Inferring Actor Communities from Videos

Sumit Negi1, Ramnath Balasubramanyan2, Santanu Chaudhury3

1IBM Research India
2Carnegie Mellon University
3Indian Institute of Technology, Delhi

sumitneg@in.ibm.com, rbalasub@cs.cmu.edu, santanuc@ee.iitd.ac.in

Abstract
In recent years there has been a growing interest in inferring social relations amongst actors in a video using audiovisual features, co-appearance features or both. The discovered relations between actors have been used for identifying leading roles, detecting rival communities in a movie plot etc. In this paper we propose an unsupervised method which uses the video’s transcript and closed caption information for discovering actor communities (group of actors or characters in a film that share a common perspective/viewpoint on an issue) from videos. The method proposed groups together actors using a topic model based approach, which jointly models actor-actor interaction (two actors interact when they share the same scene) and the topics associated with their conversations/dialogs. This joint modeling approach shows encouraging results compared to existing methods.

Index Terms: topic model, actor community, video analysis

1. Introduction

Extracting social content, for instance relationship amongst actors, from videos is a challenging task, primarily, due to the gap that exists between low-level audio-visual features and high-level social interactions. However, with improvements in object detection, tracking techniques and advancements in the area of semantic visual concept detection this gap is slowly being bridged. Inferring relations among actors in a video refers to the process of associating the information content of the video to interactions among the actors in it. The discovered relations between actors have been used for identifying leading roles, detecting rival communities in a movie plot [4], performing social-relations based story segmentation [5] etc. Our work focuses on discovering actor communities from documentary1 styled films. Documentaries often rely on narrations and interviews to capture the perspective/viewpoints of different groups on the issue being discussed. For example, a documentary on global warming would highlight the viewpoints of different groups/communities (e.g. policy makers, scientists, industry representatives, activists etc - not necessarily in any specific order) on this issue. Automatically discovering these groups or actor communities (where actor communities is defined as a group of actors or characters in a film that share a common perspective/viewpoint on an issue) from a video can aid higher order tasks such as video summarization, intelligent video retrieval etc.

Discovering such communities from videos is a challenging task for the following reasons (a) Inferring community structure based only on co-appearance of actors in the same scene2 might yield poor results. For example a mere co-appearance in the same scene(s) might not always imply that the actors share the same perspective on the issue being discussed. On the other hand actors sharing a common perspective on an issue might not always share the same scene (which is very common in documentaries). (b) Methods that use additional cues (e.g. film-editing guide-lines[3], visual and auditory information [4], [6]) for inferring relationship amongst actors have shown improved results over techniques that rely only on co-appearance based features. However all of these techniques are supervised and hence require labeled training data which might be difficult or costly to obtain.

We hypothesize that one can infer such actor communities by grouping together actors based on the topics3 characterizing their dialogs/conversations. Considering the global warm example, the dialogs/conversations of actors belonging to the policy maker actor community would be more dominated by topic(s) that can be best described by words such as - “tax”, “policy”, “regulation”, “lobbying” - and would be distinct from say the topics prevalent in the dialogs/conversations made by scientists or activists. Our proposed method groups together actors using a topic model based approach, which jointly models (a) observed actor-actor interactions (two actors interact when they share the same scene) and (b) the topics associated with their conversations/dialogs. Against this background, this paper focuses on the following problem: Given a video’s script and closed captions information along with the list of actors/characters appearing in the video we can, in an unsupervised fashion, infer actor communities from a video.

The remainder of the paper is organized as follows. Section 2 describes related work and highlights the key differences and advantages of our approach. In Section 3 we provide a detailed description of our approach. Section 4 and Section 5 present the experimental and conclusion section respectively.

2. Related Work

Many studies have been proposed to analyze movies based on audiovisual features for tasks such as genre classification [7], story segmentation [8], and video abstraction ([9], [10], [11]). In recent years there has been a growing interest in inferring social relations amongst actors in a video using audiovisual fea-

1Documentary films constitute a broad category of nonfictional motion pictures intended to document factual events.
2Scene is a group of semantically related shots, which are coherent to a certain subject or theme.
3Topic is a distribution over words
tures, co-appearance features or both. A popular approach for discovering actor communities from video uses actor-actor co-appearance features. These techniques [5] quantify character interrelationships based on the number of scenes in which both characters are present. Recent approaches ([4], [6], [3]) use additional cues such as visual concepts or film-editing guidelines to discover actor communities from videos. A key shortcoming of these techniques is their reliance on labeled training data which in many cases might be difficult or costly to obtain. The method proposed in this paper groups together actors using a topic model based approach, which jointly models actor-actor interaction (two actors interact when they share the same scene) and the topics associated with their conversations/dialogs. Our approach offers the following benefits

- Our method is completely unsupervised.
- It does not require extraction of audio-visual features/cues from videos, which is computationally expensive.
- Provides thematic indicators (in terms of topic) for the discovered actor communities.
- Allows for mixed membership (i.e. an actor can belong to multiple communities)

Moreover, our proposed approach differs from traditional community detection methods such as [18]. Unlike traditional approaches for community detection our method considers not just the actor-actor interactions but also the topics associated with the actor’s conversation when discovering community structures. This approach of finding community structure has been shown to yield better results than traditional community detection approaches where community structures are not influenced by the topics being discussed ([13], [19], [20]). Modeling approaches, such as the one presented in this paper, which consider both interactions and content tend to overcome such limitations. In our proposed approach the topics, which are inferred from the content (in our case the dialog/conversations), influence the community formation thus resulting in more accurate and meaningful community structures. To the best of our knowledge our work is the first to offer all these benefits. In the next section we provide details of our proposed approach.

3. Approach

This section describes in detail our modeling approach. In Section 3.1 we describe the pre-processing steps which involves segmenting the video into non-overlapping scenes and the construction of the scene-character relationship matrix. Section 3.2 describes in detail the proposed model for discovering actor communities.

3.1. Preprocessing

The first step is to segment a video into $D$ non-overlapping scenes $\{s^1, s^2, ..., s^D\}$, where each scene is completely described by it’s start and stop time $(t^i_{\text{start}}, t^i_{\text{end}})$ and the section of the screenplay/script associated with that scene. This is illustrated in Figure 1. This segmentation process is guided by the accompanying movie screenplay/script and closed captions. We employ the technique described in [1] for the segmentation step. The weakly supervised alignment algorithm proposed by the authors uses the screenplay and closed captions to parse a movie into a hierarchy of shots and scenes. Due to space constraints we will skip the details of the method and encourage the readers to see [1] for more details.

3.2. Jointly modeling of actor-actor interaction and dialogs

As mentioned previously our proposed approach groups together actors using a topic model based approach, which jointly models actor-actor interaction (two actors interact when they share the same scene i.e actors $c^i$ and $c^j$ interact if $M(i,j) = 1$ and $M(k,j) = 1$) and the topics associated with their conversations/dialogs. Our proposed Temporal Block-LDA model shown in Figure 2, which is inspired by the Block-LDA [17] approach, enables sharing of information between the actor co-appearance matrix on the left - that models interactions between actors - represented as edges in a graph with a block structure, and the component on the right that models a scene (document), through shared latent topics. The component on the right in Figure 2, which is an extension of LDA, models a scene (document) as sets of “bags of entities”, each bag corresponding to a particular type of entity. Every entity type has a topic wise multinomial distribution over the set of entities that can occur as an instance of the entity type. In addition, unlike the Block-LDA model, our proposed model treats each scene as a time-stamped document.

![Figure 1: Segmenting video into scenes](image-url)
with an associated real-valued attribute $s_d$. In our setting each scene contains two entity types namely script-words and actors. The script-words entity can be instantiated with words from the script vocabulary - after removing stop words and frequent words - e.g. script words={climate, ocean, glacier, ...}. As mentioned there exists a topic wise multinomial distribution over these keywords for the script words entity type. Similarly, the actors entity can be instantiated with actor/character names.

The component on the left in Figure 2 is a generative model for graphs representing actor-actor interactions with an underlying block structure. The key aspect to note in the model is the transfer of information from the scene to the (actor) actor interaction network and vice-versa through the shared topic multinomials $\beta$. This sharing causes the topic distribution in a scene to affect the cluster development in the (actor) actor network and similarly the cluster formation in the network affects the topic distributions in the scene. Additionally, it can be seen that the real-valued attribute $s_d$ associated with the scene affects the topic distribution in the scene which in turn influences the topics of the other entity types and therefore the actor cluster distribution. Therefore, we see that the actor-actor interaction network, the dialog text and time-stamp associated with the scene jointly influence each other via the shared topic parameters.

### 3.2.1. Generative Model

Let $K$ be the number of actor group/communities we wish to recover from a video. Assuming a scene consist of $T$ different types of entities (i.e. each scene/document contains $T$ bags of entities, in our setting $T = 2$), and that links in the graph are between entities of type $t_i$ (in our scenario $t_1$=actors), the generative process is as described in Figure 4. Note $e_{i1}$, $e_{i2}$, $e_{t1}$ and $s_d$ are observed random variables. Given the hyper-parameters $\alpha_L$, $\alpha_D$, $\alpha_L$, $\gamma$, $\alpha$ and $b$, the joint distribution over the documents, links, their topic distributions and topic assignments is given by

$$p(\pi_L, \theta, \beta, \alpha, e, (z_1, z_2), (e_{11}, e_{12}), \alpha_L, \alpha_D, \gamma) \propto (1)$$

$$K \prod_{k=1}^{K} \prod_{t=1}^{T} \text{Dir}(\beta_{k,t} | \gamma_t) \times K \prod_{k=1}^{K} \text{Inverse Gamma}(\mu_k, \sigma_k^2 | a, b^2) \times D \prod_{d=1}^{D} \text{Dir}(\theta_d | \alpha_D) \prod_{t=1}^{T} N_{d,t} \prod_{i=1}^{t} \text{Multinomial}(\beta_{k,t}, \alpha_L) \times$$

$$\prod_{d=1}^{D} \prod_{i=1}^{t} \text{Multinomial}(\beta_{k,t}, \alpha_D) \times \text{Dir}(\gamma | \alpha_L) \prod_{i=1}^{t} \text{Multinomial}(\beta_{k,t}, \gamma)$$

Due to the intractability of exact inference , a collapsed Gibbs sampler is used to perform approximate inference. It samples a latent topic for an entity mention of type $t$ in the text corpus conditioned on the assignments to all other entity mentions using the following expression (after collapsing $\theta_D$):

$$p(z_{t,i} = z | e_{t,i}, a, e_{11}, e_{12}, \alpha_D, \gamma, \alpha, b^2) \propto (n_{dz} + \alpha_D) \frac{n_{dz}}{\sum_{d,z} n_{dz} + |\mathcal{D}_z| \sigma_k^2} \times$$

$$\text{Dir}(\gamma | \alpha_L) \prod_{i=1}^{t} \text{Multinomial}(\beta_{k,t}, \gamma)$$

We sample a topic pair for every link conditional on topic pair assignments to all other links after collapsing $\pi_L$. One can similarly obtain the gaussian parameters $\mu_k$ and $\sigma_k^2$, details of which are presented in [2].
We use 6 English movies to evaluate our proposed method details of which are shown in Table 1. Our dataset comprises both documentary styled and fictional movies. To create the ground truth we ask three participants to identify the number of actor groups (communities) in a given movie. Once a consensus was reached on the number of actor groups in a movie, each participant was then asked to assign actors (from that movie) to one or more of these identified communities/groups. In case of any conflict the majority vote was considered. This exercise was repeated for all movies in the dataset. We compare the actor communities discovered by our method against the RoleNet [5] approach - an unsupervised method which uses the scene-character relation matrix to infer social relationship amongst actors in a movie.

To evaluate these two method the following metric (as proposed in [5]) was used. If two actors $c^1$ and $c^2$ are in the same community, the indicative values $\zeta_{ij}$ and $\zeta_{ji}$ of the pair are set to 1 otherwise, they are set to 0. There are $\binom{n}{2}$ possible actor pairs for $n$ number of actors. Among the possible $\binom{n}{2}$ pairs, we calculate how many of them are correctly labeled (as per the ground truth). The ratio of correctly labeled pairs to all possible ones is used to quantify community identification results.

Mathematically, $R = \frac{\sum_{i<j} \zeta_{ij} \delta_{ij}}{\binom{n}{2}}$, where $\delta_{ij} = 1$ if ($\zeta_{ij} = 1$ and $\zeta_{ji} = 1$) or ($\zeta_{ij} = 0$ and $\zeta_{ji} = 0$) otherwise 0. $\zeta_{ij}$ and $\zeta_{ji}$ are pair relationships as indicated by the ground truth and the method being compared (RoleNet or our proposed method) respectively. The value $\delta_{ij}$ indicates whether the identified result between the actors $c^1$ and $c^2$ is the same as the ground truth. The larger the $R$ ratio more accurate the community identification results. Please note that our method (due to its probabilistic nature) gives for each actor group the probability of an actor belonging that actor group/community. For all our experiments we consider the top 5 actors (ranked based on their probability of belonging to a group or community), $K$ i.e. the number of communities to recover from a given video is set based on the ground truth information. Figure 5 compares the $R$ ratio for the two methods. It is clear that our method shows significant improvements over the baseline (improvement ranges from 8% to 35%). As mentioned before an additional advantage of our proposed approach is that it extracts thematic indicators (in terms of topic) for the discovered actor communities. We observe that this can sometimes help in understanding the relationship that exists between actors of a given community. We show in Table 2 example of an actor community and associated topic discovered by our method. The topic words provide clues to what these actor communities represent - health professionals and farmer groups.

Table 1: Dataset description

<table>
<thead>
<tr>
<th>Legend</th>
<th>Movie Name</th>
<th>No. of actors</th>
<th>No. of groups</th>
<th>No. of Scenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>FOOD, INC</td>
<td>13</td>
<td>4</td>
<td>172</td>
</tr>
<tr>
<td>S2</td>
<td>Super Size Me</td>
<td>16</td>
<td>5</td>
<td>144</td>
</tr>
<tr>
<td>S3</td>
<td>Lake of Fire</td>
<td>21</td>
<td>5</td>
<td>139</td>
</tr>
<tr>
<td>S4</td>
<td>No End in Sight</td>
<td>26</td>
<td>4</td>
<td>122</td>
</tr>
<tr>
<td>S5</td>
<td>Babel</td>
<td>18</td>
<td>7</td>
<td>155</td>
</tr>
<tr>
<td>S6</td>
<td>Traffic</td>
<td>16</td>
<td>4</td>
<td>157</td>
</tr>
</tbody>
</table>

Table 2: Discovered actor community and associated topic for the movie Super Size Me (left) and FOOD, INC. (right)

<table>
<thead>
<tr>
<th>Actors</th>
<th>Topic</th>
<th>Top 5 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rowly</td>
<td>blind</td>
<td></td>
</tr>
<tr>
<td>Hawker</td>
<td>mummi</td>
<td></td>
</tr>
<tr>
<td>Phillips</td>
<td>heart</td>
<td></td>
</tr>
<tr>
<td>Actors</td>
<td>Topic</td>
<td>Top 5 words</td>
</tr>
<tr>
<td>Morison</td>
<td>plant</td>
<td></td>
</tr>
<tr>
<td>Roosh</td>
<td>chicken</td>
<td></td>
</tr>
<tr>
<td>Salatin</td>
<td>monsanto</td>
<td></td>
</tr>
<tr>
<td>Rajiyan</td>
<td>food</td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>farmer</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper we presented an unsupervised method for detecting actor communities from video by jointly models actor-actor interaction and the topics associated with their conversations/dialogs. Analysis on a dataset of movies consisting of both documentaries and fictional videos has validated the effectiveness of our proposed method.
6. References


