Video Streams Semantic Segmentation utilizing Multiple Channels with Different Time Granularity

Bado Lee, Ho-Sik Seok, Byoung-Tak Zhang*
Biointelligence Laboratory
School of Computer Sci. & Eng.
Seoul National University
Seoul 151-742, Korea
{bdlee, hsseok, btzhang}@bi.snu.ac.kr

Abstract

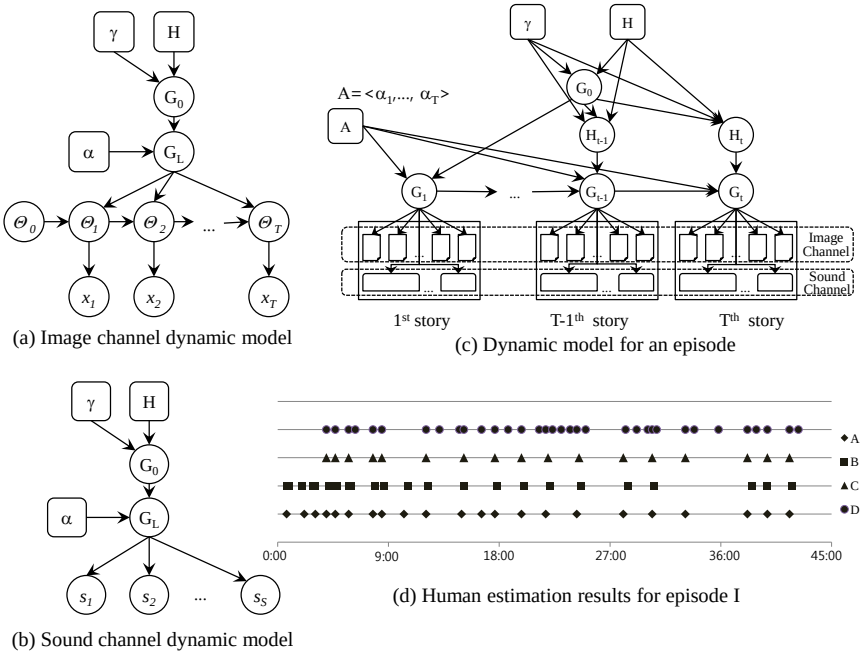
This paper deals with a semantic segmentation in video streams. The proposed method aims to detect story segments in an episode of a TV drama. For reliable story segmentation, the challenge is to build a robust model capable of capturing the underlying consistent latent states corresponding to a story segment while handling frequent changes in low-level features. We address this challenge by utilizing multiple channels inherent in a TV drama. For a given TV drama episode, it is possible to disassemble the stream into visual, sound, or textual channels. The proposed method builds a dynamic model for each modality channel and combines the resulted models. The difficulty with this approach is the difference in time granularity of each channel. In order to dissolve the difference, we introduce a hierarchical model where a common latent state generates scenes and dialogue in a story segment. Each dynamic model analyzes its own segment in the assigned channel and at a higher level the composite likelihood of a story segment change is estimated based on the each channel's estimation. We report preliminary estimation results in this paper.

1 Semantic segmenting in TV dramas and the proposed method

We discuss a semantic segmentation method for video streams. A method for partitioning a series of images in groups is already introduced in [1] and video segmentation methods using prior knowledge or repetitions in scenes are being actively researched [2]. Contrary to previous researches, our method attempts to detect inherent story segments. If a story is defined as "a topically cohesive segment of episodes that include multiple sentences and events about a single topic" (modified definition of [3]), a video stream could be interpreted as a set of stories.

The difficulties associated with semantic segmentation are as followings: (a) there dose not exist a suitable method for sematic analysis and (b) one should handle frequent changes in low-level features. In order to address these challenges, we propose a composite scheme based on the hierarchical Dirichlet process (HDP) [4]. The proposed method builds a separate dynamic model for each of the image channel and sound channel in an episode of a TV drama. An image channel dynamic model handles scene data (Fig.1 (a) and Eq. 1) and a sound channel dynamic model analyzes dialogues of each character in a video stream (Fig.1 (b) and Eq. 2). Each dynamic model is similar to the sticky HDP-HMM in [5]. The different time granularity of each channel makes analysis of video streams difficult. This difficulty is alleviated by considering likelihood of $F(x_{Lt}|G_L)$ and $F(S_{Lj}|G_L)$ (x_{Lt}: the tth scene in a story segment L, S_{Lj}: jth sound state, $F(x|\theta)$: likelihood of x given θ). When

*This work was supported by the National Research Foundation of Korea grand funded by the Korean government (MEST) (No. 2011-0016483).



(a) Image channel dynamic model

(b) Sound channel dynamic model

(c) Dynamic model for an episode

(d) Human estimation results for episode I

Figure 1: (a) A dynamic model for a group of images in an image channel. (b) A dynamic model for a set of sound states in a sound channel. (c) A composite dynamic model for story segments in an episode. (d) Human estimation results for an episode of TV drama (an American legal drama-comedy, **Boston Legal**, is used for this work).

the likelihood becomes significantly low (Eq. 3) at the same time, a story segment change point is detected. Although the resulting structure (Fig. 1 (c)) is similar to dHDP in [6], we estimates the change point based on the likelihood without assuming prior distribution. In this work, we report preliminary result (the number of estimated story segments based on various threshold).

Image channel dynamic model (Eq. 1)	Sound channel dynamic model (Eq. 2)
$G_L \vert \alpha_L, G_0 \sim DP(\alpha_L, G_0)$ $\Theta_t \vert \Theta_{t-1} \sim G_L, x_t \vert \Theta_t \sim F_{\Theta_t}$	$G_L \vert \alpha_L, G_0 \sim DP(\alpha_L, G_0)$ $S_j \vert S_{j-1} \sim G_L, S_j \sim F_{G_L}$
Composite likelihood computation (Eq. 3)	
for $t = s_0, ..., s_T$ and $j = s_0, ..., s_J$ $F(x_t, S_j \vert G_L) = (1 - I_G(\frac{F(x_t \vert G_L)}{F(x_{t-1} \vert G_L)})) + (1 - I_G(\frac{F(S_j \vert G_L)}{F(S_{j-1} \vert G_L)}))$ If $\alpha < \gamma$, $I_G(\alpha) = 1$ else $I_G(\alpha) = 0$	
Result 1: number of estimated story segments	
(low threshold) 5 $\longrightarrow$ 7 $\longrightarrow$ 15 $\longrightarrow$ 26 $\longrightarrow$ 38 (high threshold)	

2 Experimental results

- **Data and human estimations**

We generated data from an episode of TV series Boston Legal. Total play time of experimental material is approximately 42 minutes. We generated 25,443 scenes to construct the image channel using SIFT (Scale Invariant Feature Transform) method [7]. Fig. 1 (d) shows human estimation results on the experimental material. 4 Human experimenters estimated the story change points without prior information such as the number of stories or the definition of a story. From Fig. 1 (d), we can see that there does not exist a general consensus on the story structure on the given experimental material.

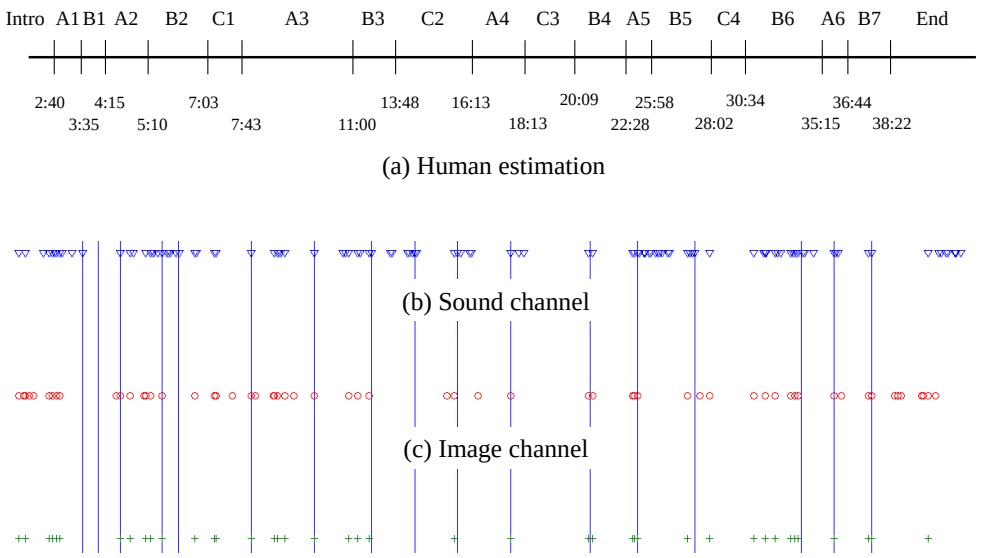


Figure 2: (a) The experimental material is composed of three stories. Story "A" is about an environmental litigation, story "B" is a sexual harassment lawsuit, and story "C" is a criminal case. In addition, there exist other segments for title role and concluding segment/ending credits. "Intro" means "Title role", "A#" is "#"th event in a story A, "B#" and "C#" is "#"th event in a story B and "#"th event in a story C. "End" means "a concluding segment and ending credit." (b) a segmentation result by the sound channel only dynamic model. (c) a segmentation result by the image channel only dynamic model. (d) a segmentation results by the composite model.

- **Experimental results and discussion**

Result 1 and Fig. 2 show experimental results. We introduce an auxiliary criteria based on a feature of color histogram in order to make advantage of the color consecutiveness in scenes. The auxiliary criteria compares color histogram of two scenes and reports the result. Prior to dynamic model implementation, the stream data is preprocessed based on the color histogram. Result 1 shows segmentation performance aided by this auxiliary criteria. Fig. 2 (a) shows the original segment structure of the experimental material. Fig. 2 (b) represents a segmenting performance of the sound channel. Fig. 2 (c) is a segmenting performance of the image channel. Fig. 2 (d) reports a segmenting performance when both of the image channel and sound channel are employed. As Fig. 2 shows, the composite method combining an image channel and a sound channel achieves more probable segmentation.

From experiments we observe the followings: (a) it is possible to approximate a semantic segment using a composite channel model, (b) it is important to obtain stable states, and (c) a single channel does not dominate during the semantic segmentation process.

Although semantical analysis of stream data is practically very difficult, it is possible to circumvent the difficulty by considering changes in visual channel and sound channel simultaneously. During approximation, each modality requires a different approach. In the case of an image channel, the approximation process is based on dominant visual features in each candidate visual segment. In the case of a sound channel, securing a stable state is very important. Contrary to an image channel, the basic computational entity of a sound channel is a recognized sentence of each character. Therefore

a sound channel dynamic model could have very diverse form depending on the recognition performance. In this work, a dynamic model for each channel detects only some kind of state changes. Semantic segments change points are included in these state change points but state change points do not constitute one-to-one correspondence with semantic change points. Because an image channel is vulnerable to camera movements and a sound channel lacks detecting capability of conversations by several characters, both channels are required for semantic segmenting.

3 Conclusion

In this paper, we explored a Bayesian nonparametric model to analyzing multimodal streams. We reported preliminary results of a video stream analysis based on the hierarchical Dirichlet process. Instead of utilizing prior knowledge on the target video streams, the proposed method employed inherent modality channels in a video stream. By introducing a hierarchical structure in which a common latent distribution generates image data and sound data for a story segment, we were able to approximate semantic changes in a stream.

The model presented herein, while enabling to explain video streams in terms of underlying variables generating each story segments, still possesses a number of limitations. In order to overcome the deviation due to camera works, we added a sound channel dynamic model. But the effect of the sound channel model is limited. For more reliable semantic segmenting, the role of textual data (a script) should be considered. Secondly, the contribution of each channel should be conditional. it would be possible to improve the segmenting performance by controlling each channel's contribution dynamically based on the reliability of a learned model. Thirdly, we considered only the image channel and sound channel in this work. However, the segmentation performance could be improved by introducing a relation channel. The co-occurrence of speakers in a conversion could be used to implement a new channel.

Acknowledgments We would like to thank Yee Whye Teh for opening up his code (http://www.gatsby.ucl.ac.uk/ ywteh/research/software.html).

References

[1] Orbanz, P., Braendle, S., & Buhmann, J.M. (2007) Bayesian Order-Adaptive Clustering for Video Segmentation. *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 334-349.

[2] Chaisorn, L., Chua, T.-S., & Lee, C.-H. (2003) A Multi-Modal Approach to Story Segmentation for New Video. *World Wide Web: Internet and Web Information Systems*, **6** (2): 187 – 208.

[3] [online]. Available: http://projects.ldc.upenn.edu/TDT/

[4] Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006) Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101** (476): 1566 – 1581.

[5] Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2011) Bayesian Nonparametric Inference of Switching Dynamic Linear Models. *IEEE Transactions on Signal Processing*, **59** (4): 1569 – 1585.

[6] Ren, L., Carin, L., &Dunson, D.B. (2008) The Dynamic Hierarchical Dirichlet Process. *International Conference on Machine Learning*.

[7] Lowe, D. G., (1999) Object Recognition from Local Scale-invariant Features, *7th IEEE International Conf. on Computer Vision*, pp. 1150 – 1157.