A Local-to-Global Approach to Multi-modal Movie Scene Segmentation

Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, Dahua Lin

CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

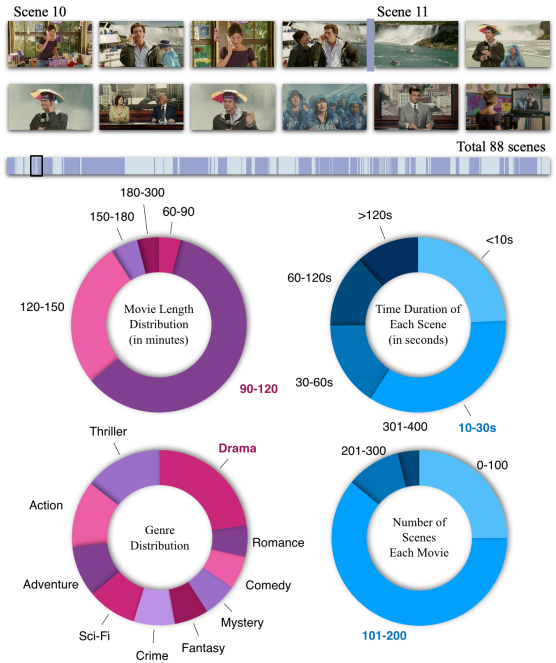
Introduction

Scene, as the crucial unit of storytelling in movies, contains complex activities of actors and their interactions in a physical location.
Scene consists of many shots, noting that a shot is an unbroken sequence of frames recorded from the same camera.
Identifying the composition of scenes serves as a critical step towards visual understanding of movies, TV episodes, entertainment shows and variety shows.

This work is going to help divide long videos into semantic continuous short videos and output a structural representation. And it also provides research opportunities towards story/plot understanding in long videos with a semantic unit.

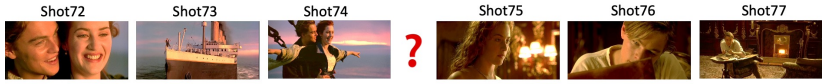
MovieScenes Dataset

MovieScenes contains **21K** scenes from **150** movies, which is **100x** larger than exiting datasets. It provides a foundation for studying the complex semantics within the scene.



Approach

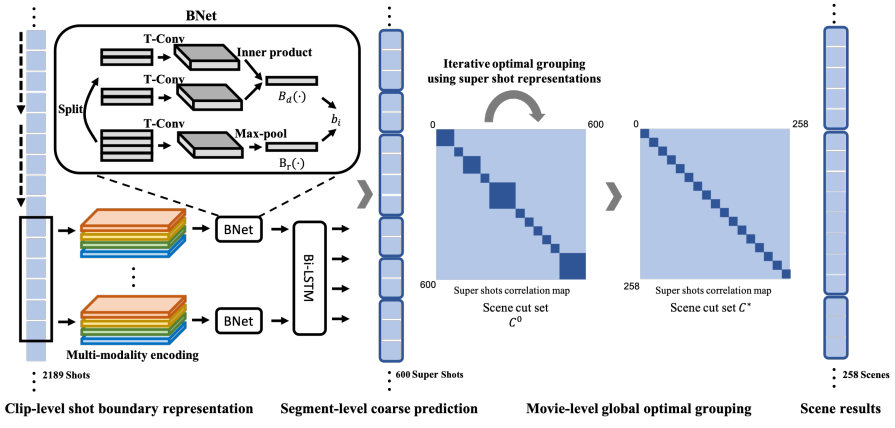
Problem formulation: binary classification



Is this shot boundary a scene boundary?

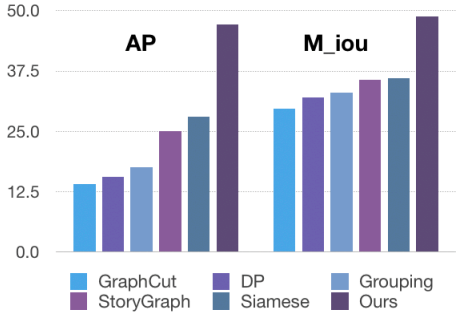
Framework: multi-modal local-to-global scene segmentation

- To cover **rich semantic information**, we extract multi-semantic elements including ***place***, ***cast***, ***action***, ***audio*** to represent a shot
- To cover **complex temporal information**, bottom-up forward and top-down guidance are implemented at clip-level, segment-level and movie-level



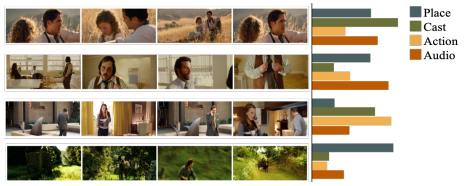
Experiments

Overall results



Ablation studies of multi-semantics

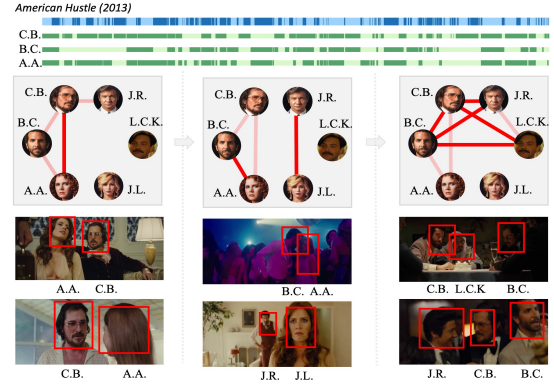
Place	Cast	Action	Audio	AP
✓				39.0
	✓			15.9
		✓		32.1
			✓	17.5
✓	✓	✓	✓	**47.1**



Applications

Human interaction graph generation

To visualize the dynamic evolution of characters' relationships over time in a movie



Cross movie scene retieval

To retrieve similar scenes in other movies given a specific scene in one movie