Humanities and Social Sciences
2017; 5(2): 91-96
http://www.sciencepublishinggroup.com/j/hss
doi: 10.11648/j.hss.20170502.17
ISSN: 2330-8176 (Print); ISSN: 2330-8184 (Online)

The Study of Subtitle Translation Based on Multi-Hierarchy Semantic Segmentation and Extraction in Digital Video

Wang Xuemei*, Jiang Yue, Wang Fang

School of Foreign Languages, Xi'an Jiaotong University, Xi'an, China

Email address:

wang.xm@mail.xjtu.edu.cn (Wang Xuemei), jiang8@mail.xjtu.edu.cnj (Jiang Yue), w9371z@mail.xjtu.edu.cn (Wang Fang)
*Corresponding author

To cite this article:

Wang Xuemei, Jiang Yue, Wang Fang. The Study of Subtitle Translation Based on Multi-Hierarchy Semantic Segmentation and Extraction in Digital Video. *Humanities and Social Sciences.* Vol. 5, No. 2, 2017, pp. 91-96. doi: 10.11648/j.hss.20170502.17

Received: March 22, 2017; **Accepted**: April 14, 2017; **Published**: April 21, 2017

Abstract: As a set of visual, auditory and text information in one integrated media information, video plays an increasingly prominent role in people's lives. This paper established a reasonable and effective multi-hierarchy semantic information descriptive model based on video segmentation and extraction technology to realize the mapping of video semantic information from the low-level feature to high-level and meanwhile carried out the subtitle translation studies based on the technical level. The study shows that the model can well describe the service practice in subtitle translation in the premise of integration of video multi-modal information, based on the right segmentation and extraction of video object, on account of the consideration of each video object synchronization as well as temporal-spatial constraints related issues.

Keywords: Video Segmentation and Extraction Technology, Multi-Hierarchy Semantic Information Descriptive Model, Temporal-Spatial Constraints, Subtitle Translation, Translation Strategies

1. Introduction

Video is a collection of visual, auditory, and text information in an integrated media. With the continuous development of information processing technology, computer technology, network technology and communication technology, video data are at explosive growth in geometric progression. It has a wide range of applications, such as intelligent transportation, robot vision, blind navigation, subtitle translation and so on. The text, audio and image information contained in the video are closely related to the video content, which can provide important clues for the understanding and retrieval of video content. Therefore, it is of great significance to detect and extract text information from the video data, including speech transcription text, which will promote the research and application in the related fields.

At present, content-based video analysis and retrieval technology has become a very active research direction in the field of multimedia technology. The video stream has been segmented according to certain rules, the feature describing the video content has been automatically extracted, and the video feature databases have been organized by using the data structure, so we can quickly browse or search in its organizational structure. Among them, Y. Zhu, D. Zhou Video did much study on browsing and retrieval based on multimodal integration [1]. H. F. Wang, Z. X. Sun focused their study on the field of semantic processing method of content-based image retrieval [2]. J. Han, X. Zhang, W. Y. Sun specialized in an implementation method of automatic segmentation and tracking of video motion object [3]. N. Dimitrova, H. J. Zhang, B. Shahraray made a specialty in applications of video-content analysis and retrieval [4]

The representative research institutes abroad include: Digital Library Project funded by the American NSF, ARPA and NASA; Multimedia Document Retrieal Project by the University of Cambridge; Microsoft news video browsing system; IBM CueVdeo system; PhotoBook system of MIT, and VisualSEEk system of Columbia university. At home we have Tsinghua University' TV-FI Video Find It program, as well as Zhejiang University' Webscope-CBVR video retrieval system, etc.

Although the video analysis and retrieval based on the content has made great progress, there is a long way for us to extract hierarchical semantic video object from diverse video resources, and establish a multi-hierarchy semantic video tree

with robust. This paper attempts to establish a reasonable and effective multi-hierarchy semantic information descriptive model based on video segmentation and extraction technology to realize the mapping of video semantic information from the low-level to high-level, and meanwhile carries on the subtitle translation studies based on the technical level.

2. Characteristics of Video Data

There are obvious differences in the structure of video data and text, image and audio data, mainly in the following three aspects:

2.1. Multi-Modal Characteristics of Video Data

Video data is essentially a combination of three polymorphic media of text, image and audio, which includes visual, audio and text information respectively. In the video system, the text, the image sequence and the synchronization of the audio signal are known as the three basic models of video data [1].

2.1.1. Text Model

The text model refers to all the text information contained in the video image sequence, including text information and speech transcripts text. Text data is a kind of data of pure character numeric structure, which does not involve the dual attributes of space (spatial) and time (temporal). The video text can be divided into scene text and artificial text. Scene text refers to the words being photographed together with the video scene, such as road signs on the road, words on the clothing and trademarks on the products, which is a part of the original natural scene, and most of the scene text has no specific meaning. Artificial text refers to the text artificially added to the video frames by using image processing tools, which contains the semantic description of the current video. The latter plays an important role in the analysis and retrieval of video content, such as captions, titles, characters dialogues.

2.1.2. Image Model

The image model refers to all the entities that can be displayed and observed in the video sequence. The image data is a kind of non-structural static data which has the property of space but no property of time.

Table 1. *Comparison of text, image, audio, and video data.*

Comparative standard	text	image	audio	video
Information content	small	rich	rich	richer
Special dimension	Static, 1 dimension	Static, 2 dimensions	Dynamic, 2 dimensions	Dynamic, 3 dimensions
Data organization	structured	unstructured	unstructured	unstructured
Data quantity	small	medium	medium	large
Data relationship	Simple, definable	Complex, indefinable	Complex, indefinable	Very Complex, indefinable

2.1.3. Audio Model

Audio model contains both the sound which can cause hearing voices including speech, music, environmental sound and the mute which cannot cause hearing voices. Audio data containing rich semantic information is a kind of unstructured data which has dynamic time attribute but no space attribute. The combination of audio information and visual information plays an important role for video content analysis.

The video data is the complex empty dimension covering the above text, image and audio information, which not only has the spatial attributes but also the temporal attribute. In other words, the video data contains both space dimension and time dimension. Spatial dimension means each frame image is two-dimensional structure, containing very complex spatial information, which is difficult to establish a clear structure. Time dimension means the video data is a flow structure composed of a plurality of image frames and distributed along the time axis. Therefore, we also call the video data 3D data. From the underlying physical characteristics, video data itself does not have any structural information, which can be seen as a sequence of image frames on the timeline. The time interval varies according to different video formats, PAL standard about 25ms, NTSC about 33ms. The temporal-spatial multi-dimensional dynamic unstructured characteristics of video data make it difficult to express and establish the data model.

2.2. Multi Granularity Characteristics of Video Data

Granularity refers to a number of relatively simple pieces of information divided by human in accordance with their respective characteristics and performance in the solution, processing and storage of large amounts of complex information in order to facilitate the processing. Each division block is considered to be a granular space [5]. The granular space of different thickness reflects the refinement and abstract levels of information and knowledge for human. The whole video as the research object itself contains a variety of attributes and characteristics. At the same time, the various parts of the video can be seen as separate objects with self attributes and characteristics, and they have certain relationships between each other. All of these attributes and relationships constitute the complete video semantic model. Among them, subtitle semantic in the video is the focus of this paper [2].

2.3. Complex Logical Structure of Video Data

Another important feature of the video data stream is its plot development. The director in the shooting gets different scenes and shots by using the concept of "split scenes" and "split shots". After that, in the post production the creative staff will combine them according to the plot development.

Finally, the video program can be obtained for the audience. Although the video data is a kind of no structural data in the form, but from the content perspective it has very strong logic structure. It expresses the specific concept information through a number of consecutive events, tasks and actions.

To sum up, because the video data is not a symbolic text, but a more vivid video language, video stream should be segmented and abstracted at different levels in order to be stored in the database for later application. Therefore, it becomes very important about how to split the structure unit and how to extract the hierarchical semantic information according to the characteristics of video data.

3. Segmentation and Extraction of Video Hierarchical Semantic Information

The video is a continuous image sequence, each frame of which (a sampling time) can be regarded as an image. Video is actually the extension of image in the time dimension. But it is not a simple extension of video image in the time dimension, which involves time redundancy and space redundancy between adjacent frames of a video sequence. The semantic information segmentation and extraction of video object involves many research areas, such as digital image processing, digital video processing, digital morphology, computer graphics and computer vision, etc. [3].

The key to the semantic information extraction of video object lies in the segmentation technology. Video segmentation technology is to complete the automatic recognition of video information in order to solve the problem of unstructured video content, and to provide a roadmap for the semantic information extraction in the next step.

3.1. Video Structure Segmentation and Extraction Technology

According to the granularity, video structure is divided into four levels: Video, Video Scene, Video Shot, Video Frame. The three corresponding major technologies include: scene detection technology, shot segmentation technology and key frame extraction technology [6].

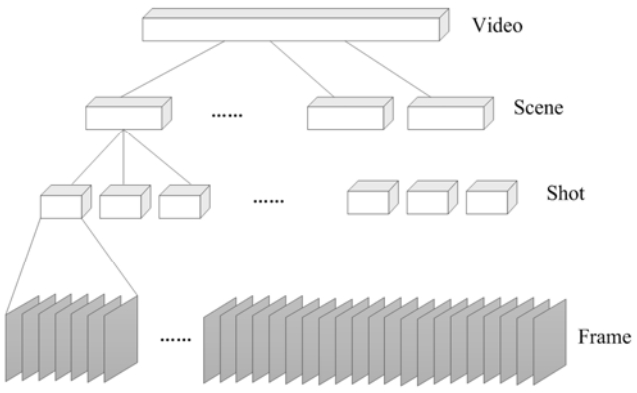

Figure 1. *Video structure model.*

3.2. Scene Detection Technology

The scene refers to a series of shots content related and time continuous. In the same scene, different shots sharing the same background combined together express relatively complete, independent high level semantic information. In contrast, the single shot usually does not consider the correlated semantic information, but the scene is usually composed of a series of relevant semantic information sequence of the same theme. It is necessary to combine the shots related in content into the higher level of the scene with complete semantic information for the convenience of video post-production.

3.3. Shot Segmentation Technology

The shot refers to the sequence of successive image frames recorded in a camera shooting action. The shot, as the basic unit of analysis of video content, up to be merged into the scene, down to be subdivided into image frames, is one of the most important link to video segmentation. If the feature between adjacent frames changes, we acknowledge that there is a shot switcher According to the different ways of connection, shot switching can be divided into two basic types: Abrupt Transition and Gradual Transition [7]. Because the image frames within a shot contains semantic information and has a strong correlation in the content, usually a shot will be regarded as a minimum physical unit for processing of the video stream. The shot segmentation is correct or not will directly affect the extraction of the key frames.

3.4. Key Frame Extraction Technology

After shot segmentation, the key frame representing the particular shot needs to be extracted. The so-called key frame extracted from the original video frames in the sequence can reflect a synopsis of the shot. Because the key frame plays an important role in video post-production, the key frame extraction technology is a hot research topic in multimedia information technology field. From the perspective of information theory, video frames with difference between each other carry more information than the video frames similar to each other, so the key frame extraction is considered more about the dissimilarity between video frames. According to the complexity of the shot content, one or more key frames from a shot can be extracted, and used to summarize the content. There are many methods to extract key frame. The first frame, the last frame or the intermediate frame can be regarded as the key frame if there is little change in a shot content. But in the case of large changes in content, extraction technology usually based on motion analysis, image information, clustering, shot activities analysis method should be accepted [8-10].

Study on the structure segmentation with different granularity characteristics of video structure has achieved great success, but how to extract multi-hierarchy semantic information object from many video resources and establish a robust multi-level semantic video information tree is an urgent research topic.

4. Multi-Hierarchy Semantic Information Descriptive Model

The traditional video compression encoding standard MPEGI/2 and H.26x adopted the technology based on image frame, which had no requirement for the segmentation of video, scene, shot, etc. This traditional technology could obtain a higher compression ratio and had been widely used in many areas. But the approach of no segmentation of video, scene, shot obviously does not support new features based on multi-hierarchy semantic information. New application calls for new multimedia encoding standard, so MPEG-4 introduces the new concept of video object (video, scene, shot, image frames) to support interactivity and scalability [11-18]. According to the MPEG-4, semantic video objects should be segmented and then hierarchical encoded afterwards.

This paper attempts to establish a reasonable and effective multi-hierarchy semantic information descriptive model based on video segmentation and extraction technology as well as MPEG-4 encoding standard to realize the mapping of video semantic information from the low-level to high-level. The model contains four layers of video structure: Video, Video Scene, Video Shot, and Video Frame. It is essentially a hierarchical semantic information tree generated from the perspective of intelligent semantics.

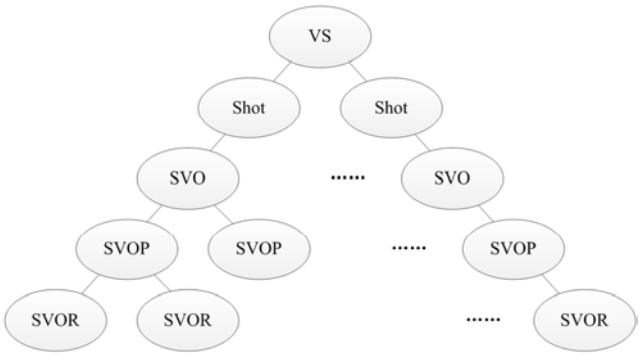

Figure 2. *Multi-hierarchy semantic information descriptive model.*

5. The Characteristics of Subtitling Translation and Its Temporal and Spatial Constraints

Caption information including the rich video semantic and dialogue subtitle, the real-time transcription of speech, play an important role in the analysis and retrieval of the original video stream. Highly summarized semantic has been provided by video caption for labeling the corresponding video stream after recognition, positioning, segmentation and extraction, then research on subtitle translation can be carried out based on hierarchical semantic video stream. The subtitle, as a special multi-symbol text, has its own characteristics, and could be restricted by some technical parameters in the process of translation.

5.1. Features of Subtitle Translation

The subtitle translation is different from other types of translation, and has a very high demand in language, culture and technology level, which has the culture characteristics of movie and television. 1) Instantaneity: it stays on the screen for the limited time, usually flashed. 2) Popularity. The subtitle is oriented to the general audience and the translation corpus is the characters of dialogue and the necessary visual information. If the subtitle translation is profound and abstruse, it will affect the audience to appreciate the picture. 3) Simplicity. Appreciation of the film, audience should not only see the subtitles, but also look at the picture; not only understand the text, but also grasp the true meaning of text images. So, the subtitle translation should be text simplified and concentrated. 4) Complementarity. The subtitles did not change the image and sound information, the caption information and acoustic information received by the audience will interact with each other, the audience can get compensation from the soundtrack if subtitles cannot be expressed or not fully expressed [19]. 5) Transfer-conformity. The subtitle translation is a special type of language conversion: from the original spoken language into a condensed written language [20].

5.2. The Temporal and Spatial Constraints of Subtitle Translation

The culture characteristics of movie and television have many special requirements and restrictions for subtitle translation. Gottlieb divided it into two kinds, one is the form (quantity) restriction, the other is a text (quality) restriction [21]. The former mainly refers to subtitle's duration time restricted by language auditory channel and language visual channel. The latter mainly refers to the spatial restriction in consideration of the subtitle's location, number and demand to meet audience's reading requirement.

6. The Application in Subtitle Translation

In the video stream the caption information, generally appeared in a continuous image frames, should last for a period of time to change, otherwise people could not see, and there will be a video interval without subtitle information between the two sections which have subtitle information in it. In particular, in the original subtitle area the plane is used tangentially with video data along the time axis according to character arrangement direction, which will form a temporal -spatial flow graph with clear fringe information, from which we can clearly see the duration of all kinds of information (Figure 3). Longitudinal difference to the temporal-spatial flow diagram, a stripe boundary can be obtained from the difference distribution. Afterwards, the start and end position about the subtitles can be calculated from the stripe's vertical direction. According to statistics, the captions usually lasts more than 6 seconds, the general dialogue subtitle will last for at least 2 seconds, which are for the PAL standard, respectively for 150 and 50 frames. Then, by using the

characteristics of primitive subtitles continuously appearing in multi frames and real-time transcription speech, we can calculate the spatial redundancy and time redundancy for subtitles, and determine the space area and time area for the corresponding translation subtitles. Subsequently, video hierarchical semantic information description model could be applied to the results of primitives subtitle to realize the mapping of video semantic information from the low-level feature to high-level. This technique is especially suitable for Subtitle amplification and Subtitle deletion.

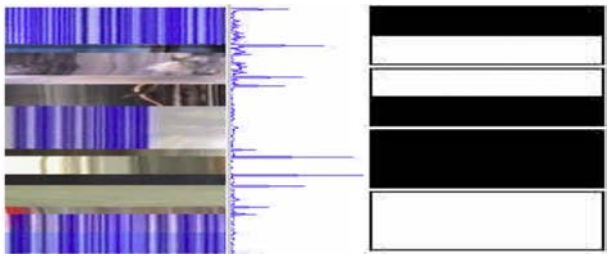

Figure 3. *Temporal -spatial flow diagram.*

6.1. Subtitle Amplification

Due to the differences in language and culture, some implicit information in language dialogue may become a semantic or cultural vacancy, which will interfere with the target audience of the film appreciation. Translators should identify and understand the hidden information and use the appropriate supplementary means to fill the gaps to ensure a deep understanding of the connotation of the film by the audience. The amplification subtitle should exist in the same duration frames with the original subtitle.

例 1 周瑜：这次你一定得考上，你看看你都超龄了。（立春）

译：This time you must pass. You are already past the age.

黄四虎：大不了我接着改户口。

译：Bull shit. I will keep changing the age on my ID card.

Huang Sihu said "hukou", as unique cultural phenomenon in China, its meaning is very broad, which can be understood as "change address", "change the name" or "change age". But according to the story, it can only be understood as "change age". Based on a thorough understanding of the original film, the translator has added the information of “keeping changing the age on my ID card”, which accurately portrayes the mental activity of Huang Sihu about keeping changing the age for many times after failures for academy of fine arts.

例 2 小妹：你不该回来。（十面埋伏）

译：You shouldn’t have come back.

金捕头：回来……为一个人。

译：I came back…*for you. My love*.

There is a dialogue between head constable Jin and his unrequited love, Xiaomei in the film of Ambush On All Sides. Everyone watching the film should know that head constable Jin’s love to her is very subtle and reserved. Even at last when Xiaomei asked him about his purpose of coming back, head constable Jin’s answer is "for a person." Although we all know that the very person is referred to Xiaomei, the target audience perhaps do not know quite clearly the probable relationship between them due to their lack of understanding about Chinese implicit expression, therefore, the hidden meaning of “for you. My love” has been added in the translation. Of course, the amplification subtitle should keep pace with the characters dialogues, actions and images involved, otherwise it will cause the delay of the information transmission.

6.2. Subtitle Deletion

We should delete the unrelated information in the limited time and space in order to highlight more correlated information. We should omit the information which is short of by recipients and cannot be added in the limited period. We also should distract the information that image and picture has provided sufficient context [12].

例 3 三年前剑法练成，飞雪执意要去刺秦，我随飞雪一同杀入秦宫。（英雄）

译：Three years ago, we perfected our skills. Flying Snow insisted on going ahead together, we stormed the palace.

It is difficult to interpret the exact meaning of “Swordsmanship”, “Emperor Qin’s Palace” which belongs to the Chinese cultural words, especially when the constraints of time and space are so limited. The translation has omitted the specific meanings of culture words which cannot be translated in the limited period by using “skills” instead of “swordsmanship”, “Palace” instead of “Emperor Qin’s Palace”, which will help the target audience cross cultural barriers and have a deep understanding of the connotation of the film.

例 4 就这么着。我还忙，那我先走了啊。（立春）

译：OK? Bye.

In order to maintain the subtitle synchronization of dialogues, actions and images, the complete sentence including the unnecessary information "I'm busy" has been deleted. Although the sentence deletion could lead to some semantic and stylistic losses, but with the information transmission from audio and visual channels, it does not affect the audience’s understanding of the plot. Instead, it speeds up the pace for audience processing the information.

7. Conclusion

This paper tries to use the existing video segmentation and extraction technology to establish reasonable and effective semantic information hierarchical descriptive model and realize the mapping of video semantic information from lower level to higher level in an attempt to further the study of film subtitle translation based on the technical level. The study shows that the model can well describe the service practice in subtitle translation in the premise of integration of video multi modal information, based on the right segmentation and extraction of video object, on account of the consideration of each video object synchronization as well as temporal-spatial constraints related issues.

Acknowledgements

Humanities and Social Sciences Foundation of Ministry of

Education (16XJAZH003); Special funding for scientific research projects of the central university (sk2016013). This thesis is the part achievements of 985 key construction disciplines of School of Foreign Languages of Xi 'an Jiaotong University.

References

[1] Y. Zhu, D. Zhou, "Video browsing and retrieval based on multimodal integration," Proceedings of IEEE/WIC International Conference on Web Intelligence, 2003: 650- 653.

[2] H. F. Wang, Z. X. Sun, "Semantic processing method of content-based image retrieval" Chinese journal of image and graphics, 2001, 6(10):945-952.

[3] J. Han, X. Zhang, W. Y. Sun, "An implementation method of automatic segmentation and tracking of video motion object" Chinese Journal of image and graphics (A), 2001, 8:732-738.

[4] N. Dimitrova, H. J. Zhang, B. Shahraray, et al., "Applications of video-content analysis and retrieval," IEEE Multimedia, 2002, 9(3):42-55.

[5] Y. P. Zhang, L. Zhang, T. Wu, "The description of different granularity---quotient space method," Chinese Journal of Computer Science, 2004, 27(3):328-333.

[6] J. Zhong, "Research on analysis technology of news video content based on multimodal information" [Ph. D. Thesis]. Tianjin: Tianjin University, 2007.

[7] X. Zhu, X. G. Lin, "Research on the method of video shot time-domain segmentation", Chinese Journal of Computer Science, 2004, 27(8): 1027-1034.

[8] Z. Y. Ye, F. Wu, Y. T. Zhuang, "A robust fusion algorithm for shot boundary detection," Journal of Computer Aided Design and Computer Graphics, 2003, 15(11):950-955.

[9] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved," IEEE Transaction on Circuits and Systems for Video Technology, 2002, 12(2):90-105.

[10] F. SH. Wang, D. Xu, W. X. Wu, "Clustering algorithm of automatic extraction of key frames based on adaptive threshold," Computer research and development, 2005, 42(10):1752- 1757.

[11] Overview of the MPEG-4 Standard[R], ISO/IECJTCI/SC29/ WG11, N1999, 3156(12).

[12] F. Gao, Y. Liu. "Acceleration of video caption retrieval algorithm based on Intel MIC," Computer engineering and Science, 2015, 37(4):634-640.

[13] Q. Sh. She, W. J. Yang, Zh. Tian, et al. "A watershed moving target detection based on motion contour marks and extraction [J],"Journal of Dalian University of Technology, 2014, 54(6):656-661.

[14] Y. Yu, M. W. Cao, F. Yue, "An improved Vibe algorithm for detecting moving objects [J]," Chinese Journal of Scientific Instrument, 2014, 35(4): 924-931.

[15] D. X. Wang, W. Ch. Xu, "Edge detection algorithm based on improved mathematical morphology [J]," Research and Exploration in Laboratory, 2014, 33(2): 89-92.

[16] Y. L. Huo, M. Qin, Zh. Qiu, "Background initialization method based on interval distribution density [J]," Computer Engineering and Science, 2015, 37(9): 67-72.

[17] J. D. Yu, X. M. Zhang, "Edge detection algorithm for lines on microscopic image," Optics and Precision Engin., 2015, 23(1): 271-281.

[18] X. Fan, Y. Cheng, Q. Fu, "Moving target detection algorithm based on Susan edge detection and frame difference," IEEE Inter. Conf. Information Science and control Engin.," 2015: 323-326.

[19] Y. X. Li, "Strategies for subtitle translation," Chinese translation, 2001, 22(4): 38-40.

[20] N. L. Birgit, "Culture-bound Problems in Subtitling [J]," Perspectives, 1993, 1(2): 207-240.

[21] Gottlieb, Henrik. Subtitling: A New University Discipline [M], Amsterdam: John Benjamins Publishing Company, 1998:80-86.