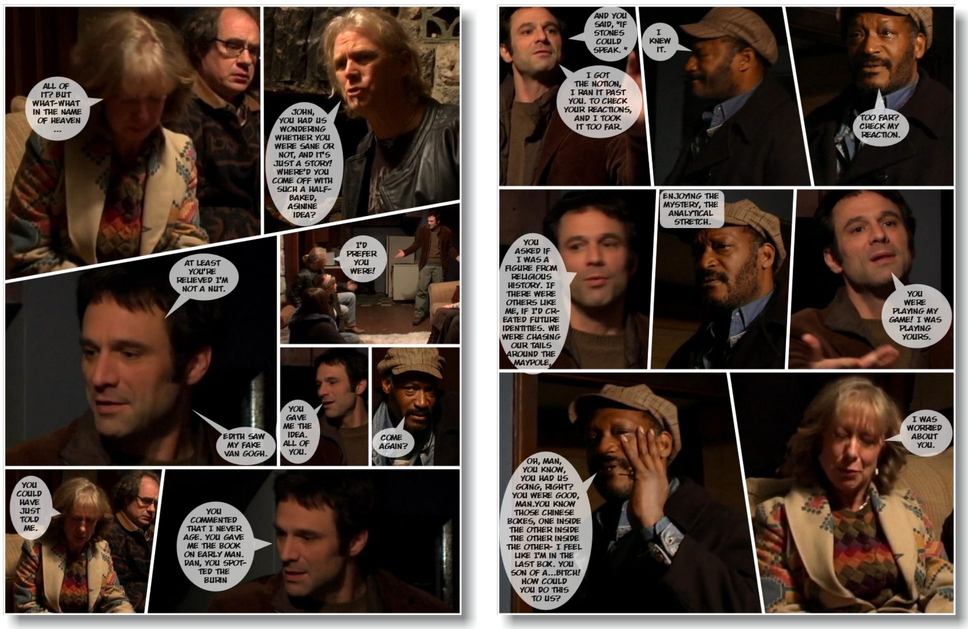

ALL OF IT? BUT WHAT-WHAT IN THE NAME OF HEAVEN ...
JOHN, YOU HAD US WONDERING WHETHER YOU WERE SANE OR NOT, AND IT'S JUST A STORY! WHERE'D YOU COME OFF WITH SUCH A HALF-BAKED, ASININE IDEA?
AT LEAST YOU'RE RELIEVED I'M NOT A NUT.
I'D PREFER YOU WERE!
EDITH SAW MY FAKE VAN GOGH.
YOU GAVE ME THE IDEA. ALL OF YOU.
COME AGAIN?
YOU COULD HAVE JUST TOLD ME.
YOU COMMENTED THAT I NEVER AGE. YOU GAVE ME THE BOOK ON EARLY MAN. DAN, YOU SPOT-TED THE BURIN
AND YOU SAID, "IF STONES COULD SPEAK. "
I KNEW IT.
I GOT THE NOTION, I RAN IT PAST YOU. TO CHECK YOUR REACTIONS, AND I TOOK IT TOO FAR.
TOO FAR? CHECK MY REACTION.
YOU ASKED IF I WAS A FIGURE FROM RELIGIOUS HISTORY. IF THERE WERE OTHERS LIKE ME, IF I'D CR-EATED FUTURE IDENTITIES. WE WERE CHASING OUR TAILS AROUND THE MAYPOLE,
ENJOYING THE MYSTERY, THE ANALYTICAL STRETCH.
YOU WERE PLAYING MY GAME! I WAS PLAYING YOURS.
OH, MAN, YOU KNOW, YOU HAD US GOING, RIGHT? YOU WERE GOOD, MAN.YOU KNOW THOSE CHINESE BOXES, ONE INSIDE THE OTHER INSIDE THE OTHER INSIDE THE OTHER- I FEEL LIKE I'M IN THE LAST BOX. YOU SON OF A...BITCH! HOW COULD YOU DO THIS TO US?
I WAS WORRIED ABOUT YOU.

Figure 4.8: Comic pages generated from movie "*The Man from Earth*".
Copyright info: ©Anchor Bay Entertainment and Shoreline Entertainment.

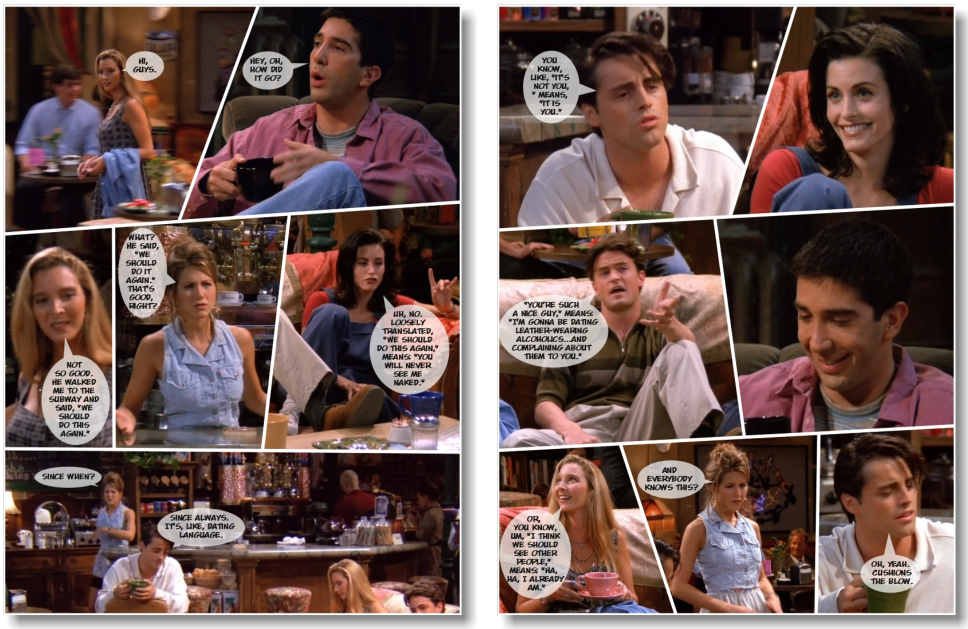

HI, GUYS.
HEY, OH, HOW DID IT GO?
WHAT? HE SAID, "WE SHOULD DO IT AGAIN." THAT'S GOOD, RIGHT?
NOT SO GOOD. HE WALKED ME TO THE SUBWAY AND SAID, "WE SHOULD DO THIS AGAIN."
UH, NO. LOOSELY TRANSLATED, "WE SHOULD DO THIS AGAIN," MEANS: "YOU WILL NEVER SEE ME NAKED."
SINCE WHEN?
SINCE ALWAYS. IT'S, LIKE, DATING LANGUAGE.
YOU KNOW, LIKE, "IT'S NOT YOU, " MEANS, "IT IS YOU."
"YOU'RE SUCH A NICE GUY," MEANS: "I'M GONNA BE DATING LEATHER-WEARING ALCOHOLICS...AND COMPLAINING ABOUT THEM TO YOU."
OR, YOU KNOW, UM, "I THINK WE SHOULD SEE OTHER PEOPLE," MEANS: "HA, HA, I ALREADY AM."
AND EVERYBODY KNOWS THIS?
OH, YEAH. CUSHIONS THE BLOW.

Figure 4.9: Comic pages generated from movie "*Friends*" TV series.
Copyright info: ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

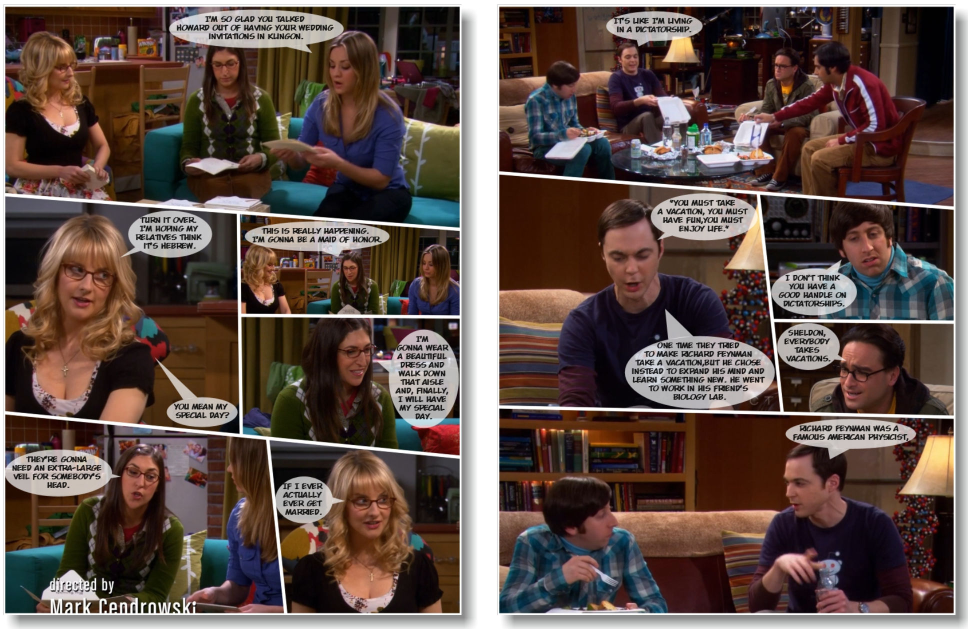

I'M SO GLAD YOU TALKED HOWARD OUT OF HAVING YOUR WEDDING INVITATIONS IN KLINGON.
TURN IT OVER. I'M HOPING MY RELATIVES THINK IT'S HEBREW.
THIS IS REALLY HAPPENING. I'M GONNA BE A MAID OF HONOR.
YOU MEAN MY SPECIAL DAY?
I'M GONNA WEAR A BEAUTIFUL DRESS AND WALK DOWN THAT AISLE AND, FINALLY, I WILL HAVE MY SPECIAL DAY.
THEY'RE GONNA NEED AN EXTRA-LARGE VEIL FOR SOMEBODY'S HEAD.
IF I EVER ACTUALLY EVER GET MARRIED.
directed by
Mark Cendrowski
IT'S LIKE I'M LIVING IN A DICTATORSHIP.
"YOU MUST TAKE A VACATION, YOU MUST HAVE FUN,YOU MUST ENJOY LIFE."
I DON'T THINK YOU HAVE A GOOD HANDLE ON DICTATORSHIPS.
ONE TIME THEY TRIED TO MAKE RICHARD FEYNMAN TAKE A VACATION,BUT HE CHOSE INSTEAD TO EXPAND HIS MIND AND LEARN SOMETHING NEW. HE WENT TO WORK IN HIS FRIEND'S BIOLOGY LAB.
SHELDON, EVERYBODY TAKES VACATIONS.
RICHARD FEYNMAN WAS A FAMOUS AMERICAN PHYSICIST,

FIGURE 4.10: Comic pages generated from "*The Big Bang Theory*" TV series. Copyright info: ©Chuck Lorre Productions and Warner Bros. Television.

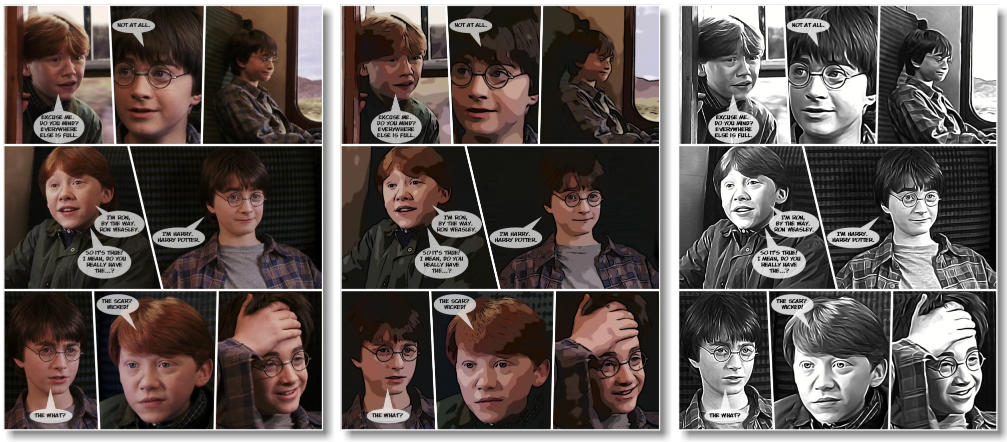

NOT AT ALL.
EXCUSE ME. DO YOU MIND? EVERYWHERE ELSE IS FULL.
I'M RON, BY THE WAY. RON WEASLEY.
I'M HARRY. HARRY POTTER.
SO IT'S TRUE! I MEAN, DO YOU REALLY HAVE THE...?
THE SCAR? WICKED!
THE WHAT?
NOT AT ALL.
EXCUSE ME. DO YOU MIND? EVERYWHERE ELSE IS FULL.
I'M RON, BY THE WAY. RON WEASLEY.
I'M HARRY. HARRY POTTER.
SO IT'S TRUE! I MEAN, DO YOU REALLY HAVE THE...?
THE SCAR? WICKED!
THE WHAT?
NOT AT ALL.
EXCUSE ME. DO YOU MIND? EVERYWHERE ELSE IS FULL.
I'M RON, BY THE WAY. RON WEASLEY.
I'M HARRY. HARRY POTTER.
SO IT'S TRUE! I MEAN, DO YOU REALLY HAVE THE...?
THE SCAR? WICKED!
THE WHAT?

FIGURE 4.11: Different stylization comic pages generated from Movie "*Harry Potter*". Besides original image style, we further produce the abstraction results with simplified color illustrations by [64] and black-white, pencil-shading effects by [110]. Copyright info: ©Warner Bros. Pictures.

done trying to summarize those long videos via a more concise style like a key-frame sequence [45], a video skim [36, 67, 74], video collage [107] and visual storylines [22]. We refer the readers to [98] for a comprehensive review of video summarization methods. Some recent methods explore the possibility to use a single static image to summarize videos, including schematic storyboard [41], or a multi-scale tapestry [10] for better navigation.

Besides converting video to comics, which can be viewed as a kind of video summarization by producing the traditional style comics which tell the story in a lively and concise manner, the output of our speaker naming work, introduced in Chapter 3, is also suitable for video summarization tasks. We can directly achieve the detailed speaking activity graph of different speakers. See Figure 4.12 for the visualization of the detailed speaking activity graph for a 3.5 minutes long video clip from "*Friends S05E05*".

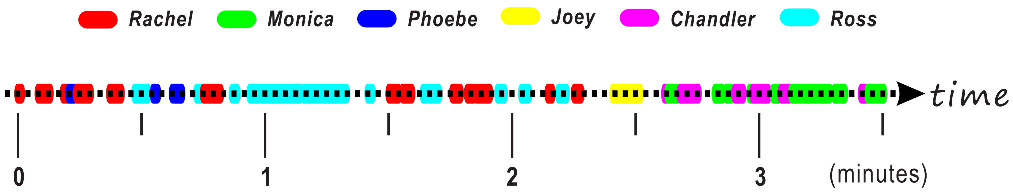


FIGURE 4.12: Visualization of the detailed speaking activity graph for a 3.5 minutes long video clip from "*Friends S05E05*". Different colors stand for different speakers.

With such visualization, we can further achieve the following two tasks, including (1) generating speaking-only part for each speaker and (2) high-level video summarization tasks.

4.3.1 Generating Speaking-only Part for Each Speaker

Speaking can be viewed as one of the most salient content in videos, especially for videos like "*Friends*" and "*The Big Bang Theory*". With the help of the detailed speaking activity graph of different speakers as in Figure 4.12, we can easily generate speaker-only part for each speaker by merging the speaking parts of the same speaker together.

Sample output can be seen from Figure 4.13. Fast speaker-based video retrieval can be further achieved on such speaking-only part video clips.

4.3.2 High-level Video Summarization

Based on the visualization of the detailed speaking activity graph, we can also achieve some high-level video summarization tasks, including

- characters conversation information, like
  - "who and who are talking to each other?"
  - "how long is one particular character speaking in total in a given time range?"
  - "when is it that at least three people are talking to each other?"
- video scene changing information (i.e., "when does scene change happen?")
- ...

We highlight several high-level video summarization examples based on our output in Figure 4.14.

4.4 Conclusion

In this chapter, through extensive experimental results, we demonstrated the ability to extend our previous multimodal speaker localization and speaker identification algorithms (refer to Chapter 2 and Chapter 3 for details) in real-life video processing tasks. Particularly, three main categories of applications were introduced, including (1) combine applying our speaker-following video subtitles work and speaker naming work to enhance video viewing experience, where a comprehensive usability study with 219 users verifies that our subtitle placement method outperformed both conventional fixed-position subtitling and another previous dynamic subtitling method in terms of enhancing the overall viewing experience and reducing eyestrain; (2) automatically convert a video sequence

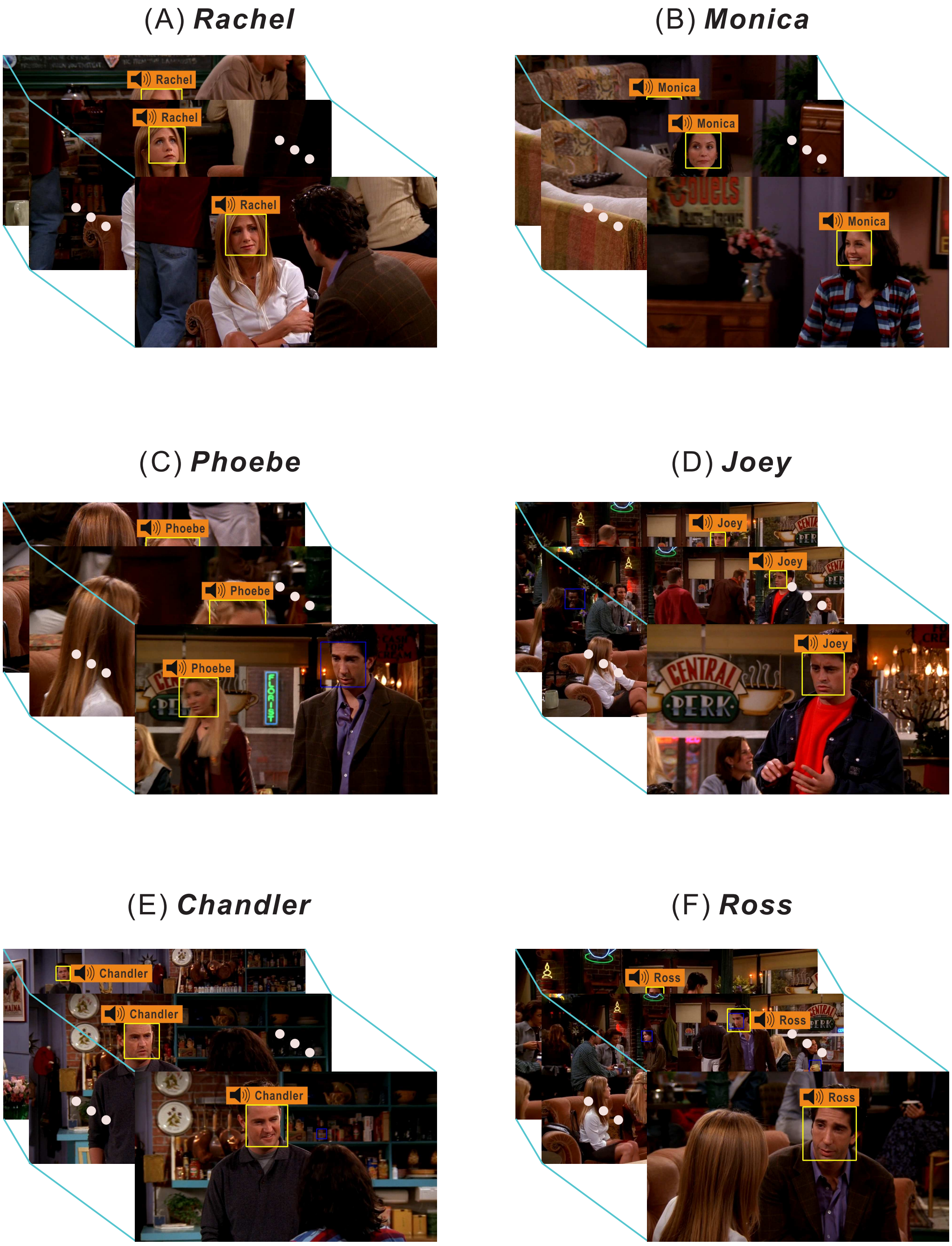


FIGURE 4.13: Generating speaking-only part for each speaker for a 3.5 minutes long video clip from "*Friends S05E05*". Copyright info: Screenshots from "*Friends*" TV series ©Bright/Kauffman/Crane Productions and Warner Bros. Television.

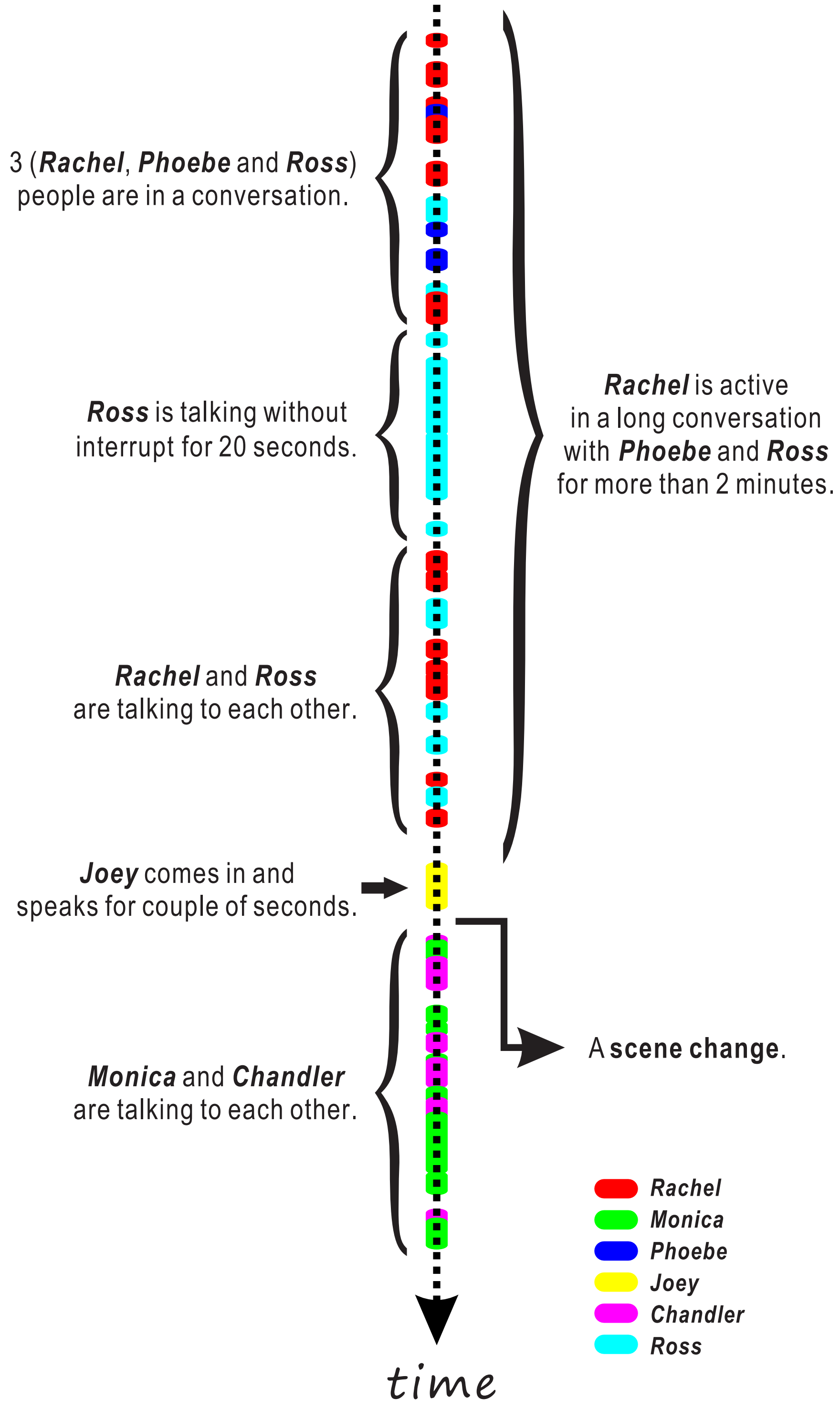

3 (Rachel, Phoebe and Ross) people are in a conversation.
Ross is talking without interrupt for 20 seconds.
Rachel and Ross are talking to each other.
Joey comes in and speaks for couple of seconds.
Monica and Chandler are talking to each other.
Rachel is active in a long conversation with Phoebe and Ross for more than 2 minutes.
A scene change.
Rachel
Monica
Phoebe
Joey
Chandler
Ross
time

FIGURE 4.14: High-level video summarization based on our speaker naming result.

into comics based on our speaker localization algorithms in the speaker-following video subtitles work and speaker naming work; and (3) extend our speaker naming work to handle real-life video summarization tasks.

Claim: the presented material in this chapter has been published in [47, 48, 54], specifically Section 4.1 is in [47], Section 4.2 in [54] and Section 4.3 is the extension work of [48].

Chapter 5

Conclusion and Future Research

5.1 Principal Contributions

In this thesis, we focus on the problem of speaker localization and identification for video processing. The contributions of the present work are as follows:

- A new algorithm for effectively detecting and localizing speakers based on multimodal visual and audio information is presented. We have introduced four new features for speaker detection and localization, including *lip motion*, *center contribution*, *length consistency* and *audio-visual synchrony*, and combine them in a cascade model. Experiments on several real-life movies and TV series indicate that *center contribution* improved the speaker detection accuracy by 0.5%–2.5%, *length consistency* improved accuracy by 1.5%–4%, and *audio-video synchrony* with better motion prediction improved accuracy by 3.5%–6.5%. Taken together, our speaker detection method outperforms previous methods by 7.5%–20.5%.
- Based on the locations of speakers, an efficient optimization algorithm for determining appropriate locations to place subtitles in order to achieve speaker-following video subtitles is proposed. This further enables us to

develop an automatic end-to-end system for subtitle placement for TV/-movies. A comprehensive usability study with 219 users verifies that our subtitle placement method outperformed both conventional fixed-position subtitling and another previous dynamic subtitling method in terms of enhancing the overall viewing experience and reducing eyestrain.

- We further propose a CNN based multimodal learning framework to tackle the task of speaker identification (*speaker naming*) in videos. Our approach is able to automatically learn the fusion function of both face and audio cues. We show that our multimodal learning framework not only obtains high face recognition accuracy but also extracts representative multimodal features which is the key to distinguish sample outliers. By combining the aforementioned capabilities, our system achieved state-of-the-art performance without introducing any face/person tracking, facial landmark localization or subtitle/transcript.
- A systematic multimodal dataset with face and audio samples collected from the real-life videos is created. The high variation of the samples in the dataset, including pose, illumination, facial expression, accessory, occlusion, image quality, scene and aging, wonderfully approximates the realistic scenarios. We show its effectness in our speaker naming task and we believe that this new dataset will also benefit other researchers to fully explore other real-life video processing applications.
- Through extensive experimental results on multiple real-life videos, we demonstrated the ability to extend our previous multimodal speaker localization and speaker identification algorithms in real-life video processing tasks. Particularly, three main categories of applications were introduced, including (1) combine applying our speaker-following video subtitles work and speaker naming work to enhance video viewing experience; (2) automatically convert a video sequence into comics based on our speaker localization algorithms in the speaker-following video subtitles work and speaker naming work; and (3) extend our speaker naming work to handle real-life video summarization tasks.

5.2 Future Research

There are many interesting problems which are worth studying. Some of these problems have been addressed in preceding chapters.

- The comprehensive usability study has validated the effectness of our speaker-following video subtitles work, i.e. outperform previous methods in terms of enhancing the overall viewing experience and reducing eyestrain. It will be interesting to accurately measure the distance of the eyes during watching such video representations. We are now working with Universität Stuttgart on this area with the help of eye-tracking equipments.

- For speaker-following video subtitles work, although our system consistently outperforms both static subtitling and a previous dynamic method, there is still room for improvement to increase user satisfaction in terms of overall viewing experience and eyestrain level. In this regard, we will further improve speaker detection accuracy especially in more challenging situations (e.g., the TV series "*ER*", in which surgeons and nurses wear face masks and move around all the time). We also aim to extend our system to cope with sound-to-text libraries in online applications such as teleconferencing and real-time video streaming where no subtitle information is available.

- One limitation of our speaker naming work is that we does not provide a holistic network to incorporate both face recognition and outlier identification. Therefore we are not able to fully accelerate our algorithm using the vectorization property of CNN in a unified manner. In our future work, we will search the possible architecture of such a network and validate it in our new dataset.

Bibliography

[1] Amazon/IMDB X-Ray. http://www.imdb.com/x-ray/.

[2] Google Play. https://play.google.com.

[3] IAM Faces Database. http://www.iam.unibe.ch/fki/databases/iam-faces-database.

[4] Richard's MIT database. http://web.mit.edu/emeyers/www/face_databases.html#richard.

[5] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(12):2037–2041, 2006.

[6] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 20(2):356–370, 2012.

[7] Ognjen Arandjelovic and Andrew Zisserman. Automatic Face Recognition for Film Character Retrieval in Feature-Length Films. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 860–867. IEEE, 2005.

[8] Enrique Bailly-Bailliére, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad Popovici,

Fabienne Porée, et al. The BANCA Database and Evaluation Protocol. In *International Conference on Audio-and Video-based Biometric Person Authentication (AVBPA)*, pages 625–638. Springer, 2003.

[9] C Barillot, D Lemoine, L Le Briquer, F Lachmann, and B Gibaud. Data Fusion in Medical Imaging: Merging Multimodal and Multipatient Images, Identification of Structures and 3D Display Aspects. *European Journal of Radiology (EJR)*, 17(1):22–27, 1993.

[10] Connelly Barnes, Dan B Goldman, Eli Shechtman, and Adam Finkelstein. Video Tapestries with Continuous Temporal Zoom. *ACM Transactions on Graphics (TOG)*, 29(4):89, 2010.

[11] Martin Bauml, Makarand Tapaswi, and Rainer Stiefelhagen. Semi-supervised Learning with Constraints for Person Identification in Multimedia Data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3602–3609. IEEE, 2013.

[12] Peter N. Belhumeur, João P Hespanha, and David Kriegman. Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):711–720, 1997.

[13] Souheil Ben-Yacoub, Yousri Abdeljaoued, and Eddy Mayoraz. Fusion of Face and Speech Data for Person Identity Verification. *IEEE Transactions on Neural Networks (TNN)*, 10(5):1065–1074, 1999.

[14] Tamara L Berg, Alexander C Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and David A Forsyth. Names and Faces in the News. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–848. IEEE, 2004.

[15] Charles Blandin, Alexey Ozerov, and Emmanuel Vincent. Multi-source TDOA Estimation in Reverberant Audio using Angular Spectra and Clustering. *Signal Processing (SP)*, 92(8):1950–1960, 2012.

[16] Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Finding Actors and Actions in Movies. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2280–2287. IEEE, 2013.

[17] Hwang Bon-Woo, Hyeran Byun, Roh Myoung-Cheol, and Lee Seong-Whan. Performance Evaluation of Face Recognition Algorithms on the Asian Face Database, KFDB. In *International Conference on Audio-and Video-based Biometric Person Authentication (AVBPA)*, pages 557–565. Springer, 2003.

[18] Ying Cao, Antoni B Chan, and Rynson WH Lau. Automatic stylistic manga layout. *ACM Transactions on Graphics (TOG)*, 31(6):141, 2012.

[19] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[20] Vassilios Chatzis, Adrian G Bors, and Ioannis Pitas. Multimodal Decision-level Fusion for Person Authentication. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans (TSMC Part A)*, 29(6):674–680, 1999.

[21] Jingdong Chen, Jacob Benesty, and Yiteng Arden Huang. Time Delay Estimation in Room Acoustic Environments: An Overview. *EURASIP Journal on Applied Signal Processing (EURASIP JASP)*, 2006(12):1–19, 2006.

[22] Tao Chen, Aidong Lu, and Shi-Min Hu. Visual Storylines: Semantic Visualization of Movie Sequence. *Computers & Graphics (C&G)*, 36(4):241–249, 2012.

[23] Hans-Christian Christiansen. *Comics & Culture: Analytical and Theoretical Approaches to Comics.* Museum Tusculanum Press, 2000.

[24] Bong-Kyung Chun, Dong-Sung Ryu, Won-Il Hwang, and Hwan-Gue Cho. An Automated Procedure for Word Balloon Placement in Cinema Comics. In *Advances in Visual Computing (AVC)*, pages 576–585. Springer, 2006.

[25] Timothee Cour, Chris Jordan, Eleni Miltsakaki, and Ben Taskar. Movie/Script: Alignment and Parsing of Video and Text Transcription. In *European Conference on Computer Vision (ECCV)*, pages 158–171. Springer, 2008.

[26] Steven Davis and Paul Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing (TASSP)*, 28(4):357–366, 1980.

[27] Anastasios Dimou, Olivia Nemethova, and Markus Rupp. Scene Change Detection for H.264 Using Dynamic Threshold Techniques. In *Proceedings of the EURASIP Conference on Speech and Image Processing, Multimedia Communications and Service (SIPMCS)*, 2005.

[28] Jon Driver. Enhancement of Selective Listening by Illusory Mislocation of Speech Sounds due to Lipe-Reading. *Nature*, 381(6577):66–68, 1996.

[29] Engin Erzin, Yücel Yemez, and A Murat Tekalp. Multimodal Speaker Identification Using an Adaptive Classifier Cascade Based on Modality Reliability. *IEEE Transactions on Multimedia (TMM)*, 7(5):840–852, 2005.

[30] Mark Everingham, Josef Sivic, and Andrew Zisserman. "Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2006.

[31] Mark Everingham and Andrew Zisserman. Automated Person Identification in Video. In *International Conference on Image and Video Retrieval (CIVR)*, pages 289–298. Springer, 2004.

[32] WAC Faernando, CN Canagarajah, and DR Bull. Scene Change Detection Algorithms for Content-based Video Indexing and Retrieval. *IOSR Journal*

of Electronics and Communication Engineering(IOSR-JECE), 13(3):117–126, 2001.

[33] Marcos Faundez-Zanuy. Data Fusion in Biometrics. *Aerospace and Electronic Systems Magazine (AESM)*, 20(1):34–38, 2005.

[34] Andrew Fitzgibbon and Andrew Zisserman. On Affine Invariant Clustering and Automatic Cast Listing in Movies. In *European Conference on Computer Vision (ECCV)*, volume 3, pages 304–320. Springer-Verlag, 2002.

[35] Gerald Friedland, Chuohao Yeo, and Hayley Hung. Visual Speaker Localization Aided by Acoustic Models. In *Proceedings of the ACM international conference on Multimedia (ACMMM)*, pages 195–202. ACM, 2009.

[36] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-view Video Summarization. *IEEE Transactions on Multimedia (TMM)*, 12(7):717–729, 2010.

[37] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The CAS-PEAL Large-scale Chinese Face Database and Baseline Evaluations. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans (TSMC Part A: SH)*, 38(1):149, 2008.

[38] Athinodoros S Georghiades, Peter N Belhumeur, and David Kriegman. From Few to Many: Generative Models for Recognition under Variable Pose and Illumination. In *IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 277–284. IEEE, 2000.

[39] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. The repere corpus: a multimodal corpus for person recognition. In *LREC*, pages 1102–1107, 2012.

[40] Herbert Gish and Michael Schmidt. Text-independent Speaker Identification. *IEEE Signal Processing Magazine (SPM)*, 11(4):18–32, 1994.

[41] Dan B Goldman, Brian Curless, David Salesin, and Steven M Seitz. Schematic Storyboarding for Video Visualization and Editing. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 862–871. ACM, 2006.

[42] Mihaela Gordan, Constantine Kotropoulos, and Ioannis Pitas. A Support Vector Machine-Based Dynamic Network for Visual Speech Recognition Applications. *EURASIP Journal on Applied Signal Processing (EURASIP JASP)*, 2002(1):1248–1259, 2002.

[43] Daniel B Graham and Nigel M Allinson. Characterizing Virtual Eigensignatures for General Purpose Face Recognition. *Face Recognition: From Theory to Applications; NATO ASI Series F, Computer and Systems Sciences (FRFTA)*, 163:446–456, 1998.

[44] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal Semi-supervised Learning for Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 902–909. IEEE, 2010.

[45] Alan Hanjalic and HongJiang Zhang. An Integrated Scheme for Automated Video Abstraction based on Unsupervised Cluster-validity Analysis. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 9(8):1280–1289, 1999.

[46] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 421–430. ACM, 2010.

[47] Yongtao Hu, Jan Kautz, Yizhou Yu, and Wenping Wang. Speaker-Following Video Subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(2), 2014.

[48] Yongtao Hu, Sijie Ren, Jingwen Dai, Chuang Yuan, and Wenping Wang. Deep Multimodal Speaker Naming. In *Proceedings of the ACM international conference on Multimedia (ACMMM)*. ACM, 2015. Submitted.

[49] Yongtao Hu, Sijie Ren, Jingwen Dai, Chuang Yuan, and Wenping Wang. MUD4SN: A New Multimodal Database for Speaker Naming. In *Proceedings of the ACM international conference on Multimedia (ACMMM)*. ACM, 2015. Submitted.

[50] Gary B Huang, Marwan Mattar, Tamara Berg, Eric Learned-Miller, et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *University of Massachusetts, Amherst, Technical Report (MAU Tech. Rep.)*, 2008.

[51] Won-Il Hwang, Pyung-Jun Lee, Bong-Kyung Chun, Dong-Sung Ryu, and Hwan-Gue Cho. Cinema comics: Cartoon generation from video stream. In *International Conference on Computer Graphics Theory and Applications (GRAPP)*, pages 299–304, 2006.

[52] Gaël Jaffré, Philippe Joly, et al. Costume: A New Feature for Automatic Video Content Indexing. In *Proceedings of Recherche d' Informations Assistee par Ordi- nateur - Computer Assisted Information Retrieval (Proc. RIAO)*, pages 314–325, 2004.

[53] Oliver Jesorsky, Klaus J Kirchberg, and Robert W Frischholz. Robust Face Detection using the Hausdorff Distance. In *International Conference on Audio-and Video-based Biometric Person Authentication (AVBPA)*, pages 90–95. Springer, 2001.

[54] Guangmei Jing, Yongtao Hu, Yanwen Guo, Yizhou Yu, and Wenping Wang. Content-Aware Video2Comics with Manga-Style Layout. *IEEE Transactions on Multimedia (TMM)*, 2015. To appear.

[55] Marcel Adam Just and Patricia Ann Carpenter. *The Psychology of Reading and Language Comprehension.* Allyn & Bacon, 1987.

[56] Tomi Kinnunen, Evgeny Karpov, and Pasi Franti. Real-time Speaker Identification and Verification. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 14(1):277–288, 2006.

[57] Brendan F Klare. Spectrally Sampled Structural Subspace Features (4SF). *Michigan State University Technical Report MSU-CSE-11-16 (MSU Tech. Rep.)*, 2011.

[58] Joshua C Klontz, Brendan F Klare, Scott Klum, Anil K Jain, and Mark J Burge. Open Source Biometric Recognition. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2013.

[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information and Processing Systems (NIPS)*, pages 1097–1105, 2012.

[60] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 365–372. IEEE, 2009.

[61] Cheng-Hao Kuo, Chang Huang, and Ram Nevatia. Multi-Target Tracking by On-Line Learned Discriminative Appearance Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 685–692. IEEE, 2010.

[62] David Kurlander, Tim Skelly, and David Salesin. Comic Chat. In *Proceedings of SIGGRAPH*, pages 225–236, 1996.

[63] Akash Kushal, Mandar Rahurkar, Li Fei-Fei, Jean Ponce, and Thomas Huang. Audio-visual Speaker Localization using Graphical Models. In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 291–294. IEEE, 2006.

[64] Jan Eric Kyprianidis and Jürgen Döllner. Real-time Image Abstraction by Directed Filtering. *ShaderX7–Advanced Rendering Techniques, Charles River Media*, pages 285–302, 2009.

[65] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE (Proc. IEEE)*, 86(11):2278–2324, 1998.

[66] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding Facial Expressions with Gabor Wavelets. In *IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 200–205. IEEE, 1998.

[67] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. A Generic Framework of User Attention Model and Its Application in Video summarization. *IEEE Transactions on Multimedia (TMM)*, 7(5):907–919, 2005.

[68] Elzbieta Marszalec, Birgitta Martinkauppi, Maricor Soriano, Matti Pietika, et al. Physics-based Face Database for Color Research. *Journal of Electronic Imaging (JEI)*, 9(1):32–38, 2000.

[69] Aleix M Martinez. The AR Face Database. *CVC Technical Report (CVC Tech. Rep.)*, 24, 1998.

[70] George W McConkie, Paul W Kerr, Michael D Reddix, David Zola, and Arthur M Jacobs. Eye Movement Control during Reading: II. Frequency of Refixating a Word. *Perception & Psychophysics (P&P)*, 46(3):245–253, 1989.

[71] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. XM2VTSDB: The Extended M2VTS Database. In *International Conference on Audio and Video-based Biometric Person Authentication (AVBPA)*, volume 964, pages 965–966. Citeseer, 1999.

[72] Stephen Milborrow, John Morkel, and Fred Nicolls. The MUCT Landmarked Face Database. *Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 201(0), 2010.

[73] Gianluca Monaci. Towards Real-Time Audiovisual Speaker Localization. In *Proceedings of European Conference on Signal Processing (EUSIPCO)*, 2011.

[74] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. Video Summarization and Scene Detection by Graph Modeling. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 15(2):296–305, 2005.

[75] Harriet J Nock, Giridharan Iyengar, and Chalapathy Neti. Speaker Localisation Using Audio-Visual Synchrony: An Empirical Study. In *International Conference on Image and Video Retrieval (CIVR)*, pages 488–499. Springer, 2003.

[76] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1717–1724, 2014.

[77] Soo-Hyun Park, Seung-Hyun Ji, Dong-Sung Ryu, and Hwan-Gue Cho. A Smart and Realistic Chatting Interface for Gaming Agents in 3-D Virtual Space. In *International Conference on Games Research and Development (CyberGames)*, 2008. To appear.

[78] Soo-Hyun Park, Seung-Hyun Ji, Dong-Sung Ryu, and Hwan-Gue Cho. A Smart Communication System for Avatar Agents in Virtual Environment. In *International Conference on Convergence and Hybrid Information Technology (ICHIT)*, pages 119–125. IEEE, 2008.

[79] Vyacheslav Parshin, Aliaksandr Paradzinets, and Liming Chen. Multimodal Data Fusion for Video Scene Segmentation. In *Visual Information and Information Systems (VIIS)*, pages 279–289. Springer, 2006.

[80] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and J Gowdy. CUAVE: A New Audio-visual Database for Multimodal Human-computer Interface Research. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages II–2017. IEEE, 2002.

[81] Peter Peer, Borut Batagelj, and Jure Kovac andFranc Solina. Color-based Face Detection in the 15 Seconds of Fame Art Installation. In *International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications (MIRAGE)*, volume 1, 2003.

[82] Patrick Perez, Jaco Vermaak, and Andrew Blake. Data Fusion for Visual Tracking with Particles. *Proceedings of the IEEE (Proc. IEEE)*, 92(3):495–513, 2004.

[83] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The FERET Evaluation Methodology for Face-recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(10):1090–1104, 2000.

[84] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent Advances in the Automatic Recognition of Audio-Visual Speech. *Proceedings of the IEEE (Proc. IEEE)*, 91(9):1306–1326, 2003.

[85] Jacqueline Preu and Jörn Loviscach. From Movie to Comic, Informed by the Screenplay. In *ACM SIGGRAPH posters*, page 99. ACM, 2007.

[86] Deva Ramanan, Simon Baker, and Sham Kakade. Leveraging Archival Video for Building Face Datasets. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.

[87] Keith Rayner. The Perceptual Span and Peripheral Cues in Reading. *Cognitive Psychology (CP)*, 7(1):65–81, 1975.

[88] Kate Saenko, Trevor Darrell, and James R Glass. Articulatory Features for Robust Visual Speech Recognition. In *Proceedings of International Conference on Multimodal Interfaces (ICMI)*, pages 152–158. ACM, 2004.

[89] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James Glass, and Trevor Darrell. Visual Speech Recognition with Loosely Synchronized Feature Streams. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1424–1431. IEEE, 2005.

[90] Ferdinando S Samaria and Andy C Harter. Parameterisation of a Stochastic Model for Human Face Identification. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 138–142. IEEE, 1994.

[91] Mehmet Emre Sargin, Engin Erzin, Yücel Yemez, and A Murat Tekalp. Multimodal Speaker Identification using Canonical Correlation Analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I–I. IEEE, 2006.

[92] Ishwar K Sethi and Nilesh V Patel. A Statistical Approach to Scene Change Detection. In *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology (EIST)*, pages 329–338. International Society for Optics and Photonics, 1995.

[93] Ariel Shamir, Michael Rubinstein, and Tomer Levinboim. Generating Comics from 3D Interactive Computer Graphics. *Computer Graphics and Applications (CGA)*, 26(3):53–61, 2006.

[94] Terence Sim, Simon Baker, and Maan Bsat. The CMU Pose, Illumination, and Expression Database. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(12):1615–1618, 2003.

[95] Josef Sivic, Mark Everingham, and Andrew Zisserman. Who are you?-Learning Person Specific Classifiers from Video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1145–1152. IEEE, 2009.

[96] Makarand Tapaswi, M Bauml, and Rainer Stiefelhagen. Knock! Knock! Who is it? Probabilistic Person Identification in TV-Series. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2658–2665. IEEE, 2012.

[97] Roberto Togneri and Daniel Pullella. An Overview of Speaker Identification: Accuracy and Robustness Issues. *IEEE Circuits and Systems Magazine (CSM)*, 11(2):23–61, 2011.

[98] Ba Tu Truong and Svetha Venkatesh. Video Abstraction: A Systematic Review and Classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1):1–37, 2007.

[99] Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience (JCN)*, 3(1):71–86, 1991.

[100] Michal Uřičář, Vojtěch Franc, and Václav Hlaváč. Detector of Facial Landmarks Learned by the Structured Output SVM. In *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 547–556, 2012.

[101] Himanshu Vajaria, Tanmoy Islam, Pranab Mohanty, Sudeep Sarkar, Ravi Sankar, and Rangachar Kasturi. Evaluation and Analysis of a Face and Voice Outdoor Multi-biometric System. *Pattern Recognition Letters (PRL)*, 28(12):1572–1580, 2007.

[102] Paul Viola and Michael Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–511. IEEE, 2001.

[103] Paul Viola and Michael J Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2004.

[104] Mahesh Viswanathan, Homayoon SM Beigi, Alain Tritschler, and Fereydoun Maali. Information Access using Speech, Speaker and Face Recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pages 493–496. IEEE, 2000.

[105] Mark T Wallace, GE Roberson, W David Hairston, Barry E Stein, J William Vaughan, and Jim A Schirillo. Unifying Multisensory Signals across Time and Space. *Experimental Brain Research (EBR)*, 158(2):252–258, 2004.

[106] Meng Wang, Richang Hong, Xiao-Tong Yuan, Shuicheng Yan, and T-S Chua. Movie2comics: Towards a Lively Video Content Presentation. *IEEE Transactions on Multimedia (TMM)*, 14(3):858–870, 2012.

[107] Tang Wang, Tao Mei, Xian-Sheng Hua, Xue-Liang Liu, and He-Qin Zhou. Video Collage: A Novel Presentation of Video Sequence. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1479–1482. IEEE, 2007.

[108] Wikipedia. Comics. http://en.wikipedia.org/wiki/Comics.

[109] Wikipedia. Vision Span. http://en.wikipedia.org/wiki/Vision_span.

[110] Holger Winnemöller, Jan Eric Kyprianidis, and Sven C Olsen. XDoG: An Extended Difference-of-Gaussians Compendium including Advanced Image Stylization. *Computers & Graphics (C&G)*, 36(6):740–753, 2012.

[111] Lior Wolf, Tal Hassner, and Itay Maoz. Face Recognition in Unconstrained Videos with Matched Background Similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534. IEEE, 2011.

[112] Chao Xiong, Guangyu Gao, Zhengjun Zha, Shuicheng Yan, Huadong Ma, and Tae-Kyun Kim. Adaptive Learning for Celebrity Identification With Video Context. *IEEE Transactions on Multimedia (TMM)*, 16(5):1, 2014.

[113] A Daniel Yarmey. Earwitness speaker identification. *Psychology, Public Policy, and Law (PPPL)*, 1(4):792, 1995.

[114] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

[115] Lei Zhang, Longbin Chen, Mingjing Li, and Hongjiang Zhang. Automated Annotation of Human Faces in Family Albums. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 355–358. ACM, 2003.

[116] Shiliang Zhang, Ming Yang, Xiaoyu Wang, Yuanqing Lin, and Qi Tian. Semantic-aware Co-indexing for Image Retrieval. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1673–1680. IEEE, 2013.

[117] Bin Zhao and Eric P Xing. Quasi Real-time Summarization for Consumer Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2513–2520. IEEE, 2014.

[118] Xiaotao Zou and Bir Bhanu. Tracking Humans using Multi-modal Fusion. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (CVPR Workshops)*, pages 4–4. IEEE, 2005.