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

Anyi Rao[1], Linning Xu[2], Yu Xiong[1], Guodong Xu[1], Qingqiu Huang[1], Bolei Zhou[1], Dahua Lin[1]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2]The Chinese University of Hong Kong, Shenzhen
{anyirao, xy017, xg018, hq016, bzhou, dhlin}@ie.cuhk.edu.hk, linningxu@link.cuhk.edu.cn

1. Details of Global Optimal Grouping

Details of iterative optimization The detailed algorithm of the iterative optimization in section 4.4 of the main part of the paper [2] is illustrated in Algorithm 1. Here, Iter # stands for the iteration number of the optimization, which are got from validation set. In our experiments, the optimal scene segments set $\mathbf{C}^\star$ usually get converged after Iter # $= 5$.

Super shots representation refinement. In the p-th step of the iteration, the representation of k-th super shot comes from the weighted sum of the shots consist of it, which is defined as $\mathcal{C}_k^p = W_k^p S_k^p$, and S_k^p is the shots that constitute of the super shot $\mathcal{C}_k^p$ [1], and W_k^p are parameters. With the maximum value $F(W)$ achieved in DP, we update W with gradient decent. Through this process, we update the representation of super shots $\mathcal{C}_k^p = W_k^p S_k^q$.

Algorithm 1 Global optimization algorithm

Input: coarse scene cut set $\mathbf{C}$ from local segmentation and Iter #

Output: optimal scene cut set $\mathbf{C}^*$

for $p \leftarrow 1$, Iter # **do**
  Get maximum scene cut score F and merged scenes $\mathbf{\Phi}$ with dynamic programming.
  Update super shots $\mathbf{C}$ representation according to F. (See below for the details of super shots representation.)
  $\mathbf{C} \leftarrow \mathbf{\Phi}$.
end for
return $\mathbf{C}^* \leftarrow \mathbf{C}$

2. MovieScenes Dataset

Table 1 shows some basic statistics about our dataset. MovieScenes consists a total of $272,301$ binary decisions from 150 movies.

Table 1. Statistics of the *MovieScenes* annotation set.

	Train	Val	Test	Total
Number of Movies	100	20	30	150
Number of Scenes	14,389	2,338	4,701	21,428
Number of Shots	188,892	23,549	58,009	270,450
Avg. Dur. of Movie (h)	2.01	1.74	2.02	1.98
Avg. Dur. of Scene (s)	50.00	53.33	46.08	49.50
Avg. Dur. of Shot (s)	3.84	5.31	3.78	3.95

2.1. Details of Dataset Diversity

Our dataset covers movies in different length and genres as shown in the left column of Figure 1, Most movies in our dataset have time duration between 90 to 120 minutes. A wide range of genres is covered, 10 genres covered in the dataset. The statistical information about scenes is shown in the right column of Figure 1. Each movie holds different number of scenes, and most of them contain 100 to 200 scenes, derived from $1,000$ to $2,000$ shots. The length of scenes varies significantly, ranging from less than $10s$ to more than $120s$, and most of them last for $10 \sim 30s$.

2.2. Details of Dataset Consistency

As shown in the Table 1 in the main part of paper [2]. We divide all annotations into three categories: (1) *high consistency cases*, *i.e.* those that received same results from three annotators in the first round or those that received same results from four of the five annotators after the second round. (2) *low consistency cases*, *i.e.* those that received same results from three of the five annotators after the second round. They are hard cases for human since annotators achieve low consistency; (3) *unsure cases*, *i.e.* those are bad cases for human since annotators cannot achieve consistency. We discard this category in the our experiments.

[1]Recall that in a video, shots constitute super shots, and super shots constitute scenes.

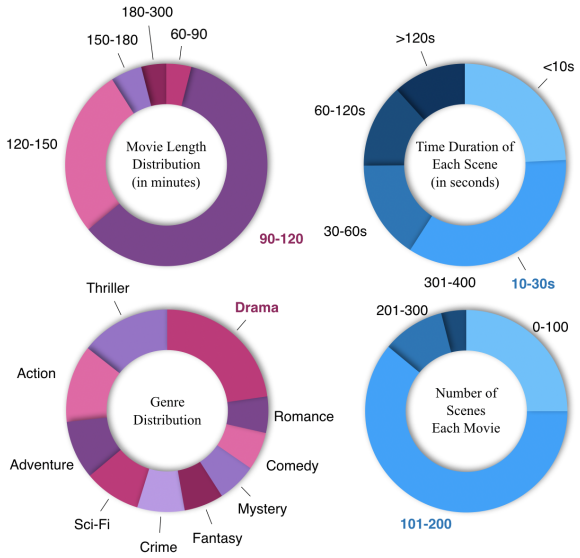

Figure 1. Four selected characteristics of *MovieScenes*. *Left top:* movie length distribution. *Left bottom:* movie genera distribution. *Right top* and *right bottom* tell about the general statistical information of scene summarized from the dataset.

2.3. Annotation Interface

The annotation interface is shown in Figure 2. Annotators click *MOVIE LIST* to chose a clip to annotate. *MANUAL* shows instructions. *LANGUAGE SWITCH* switches interface language.

If the central shots pair is a scene transition, annotators press *X* to annotate *Transit*. If the central shots pair is not a scene transition, annotators press *Z* to annotate *Continue*. If annotators are unsure about their decision, they press *C* to annotate *Skip*.

The center of the interface are the videos of a shots pair. And the left and right side are two still images. If the annotation is continue, a blue arrow will show between two shots. If the annotation is transit, a red vertical bar will show between two shots. At the top of each shot, there is a indicator that shows the number. At the bottom, there shows the progress of annotations.

3. More Qualitative Results

More qualitative results are shown in Figure 3 and 4. Figure 3 and 4 show non-transition and transition cases respectively, in both of which our model make right predictions. Here non-transition means the shot boundary is not a scene boundary, and vice verse.

Although, in Figure 3, four consecutive shots from one scene seemingly have different semantic information, our model can still predict non-transition through multiple semantic elements cues, such as places, role's appearance from different views and relationship between the parts and the whole of an object, and the environmental sound.

Table 2. Comparison of accuracy on the hard cases of character recognition. 1) Rand: Random guess; 2) Char (shot): Character-character relationship within shot; 3) Char: Character-character relationship within scene; 4) Scene: Character-scene relationship; 5) Both: use both character-scene relationship and character-character relationship.

Method	Rand	Char (shot)	Scene	Char	Both
Accuracy	9.2	20.1	29.5	28.4	33.6

Scene transition indicates salient semantics change, thus usually contains sharp visual and audio features change including places, light conditions, characters, action, environmental sound and background music. Figure 4 shows some transition cases and our successful predictions.

4. Applications with the Help of Scenes

4.1. Improving Character Recognition

Character recognition in a movie is a challenging task since a movie contains lots of shots where characters do not show up their full faces, as shown in Figure 7.

With the help of scene, we are able to establish two strong priors to handle these ambiguous cases, *i.e.* 1) *character-scene* relationship and 2) *character-character* relationship. For the first prior, it is known that the characters appeared within one scene must be the same. The character without full faces is likely to be the one shown up with faces in the rest part of the same scene. Thus we take advantages of character-scene relationship to infer those characters without faces. For the other prior, along with the whole movie, character-character relationship can be got from scenes. We are able to know which group of characters is more likely to appear in the same scene. Therefore, in the case that there are two people appearing in one shot with one people showing face and the other one only showing the back, we can leverage the character-character relationship to make prediction for those characters without clear faces.

We conduct experiments on our test set using character label from [1] and pick out the shots where no faces are detected. For each movie, 10 casts are annotated and have annotation in each shot. 3,000 shots are picked out from 30 movies. These shots are hard cases for traditional character recognition methods, since there are no clear faces for characters.

Character-scene relationship and character-character relationship are used as priors to infer character identity. We also build character-character relationships based on shots (Char (shot)) for comparison. All the results as shown in Table 2. Since we are only concerned about the 10 casts. So random guess is at 9.2 accuracy on character recogni-

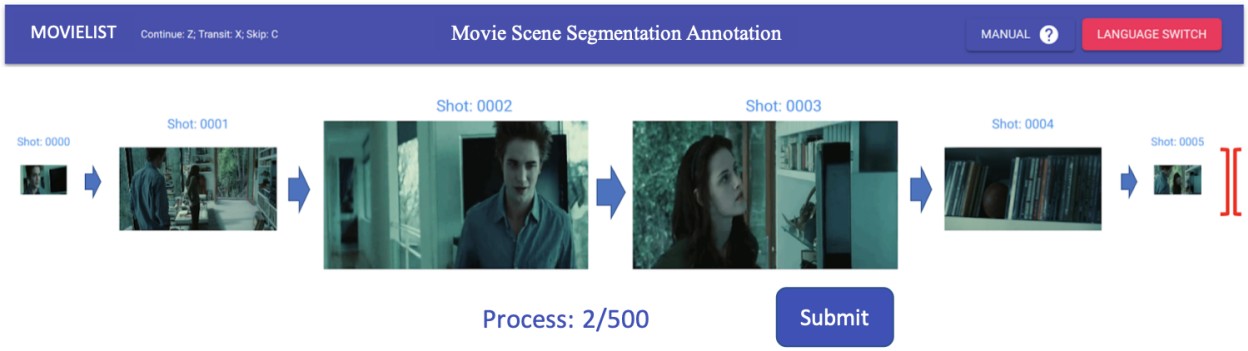

Figure 2. Annotation interface.

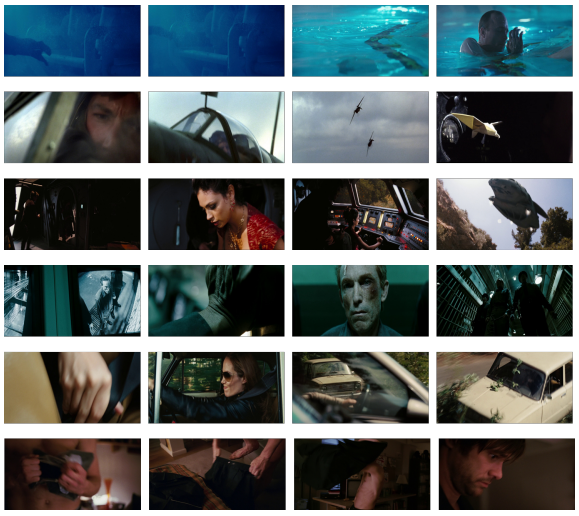

Figure 3. Non-transition cases. Each row represents four consecutive shots from one scene. These scenes from top to bottom are underwater diving, driving fighters, driving helicopters, escorting prisoners, driving cars and dressing. Our model is able to find internal details correlation and give right predictions.

Figure 4. Transition cases. Each row represents four consecutive shots from two adjacent scenes. The scene boundary is in the middle of the second and the third shot. Our model is able to recognize salient semantics change and give right predictions.

tion, while our best method, which use the above two priors, achieve 33.6 accuracy. It is much better than random guess and shot-based character interaction. It is shown that Scenes help build up a better character-character relationship and combining it with character-scene relationship help to improve performance on the hard cases of character recognition.

4.2. Generating Human Interaction Graph

After recognizing characters in each shot [1], we can group these characters into scenes according to shot-scene relationship coming from the segmented scenes. Figure 5 visualizes each character occurrence scenes. We derive a novel human interaction graph from all characters occurrence, to visualize the evolution of characters' relationships over time. It clearly shows that the characters' interaction develop over time. Furthermore, based on the segmented Scenes, we can count the total character interaction time, *e.g.* C.B and A.A are with each other fifty minutes in the two hour *American Hustle*. [2]

4.3. Cross Movie Scene Retrieval

It is still an open question to represent a long video such as a scene. A collection of well segmented scenes can serve as good samples for studying how to organize high-level semantic video representations as well as facilitate the development of methodologies for a number of salient tasks related to semantics understandings, *e.g.* scene retrieval, language query retrieval and movie summarization.

Consider the cross movie scene retrieval, where we are given a specific scene and asked to retrieve similar ones in other movies. For example, given a scene of person chatting

[2]Shot-based character occurrence counting is much less accurate than the scene-based character occurrence counting, since not all the consecutive shots contain the full face of a character though he/she is inside the scene.

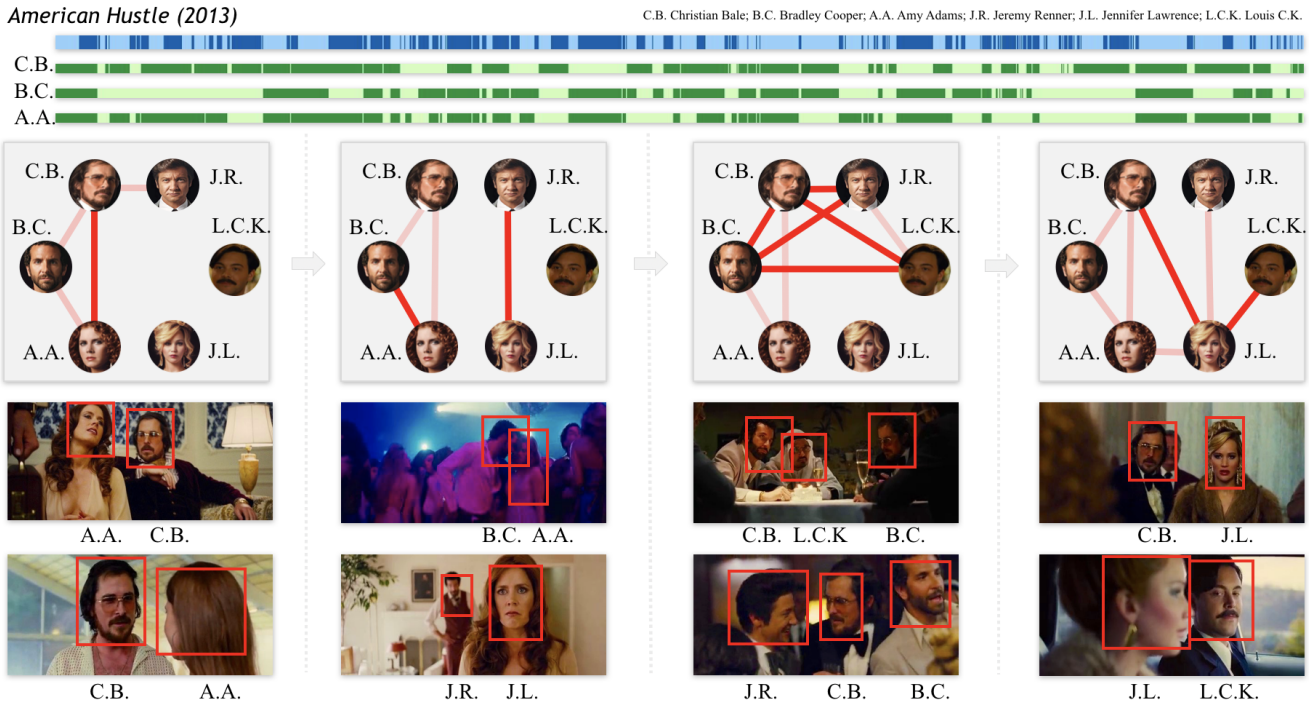

Figure 5. Human Interaction Graph. The first line is *American Hustle* scene segmentation coming from scene detection, where dark blue and light blue intertwine with each other to represent different Scenes. The second to fourth line corresponds to *C.B. Christian Bale; B.C. Bradley Cooper; A.A. Amy Adams* Scenes occurrence time lines in this movie respectively, where the dark green means occurring while light green does not. The graph below represents their interaction over the story line. The dark red represents a closer relationship while the light red represents a far-away relationship and two demo pictures are shown for closer relationships.

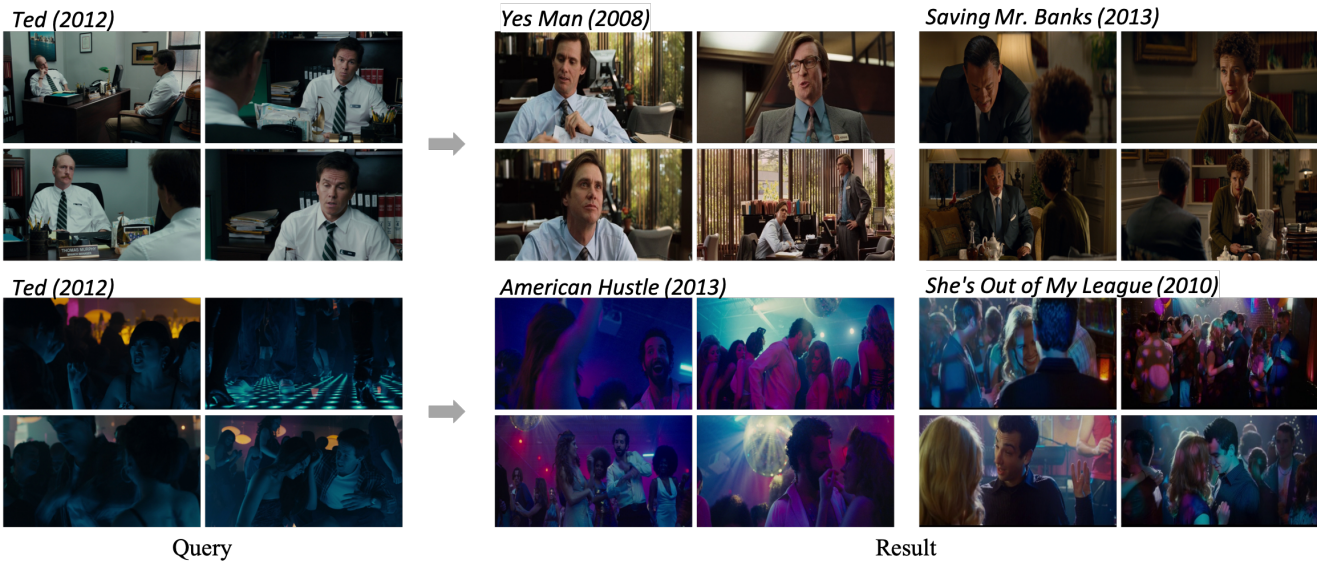

Figure 6. Cross movie scene retrieval. We choose a conversation scene and a party scene from *Ted (2012)* as query, then retrieved scenes from other movies.

in a room, we would like to retrieve all the similar scenes from all the other movies. The task is of great practical interest and has a variety of real-world applications, *e.g.* personalized marketing and intelligent searching. It demands a deep analysis of videos that goes beyond recognizing the visual appearance or simple action pattern. Compared to a single shot, the segmented scene allows us to extract rich correlated features (*e.g.* characters, specific objects, action,

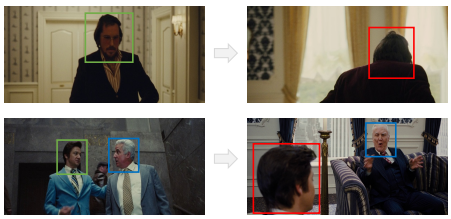

Figure 7. Character Recognition. The first row uses the character-scene relationship to recognize the character in red bounding box from the one in green bounding box. The second row uses the character-character relationship to recognize the character in red bounding box using the relationship between the one in green bounding box and the one in blue bounding box.

places, audio) in a self-contained semantic segment.

We show some qualitative results here. We extract multiple semantic elements features (place, cast, action, and audio) and compute similarity for every scenes pair. We take the top similar scenes as results. Some example results of the cross movie scene is shown in Figure 6. The retrieved scenes do contain the same semantic meaning as the query scene though they differs in the low-level visual and audio cues.

References

[1] 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, 3

[2] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1