CMPUT414 Project Literature Review, Winter 2019

Group name: Segmentpie
Group members: Bowei Wang, Qingyang Zhang

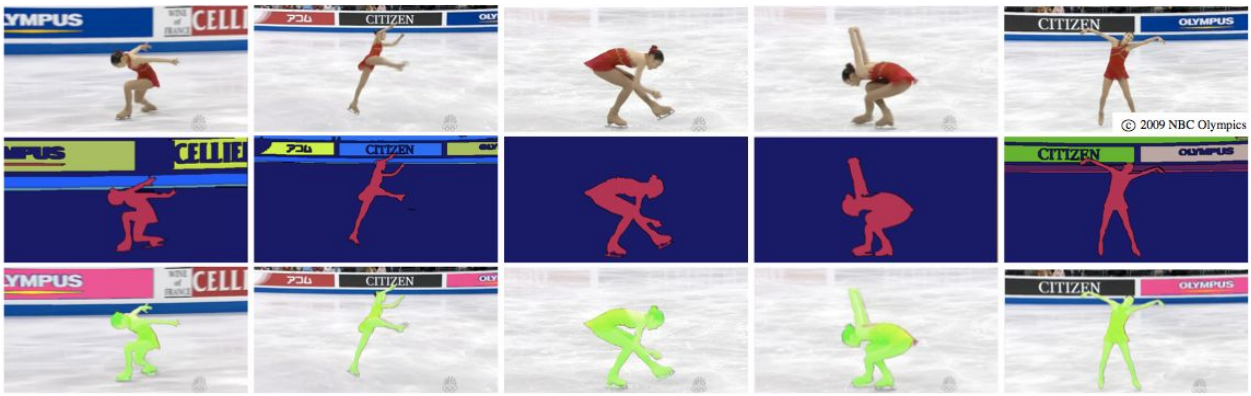
Content-based Video Segmentation review

Introduction :

Content-based video segmentation plays an important role in video compression and manipulations. [2] It is also very useful in our daily lives. Video segmentation makes it easier for users to browse video content. It also provides the user with the flexibility of accessing the video instead of watching videos in a normal sequence. Content-related functionalities is also an important request on new era's multimedia applications. That needs software segments videos into different contents correctly. [3]

There are already several methods of achieving video segmentation. The traditional method of implementing is very simple. The segmentation problem is simplified to a two-class classification problem. That is if the content of the current frame is the same as the content of previous frames or they are different. After setting a threshold, the classification will be processed automatically. The basic idea of comparing frames is comparing their color content by subtraction on three dimensions YUV.[6] Several challenging points on video segmentation are raised: Temporal coherence, Automatic processing, and Scalability. Most of these problems are perfectly solved by M. Grundmann, V. Kwatra, M. Han and I. Essa in 2010. It is regarded as one of the modern state-of-the-art video segmentation methods. However, the sensitivity to MPEG encoding artifacts shows that it still has some room for improvement. [4] There are also some other papers that are valuable to be referenced. S. Minaee and Y. Wang proposed an algorithm for separating the foreground and background. That method is very useful especially when we need to handle some videos that the background changes swiftly.[1]

Further development of content-based video segmentation is video-content analysis. The video-content analysis focusing on analyzing the emotion under the content of videos. Conventional content-based video analysis focuses on specified objects, such as news or sports[7]. Now people are paying more attention to the emotion affection of video contents. By analyzing people's feeling about the content of videos, we can collect the the most fascinating part of a movie or a sports game. Besides, video-content analyses can also improve the quality of personalizing video push. Users can get in touch with more videos that match their taste.[7]



The processing result of "Efficient hierarchical graph-based video segmentation"[4] Top: original image. Middle: Segmentation result computed in 20 min. Bottom: User-selected regions.

Review of Existing Content-based video Segmentation Algorithms:

The traditional and classical method of content-based video segmentation is proposed by B. Gunsel and A. M. Tekalp in 1998. It will trace and analyze the whole video frame by frame. Basically, it will subtract the YUV value of the current frame with the previous frames. Then, the difference between the current frame and precious frames can be figured out. After measuring the difference with a two-class classification function which has classes “same” and “different”, the video will be segmented into different parts. The difference, which is the key part of this algorithm is measured by two elements: f1 and f2. f1 represents the difference between the current frame and the last frame. f2 represents the difference between the current frame and the average of previous frames. By analyzing these two elements, whether the current frame is showing the same content with the previous frames will be figured out. [6]

Another key point of this method is the automatic threshold selection. While processing the classification, a threshold is needed to divide the frame into “same” and “different” classes. Obviously, setting a stable threshold is not acceptable. Videos are dynamic objects, so the threshold for measuring the video segments should also be a dynamic value. Otsu method is applied to solve this problem. Otsu aims to minimize the weighted sum of class variances to specify the best threshold. Therefore, the threshold is defined to be the T that can minimize the variance between classes divided by the variance inside one class. That is, the threshold will focus on the frames with different content and ignore the frames with the same content.[6]

However, there are still many disadvantages to this algorithm. Since the subtraction of two frame matrixes is processed without compression, the cost of calculation may be very high. Especially when high-resolution videos (eg. 4K videos) are handled. Secondly, directly subtraction is not efficient. It may generate many integers, instead of binary values 0 or 1. There is still much progress can be made based on this classical method.

Instead of focusing on basic content-based video segmentation, some more complex problems are raised. M. Grundmann, V. Kwatra, M. Han and I. Essa implemented a method for gathering pixels together by their continuous in 2010. This method can be used on not only simple video segmentation but also the motion analysis of videos. They had raised three challenging factors of this problem: Temporal coherence, Automatic processing, and Scalability. Temporal coherence referring that even two neighboring frames cannot be expressed by a continuous function, because the changing of frames is not continuous. Automatic processing represented that the region to be segmented is neither stable nor already known. We need to keep tracing the content that is segmented. Scalability indicated that directly segmentation on video is slow and may take much space, so it is necessary to optimization the segmentation process decrease the segmentation cost.[4]

In 1995, Hongjiang Zhang, Stephen W. Smoliar and JIan Hua Wu proposed video parsing algorithms use for video browsing tools based on content. The video parsing algorithms are compounds of video segment boundaries and abstraction information. In this system, they implemented the algorithms that make it can automatically detect boundaries based on the previous search. It points at the qualitative difference between successive frames. The first approach is by means of pair-wise comparison. Using pixel on coordinated in frames, and counts the change on the coordinates. When the change condition satisfies, it will be denoted as different content. Also, it has a formula for histogram comparison which is based on the change of RGB. The second approach is based on ranges of Hue and Chroma and makes the content into a two-dimensional histogram. This is a quantitative-based method. Although using this method works good at transitional contents, the same contents with different camera movements also are considered as different contents. For the browsing tool, it has two different methods which lead to sequential access and random access to a video.[10]

Video	Length (sec.)	N_d	N_m	N_f	N_k
Stock footage 1	451.8	35	1	1	116
Stock footage 2	1210.7	78	1	5	271
Singapore	173.8	31	1	0	71
Dance	2109.1	90	17	4	205

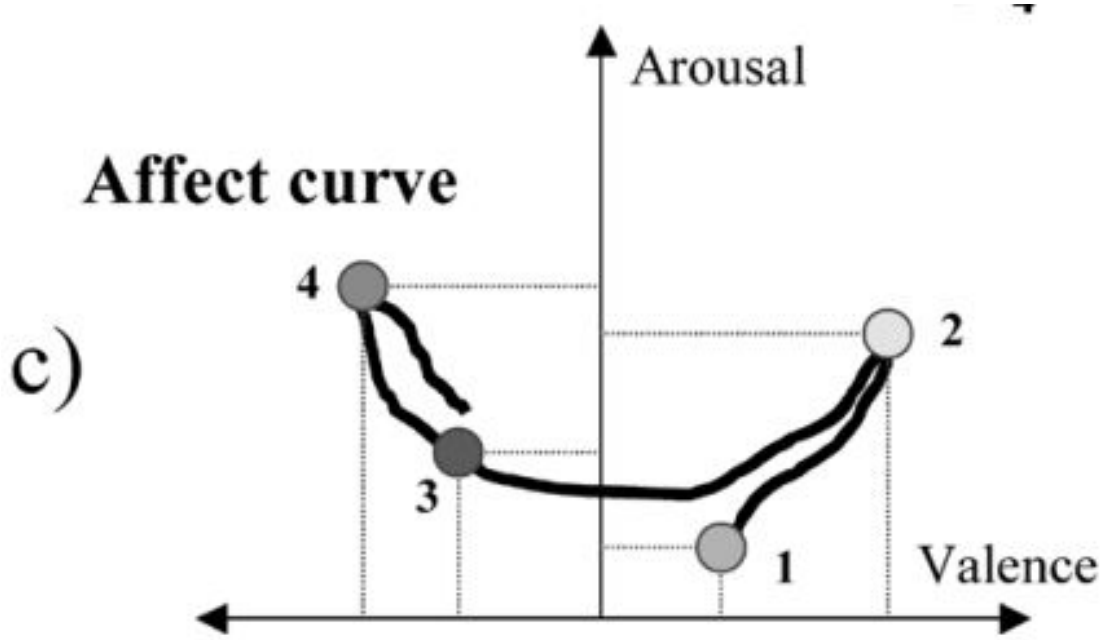
Table: “ Video segmentation and key frame extraction results: Nd: number of shots correctly detected; Nm: number of shots missed by the detection algorithm; N,< false detection of shot boundaries; Nk: number of keyframes extracted.”[10]

Jenq-Neng Hwang who is a professor from the University of Washington and Changick Kim who is from Korea Advanced Institute of Science and Technology proposed an algorithm regarding content-based segmentation and video object planes (VOPs) for extraction in 2002. Before they did this work, they found that the traditional algorithm lacked interpretation for video contents in high-level. Therefore, the new algorithm is highly efficient for MPEG-4.There three main steps concerning extraction under three different circumstances respectively. The first situation is for moving edge (ME) map. Find edge map, DEn of difference between previous and current gray level images which are two consecutive frames. Then find moving object edge, MEn of current frame n using background edge map Eb and current edge map En. It will calculate and compare the difference to generate the change. Scattered noise might be shown on MEn, if it exists, scattered noise should be withdrawn before the next processing. Finally, combine maps. The second situation is for VOP. After the first situation, it got DEn, so it satisfied to extract VOP. Find the intersection region of horizontal and vertical VOP candidates by the means of logical operator AND. The third situation is for extracting single VOP from multiple objects sense. This is much more complicated through square pixels, differences vector, average magnitude and more. It can archive the real-time VOP so it is useful for multimedia applications. This algorithm has three main advantages those are effectiveness, real-time processing, and open framework.[3]

Future development of content-based video segmentations:

Instead of using the traditional mathematical method to implement video segmentation, there are also several new technologies to solve this problem.

Simply content-based video segmentation cannot feed people's requirement now. A new approach to video content analyses on video content. A representative work is "Affective Video Content Representation and Modeling" by A. Hanjalic and Li-Qun Xu.[9] The framework of direct video affective content analysis mainly consists of two parts: video feature extraction and emotion classification/regression[7]. That corresponds to two levels in the method: Cognitive level and Affective level. Cognitive level referring to objective items, like news and weather forecast. Affective level means the factors that may affect people's emotion, like the romantic atmosphere. The cognitive part is not hard to be analyzed, so the method focuses on analyzing the effective level. Affect is measured in three dimensions Valence, Arousal and Control based on the existing study at that period. Because control is not an important factor in characterizing various emotional states, we focus on the effect of Valence and Arousal. Valence representing the type of emotion and arousal stands for the intensity of emotion. By combining the curve of arousal-time and valence-time, an affect curve can be generated.[9]



Example of affect curve, from [9]
By analyzing the affect curve, emotion of the arbitrary time of this video can be analyzed.

Machine learning is a very popular and effective method. Many machine learning algorithms can be used on video segmentation and video content analysis. Including support vector machines, multiple layer perceptions, k-nearest neighbor, hidden models, dynamic Bayesian networks, and conditional random fields. The neural network is widely used recently. However, the neural network still has a problem while handling these questions. Processing of the neural network is a black box. All the process under hidden-layer is unknown, thus people don't know it's internal working mechanism. GMM, which refers to the Gaussian mixture model is also a possible approach to solve this problem. Since GMM can be used to representing all the continuous function on a limited domain. The emotion function of a video can also be learned are represented by this method.[7]

Another possible direction of content analysis of videos is about audios. Some study shows that audio features usually contains more information than visual features on emotional analyses of video content.[7]However, comparing to other methods of video content analyze, video content analysis based on audio is more challenging because the audio can be a mixture of many different voices. Hee Lin Wang and Loong-Fah Cheong have implemented a good method on it by dividing the audio analysis with some prior.[11]

References

[1]S. Minaee and Y. Wang, "Screen content image segmentation using least absolute deviation fitting," *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, 2015, pp. 3295-3299.
doi: 10.1109/ICIP.2015.7351413

[2]D. Zhong and S. -. Chang, "Video object model and segmentation for content-based video indexing," *Proceedings of 1997 IEEE International Symposium on Circuits and Systems. Circuits and Systems in the Information Age ISCAS '97*, Hong Kong, 1997, pp. 1492-1495 vol.2.
doi: 10.1109/ISCAS.1997.622202

[3]H. Li and K. N. Ngan, "Automatic video segmentation and tracking for content-based applications," in *IEEE Communications Magazine*, vol. 45, no. 1, pp. 27-33, Jan. 2007.
doi: 10.1109/MCOM.2007.284535

[4]M. Grundmann, V. Kwatra, M. Han and I. Essa, "Efficient hierarchical graph-based video segmentation," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 2141-2148.
doi: 10.1109/CVPR.2010.5539893

[5]Naveen Shankar Nagaraja, Frank R. Schmidt, Thomas Brox; "Video Segmentation With Just a Few Strokes"The IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3235-3243

[6]B. Gunsel and A. M. Tekalp, "Content-based video abstraction," Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), Chicago, IL, USA, 1998, pp. 128-132 vol.3.
doi: 10.1109/ICIP.1998.727150

[7]S. Wang and Q. Ji, "Video Affective Content Analysis: A Survey of State-of-the-Art Methods," in *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410-430, 1 Oct.-Dec. 2015.
doi: 10.1109/TAFFC.2015.2432791

[8]Warnick, J., Ferman, A. M., Gunsel, B., Naphade, M. R., & Mehrotra, R. (2001). *U.S. Patent No. 6,195,458*. Washington, DC: U.S. Patent and Trademark Office.

[9]A. Hanjalic and Li-Qun Xu, "Affective video content representation and modeling," in *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143-154, Feb. 2005.
doi: 10.1109/TMM.2004.840618

[10]Zhang, H., Smoliar, S. W., & Wu, J. H. (1995, March). Content-based video browsing tools. In *Multimedia Computing and Networking 1995* (Vol. 2417, pp. 389-399). International Society for Optics and Photonics.

[11]Hee Lin Wang and Loong-Fah Cheong, "Affective understanding in film," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 6, pp. 689-704, June 2006.
doi: 10.1109/TCSVT.2006.873781