드라마 동영상의 스토리 분석을 위한 계층적 은닉변수 모델

이바도O, 석호식, 장병탁
서울대학교 컴퓨터공학부
{bdlee, hsseok, btzhang}@bi.snu.ac.kr

Hierarchical Latent Variable Models for Story Analysis of TV Dramas

Bado LeeO, Ho-Sik Seok, Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University

요 약

통계기법을 이용한 기계학습 연구가 활발히 진행되면서 시간 정보가 포함된 동적 스트림(stream) 분석에 기계학습 기법을 적용하려는 시도가 주목 받고 있다. 특히 이미지, 텍스트, 음성 등 다양한 특성이 결합된 멀티모달 동영상을 지능적으로 분석하여 스토리 구간을 유추하려는 시도가 다양한 연구자에 의해 진행되어왔다. 그러나 기존 연구는 동일 이미지 반복이라는 사전 지식을 이용하여 이미지 구간을 분리하였으며 각 스토리 구간을 특정 짓는 이미지/텍스트의 분포가 뚜렷하게 구분되는 데이터를 대상으로 연구를 진행했기 때문에, 다양한 동영상 데이터에 적용하기에는 적합하지 않았다. 본 논문에서는 이미지 반복 등의 사전 지식을 이용하지 않고 비디오 스트림을 설명할 수 있는 생성 모델(Generative Model)을 구성한 후 구성된 모델이 관찰한 장면(scene)을 만들어 낼 수 있는 가능성(Likelihood)에 기반 하여 주어진 드라마의 스토리 구간을 추정할 수 있는 방법을 소개한다. 제안 방법은 드라마 스트림에 존재하는 스토리 구간의 은닉 구조(Latent Structure)를 가정한 후 해당 구간의 이미지 및 텍스트 분포를 추정하는 방법으로 첫째, 기존 방법과 달리 스토리 구간 분포를 가정하지 않고도 여러 개의 스토리 구간을 구분할 수 있으며, 둘째, 분석 대상 드라마 스트림이 온라인으로 입력되는 상황을 처리할 수 있고, 셋째, 분석 대상 데이터에 대한 사전 지식이 필요하지 않다는 장점이 있다. 본 논문에서는 특히 각 스토리 구간의 은닉 구조 설명 모수 비교가 어렵다는 난점 해결을 위해 스토리 구간의 은닉 구조가 주어졌을 때 데이터를 설명할 수 있는 가능성을 계산하는 방법을 이용하여 스토리 구간을 추정하는 방법을 제안한다. 토픽이 스팬(span)하는 공간이 스토리를 설명한다는 가정 하에 새로운 데이터를 설명하는 스토리 공간이 기존의 공간과 같지 않으면 스토리가 변화했다고 가정하였다. 본 논문에서는 드라마 동영상에 제안 방법을 적용하여 획득된 추정 결과를 인간 실험자의 스토리 구분 결과와 비교하여 제안 방법의 성능을 실험적으로 제시하였다.

1. 서론

우리가 접하는 세상은 시간성을 띄고 있는 대규모 데이터로 이루어져 있기 때문에[1] 실제 발생하는 데이터를 이해할 수 있으려면 시간성을 띄고 있는 데이터를 지능적으로 처리할 수 있는 방법이 필요하다. TDT (Topic Detection and Tracking) 연구 그룹에서는 이미 동영상 데이터를 분석하여 스토리를 구분하고 주제를 감지하는 연구를 진행하여 왔다[2]. 그러나 기존 연구자들[3, 4]은 사전에 구성해 놓은 이미지 인식을 이용하거나, 동일한 이미지가 반복되고 스토리에 따라 등장하는 단어가 확연히 구분되는 데이터를 대상으로 연구를 진행했기 때문에, 데이터에 잠재된 시간 정보를 감안하여 동적으로 데이터를 분석하기에는 한계가 있었다.

데이터에 대한 사전 지식을 이용하지 않고 시간 변화에 따른 데이터 구조(분포) 변화를 추정할 수 있도록 X. Wang 등은 TOT (Topics over Time) 모델을 소개하였으며[5], X. Wei 등은 데이터의 시간 정보를 고려한 동적 혼합 모델 DMM (dynamic mixture model)을 소개하였다[6]. 기존 연구 방법들은 시간성을 고려한 동적 혼합 모델을 생성할 수 있지만 비디오 스트림과 같은 데이터 설명에는 아직 어려움이 많다. 예를 들어 D. Blei 등의 DTM (Dynamic Topic Modeling)[8] 방법의 경우 순차적인 데이터가 군집화 되어있기 때문에 진정한 동적 모델이라고 하기 어렵다. 군집 내에서의 문서들은 서로 그 순서가 교환가능 하기 때문에(exchangeable) 문서간의 시간적인 변화를 나타내기에 부족하다. 또한 토픽 공간의 시간적인 연결이 정규적으로 진행되기 때문에 큰 변화를 나타내지 못한다. DMM (Dynamic Mixture Model)의 경우 각각의 문서를 개별화 하지만 토픽을 고정함으로써 그 표현력이 현저히 떨어지고, 시멘틱의 변화를 추정하려는 우리 문제에 적합하지 않다.

본 논문에서는 순차 데이터 중에서도 특히 다양한 모달리티로 구성된 동영상 데이터를 처리할 수 있는 동적 혼합 모델을 제안한다. 제안 모델은 DTM (Dynamic Topic Model)에 기반한 것으로 Topic Mixture 모델은 주어진 데이터의 은닉 인자의 형태를 정확히 알 수 없을 때 특히 유용하다.

본 논문에서 제안하는 모델은 각 시점에서 관찰되는 이미지 정보 및 텍스트 정보의 변화와 현재 스토리 구간에 대하여 생성된 스토리 모델이 현 시점의 데이터를 생성할 우도(Likelihood)를 모두 감안하여 스토리 변화를 추정하는 방법으로 본 논문에서는 간략화 된 추론 모델을 이용하여 스토리 구간 모델을 생성하고 우도를 계산한다. 스토리

구간을 구분하기 위해 연속된 구간의 은닉 구조를 설명하는 인자를 직접 비교하는 방법은 여러 가지 어려움이 있다. 우리는 이런 어려움을 해결하기 위하여 은닉 구조 설명 인자를 비교하지 않고, 은닉 구조가 새로운 데이터를 생성할 수 있는 가능성을 계산하여 스토리 구간을 추정하는 방법을 제안하였다. 제안 방법론을 미국 드라마 동영상 파일에 적용하여 해당 파일에 존재하는 스토리 변환 지점을 추정하였으며, 인간 실험자의 스토리 변환 지점 판단 결과와 비교하여 제안 방법의 성능을 확인하였다.

2. 스토리 변환시점 추정

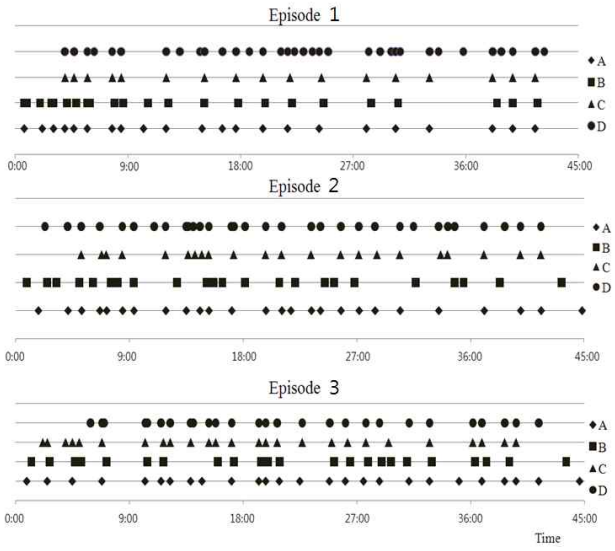

그림 1 인간 실험자에 의한 스토리 지속 구간 추정 결과. 4명의 피실험자가 3개 에피소드에 대하여 실험.

표 1 인간 실험자가 판단한 스토리 변환점의 개수

	Episode 1	Episode 2	Episode 3
실험자 A	23	26	26
실험자 B	24	23	23
실험자 C	17	22	26
실험자 D	34	33	27

TDT Community에서 스토리는 “a topically cohesive segment of news that includes two or more DECLARATIVE independent clauses about a single event” 라고 정의된다[9]. 본 논문에서는 스토리에 대한 인간 실험자의 개념을 제한하지 않기 위하여 스토리의 정의를 사전에 인간 실험자에게 제공하지 않고 주어진 동영상[1]에서 한 스토리가 다른 스토리로 변환되는 시점을 판단하도록 요청하였다. 인간 실험자의 판단 결과는 표 1과 2 및 그림 1에 정리되어 있다.

그림 1과 표1에서 알 수 있듯 동일한 동영상임에도 불구하고 스토리 변환 시점 판단 결과가 매우 다양하며, 표 2에서는 모든 인간 실험자가 공통되게 판단한 변환 시점의 수도 많지 않다는 사실을 알려준다. 본 논문에서는 인간 실험자가 판단한 모든 변환 지점을 유효한 스토리 변환 시점으로 간주하고 해당 변환 시점을 제안 방법론을 통해 추정하였다.

표 2 인간실험자가 공통되게 판단한 스토리 변환점의 수

	공통점 수 2	공통점 수 3	공통점 수 4
Episode 1	5	10	6
Episode 2	10	19	0
Episode 3	10	10	5

3. 스토리 지속구간 추정 방법 및 결과

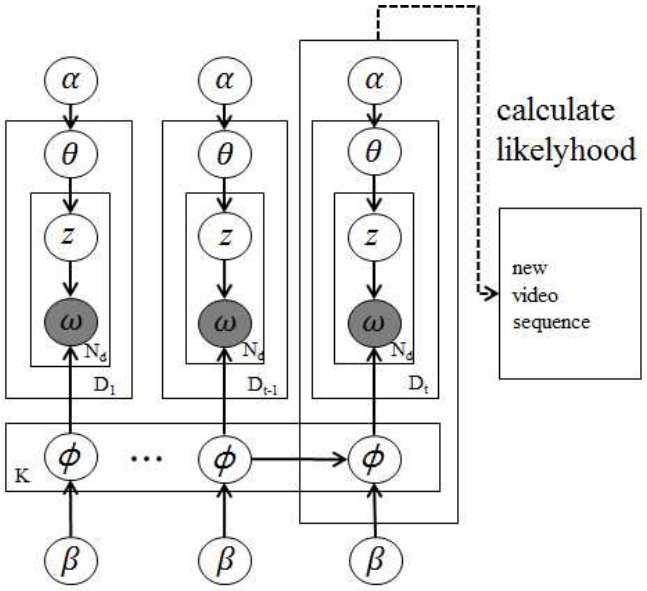

그림 2 제안 방법론의 도시. 데이터 변화 및 데이터 생성 우도(likelihood)에서의 유의미한 변화 발생 여부 추정 각 스토리 에 대한 모델. θ: 각 이미지에 대한 토픽 혼합 결정 인자, ω: 시점 t에서 관찰한 이미지+텍스트, α: Dirichlet 분포 초기화 결정 인자, ϕ: 각 토픽의 워드 분포, K: 토픽의 수, D: 이미지 개수, N_d: 각 이미지 안에 포함된 시각 단어의 개수, z: 토픽

그림 2에서 제안 방법론을 도시하였다. 제안 방법에서는 관찰 데이터 X의 집합으로 스토리를 구성하고 해당 스토리 구간을 설명할 수 있는 확률 모델을 토픽 모델에 기반 하여 추정한다. 추정된 확률 모델은 현재시점 t에서 관찰한 데이터 생성 가능성(우도) 계산에 사용하는데, 제안 방법에서는 각 특성 그룹(토픽)이 스토리의 기저 벡터를 형성한다고 가정하였다. 새로운 데이터가 들어왔을 때 이 데이터가 기존 데이터의 스토리의 연장선상에 있다고 하면, 이 데이터는 기존의 파라미터로써 표현 될 수 있어야 한다. 각각의 토픽의 경우 시간 구간 마다 그 형태가 변하게 되는데, 이의 변화를

1) 본 논문에서는 20세기 폭스 텔레비전이 ABC를 위해 제작한 미국의 법률 드라마인 보스턴 리걸(Boston Legal)의 에피소드 3개(시즌 1의 에피소드 1, 2, 3)를 실험용 동영상으로 사용하였다.

감지하는 것은 두 가지 이유에서 불가능 하다. 첫째, 토픽 모델은 확률적인 모델로써 매번 추정 할 때마다 그 형태가 달라질 가능성이 있다. 둘째, 형태를 고정하는 것이 가능하다고 하더라도 각 시간 구간에서의 토픽들을 서로 비교하는 것이 불가능 하다. 이와 같은 난점을 해결하기 위해 제안 방법에서는 이제까지 학습한 토픽 모델에 기반 하여 새로 관측된 장면(Scene)의 우도를 계산함으로써 스토리 공간의 변화를 추정한다. 이 때, 제안 모델에서 사용하는 토픽은 사전에 정의한 토픽이 아니라 특성의 결합을 통해 실험 과정에서 추정되는 특성 집합을 의미하며 LDA (Latent Dirichlet Allocation)[10]모델에서의 토픽과 같다.

본 논문에서는 에피소드 1에 대하여 수행한 스토리 구간 변화 추정 결과를 소개한다. 에피소드 1은 총 42분의 동영상으로써 초당 10장의 이미지를 샘플링 하여 총 25443장의 연속된 이미지로 변환하였다. SIFT (Scale Invariant Feature Transform)[11]을 이용하여 각각의 이미지에서 시각 특성을 추출하였고, 추출된 시각 특성으로 1000개의 시각 단어를 구성하여 각각의 이미지를 시각 단어의 히스토그램으로 변환 하였다. 각각의 이미지는 토픽 모델에서의 문서(Document) 에 해당한다. 토픽 모델 기법을 사용하려면 문서를 문서군 으로 다룰 필요가 있는데 다음과 같은 가정을 사용하여 전체 이미지를 군집화 하였다. 첫째, 연속된 이미지의 칼라 히스토그램의 거리 값이 적을 경우 연속된 이미지로 판단한다. 둘째, 첫째 가정에 의해 군집화 된 이미지의 경우 한 공간 안에서 카메라촬영 때문에 ABA'와 같이 곧바로 같은 이미지로 돌아오는 상황이 빈번히 발생한다. 따라서 A의 마지막 이미지와 A'의 시작 이미지의 칼라 히스토그램의 거리 차가 적을 경우 ABA'를 하나의 새로운 군집 A로 재정의 하였다. 첫 번째 가정으로 25443개의 후보 구간을 518개로 줄일 수 있고, 두 번째 가정으로 다시 153개의 후보 구간으로 축소할 수 있다.

현재 관찰데이터 X_t는 이미지 데이터와 텍스트 데이터로 구성된 복합 데이터로 이미지 데이터는 현재 시점에서 동영상 데이터의 스크린샷(Screen shot)이며 텍스트 데이터는 현재 시점 해당하는 대사에 해당한다. 이미지 데이터와 텍스트 데이터는 각 모달리티에 해당하는 토큰(Token)으로 구성되며 인자 θ_t를 이용하여 현재 시점 t에 해당하는 복합 데이터를 구성하는 토큰의 구성을 설명한다.

X. Wei 등은 이런 상황에서 연속 데이터의 동적 혼합 모델(Dynamic Mixture Model) 구성에 사용할 수 있는 DMM (Dynamic Mixture Model)을 제안하였다[6]. 그러나 DMM에서는 θ_t를 θ_{t-1}에 기반 하여 추정하기 때문에 연속된 스토리 구간에 대한 생성 모델을 구성할 수는 있지만 본 논문에서 의도하는 스토리 변환 판단에는 사용할 수 없다. 표 3에서 스토리 변환 판단을 위해 제안된 방법론을 설명한다.

표 3. 스토리 지속구간 추정

- 입력: 전처리된 멀티 모달 스트림 데이터(드라마 동영상)
- 출력: 스토리 변화 발생 추정점

1. 초기화
   - 새로운 스토리 시작점에 대하여
2. 군집 t에서의 혼합 분포(mixture distribution) 인자 추정: 각 군집에서 서로 다른 ϕ값을 가진다.
   - $\hat{\phi}_{z,w} = \frac{m_{z,w} + \beta_w}{\sum_{w=1}^{V} m_{z,w} + \beta_w}$ ...(1)
3. 은닉 컴퍼넌트 혼합 분포 인자 추정
   - $\hat{\theta}_{t,z} = \frac{n_{t,z} + \alpha}{\sum_{z'=1}^{K} n_{t,z'} + \alpha}$ ...(2)
4. 새로운 스토리 시작점 추정
   - $P(D_{t+1}|\phi_t, \theta_t)$ ≒ 군집 t에서 다음 스탭인 t+1의 단어를 생성할 우도. D_t의 크기와 D_{t+1}크기의 곱으로 정규화한다. 우도의 크기에 따라서 분석 구간의 크기가 달라지게 된다.

□ 표기

$\hat{\theta}_{t,z}$:	시점 t에서 관찰된 데이터X_t의 토큰 혼합 분포 인자
$m_{z,w}$:	은닉 컴퍼넌트 z에서 관찰되는 토큰 w의 수
$n_{t,z}$:	X_t의 토큰 중 은닉 컴퍼넌트 z에 속한 토큰의 수
α :	토큰 혼합분포 인자의 하이퍼인자
β_w:	해당 스토리구간의 하이퍼인자로 해당 이미지에서의 집중도 결정
ϕ:	토픽의 단어 분포
V:	토큰 집합의 크기
K:	은닉 컴퍼넌트 집합의 크기

수식 (1)에서 $m_{z,w}$는 토픽 z에 할당된 단어 w의 수이다. 깁스 샘플링을 통하여 반복적으로 ϕ 값을 갱신할 수 있다. 수식 (2)는 각 토픽에 대한 단어의 분포 확률로써, $n_{t,z}$는 문서 t 내에서의 토픽 z에 할당된 토큰의 수이다. 이 또한 깁스 샘플링을 통하여 추정 될 수 있다.

본 논문에서 제안한 방법은 은닉 컴포넌트에서 관찰되는 토큰의 수와 데이터 X_t의 토큰이 속한 은닉 컴포넌트 정보를 이용하여 획득된 은닉 컴포넌트들의 혼합비를 이용하여 스토리 구간을 설명하는 모델을 생성한다.

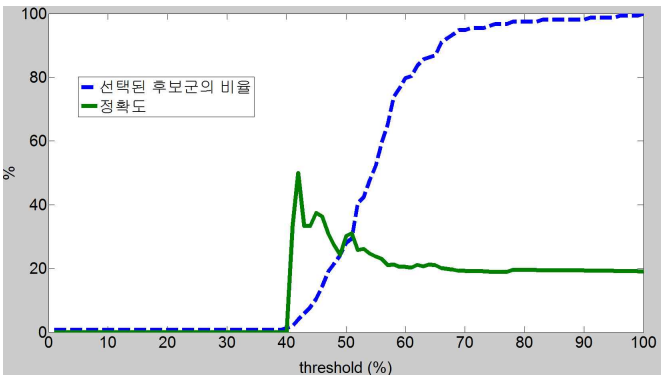

그림 3 우도 기준 값에 따른 후보군 선택 비율과 그 정확도 그래프. 정확도는 ±1.5초의 구간에서 겹칠 경우 참으로 간주

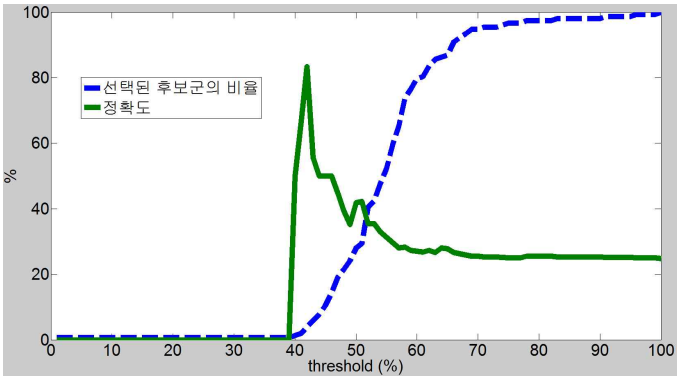

그림 4 우도 기준 값에 따른 후보군 선택 비율과 그 정확도 그래프. 정확도는 ±2초의 구간에서 겹칠 경우 참으로 간주

그림 3과 그림 4에서 스토리 변환 추정 결과를 제시하였다.

$$\text{선택된 후보군의 비율} = \frac{\#(\text{모델 적용 후의 후보군})}{\#(\text{모델 적용 전의 후보군})}$$

$$\text{정확도} = \frac{\#(\text{인간 실험자 결과와 일치하는 후보군})}{\#(\text{모델 적용 후의 후보군})}$$

우도 기준값(threshold) : 각 군집이 다음 군집을 생성할 수 있는지 판단하는 기준

사람이 판별한 구간을 모두 참값으로 가정하고 모델을 통하여 찾은 값이 이 시점과 ±1.5초 또는 ±2초 안에 일치할 경우 값을 찾았다고 판별 하였다. 각 그림에서 점선은 선택된 후보군의 비율로써, 전처리를 통하여 뽑아낸 153개의 후보군 중에서 우도의 기준 값을 변화 시키면서 찾아진 후보군이 어떤 비율로 생성되는지를 나타낸다. 기준 값이 작을 때 (0%~40%) 는 각 군집별로 다음 군집을 생성할 우도가 매우 작더라도 무조건 전 군집으로부터 다음 군집이 생성 되었다고 판별하기 때문에 후보군이 나오지 않게 된다(모든 군집이 하나로 군집화 되고, 스토리 구간을 나누지 못하게 된다). 실선은 정확도로써, 모델에 의해 선정된 후보군 중 실제 인간 실험자의 결과와 일치하는 후보군의 비율이다. 우도 기준값 을 늘리기 시작하면, 뽑힌 후보군의 크기는 점점 커지기 시작하지만, 뽑힌 값의 정확도는(실선) 점점 낮아지게 된다.

4. 결론 및 토의

본 논문에서는 동영상 에피소드를 구성하는 각 스토리 구간을 Dirichlet 분포에 기반하여 모델링할 수 있다는 가정 하에 스토리 변환 구간을 추정하는 방법론을 제안하였다. 동영상 데이터에 대한 사전지식(토픽)에 기반 하여 연관된 데이터가 연속된 경우만을 가정했던 기존 방법론과 달리, 제안 방법은 연관된 데이터가 일정 시간 지속된 후 새로운 종류의 데이터(스토리)가 등장하는 상황을 처리할 수 있는 방법이다. 인간 실험자의 스토리 변환 판단 결과와 비교하여 제안 방법의 성능을 평가하였다. 스토리 변환이라는 문제는 인간 실험자들조차 공통된 변환점을 판단하지 못하는 어려운 문제지만, 인간 실험자의 변환점 판단 결과를 정답이라고 간주하고 비교한 결과를 측정할 수 있는 새로운 방법을 정의하여 추정 결과를 수치적으로 분석한 결과는 다음과 같다. 첫째, 생성 모델에 기반 한 우도 계산으로 실제 스토리 구분점을 찾을 수 있음을 확인하였다. 둘째, 스토리 구분점 후보군의 규모를 조절할 수 있는 수단을 확보하였다.

우리는 Dirichlet 분포를 이용하여 스토리 구간을 모델링할 수 있다고 가정했을 뿐 아니라 그 외에도 이미지 토픽의 존재를 가정하여 변환 지점을 판단하였다. 토픽의 개수의 한정은 추정 정확성에 상당한 영향을 미칠 수 있는 제약이 될 수 있다. 추후 연구에서는 이런 문제를 해결할 수 있도록 토픽의 수를 사전 한정하지 않는 모델링 방법을 제안하고자 한다.

감사의 글

이 논문은 교육과학기술부의 재원으로 국가연구재단의 지원을 받아 수행된 연구(2011-0016483, Videome)이며, 한국학술진흥재단 (314-2008-1-D00377, Xtran) 및 교육과학기술부의 BK21-IT 사업에 의해 일부 지원되었음.

참고문헌

[1] G. Laurent, "Sequence Coding and Learning", Dynamic Coordination in the Brain, C. v. d. Malsburg, W. A. Phillips, and W. Singer (Eds), MIT Press, pp. 35-42, 2010.

[2] [Online] http://projects.ldc.upenn.edu/TDT/

[3] J.-P. Poli, "An Automatic Television Stream Structuring System for Television Archives Holders", *Multimedia Systems*, vol. 14, pp. 255-275, 2008.

[4] G. Manson and S.-A. Berranim "Automatic TV Broadcast Structuring", *International Journal of Digital Multimedia Broadcasting*, vol. 2010, Article ID 153160, 2010.

[5] X. Wang and A. McCallum, "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends", 12th ACM SIGKDD, pp. 424-433, 2006.

[6] 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.

[7] Y.-W. Teh, "An Introduction to Bayesian Nonparametric Modelling", Machine Learning Summer School 2009.

[8] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian Nonparametric Models with Applications", Bayesian Nonparametrics: Principles and Practice, N. Hjort, C. Holmes, P. Mueller, and S. Walker (Eds.), Cambridge University Press, 2010.

[9] [Online] http://www.lsv.uni-saarland.de/Vorlesung/snlp/summer06/snlp06_chap9.pdf

[10] D. Blei, A. Ng, and J. Michael, Latent Dirichlet allocation, JMLR, vol.3, pp.993-1022, 2003.

[11] D.G Lowe, Object recognition from scale-invariant features, ICCV, vol.2, pp.1150, 1999.