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
{anyirao, xy017, xg018, hq016, bzhou, dhlin}@ie.cuhk.edu.hk, linningxu@link.cuhk.edu.cn

Abstract

Scene, as the crucial unit of storytelling in movies, contains complex activities of actors and their interactions in a physical environment. Identifying the composition of scenes serves as a critical step towards semantic understanding of movies. This is very challenging – compared to the videos studied in conventional vision problems, e.g. *action recognition, as scenes in movies usually contain much richer temporal structures and more complex semantic information. Towards this goal, we scale up the scene segmentation task by building a large-scale video dataset MovieScenes, which contains* $21K$ *annotated scene segments from* 150 *movies. We further propose a local-to-global scene segmentation framework, which integrates multi-modal information across three levels,* i.e. *clip, segment, and movie. This framework is able to distill complex semantics from hierarchical temporal structures over a long movie, providing top-down guidance for scene segmentation. Our experiments show that the proposed network is able to segment a movie into scenes with high accuracy, consistently outperforming previous methods.* [1]

1. Introduction

Imagine you are watching the movie *Mission Impossible*: In a fight scene, Ethan leaps onto a helicopter's landing skid and attaches an exploding gum to the windshield to destroy the enemy. Suddenly, the story jumps into an emotional scene where Ethan pulled the trigger and sacrificed his life to save his wife. Such a dramatic change of scenes plays an important role in the movie's storytelling. Generally speaking, a movie is composed of a well-designed series of intriguing scenes with transitions, where the underlying storyline determines the order of the scenes being presented. Therefore recognizing the movie scenes, including the detection of scene boundaries and the understanding of the scene content, facilitates a wide-range of movie understanding tasks such as scene classification, cross movie scene retrieval, human interaction graph and human-centric storyline construction.

It is worth noting that *scenes* and *shots* are essentially different. In general, a *shot* is captured by a camera that operates for an uninterrupted period of time and thus is *visually continuous*; while a *scene* is a semantic unit at a higher level. A scene comprises a sequence of shots to present a *semantically coherent* part of the story. Therefore, whereas a movie can be readily divided into shots based on simple visual cues using existing tools [12], the task of identifying those sub-sequences of shots that constitute scenes is challenging, as it requires semantic understanding in order to discover the associations between those shots that are semantically consistent but visually dissimilar.

There has been extensive studies on video understanding. Despite the great progress in this area, most existing works focus on recognizing the categories of certain activities from short videos [15, 8]. More importantly, these works assume a list of pre-defined categories that are visually distinguishable. However, for movie scene segmentation, it is impossible to have such a list of categories. Additionally, shots are grouped into scenes according to their semantical coherence rather than just visual cues. Hence, a new method needs to be developed for this purpose.

To associate visually dissimilar shots, we need semantical understanding. The key question here is *"how can we learn semantics without category label?"* Our idea to tackle this problem consists in three aspects: 1) Instead of attempting to categorize the content, we focus on scene *boundaries*. We can learn what constitute a boundary between scenes in a supervised way, and thus get the capability of differentiating between within-scene and cross-scene transitions. 2) We leverage the cues contained in multiple semantic elements, including *place*, *cast*, *action*, and *audio*, to identify the associations across shots. By integrating these aspects, we can move beyond visual observations and establish the semantic connections more effectively. 3) We also explore the top-down guidance from the overall understanding of the movie, which brings further performance gains.

Based on these ideas, we develop a local-to-global framework that performs scene segmentation through three

[1] The dataset will be published in compliance with regulations. https://anyirao.com/projects/SceneSeg.html

stages: 1) extracting shot representations from multiple aspects, 2) making local predictions based on the integrated information, and finally 3) optimizing the grouping of shots by solving a global optimization problem.

2. MovieScenes Dataset

To facilitate the scene understanding in movies, we construct *MovieScenes*, a large-scale scene segmentation dataset that contains $21K$ scenes coming from 150 movies, which provides a foundation for long video understanding.

2.1. Definition of Scenes

Following previous definition of *scene* [9, 2, 5, 13], a scene is a plot-based semantic unit, where a certain activity takes place among a certain group of characters. While a scene often happens in a fixed place, it is also possible that a scene traverses between multiple places continually, *e.g*. during a fighting scene in a movie, the characters move from indoor to outdoor. These complex entanglements in scenes cast more difficulty in the accurate detection of scenes which require high-level semantic information. Figure 1 illustrates some examples of annotated scenes in *MovieScenes*, demonstrating this difficulty.

The vast diversity of movie scenes makes it hard for the annotators complying with each other. To ensure the consistency of results from different annotations, during the annotation procedure, we provided ambiguous examples with specific guidance to clarify how such cases should be handled. Moreover, all data are annotated by different annotators independently for multiple times. In the end, our multiple times annotation with the provided guidance leads to highly consistent results, *i.e*. 89.5% consistency in total.

2.2. Annotation Statistics

Large-scale. Table 1 compares *MovieScenes* with existing similar video scene datasets. We show that *MovieScenes* is significantly larger than other datasets in terms of the number of shots/scenes and the total time duration. Furthermore, our dataset covers a much wider range of diverse sources of data, capturing all kinds of scenes, compared with short films or documentaries.

Diversity. Most movies in our dataset have time duration between 90 to 120 minutes, providing rich information about individual movie stories. A wide range of genres is covered, including most popular ones such as dramas, thrillers, action movies, making our dataset more comprehensive and general. The length of the annotated scenes varies from less than $10s$ to more than $120s$, where the majority last for $10 \sim 30s$. This large variability existing in both the movie level and the scene level makes movie scene segmentation task more challenging.

Table 1. A comparison of existing scene datasets.

	#Shot	#Scene	#Video	Time(h)	Source
OVSD [11]	10,000	300	21	10	MiniFilm
BBC [1]	4,900	670	11	9	Docu.
MovieScenes	270,450	21,428	150	297	Movies

3. Local-to-Global Scene Segmentation

As mentioned above, a scene is a series of continuous shots. Therefore, scene segmentation can be formulated as a binary classification problem, *i.e*. to determine whether a shot boundary is a scene boundary. However, this task is not easy, since it requires the recognition of multiple semantic aspects and the usage of complex temporal information.

To tackle this problem, we propose a Local-to-Global Scene Segmentation framework (LGSS). The overall formulation is shown in Equation 1. A movie with n shots is represented as a shot sequence $[\mathbf{s}_1, \cdots, \mathbf{s}_n]$, where each shot is represented with multiple semantic aspects. We design a three-level model to incorporate different levels of contextual information, *i.e*. clip level ($\mathcal{B}$), segment level ($\mathcal{T}$) and movie level ($\mathcal{G}$), based on the shot representation $\mathbf{s}_i$. Our model gives a sequence of predictions $[o_1, \cdots, o_{n-1}]$, where $o_i \in \{0, 1\}$ denotes whether the boundary between the i-th and $(i+1)$-th shots is a scene boundary.

$$\mathcal{G}\{\mathcal{T}[\mathcal{B}([\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n])]\} = [o_1, o_2, \cdots, o_{n-1}] \quad (1)$$

In the following parts of this section, we will first introduce how to get $\mathbf{s}_i$, namely how to represent the shot with multiple semantic elements. Then we will illustrate the details of the three levels of our model, *i.e*. $\mathcal{B}$, $\mathcal{T}$ and $\mathcal{G}$.

3.1. Shot Representation with Semantic Elements

Movie is a typical multi-modal data that contains different high-level semantic elements. A global feature extracted from a shot by a neural network, which is widely used by previous works [1, 13], is not enough to capture the complex semantic information.

A scene is a sequence of shots sharing some common elements, *e.g*. place, cast, *etc*. Thus, it is important to take these related semantic elements into consideration for better shot representation. In our framework, a shot is represented with four elements that play important roles in the constitution of a scene, namely *place*, *cast*, *action*, and *audio*.

To obtain semantic features for each shot $\mathbf{s}_i$, we utilize 1) Places [17]-pretrained ResNet50 [6] on key frame images to get *place* features, 2) CIM [7] pretrained Faster-RCNN [10] to detect cast instances and PIPA [16] pretrained ResNet50 to extract *cast* features, 3) AVA [4] pretrained TSN [15] to get *action* features, 4) AVA-ActiveSpeaker [4] pretrained NaverNet [3] to separate speech and background sound, and stft [14] to get their features respectively, and concatenate them to obtain *audio* features.

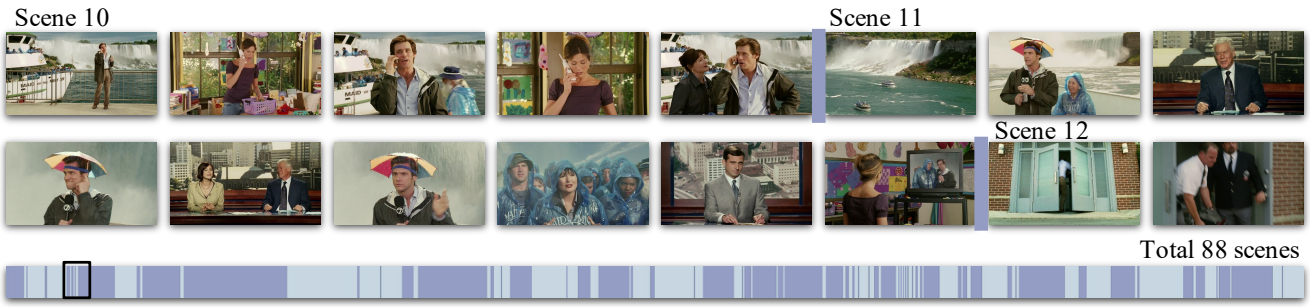

Figure 1. Example of the annotated scenes from movie *Bruce Almight (2003)*. The blue line in the bottom is the whole movie timeline where the dark/light blue regions represent different scenes. In Scene 10, the characters are having a phone call in two different places, which requires a semantic understanding to avoid categorizing into different scenes. In Scene 11, the task becomes more difficult, as the live broadcasting scene involves more than three places and groups of characters. In this case, visual cues only are likely to fail, thus the inclusion of other aspects such as the audio cues becomes critical.

3.2. Shot Boundary Representation at Clip Level

As we mentioned before, scene segmentation can be formulated as a binary classification problem on shot boundaries. Therefore, how to represent a shot boundary becomes a crucial question. Here, we propose a Boundary Network (BNet) to model the shot boundary. As shown in Equation 2, BNet, denoted as $\mathcal{B}$, takes a clip of the movie with $2w_b$ shots as input and outputs a boundary representation $\mathbf{b}_i$. Motivated by the intuition that a boundary representation should capture both the *differences* and the *relations* between the shots before and after, BNet consists of two branches, namely $\mathcal{B}_d$ and $\mathcal{B}_r$. $\mathcal{B}_d$ is modeled by two temporal convolution layers, each of them embeds the shots before and after the boundary respectively, following an inner product operation to calculate their differences. $\mathcal{B}_r$ aims to capture the relations of the shots, it is implemented by a temporal convolution layer followed a max pooling.

$$\begin{aligned}\mathbf{b}_i &= \mathcal{B}([\mathbf{s}_{i-(w_b-1)}, \cdots, \mathbf{s}_{i+w_b}]) \quad \text{(window size } 2w_b\text{)} \\ &= \begin{bmatrix} \mathcal{B}_d([\mathbf{s}_{i-(w_b-1)}, \cdots, \mathbf{s}_i], [\mathbf{s}_{i+1}, \cdots, \mathbf{s}_{i+w_b}]) \\ \mathcal{B}_r([\mathbf{s}_{i-(w_b-1)}, \cdots, \mathbf{s}_i\ ,\ \mathbf{s}_{i+1}, \cdots, \mathbf{s}_{i+w_b}]) \end{bmatrix}\end{aligned} \tag{2}$$

3.3. Coarse Prediction at Segment Level

After we get the representatives of each shot boundary $\mathbf{b}_i$, the problem becomes predicting a sequence binary labels $[o_1, o_2, \cdots, o_{n-1}]$ based on the sequence of representatives $[\mathbf{b}_1, \cdots, \mathbf{b}_{n-1}]$, which can be solved by a sequence-to-sequence model. However, the number of shots n is usually larger than 1000, which is hard for existing sequential models to contain such a long memory. Therefore, we design a segment-level model to predict a coarse results based on a movie segment that consists of w_t shots ($w_t \ll n$). Specifically, we use a sequential model $\mathcal{T}$, *e.g.* a Bi-LSTM, with stride $w_t/2$ shots to predict a sequence of coarse score $[p_1, \cdots, p_{n-1}]$, as shown in Equation 3. Here $p_i \in [0, 1]$ is the probability of a shot boundary to be a scene boundary.

$$[p_1, \cdots, p_{n-1}] = \mathcal{T}([\mathbf{b}_1, \cdots, \mathbf{b}_{n-1}]) \tag{3}$$

Then we get a coarse prediction $\bar{o}_i \in \{0, 1\}$, which indicates whether the i-th shot boundary is a scene boundary. By binarizing p_i with a threshold τ, we get

$$\bar{o}_i = \begin{cases} 1 & \text{if } p_i > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

3.4. Global Optimal Grouping at Movie Level

The segmentation result $\bar{o}_i$ obtained by the segment-level model $\mathcal{T}$ is not good enough, since it only considers the local information over w_t shots while ignoring the global contextual information over the whole movie. In order to capture the global structure, we develop a global optimal model $\mathcal{G}$ to take movie-level context into consideration. It takes the shot representations $\mathbf{s}_i$ and the coarse prediction $\bar{o}_i$ as inputs and make the final decision o_i as follows,

$$[o_1, \cdots, o_{n-1}] = \mathcal{G}([\mathbf{s}_1, \cdots, \mathbf{s}_n], [\bar{o}_1, \cdots, \bar{o}_{n-1}]) \tag{5}$$

The global optimal model $\mathcal{G}$ is formulated as an optimization problem. Before introducing it, we establish the concept of super shots and objective function first.

The local segmentation gives us an initial rough scene cut set $\mathbf{C} = \{\mathcal{C}_k\}$, here we denote $\mathcal{C}_k$ as a *super shot*, *i.e.* a sequence of consecutive shots determined by the segment-level results $[\bar{o}_1, \cdots, \bar{o}_{n-1}]$. Our goal is to merge these super shots into j scenes $\mathbf{\Phi}(n = j) = \{\phi_1, \ldots, \phi_j\}$, where $\mathbf{C} = \biguplus_{k=1}^{j} \phi_k$ and $|\phi_k| \geq 1$. Since j is not given, to automatically decide the target scene number j, we need to look through all the possible scene cuts, *i.e.* $F = \max_{j, j<|\mathbf{C}|} F(n = j)$. With fixed j, we want to find the optimal scene cut set $\mathbf{\Phi}^\star(n = j)$. The overall optimization

problem is as follows,

$$
\begin{aligned}
F^{\star} &= \max_{j} F(n=j) \qquad (6)\\
&= \max_{j} \left(\max_{\mathbf{\Phi}} \sum_{\phi_k \in \mathbf{\Phi}} g(\phi_k) \right), \\
&\text{s.t.} \quad j < |\mathbf{C}|, |\mathbf{\Phi}| = j.
\end{aligned}
$$

Here, $g(\phi_k)$ is the optimal scene cut score achieved by the scene ϕ_k. It formulates the relationship between a super shot $\mathcal{C}_l \in \phi_k$ and the rest super shots $\mathcal{P}_{k,l} = \phi_k \backslash \mathcal{C}_l$. $g(\phi_k)$ constitutes two terms to capture a global relationship and a local relationship, $F_s(\mathcal{C}_k, \mathcal{P}_k)$ is similarity score between $\mathcal{C}_k$ and $\mathcal{P}_k$, and $F_t(\mathcal{C}_k, \mathcal{P}_k)$ is an indicate function that whether there is a very high similarity between $\mathcal{C}_k$ and any super shot from $\mathcal{P}_k$ aiming to formulate shots thread in a scene. Specifically,

$$
g(\phi_k) = \sum_{\mathcal{C}_k \in \phi_k} f(\mathcal{C}_k, \mathcal{P}_k) = \sum_{\mathcal{C}_k \in \phi_k} (F_s(\mathcal{C}_k, \mathcal{P}_k) + F_t(\mathcal{C}_k, \mathcal{P}_k)),
$$

$$
\begin{aligned}
F_s(\mathcal{C}_k, \mathcal{P}_k) &= \frac{1}{|\mathcal{P}_k|} \sum_{\hat{\mathcal{C}}_k \in \mathcal{P}_k} \cos(\mathcal{C}_k, \hat{\mathcal{C}}_k), \\
F_t(\mathcal{C}_k, \mathcal{P}_k) &= \sigma(\max_{\hat{\mathcal{C}}_k \in \mathcal{P}_k} \cos(\mathcal{C}_k, \hat{\mathcal{C}}_k)).
\end{aligned}
$$

DP. Solving the optimization problem and determining target scene number can be effectively conducted by dynamic programming (DP). The update of $F(n=j)$ is

$$
\max_{k} \{F^{\star}(n = j-1 | \mathbf{C}_{1:k}) + g(\phi_j = \{\mathcal{C}_{k+1}, \ldots, \mathcal{C}_{|\mathbf{C}|}\})\},
$$

where $\mathbf{C}_{1:k}$ is the set containing the first k super shots.

References

[1] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. A deep siamese network for scene detection in broadcast videos. In *23rd ACM International Conference on Multimedia*, pages 1199–1202. ACM, 2015. 2

[2] Vasileios T Chasanis, Aristidis C Likas, and Nikolaos P Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *IEEE transactions on multimedia*, 11(1):89–100, 2008. 2

[3] Joon Son Chung. Naver at activitynet challenge 2019–task b active speaker detection (ava). *arXiv:1906.10555*, 2019. 2

[4] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 2

[5] Bo Han and Weiguo Wu. Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In *2011 IEEE International conference on multimedia and expo*, pages 1–6. IEEE, 2011. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] Qingqiu Huang, Yu Xiong, and Dahua Lin. Unifying identification and context learning for person recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2217–2225, 2018. 2

[8] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[9] Zeeshan Rasheed and Mubarak Shah. Scene detection in hollywood movies and tv shows. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–343. IEEE, 2003. 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 2

[11] Daniel Rotman, Dror Porat, and Gal Ashour. Optimal sequential grouping for robust video scene detection using multiple modalities. *International Journal of Semantic Computing*, 11(02):193–208, 2017. 2

[12] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, 2011. 1

[13] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Storygraphs: visualizing character interactions as a timeline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 827–834, 2014. 2

[14] Srinivasan Umesh, Leon Cohen, and D Nelson. Fitting the mel scale. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 217–220. 2

[15] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 2

[16] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4804–4813, 2015. 2

[17] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2018. 2