Bridging Social Graphs with Character-Centered Story Contexts in Videos

Jiang Gao
Samsung Research America
Mountain View, CA, USA
jiang.gao@samsung.com

ABSTRACT
Humans implicitly understand social connections involving character roles in a storyline, either by reading a novel, or watching a movie. We present a project for bridging the social graphs obtained from verbal stories, such as a novel, with the social connections extracted directly from a video. We present the system architecture, and outline the current progress, including extracting main characters from videos, and building story contexts to represent character presence patterns, in order to match with time-varying social graphs. We study several direct benefits of this approach, and expect the individual graphs extracted can be combined toward building a social knowledge graph covering a broader context and a range of character roles. This project is being conducted in Samsung Research AI Center, and utilizes Bixby services for communications between the NLU backend, the Characters database, and an interactive frontend.

CSC CONCEPTS
• Information system–Data mining • Computing methodologies–Visual content-based indexing and retrieval • Computing methodologies–Active learning setting

KEYWORDS
Knowledge representation, social graph, video analysis, social context, active learning, visual analytics

1 Introduction
The ability to align storylines with video concepts enables human-like understanding of video contents. Different from the other approaches for storyline analysis, our project emphasizes on the social graphs involving the main characters in the stories, and match them with the character presence patterns in the videos. There are public datasets for archive videos and their text descriptions, such as the Movie Descriptions Dataset [1], the MovieBook and BookCorpus Datasets [2]. In addition to focusing on the archive datasets, we also aim to expand the existing video datasets and their descriptions, and use the learned system components to build storylines for videos with only partial coverage from verbal stories or without verbal stories at all [3].

Using the character-centered approach for video analysis poses several major challenges to the project. First, there is a large amount of video data in the archives, with many different character roles involved. It is impractical to prebuild models for classifying a majority of the characters in the video archives. Therefore, we designed our system to start without knowledge about the character identities and their facial images, and to build individual character models as the video frames stream in. The system should be able to infer the identities, once the alignment is established based on the character presence patterns. To cold start this modeling process, we designed an active learning algorithm on top of metric learning. We outline the main measures and components of our active learning system in Section 2.2. There is also an interactive user interface, which we will show in our demo. The interactions with the system are multimodal, using voices, touches, and gestures.

Second, social graphs in a story can be represented in different formats. The easiest is using a static social graph to summarize the connections between characters in a story. However, time-varying social graphs [4] could be a better representation, if we are concerned with the evolving character social connections in a mostly sequential storyline. We designed a system pipeline to process videos as the frames stream in, and it is a one-pass efficient process. We also define several time-varying character presence patterns in videos as Story Context in Section 2.1.
2 System Architecture and Progress

Fig. 2 shows our system architecture. The major components of the system include the Natural Language Understanding (NLU) and Face Backend Services, which extract characters from texts and video frames, respectively. The Character Database is where the character presence data are saved and queried by other components in the system. The input component processes both real-time video input and video archives, and support multimodal user interactions. Bixby Services is responsible for voice communications between the Character Database, NLU Backend, and the user/input interface.

![Image 2: The system architecture.](image)

2.1 Social Graphs and Story Contexts

To accurately align with the time-varying social graphs obtained from storylines, we define Story Contexts as the following time-varying character presence patterns in videos:

1. Visual consistency: character instances should be visually similar to one of the gallery instances with the same character label \( l_b \).
2. Co-Presence: character \( l_a \) and character \( l_b \) appear together in a frame concurrently.
3. Sequential: character \( l_a \) appears in frame \( b \), after character \( l_a \) appears in previous frame \( a \), in a sequential order.

Our story/video matching objective function based on storylines leverages on the sequential and co-presence patterns between each character. The occurrence probability of these patterns is estimated based on the optimal alignment between a video clip and its storyline up till the current frame.

2.2 Character Modeling using Active Learning

It is challenging to establish character models, especially in videos where characters could go through a lifetime in a clip, with their facial appearances change dramatically between scenes, and frequently occluded. Our learning algorithm is different from the setup of a pool-based active learning systems [5]. First, our data is a stream due to the sequential nature of video clips, and it is desirable to learn and ground the faces as the frames stream in. Second, since we do not have initially labeled face models, it is a cold-start process to build the gallery samples. Third, different from conventional active learning algorithms, our query proposal function depends not only on visual appearance, but also on the story contexts of these characters. Fig. 3 shows the main measures and components in our active learning system. Fig. 4 shows our learning algorithm proposing query character faces to users.

![Image 3: The active learning system for proposing queries to users, based on both visual similarity and story context.](image)

3 The Demo System

User can choose real-time video input to cold-start character classification of conference attendants, and build story contexts for each attendant at the same time. The user can also choose to run an archive video, and build character models and their social contexts along the way. Fig. 1 shows a snapshot from the demo system.

![Image 4: Query proposing based on active learning.](image)

References