멀티채널 기반 드라마 동영상 의미 분절화를 위한 비모수 베이지안 방법

석호식, O 이바도, 장병탁

서울대학교 컴퓨터공학부

{hsseok, bdlee, btzhang}@bi.snu.ac.kr

Nonparametric Bayesian Approach for Multichannel based Semantic Segmentation of TV Dramas

Ho-Sik Seok, O Bado Lee, Byoung-Tak Zhang

Department of Computer Science and Engineering

Seoul National University

요 약

본 논문에서는 드라마 동영상의 의미 분절화(Semantic segmentation)를 위한 멀티 채널 기반 비모수적 베이지안 방법론을 소개한다. 기존 방법론은 매우 한정적인 특징만을 이용하여 분절화를 시도하거나 이미지 채널이나 오디오 채널과 같은 단일 채널에서만 유효한 방법론을 이용하여 데이터 분석을 시도하였기에, TV 드라마와 같이 예측할 수 없는 변화를 보여주는 스트림 데이터에 적용하기에는 어려움이 많았다. 이와 같은 단점을 극복하기 위해 우리는 주어진 동영상을 단일 모달리티의 채널로 분할한 후 각 채널 별로 분절화를 시도하고 각 채널의 분절 결과를 동적으로 결합하여 주어진 동영상에서의 의미 분절화를 근사하는 방법을 개발하였다. 제안 방법은 실제 TV 동영상의 의미 분절화에 적용되었으며 인간 평가자에 의한 의미 변화 구간과의 비교를 통해 그 성능을 확인하였다.

1. 서 론

실세계 데이터를 처리할 때는 시간적 특징을 고려하는 것이 매우 중요하다[1]. 현재 각광 받고 있는 통계적 기계학습방법들에서 시간성을 감안한 많은 방법론을 소개하고 있지만[2,3], 기존 방법에서는 숏(Shot) 및 프레임(Frame) 사이의 전이 확률 분포를 가정하거나 스트림에서 반복되는 프레임의 존재와 같은 매우 한정적인 사전 지식을 활용하기 때문에[2] TV 드라마와 같이 예측할 수 없는 변화를 보이는 데이터에 적용하는 것은 쉽지 않다.

본 논문에서는 쉽게 확보하기 어렵거나 많은 비용을 필요로 하는 태그 데이터나 한정된 패턴에 의존하지 않고 TV 드라마의 의미 분절화를 근사 하는 방법론을 소개한다. TV 드라마는 시간성과 거의 무제한에 가까운 이미지 특성, 객체, 단어, 문장 등으로 이루어진 개념적 계층구조와 같은 독특한 특성을 지니고 있다[4]. 제안 방법은 이와 같은 계층구조를 활용하는 것으로, 의미 세그먼트 근사를 위하여 이미지 채널에서의 분절 및 오디오 채널에서의 분절을 추정한 후 각 채널의 분절 구조를 동적으로 통합하여 전체 동영상에서의 의미 변화를 추정한다. 이전 연구에서는 사전 확률 분포를 가정하여 스트림 내부에서의 의존성을 근사하고 분절화의 확률적 모델을 제시하였다[5, 6]. 그러나 의미적 분절화가 사전 분포에 의해서 전적으로 설명될 수 있다고 가정하는 것은 현실적이지 못하기에 우리는 새로운 근사 방법론을 개발하였다.

우리는 사전 분포를 한정한 후 연관된 인자를 추정하는 방법이 아니라, 비모수적 접근법을 통해 주어진 데이터를 분석하는 방법을 사용한다. 제안 방법은 특히 HDP (Hierarchical Dirichlet process)[7]를 이미지 채널 분석에, MFCC (Mel-frequency cepstral coefficients) 방법을 사운드 채널 분석에 적용하였다.

제안 방법은 실제 TV 드라마 [1] 에피소드에 적용되었으며 인간 평가자가 수행한 의미 변화 평가 결과와 비교하여 그 성능을 확인하였다.

2. 제안 방법

그림 1과 알고리즘 1에서 제안 방법을 설명하였다. 우리는 TV 드라마에서 관찰되는 연속 프레임 및 프레임 그룹을 설명할 수 있도록 각 모달리티(Modality)에서의 연속성 및 주어진 프레임이 현재 모델에 의해 생성될 우도(Likelihood)를 활용하였다. 사운드 채널에는 MFCC 방법론을 적용하였다. 본 논문에서는 화자 인식까지 활용하지는 않았으며, MFCC 특성을 이용하여 대화 구간과 무음 구간을 구분하여 사운드 채널을 분절화하였다.

[1] 미국의 법률 드라마-코메디인 Boston Legal을 활용

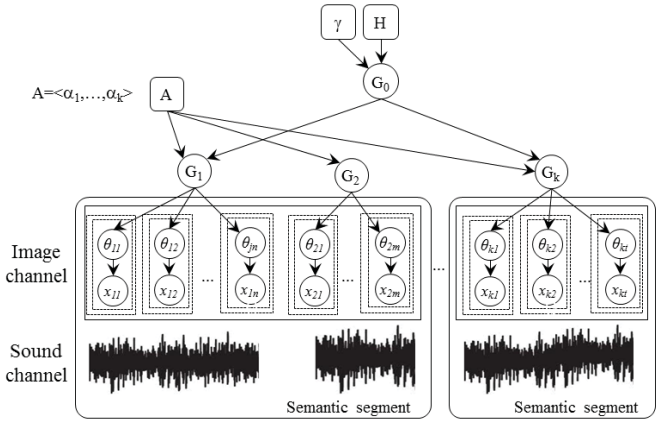

그림 1 이미지 은닉변수 모델의 분절화 결과와 사운드 채널 분절화 결과를 동적으로 결합하여 변환점 추정

HDP 방법론은 데이터 그룹을 설명하기 위한 비모수적 방법으로 큰 가능성을 지니고 있지만 급격하며 빈번한 변화에는 약점을 갖고 있다[8]. 이와 같은 변화에 대응하기 위하여 우리는 컬러 히스토그램(Color histogram)을 이용하여 이미지 구간의 후보를 생성한 후 해당 구간에 대하여 HDP 모델을 구축하는 방법을 고안하였다. HDP 모델은 다음과 같이 표현된다[7].

$$G_0|\gamma, H \sim DP(\gamma, H)$$
$$G_L|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$$

j 번째 세그먼트의 i 번째 프레임, y_{ji} 를 생성하는 인자를 θ라고 표시하면 아래 식으로 설명할 수 있다. [9].

$$y_{ji}|\{\theta_k\}_{k=1}^{\infty}, \{k_{jt}\}_{t=1}^{\infty}, t_{ji} \sim F\left(\theta_{k_{jt_{ji}}}\right) \dots (6)$$

$$t_{ji}|\pi_j \sim \tilde{\pi}_j$$
$$k_{jt}|\beta \sim \beta$$

(6)에서 t_{ji}와 k_{jt}는 표식 역할을 하는 변수이며 π는 혼합 가중치(Mixture weight)로 Dirichlet 분포를 따르고 β는 프레임을 구성하는 비주얼 워드의 활용도 표시 인자이고 $F(\cdot)$ 는 사전에 가정한 분포함수이다. 각 인자의 분포는 다음을 따른다[9].

$$\mathrm{p}(t_{ji}|t_{j1}, \dots, t_{ji-1}, \alpha) \propto \sum_{t=1}^{T_j} \tilde{n}_{jt}\delta(t_{ji}, t) + \alpha\delta(t_{ji}, T_j + 1)$$

$$\mathrm{p}(k_{jt}|\underline{k}_1, \underline{k}_2, \dots, \underline{k}_{j-1}, k_{j1}, \dots, k_{jt-1}, \gamma) \propto \sum_{k=1}^{K} m_{\cdot k}\delta(k_{jt}, k) + \gamma\delta(k_{jt}, K+1)$$

$$m_{\cdot k} = \sum_j m_{jk}, \underline{k}_j = \{k_{j1}, \dots, k_{jT_j}\}$$

$\tilde{\boldsymbol{n}}_{\boldsymbol{jt}}$는 특정 그룹에 할당된 비주얼 워드의 수를 의미하며, $\boldsymbol{m}_{\boldsymbol{jk}}$는 세그먼트에서 인식자 $\boldsymbol{k}$를 부여 받은 그룹의 수이다. $\mathbf{K}$는 총 식별자 수, $\boldsymbol{T}_{\boldsymbol{j}}$는 이미지 그룹의 총 수를 표시한다. 또한 $\boldsymbol{\delta}(\cdot,\cdot)$는 Kronecker delta를 의미한다.

표 1. 알고리즘

표기법

- $g(\tau_j|y_{s:t-1}, y_t, G_L)$: 비디오 스트림 $y_{s:t-1}$을 관찰한 상태이며 현재 세그먼트 모델이 G_L일 때 현재 프레임 y_t에서 의미 변화가 발생할 가능성 계산 함수
- $f(y_t, D_k|G_L)$: 프레임 y_t와 대화 구간 D_k가 G_L에 의해 생성되었을 가능성 계산 함수
- $f(y_t|G_L)$: 프레인 y_t가 G_L에 의해 생성되었을 가능성 계산 함수
- $G_{L,I}$: 현재 세그먼트에 대한 이미지 채널 모델 의미
- $G_{L,S}$: 현재 세그먼트에 대한 사운드 채널 모델 의미

입력: $y_{1:t}$ → 1~t번째 프레임으로 구성된 스트림
출력: m개의 변환점 $\tau_1, \dots, \tau_m$
가정: τ_1 에서 τ_{j-1} 까지의 변환점이 예측되었으며 τ_{j-1}이후의 프레임 집합 $y_{s:t-1}$를 관찰

$$g(\tau_j|y_{s:t}, D_k, G_L) = g(\tau_j|y_{s:t-1}, D_k, y_t, G_L) \propto R(y_t, D_k|G_L) \dots (1)$$

시작

새로운 프레임 y_t 및 대화 인터벌 D_k에 대하여

$$R(y_t, D_k|G_L) = \omega \cdot \Delta L_t + (1-\omega) \cdot I(D_k) \dots (2)$$

$$\Delta L_t = \begin{cases} 1 \; if \; F(x_t|G_L) - F(x_{t-1}|G_L) < 0 \\ 0 \; if \; F(x_t|G_L) - F(x_{t-1}|G_L) > 0 \end{cases} \cdots (3)$$

$$I(s_J) = \begin{cases} 0 \text{ 대화} \\ 1 \text{ 무음} \end{cases} \cdots (4)$$

만약 $R(y_t, D_k|G_L) < \vartheta$이면 τ_k를 출력하고 새로운 모델 G_{L+1}을 구축
만약 $R(y_t, D_k|G_L) > \vartheta$이면 현재 모델 G_L 강화

의미적 분절화를 위한 핵심은 이미지 채널에서의 분절화 지점과 사운드 채널에서의 분절화 지점의 결합이다. 현재 세그먼트의 은닉 변수 모델이 주어졌을 경우 현재 관찰한 프레임 y_t와 대화구간 D_k에 기반하여 변환점 추정을 대체할 수 있다. 자세한 결합 방법은 식(2)~식(4)에서 설명하였다. 실험 데이터로부터 우리는 현재 모델 G_L이 t번째 프레임 x_t와 $t-1$번째 프레임 x_{t-1}을 설명하는 우도의 차이가 의미 구간 설명에 매우 중요한 역할을 한다는 것을 발견하였다. 이와 같은 관찰을 반영하기 위해 각 프레임에 대한 우도 차와 대화 유무 여부를 고려하여 채널 별 분절화 결과를 통합하였으며 식 (1)의 계산 결과에 실험적으로 파악한 기준점을 적용하여 의미 구간을 근사하였다.

3. 실험 결과

19명의 평가자(학부생 15명, 대학원생 4명)에게 실험에 사용한 에피소드 3개에 대한 의미 구간 변화를 판정하도록 의뢰하였다(표 2). 대부분의 인간 평가 결과가 유사했으나, 정확한 변환시점에서는 일치하지

않기 때문에 인간 평가자의 변환 지점을 하나의 구간으로 변환한 후, 제안 방법론의 추정지점이 변환된 구간에 속할 경우 정확한 추정점으로 판단하였다.

표 2. 인간 평가자의 평가 결과

	에피소드 1	에피소드 2	에피소드 3
변화 횟수	156	186	176
인터벌 개수	43	39	39

Precision, Recall 측도로 성능을 측정하였다.

$$\text{Precision} = \frac{\#(\textit{정확하게 추정된 변환점})}{\#(\textit{총 변환점의 수})}$$

$$\text{Recall} = \frac{\#(\textit{정확하게 추정된 변환점})}{\#(\textit{인간평가자의 총 변환 구간 수})}$$

표 3. Precision과 Recall

Precision이 최고 성능일 때			
	에피소드 1	에피소드 2	에피소드 3
변환점 수	8	6	9
Precision	0.62	0.38	0.82
Recall	0.15	0.13	0.20
Recall이 최고 성능일 때			
	에피소드 1	에피소드 2	에피소드 3
변환점 수	52	45	41
Precision	0.059	0.060	0.057
Recall	0.96	0.96	0.93

표 3에서 제안 방법의 성능을 Precision 및 Recall 측면에서 분석하였다. Precision이 가장 우수할 때 제안 방법은 각 에피소드에 대하여 각각 0.62, 0.38, 0.82의 Precision을 달성하였고, Recall이 가장 우수할 때 제안 방법은 각 에피소드에 대하여 각각 0.96, 0.96, 0.93의 Recall을 해당 경우에서 0.059, 0.060, 0.057의 Precision을 달성하였다.

4. 결론

본 논문에서는 비디오 스트림의 의미적 변화를 근사하기 위하여 이미지 채널 분절화 결과와 사운드 채널 분절화 결과를 동적으로 통합하는 방법을 제안하였다. 각 의미 구간(Scene) 사이의 전이 확률을 학습하려고 노력하는 대신 우리는 유사한 비주얼 워드가 존재하면서 동시에 대화가 진행되는 구간을 선택하여 의미 변화를 추정하는 방법을 제안하였다. 기본 방법과 비교했을 때 제안 방법은 비용이 많이 소모되는 사전 태깅 혹은 의미 사전을 요구하지 않으며, 급격한 변화에 강건하기 때문에 더욱 양호한 분절화 결과를 달성할 수 있다.

제안 방법을 이용하면 각 의미구간에 대한 모달리티 모델을 설정하여 주어진 스트림을 설명하는 새로운 인식자를 생성할 수 있다. 우리는 추후 연구에서 새로운 인식자를 생성한 후 생성된 결과를 비디오 추천에 활용하는 방법에 대하여 연구할 것이다.

감사의 글

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(No. 2012-0005643, Videodome), 정부(지식경제부)의 재원으로 한국산업기술평가관리원의 지원(10035348, mLife) 및 교육과학기술부의 BK21-IT 프로그램에서 일부 지원되었음.

참고문헌

[1] X. Wang and A. McCallum, “Topics over time: A non-markov continuous-time model of topical trends”, *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 424-433, 2006.

[2] J.-P. Poli, “An automatic television stream structuring system for television archives holders”, *Mulitimedia Systems*, Vol. 14, pp. 255-275, 2008.

[3] X. Wei, J. Sun, and X. Wang, “Dynamic mixture models for multiple time series”, *20th International Joint Conference on Artificial Intelligence*, pp. 2909-2914, 2007.

[4] A. Mittal, “An overview of multimedia content-based retrieval strategies”, *Informatica*, Vol. 30, pp. 347-356, 2006.

[5] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, “Temporal video segmentation to scenes using high-level audiovisual features”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 8, pp. 1163-1177, 2011.

[6] V. Parshyn and L. Chen, “Video segmentation into scenes using stochastic modeling”, Technical report, Lab. LIRIS, Ecole Centrale de Lyon, 2006.

[7] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes”, Technical report, Dept. of Computer Science, National University of Singapore, 2005.

[8] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “An HDP-HMM for Systems with State Persistence”, *25th International Conference on Machine Learning*, pp. 312-319, 2008.

[9] E. B. Fox, “Bayesian Nonparametric Learning of Complex Dynamical Phenomena”, PhD thesis, Massachusetts Institute of Technology, 2009.