

Curtin at Trecvid 2006 - Rushes Summarization

Ba Tu Truong and Svetha Venkatesh

Department of Computing
Curtin University of Technology
Perth, Western Australia
{truongbt,Svetha}@curtin.edu.au

1. Introduction

The focus of this paper is on the summarization of Rushes data set. This data set is best described as:

“Rushes are the raw material (extra video, B-rolls footage) used to produce a video. 20 to 40 times as much material may be shot as actually becomes part of the finished product. Rushes usually have only natural sound. Actors are only sometimes present. So very little if any information is encoded in speech. Rushes contain many frames or sequences of frames that are highly repetitive, e.g., many takes of the same scene redone due to errors (e.g. an actor gets his lines wrong, a plane flies over, etc.), long segments in which the camera is fixed on a given scene or barely moving, etc. A significant part of the material might qualify as stock footage - reusable shots of people, objects, events, locations, etc ” (Trecvid)

In this work, we propose an approach to the summarization of Rushes data by exploiting following features:

- Shot clusters – formed by SIFT feature matching.
- Interview shots – explicitly detected.
- Shots with dominant faces – using a face detector.
- Optical flow vectors.

Our summarization algorithm aims to extract a set of frames to represent a scene in the original clip. One clip often contains 1 to 3 scenes. Our approach is based on the assumption that a good summary, for the browsing purpose, should:

- *Concise*. We consider 1 to 10 frames to represent a scene in Rushes data is reasonable.
- *‘Stable’*. It should not contain frames that contain high motion, people moving in front of the camera, intermedia camera transitions.
- *Help the user identify following elements*.
 - Characters that appear in the scene.
 - Activity/events unfolding in the scene.
 - Place/location/setting.
 - Relative importance of each element in the scene

2. Implementation

2.1. Data-Preparation

Shot segmentation. A shot is the basic syntactic unit of a video sequence. In this work, shot boundaries are detected by a simple method of applying an adaptive-threshold on the discontinuity curve. For video sequences in Rushes data set, this simple method is highly reliable, giving only a couple of false detection and missed boundaries.

Keyframe extraction. We use a simple, efficient method for extracting representative frames of a shot. In this method, the first frame is always a keyframe, and the current frame is selected as a keyframe if its visual appearance significantly differs from that of the last keyframe. We also force the last frame of the shot as the keyframe.

Session/scene boundaries identification. A scene or a session in raw videos is delimited by temporal and/or spatial discontinuities. It is defined as a collection of consecutive shots captured at the same place and at the same time. Therefore, we can safely assume that the appearance (e.g. face, hair, clothing features) of the people in each session does not change across multiple shots. This enables the linking of the same subject in different shots/frames via the use of SIFT features as the appearance model. Here, we assume session boundaries are available as they can be marked by the user when filming using the built-in feature of the camera or by power on/off operations. Alternatively, given the time information associated with each shot we can easily define an effective classifier to automatically locate session boundaries.

Shot Clustering. In our recent work [1], robust shot clusters can be extracted via the use of SIFT features. Each cluster generally corresponds to one view of the action, possibly with different shot distances or camera focal lengths, and they often lie in the same side of a 180-degree axis. From summarization perspective, shots from different clusters should be used since they represent different viewpoints of the event unfolding.

2.2. Shot/Keyframe Characterization

2.2.1. Interview Shot Detection.

Interview shots in Rushes data set have following characteristics.

- *Long duration.* Since interviewer may ask a lot of questions and there are period of silences and preparation, the shot tend to be very long. Some last for the entire duration of the clip.
- *Faces.* The nature of interview shots means a large face is present in the shot. We have tried two face detectors. OpenCV and CMU NeuralNet. The first one detect all faces in the interview shots or at least one of its keyframe, with a lot of false positives while the second miss a lot of faces with better precision. We decided to use OpenCV detector and work on the distribution of the face position and size to differentiate between interview shots and non-interview shots.
- *No camera movements.* Interview shots barely contain any camera movement. We characterize movement of the camera for interview shots by computing the motion vector of the 20% of features with the smallest displacements. This is an easy way to avoid including local motion in the estimation.
- *Speech dominance.* Although some non-interview shots may contain certain amount of speech, it is much less dominant and less clean comparing to interview shots. Currently, we do not use any audio feature.

2.2.2. Relative Shot Distance in a Cluster.

Since all shots in a cluster produced by SIFT matching can be align directly to at least one of the shot in the cluster. We align all keyframes to one frame as a tree and use the alignment parameters to estimate the relative camera distance.

2.2.3. Face Detection.

In order to detect interview shots, we use the results of OpenCV for its high recall. For identifying shots with dominant faces for summarization, we use the CMU Neural Net face detector for its high precision.

2.2.3. Detecting Unstable Frames.

Some keyframes are not good to be included in the summary, we detect shaky, transition, people walking in front of the camera by looking at the number of optical flow vectors with the length above a certain threshold. While it is not accurate for the task, it gets rid of most unwanted keyframes.

2.3. The Summarization Algorithm.

The summarization is a process of selecting which shot and which keyframe from each shot to be included in the summary. The selection rule is a simplified version of utility-based framework in [2] and can be explained as follows.

1. If the shot does not belong to any cluster (ie., the shot is a cluster of itself), exclude it from the summary, unless:
 - a. It is an interview shot.
 - b. It contains and face is not found in any cluster.

In the case where all clusters contain single shot, just pick maximum 3 shots at random. The institution behind ignoring single-shot clusters is that shots of sufficient importance tend to be captured at different camera configurations.

2. Pick only one keyframe (shot) from each shot cluster. This is because if we pick more than one keyframes from each shot cluster, some of them tend to be redundant.
3. If the cluster contains no faces, select the keyframe (shot) with the longest camera distance (relative). This is because the cluster tends to present the setting/location, and for this we want to have the best overview of the scene.
4. If the cluster contains a large faces, select the most dominant one. Larger faces mean it is easier to identify characters.
5. The selected keyframes are ordered by the size of the cluster it belongs. However, the interview shot is always ordered first.

3. Evaluation

3.1. The Setup

For the summary of each scene in Rushes, we define following criteria for the evaluation:

1. Does the summary contain all main characters or how many out of the total number of main characters are captured in the summary?
2. Does the summary tell us about activity unfolding in scene? (out of 5)
3. Does the summary give us a sense about the location where the scene is captured? (out of 5)
4. Does the order of summary frames correspond to the relative importance of characters/objects and setting in the original video sequence.

5. How many frames are considered redundant and can be removed without affecting its effectiveness?
6. How many frames are considered missing and should be added to the summary to improve its effectiveness.
7. The subjective judgement of the summary quality (out of 5)

For now, we only manage to get four people to evaluate our summarization algorithm. They are all unfamiliar with the field and are given brief verbal explanation of the purpose of the summary generated and the data. (Ideally, we would like have someone who is working on this kind of data (production team, etc) to evaluate the effectiveness). They are presented with the original footage and a list of keyframes (first and last one of every shot) and the summary, and ask to judge the summary according to above criteria.

At the moment, we only process the training set of Rushes data. However, since we do not specifically tune the algorithm based on the data, the performance on the testing sequence should be very similar.

3.2. Summary Examples

Here we show the summary examples and our evaluation of the summary based on above criteria. We picked the scene from different classes that are typical in Rushes data set:

1. **Setting/Scenery.** All shots are setting shots, captured a building, city view and scenery from various angles and camera-distances.
2. **Interview.** All shots are interview shots of various people captured roughly at the same location and likely for the same topic.
3. **Object shots.** The cameraman tries to get different versions of the same object such as flags, signboards, statues, etc.
4. **Activity.** The purpose of the scene is to capture an activity, such as people working in an office, the street, the market.
5. **Acting sequence.** In these scenes, shots are retaken and characters act in the scene to create different versions for final production.

Many scenes, however, contain a mixture of above elements. For example, a scene can contain an activity and then the interview with one of characters that participate in the activity.

Each figure presented shows shot clusters (excluding single-shot clusters) and the summary. The average score for each scene is also presented, (-) mean the criteria is not applicable to the scene.

1. FRANC125 – Interview

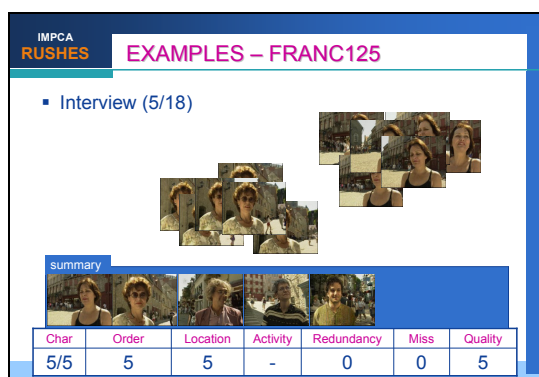
This scene consists purely of interview shots with 5 subjects.

The summary is perfect, containing one shot for each subject with correct ordering. There are three interviewers who are not in any cluster but still included in the summary, because they are detected as interview shots.

2. FRANCO78 – Food and Chief

This is a short scene which contains a few shots of foods and interviews with the chief in the room and outside.

The summary is not very good, since it does not recognize the same chief that appear in two separate shots. The perfect summary should contain only one shot which capture both the chief and the food on the table.



FRANC25 - interview



FRANCO075 – interview

3. FRANC125 – City view



This scene consists purely of setting shots, the city view. The city is captured with 17 shots at different camera configurations and time of the day.

The summary consists of only one shot. While two testers see that as sufficient, two testers think that it should include the city shot of the sun-set too.

4. FRANC105 – Buildings

Similarly to above scene, this scene also consists of purely setting shots. It capture various buildings.

Testers think that we should have only one shot to represent the interior of the building.

<p>IMPCA RUSHES</p> <p>EXAMPLES - FRANC125</p> <ul style="list-style-type: none"> Scenery/setting (1/27)  <p>summary</p> <table border="1"> <thead> <tr> <th>Char</th> <th>Order</th> <th>Location</th> <th>Activity</th> <th>Redundancy</th> <th>Miss</th> <th>Quality</th> </tr> </thead> <tbody> <tr> <td>0/0</td> <td>-</td> <td>5</td> <td>-</td> <td>0</td> <td>0.5</td> <td>4.75</td> </tr> </tbody> </table> <p>FRANCO72 – City building</p>	Char	Order	Location	Activity	Redundancy	Miss	Quality	0/0	-	5	-	0	0.5	4.75	<p>IMPCA RUSHES</p> <p>EXAMPLES - FRANC105</p> <ul style="list-style-type: none"> All scenery/building shots, many repeats  <p>summary</p> <table border="1"> <thead> <tr> <th>Char</th> <th>Order</th> <th>Location</th> <th>Activity</th> <th>Redundancy</th> <th>Miss</th> <th>Quality</th> </tr> </thead> <tbody> <tr> <td>0/0</td> <td>5</td> <td>5</td> <td>-</td> <td>0.75</td> <td>0.25</td> <td>4.75</td> </tr> </tbody> </table> <p>FRANC115 – City view</p>	Char	Order	Location	Activity	Redundancy	Miss	Quality	0/0	5	5	-	0.75	0.25	4.75
Char	Order	Location	Activity	Redundancy	Miss	Quality																							
0/0	-	5	-	0	0.5	4.75																							
Char	Order	Location	Activity	Redundancy	Miss	Quality																							
0/0	5	5	-	0.75	0.25	4.75																							

5. FRANCO72 – In banana farm


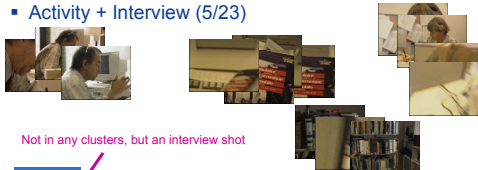
This scene shows activity of a person in a banana farm.

The summary contains 8 shots (from 45 shots), showing the farm setting, bananas and the main character. While the character is not belong to any cluster, it is included in the summary (and correctly so) because it is the only shot with dominant face. However, some consider the summary contains too many shots of bananas.

6. FRANC125 – In the office

This scene is an activity scene with one interview shot. It shows people working in an office.

5 frames are selected to summarize the original sequence of 23 shots. The summary is very good, capturing all essence of the scene. Some evaluators consider some shots (of a book close-up and the bookshelf) are redundant

<p>IMPCA RUSHES</p> <p>EXAMPLES - FRANCO72</p> <ul style="list-style-type: none"> Activity (2/6)  <p>summary</p> <table border="1"> <thead> <tr> <th>Char</th> <th>Order</th> <th>Location</th> <th>Activity</th> <th>Redundancy</th> <th>Miss</th> <th>Quality</th> </tr> </thead> <tbody> <tr> <td>1/1</td> <td>4.5</td> <td>5</td> <td>4.5</td> <td>2.25</td> <td>0.5</td> <td>3.5</td> </tr> </tbody> </table> <p>FRANCO72 – In banana farm</p>	Char	Order	Location	Activity	Redundancy	Miss	Quality	1/1	4.5	5	4.5	2.25	0.5	3.5	<p>IMPCA RUSHES</p> <p>EXAMPLES - FRANC115</p> <ul style="list-style-type: none"> Activity + Interview (5/23)  <p>Not in any clusters, but an interview shot</p> <p>summary</p> <table border="1"> <thead> <tr> <th>Char</th> <th>Order</th> <th>Location</th> <th>Activity</th> <th>Redundancy</th> <th>Miss</th> <th>Quality</th> </tr> </thead> <tbody> <tr> <td>3/3</td> <td>4.5</td> <td>4.75</td> <td>5</td> <td>1</td> <td>0</td> <td>4.75</td> </tr> </tbody> </table> <p>FRANC115 – In the office</p>	Char	Order	Location	Activity	Redundancy	Miss	Quality	3/3	4.5	4.75	5	1	0	4.75
Char	Order	Location	Activity	Redundancy	Miss	Quality																							
1/1	4.5	5	4.5	2.25	0.5	3.5																							
Char	Order	Location	Activity	Redundancy	Miss	Quality																							
3/3	4.5	4.75	5	1	0	4.75																							

7. FRANC103 – Gathering in the park

This scene shows activity of a group of people gathering and walking in the park, it contains interview shots of various people in the group.

The summary contains 8 shots (from 37 shots). These shots together depict the activity, the park setting, gathering and characters in the scene well. There is one character not included in the summary, because it belongs to the same cluster as another interview shot but of a different character (which is picked to be in the summary).

8. FRANCO78 – Island Tour

This scene is a composite scene of many elements. It contains establishing shots of an island and sea. It also contains shots of various activities of tourists such as on the boat, enter the car, walking, chatting. There is a central character in the scene, who was interviewed.

This is the scene where the strength of the summarization algorithm clearly shows. From the original of 42 shots, 6 shots are picked to be in the summary. The summary clearly shows the character and island and tourist setting. Some testers consider the shot in which people walk off the boat is redundant.

IMPCA RUSHES EXAMPLES - FRANCO103

- Activity + Interview (8/37) miss

summary

Char	Order	Location	Activity	Redundancy	Miss	Quality
5/6	5	4	4.75	0.5	1	4.5

FRANCO 103 – Gathering ..

IMPCA RUSHES EXAMPLES - FRANCO78

- Activity + setting + interview(6/42)

summary

Char	Order	Location	Activity	Redundancy	Miss	Quality
1/1	5	4.5	5	0.75	0	4.75

FRANCO78 – Island tour

9. FRANCO123 – Camping, eating on the hill

This scene shows a group of people camping and eating on a hill top. It also contains interview shots with a number of people in the group.

The summary is very good in terms of showing the location, activity and characters. However, the indication of the main character may be not correct, since the man with black/white shirt should be the main character since he appears in other shots as well. One shot of group eating can be considered redundant.

10. FRANCO104 – Hotel reception

This scene is purely an acting sequence with two main characters the guest and the hotel receptionist. Many shots are retaken and therefore redundant.

The summary contains 3 shots (from 17 shots). These shots summary the scene perfectly, it clearly shows the hotel setting and two main characters.

IMPCA RUSHES EXAMPLES - FRANCO123

- Activity + Interview (6/23)

summary

Char	Order	Location	Activity	Redundancy	Miss	Quality
3/3	4	5	5	0.5	0	4.5

FRANCO 123 – Camping,..

IMPCA RUSHES EXAMPLES - FRANCO115

- Acting (3/17)

summary

Char	Order	Location	Activity	Redundancy	Miss	Quality
2/2	5	5	5	0	0	5

FRANCO115 – Hotel reception

11. FRANCO76 – Market Scene

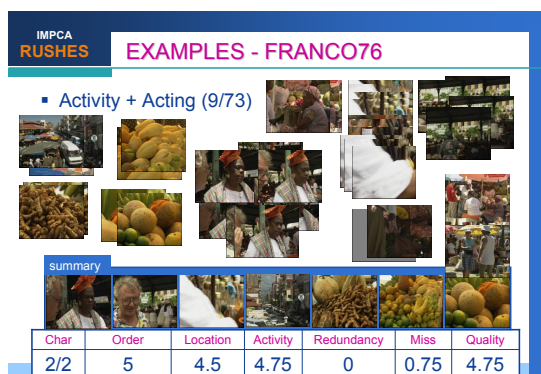
This scene captures activity of a marketplace, including setting shots, close-up shots of fruits and items on sale, sellers, buyers etc. There is a long acting sequence in this scene that models a conversation between a seller and a buyer as two main characters in the scene.

The summary, containing 9 shots out of a total of 73 shots, picks up these two main characters and correctly order them. Some grocery shots are also included together with setting shots enough to depict the entire scene. In our view, one more setting shot should be added to the summary to show the marketplace more clearly. Consider the complexity of the scene, the overall quality of the summary is very good.

12. FRANCO96 – Car broken down

This is a purely acting scene where an actor car is broken down and she phone for help.

While the summary correctly captures the main character, it fails to capture the context effectively. The close-up shot of the phone is likely redundant, whilst the shot of the car broken down and the actor sitting in the car should be included to realize the story better.



FRANCO76 – In the market



FRANCO96 – Car broken down

3.3. Comments

Overall the static summary produced by our technique works well for the Rushes data set. It hides most of repetitive and redundant shots and only present core elements of the scene. Testers comment that whilst it is easy to identify redundant frames, it is harder to identify missing frames from the summary set. Also, for interview sequences, the technique has problem of separating different interviewee if the interview location is the same. The technique work best of interview and acting sequences.

4. Conclusions

The proposed approach is very promising however this work is still very primitive. We plan to use a more robust face detector on Rushes data so that SIFT-based face clusters can further help the summarization through the utility-based framework we develop for home videos (to be present in Multimedia Modelling Conference). The scale of our evaluation needs to be extended and we are looking at the use of Flickr and YouTube as the evaluation platform and hopefully getting more users to participate in the process. Our work can be extended to produce moving summaries instead of static ones. For this, audio feature may be exploited and activity level within each segment (boat is sailing, people is walking) should to be an important factor in segment selection.

References

[1] Ba Tu Truong and Svetha Venkatesh (2007). Linking Identities and Viewpoints in Home Movies Based on Robust Feature Matching. Accepted for MMM07
 [2] Ba Tu Truong and Svetha Venkatesh (2007). Utility-Based Summarization of Home Videos. Accepted for MMM07.