Mid-Scale Shot Classification for Detecting Narrative Transitions in Movie Clips

Bipeng Zhang and **Arnav Jhala**
Computational Cinematics Studio
Department of Computer Science
University of California, Santa Cruz
{bizhang, ajhala}@ucsc.edu

Abstract

This paper examines classification of shots in video streams for indexing and semantic analysis. We describe an approach to obtain shot motion by making use of motion estimation algorithms to estimate camera movement. We improve prior work by using the four edge regions of a frame to classify *No Motion* shots. We analyze a neighborhood of shots and provide a new concept, middle-scale classification. This approach relies on automated labeling of frame transitions in terms of motion across adjacent frames. These annotations form a sequential scene-group that correlates with narrative events in the videos. We introduce six middle-scale classes and the corresponding likely sequence content from three clips of the movie *The Lord of the Rings : The Return of the King*, demonstrate that the middle-scale classification approach successfully extracts a summary of the salient aspects of the movie. We also show direct comparison with prior work on the full movie *Matrix*.

Introduction

Characterization of camera movement and movement of salient content within the frame play a critical role in content-based video indexing. Rhythm of camera movements extracted from video segments could provide interesting information about the nature of narrative events in the video from low-level features. Therefore, the estimation of camera movement and analyzing the content changes between adjacent frame is important for video analysis task.

Many algorithms have been developed on characterizing camera movement by a series of equations extracted with low-level frame analysis for estimating camera parameters (Andrey et al 2012, Xingquan et al 2002, Srinivasan et al 1997). The low-level visual features provide an elegant set for predicting camera movement without other semantic information about the content. For different types of low-level shots visuals in the frame change systematically over a number of frames within a neighborhood. The changing pixels indicate that there is movement in the frame. For detecting whether there is movement within the frame, or outside the frame, it is important to track the edge region of the frame to judge whether the camera is moving. Intuitively, another reason for tracking the edges rather than the whole frame is that under most circumstances the director may put the characters or meaningful scenes in the middle of the screen. Thus, to some degree, only tracking the edge regions exclude the interference arising from the movement of the characters in the middle. There are two methods for estimating changes of the edges in a frame, both of which are based on motion tracking analysis. Once we track the content in the frame, a model of the content's movement can be extracted with a movement vector by observing pixel changes in the neighborhood. The first method makes use of the sum of normal of every movement vectors. And the other one takes advantage of the sum of absolute value of pixels' changes for a same object. The former technique focuses on the relevant distance of the same object across adjacent frames, and the latter describes changes of the object within the frame.

The block motion estimation technique (Renxiang and Zeng 1994) searches for a motion vector between two adjacent frames that yields the minimum block distortion measure (BDM) within a neighborhood area. The picture is divided into small blocks. The algorithm runs for each block and traverses neighboring blocks according to a similarity metric. In this paper, we adopt an exhaustive search to achieve best match between the two frames. It means that the algorithm traverses all points within a

certain scope. Although the method needs more computation, for this work, we focus on accuracy of results rather than performance bottlenecks of the algorithm.

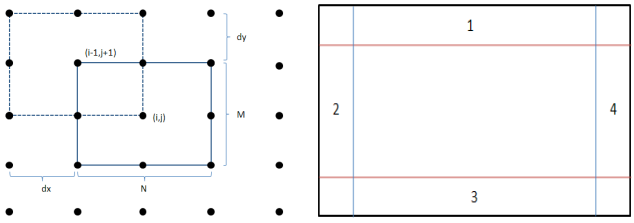
Adams et al (2001), followed this approach and introduced three shot motion classes: *No motion, Fluid and Staccato.* They demonstrated that these are useful in classification of shots in movies (Adams et al 2001). They defined a parameter **m** which denotes shot motion. First, based on a given threshold of the average **m** in the whole shot, they labeled shots with *No Motion* and *Motion*. Then each motion shot was classified as *Fluid* or *Staccato* by a threshold of the average of the derivative of **m** for the whole shot length. They chose three movies, Lethal Weapon2, Color Purple and Titanic, to test this method. Finally, they provided seven classes to distinguish a group of shots and presented the likely sequence content and corresponding causes of them from film grammar.

In this paper, we focus our analysis on *The Lord of the Rings: The Return of the King*, and analyze several clips from the movie. Choice of this movie is justified due to the variety of visual variations and narrative content. In terms of pacing, the movie has faster action segments as well as slower dramatic dialogue segments. It provides us with rich content including long dialogue, battle scene, smaller skirmishes, variety of lenses and transitions.

First, we describe our approach for achieving motion parameter **m** and provide a method to distinguish immediate transitions during a clip. Then we classify shots from the movie and introduce our enhancement, Middle-Scale classification, by analyzing three clips of the movie. Finally, we provide six middle-scale classes and show their likely sequence content in the table.

Background and Features

Motion tracking



(a) *(b)*

Figure 1: (a) the diagram to show Full Search Method (b) four edge regions of a frame

First, we introduce Full Search Method briefly (Figure 1(a)). The whole frame is divided into several M*N blocks. Choose one point in a block and define searching scope dx and dy in horizontal and vertical direction respectively. In the whole scope, we need to find (i_0,j_0) to achieve minimum Mean Absolute Deviation (MAD).

$$MAD(i_0,j_0)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|f_k(m,n)-f_{k-1}(m+i_0,n+j_0)\right|$$

where f_{k-1}, f_k refer to the blocks in previous frame and current frame that are to be compared.

Once we find such (i_0,j_0), it means that point A(i,j) in the current frame and point B($i+i_0$,$j+ j_0$) in the previous frame are the same point. Then we can also define the pixel changes of the same points in the adjacent frames by

$$DIFF(i,j)=f_k(i,j)-f_{k-1}(i+i_0,j+j_0)$$

where i_0 and j_0 may change for different blocks.

The DIFF defines changes with regardless of the distance between such two points. For each frame, we also calculate SDIFF, the sum of absolute value of each element in DIFF. The greater SDIFF means that there is a more obvious change for the same object in a pair of frames. The SDIFF(N), where N is the number of a frame in the whole shot, can define such trend in the whole shot. If we want to evaluate the camera movement, we should define a coefficient SNV which is the sum of normal of the vectors like $\overrightarrow{AB}$.

$$SNV=\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\left|\overrightarrow{AB}\right|/(dx^2+dy^2)^{1/2}\right)$$

where $\left|\overrightarrow{AB}\right|$ refers to the distance of the same blocks in the two adjacent frames.

The greater SNV means the object or the camera move faster. For the four edge regions (Figure 1(b)), we calculate four SNV_i(i=1,2,3,4). The camera will be considered moving only if the SNV_i (i=1,2,3,4) satisfy the corresponding threshold T_i(i=1,2,3,4). Like DIFF, we define the sum of normal of vectors in edge regions by SSNV, which is the sum of SNV_i (i=1,2,3,4) with subtracting the value of four corner regions. Also, the SSNV(N) can define the camera movement in the whole shot.

Otherwise, we take advantages of the derivative of SDIFF(N) and SSNV(N) to measure the rate of content change and camera movement in edge regions. We choose the average of SSNV(N) to judge whether there is camera movement in the frame and the derivative of SSNV(N) can help to judge whether the shot is Fluid or Staccato.

Detecting cuts with the derivative of SDIFF(N)

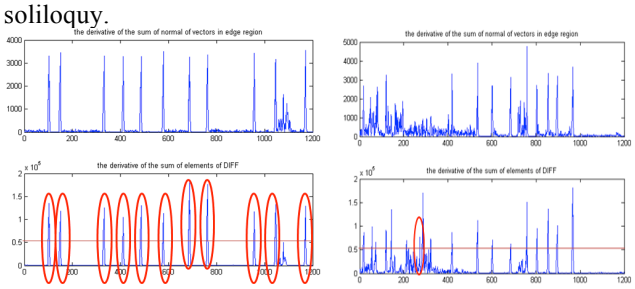
We use information from the derivative of SDIFF(N) to decide whether there is a cut. First, by a given threshold, choose pulses with greater values from the derivative of SDIFF(N). Those big values indicate that the content in the edge regions has changed a lot and such changes may result from the immediate transitions. For each pulse, we find its peak and assume that there is a transition between the corresponding frame and the previous frame.

Experimental Results

We chose six clips with varied narrative content from the movie. Each clip used in this analysis contained 1200 frames. Cut detection is the first step, followed by shot classification. Finally, we show how information from these steps can be used to analyse narrative content in a film.

Cut Detection

We illustrate our method for cut detection with two clips. The first clip (Figure 2(a)) shows that Gandalf and Pippin just came to Minas Tirith and talked to Lord Denethor. In the scene, the camera switches between the three people often. The second clip (Figure 2(b)) shows that Faramir defends orcs beside the river in the north of Minas Tirith. In this clip, there is a fierce fight at first. Then the humans are defeated and finally the leader of the orcs has a short soliloquy.



(a) (b)

Figure 2: in both (a) and (b) the picture above is the derivative of SSNV(N) and the picture below is the derivative of SDIFF(N).The line indicates the threshold to detect cuts.(a) the cuts detected are labeled with circles in the clip containing dialogue mainly (b) the wrong cut detected labeled with the circle in the clip containing battle and dialogue

In Figure 2(a), we show transitions we detect with the method. In Figure 2(b), there is a false positive as the labeled cut does not exist in the second clip. The reason is that there is a big change between the two adjacent frames during the shot. So the labeled pulse results from the changes of the content rather than camera movement. Otherwise, there are some pulses which occur in the second clip but we could not find corresponding pulses in SSNV(N) (For each of other pulses in SDIFF(N), we always could find a corresponding pulse in SSNV(N)). A possible reason is that the tracking algorithm doesn't work well in those frames as we only try to find a best match block within a certain scope and such block can always be found based on the algorithm, however it may not be what we really want. Such circumstances usually happen when the camera moves very fast. Table 1 shows our results of detecting the cuts in six clips.

Table 1 the results of detecting cuts in the six clips

Clips' type	correct	incorrect	total
dialogue	20	0	20
Battle with dialogue	17	1	18
Long lens with fast moving	12	4	16
Character moving with dialogue	25	0	25
total	74	5	79

Shot Classification

As described in Adams et al (2001), we use the same low-level features to classify camera movement (Adams et al 2001). In this technique the average motion parameter **m** is estimated using and *No Motion* and *Motion* categories are selected based on a threshold. Within the shots classified in the Motion category, an average of derivative of the motion parameter **m** is used to further classify *Fluid* and *Staccato*.

In this paper, based on Brett's principle, we propose a improved principle with our approach to achieve motion parameter **m**. The main intuition for this enhancement is to add a preprocessing before calculating the average of shot motion that provides more information about the next classification stage. We consider only the four edge regions of a frame. Regardless of particular camera moves, the four edge regions change whenever there is camera movement. We divide the frame into blocks and calculate both the displacement of the block and content changes within each block. We take advantages of the derivative of SDIFF(N) to decide whether there is an immediate transition in a clip. The property of function SDIFF(N) is its range, which results in a clear distinction between the extreme changes in shots. However, it is inappropriate for choosing a threshold in those small values to classify whether the camera moves or not because of its limited range in the small values.

We classify the camera movement by following criterion:

1) By a given small threshold T1, judge whether objects in all four edges region of a frame changes.

2) Calculate the average of SSNV(N).

3) By a given threshold T2, decide whether the shot is No Motion or Motion.

4) Calculate the average the derivative of SSNV(N).

5) By a given threshold T3, decide whether the shot is Fluid or Staccato.

The first criterion helps in some circumstances that the camera does not move however there is some object moving in the screen. If the object is not big enough to make changes in all four edges, then we can judge there is no camera motion in the frame. Here we consider that the camera may have a very small movement though it is

stable and no object moves obviously in the frame. Therefore, we assume a threshold T1 =T_i (i=1,2,3,4)=20 in this paper to avoid misjudgement, which means that if one of SNV_i(i=1,2,3,4) is smaller than T1, we still consider it as zero.

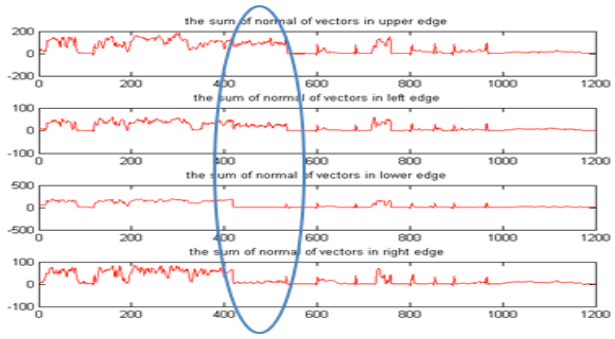


Figure 3: the SSNV(N) in four edge regions in a frame and the labeled circle indicates there are two regions changing a lot and the other two ones do not.

Figure 3 shows SNV_i(i=1,2,3,4) in the four edge regions of a clip of the movie and it is an good example to illustrate the criterion 1). This is the clip in which Faramir defends from orcs beside the river in the north of Minas Tirith. The labeled close-up shot shows one of soldiers down and dying and a team of orcs are walking towards him. The camera does not move but due to object movement within the upper and left region, the SSNV in the shot changes a lot. Now, due to the first criterion, This shot is classified as *No Motion*. However, if we only consider the SSNV of the whole edge region, the shot should be classified as Motion, which is a misjudgement.

If all four SNV exceed T1, for specifying No Motion and Motion, we calculate the average of SSNV(N) and give the threshold T2 based on experiments.

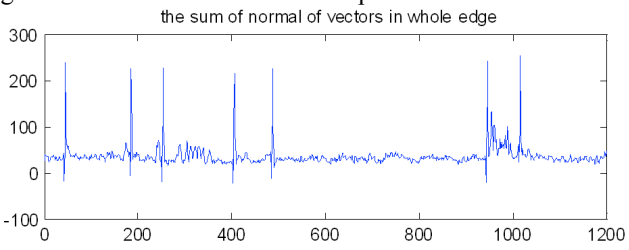


Figure 4: the SSNV(N) in the whole edge region of a group of No Motion shots with unstable background in the clip

Figure 4 shows a clip, which is a dialogue between Gollum and his shadow in the river. Due to the flow of the river and the movement of the character's head, all the SNV_i(i=1,2,3,4) are above the threshold T1. However, truth is that the camera does not move. Thus, we need an another threshold T2 to classify Motion and No Motion by the average of SSNV(N) in criterion 3). From the figure (d), we assume the threshold T2 is 60.

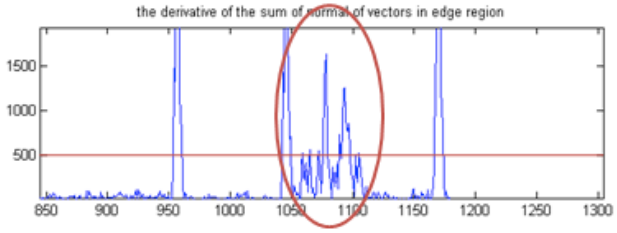


Figure 5: the derivative of SSNV(N) in the whole edge region and the labeled circle indicates a Fluid shot

Figure 5 shows a segment of another clip and the labeled shot indicates a camera lens from far and near, which can be classified as Fluid. Therefore, we assume 500 as the threshold T3 to distinguish Fluid and Staccato in criterion 5).

Table 2 shows the results of classifying all shots in the six clips. The correct classification for No motion, Fluid and Staccato, is 93%, 87% and 67% respectively. Both too fast and too slow camera movement have a negative effect on the classification.

Table 2 the results to classify all shots in six clips

Clips' type	No Motion	Fluid	Staccato
dialogue	19	1	0
Battle with dialogue	8	5	5
Long lens with fast moving	4	7	5
Character moving with dialogue	13	10	2
total	44	23	12

The middle-scale classification and scene content analysis

Adams provided seven classes based on the three basic classifications, No Motion (N), Fluid (F) and Staccato (S), to explain films' rhythm. He distinguished the groups of shots by the seven classes, N, F, S, NF, NS, FS and NFS (showing in Table 3), with considering what kind of basic classes (N,F,S) involve in the neighborhoods of shots and the number of basic classes in the group. It can be considered as he analyzes the rhythm of movies in a big scale as every groups may contain hundreds of shots based on his method in a whole movie. Such big-scale classification ignores the order of three basic classes. In fact, such order may convey a certain meaning from a filmmaker and the ordered combinations can be considered as a sort of sense-group which provides us a middle-scale (several shots with certain meaning) between the small-scale (only focus on one shot) and big-scale (regard hundreds of shots as a group) to analyze rhythm of films and index them.

Here we provide six middle-scale classes with N, S, F

under a certain order, like NFN, SFS, FNF, SNS, NSN and FSF(showing in Table 4). When we consider about the middle-scale classification, for the six classes, they should satisfy one pre-condition that the three fundamental classes(N, S, F) in each of them should have casual constrains between each other. In detail, we predict that the class N shots extracted from middle-scale classes FNF and SNS may reflect the results of the group of shots or clues of plots afterward. The class F or S extracted from middle-scale classes NFN and NSN will describe scenes and atmosphere or the potential meaning of the movie. And the middle F or S extracted from SFS and FSF indicates the pace changes in a neighborhood of shots.

Table 3 the seven classes in Brett's paper[(Adams et al 2001)]

Class	Likely Sequence Content	Cause (Film Grammar)
N	Dialog, small location	No natural motion, and no need to contrive it
F	Long, progressive establishing scenes	Continual fluid movement requires a continual transfer of attention, rare requirement. Extended periods rare
S	Violent/extreme sequences	Extreme periods rare, very taxing
N/F	Buildups, establishers, cinematic pieces, ...	E.g. Establishing: fluid motion allows audience examination of location
N/S	Fights, emotionally charged or exuberant sequences, ...	E.g. Split perspective, action vs. reaction shots
F/S	Frenetic pieces, action sequences, chases, ...	Dual cinematic perspective (e.g. in car (environmental) – tracking/stationary shot) (plus a need to not tax the viewer by inundating them with constant staccato/environmental motion...
All	Transitions, between both scenes and events, ...	Transition sequences involve movement across different locations, expanded angle/shooting possibilities and the need to convey info about the path taken (fluid/establishers), ... (also, Artifact of window grouping of abrupt changes)

Table 4 The six middle-scale classes and corresponding sequence content

Middle-scale class	Possible sequence content
SNS,FNF	During long progressive scene or extreme sequence, for showing the result or clue of some movement or plot, which helps to push story .
NSN,NFN	During dialogue or stable sequence, for revealing characteristics of roles or potential meaning , which riches connotation and expression in a movie.
SFS,FSF	During a battle or a long progressive scene, for showing details or displaying magnificent scenes, which changes the pace of movie.

Based on the specific scene-group indicated by middle-scale classes, we make use of them to analyze movies' rhythm. The basic thought is to extract every middle-scale classes during a movie. In Figure 5, we give a completed algorithm from cuts detection to post-processing. In the post-processing part, mainly we give each of six middle-scale class a weight. By such processing, we not only can distinguish different camera motion by the corresponding weights but also can achieve results with different meanings by different weight strategy.

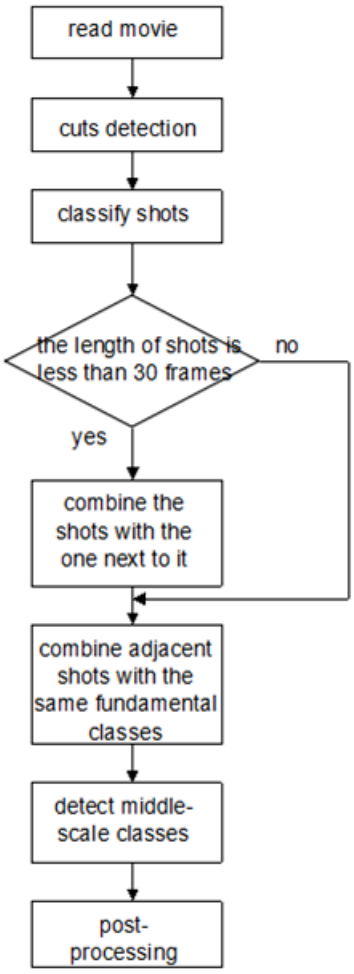


Figure 5: the completed algorithm for analyzing the rhythm of movies

For testing our Middle-scale classification theory, we did two experiments. Our goal for the pilot experiment was to describe how middle-scale features help in analysis of rhythm of movies. The second experiment directly offers comparison with Adams et al's results to show improvement due to mid-scale features.

In our first experiment, we analyze the three clips from the movie, *The Lord of the Rings: The Return of the King,* with the middle-scale theory. The Figure 6 shows the derivative of SSNV(N) of each of three clips with labeled

the Middle-scale classes we defined before.

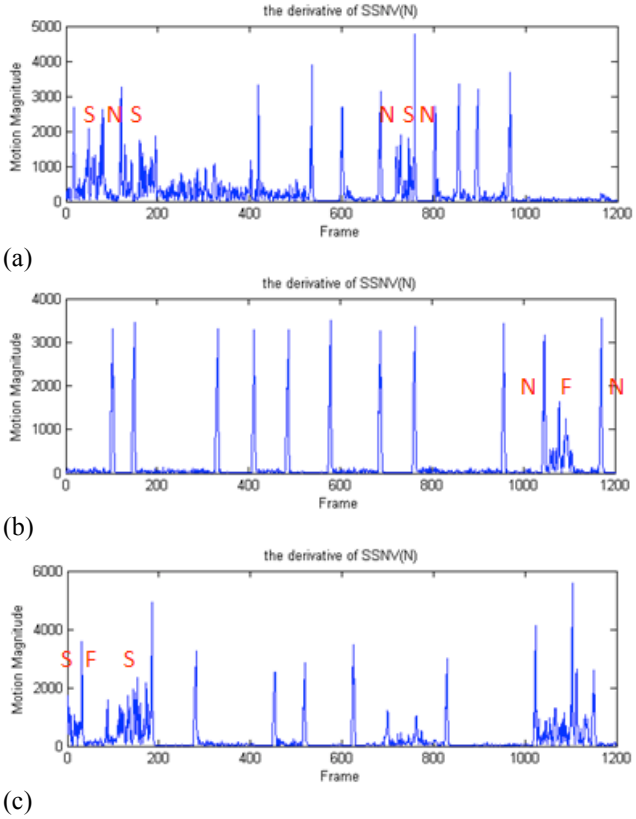


Figure 6: (a) the clip contains battle and dialogue; (b) the clip contains dialogue and close-up lens; (c) the clip contains long progressive scenes

The clip in Figure 6(a) describes the scene where humans are defeated in a battle againts orcs. The discontinuous shots in this segment express fear and helplessness against a strong enemy. The *No Motion* shots here mainly exert the function of pushing the story. The first labeled group of shots is classified as SNS. During the series of shots, there is a *No Motion* shot describing a human solider down, which paves the way for the later plot that the solider will be killed. The class SNS has particular characteristics for filming such a sequence. The N shot is important under this circumstance as it embodies some intention of the filmmaker.

For the class NSN, the important part is the S. The S between two No Motion plots should reveal some intentions of the director, which may have causal relationship with the two N shots next to the class S. In this neighborhood of shots, the first N shot describes a team of orcs have come to an injured human solider. The second close-up N shot shows the hands of the orc and a spear next him. And in the S shot the director pans the lying solider to express his despair and at that time the lying solider is an epitome of the human beings. Such scene-group is just like NFN, for which we could find an example in Figure (6)(b). The F shot is from far and near to express the Lord Denethor's anger. And such his behavior has direct casual connection with the two No Motion plots next to which is Gandalf's opinion and Denethor's refutation.

And the last clip in Figure (6)(c) gives us an example of the class SFS. The shots describe that Piping and Gandalf ride through the field on the way to Minas Tirith. The middle F shot slows down the pace of the group of shots. Both of S shots convey a sense of hurry, however, the director inserts a F shot into them, which makes the group of shots distinguish from some extreme sequence like SSS.

In the second experiment, we applied our method to the whole movie, The Matrix. After the detection, we use 1,2,3 referring to three fundamental classes and use other six numbers to indicate middle-scale classes. The bigger number means more fierce camera motion in corresponding shots. For comparing with the result in Adams's paper, we use four red circles to show the similar parts and one rectangle to show mainly different part in both pictures in Figure 7. From macroscopic point, we could achieve similar results to prior work.

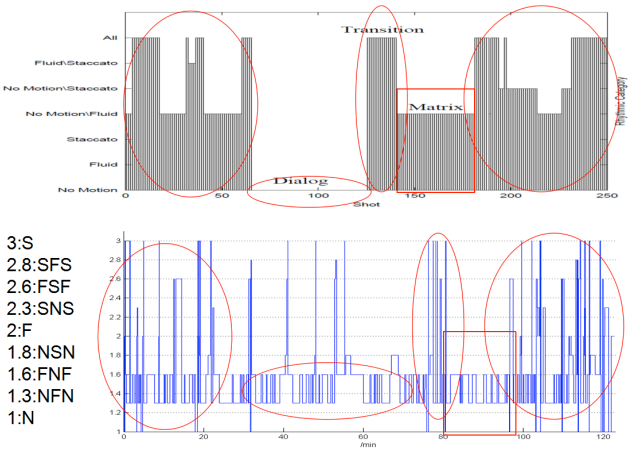


Figure 7: the rhythm of the movie, The Matrix, coming from Adams's method and our method respectively. The first picture is Adams's result. The second picture results from Middle-scale classification theory. The four red circles in each of pictures indicate the same parts in the movie. The two rectangle shows mainly different results coming from the two methods.

On further analysis, the mid-scale approach reveals more details. For example, for the dialog part in Adams et al's result, they judged the neighborhood of the shots is No Motion class. In our result, we can detect several extreme camera motion during this episode. Also, the results arising from our method is more consistent. Here the notion of consistency is interpreted as the classification of similar

narrative segments (such as conversations) in the same class. In the Matrix part labeled in the first picture in Figure 7, the scenes include the traitor is wiped out and agent Smith interrogates Morpheus, both of which are formed by several long conversations. From the view of content of a plot, the labeled Matrix part should be the same as labeled Dialogue part. But with Adams's method, the two parts are divided into different classes.

The reason for this is that Adams et al used the difference in shots within a sliding window to generate different phases during a whole movie. However, in some scenes, there may be a very long conversation, just like the Dialogue part in The Matrix, with many cuts resulting from the successive camera switch between each characters. The increasing number of such cuts makes the detection not classify the action part due to the very limited number of action shots. Furthermore, in the Matrix part, due to excessive camera motion, the method just counts the number of shots with fundamental classes instead of considering the casual constrains, which results in the different judgment from the Dialogue part. On the other hand, our method considers the relationship between the adjacent shots, which means it does not focus on the number of a certain class but detect what kinds of classes are involved in sequential shots. Because in both Dialogue part and Matrix part, most shots in the two parts belong to class N and class F, based on our method, they belong to the same genre.

For Adams's method, although we could decrease the width of sliding window to achieve more details in the dialogue part, it is at the cost of accuracy on other parts. Also, too narrow window improves recall but affects precision.

Our method provides more flexible control to visualize results. We can do different post-processing to approach the goal for different purpose. For example, in Figure 8, we assume the weight of both middle-scale class NFN and fundamental class N as number 1 to achieve more clear view. The reason that we can change the weight of class NFN is that fundamental class N can be regarded as NNN which is similar with middle-scale class NFN. For another example, when class N and F appear alternatively, due to the initialization of the algorithm, class FNF and NFN can be regarded as the same class. Thus we also can emphasis the action part in the movie by setting the weight of middle-scale class FNF as 1 (Figure 9). Until now the weights of classes are in accord with the average camera motion in the shots belonging to the corresponding classes. We also can reverse the weight to emphasis slow phase in the movie. Therefore, based on user's goals, we could give different values to each of six middle-scale classes to discover rhythm of movies from different views.

Here is another example. In Figure 10, it is the part result of the movie, The Lord of Rings: The Return of The King. In the movie, from the 90th minute, most scenes describe the biggest battle in the movie, defense of the Minas Tirith. We could see there are many extreme camera movements in this part.

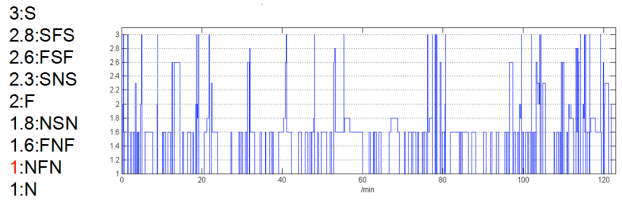


Figure 8: Showing the rhythm result of The Matrix, which arises from different weight strategy. In this picture, we assume the weight of class NFN equals that of class N

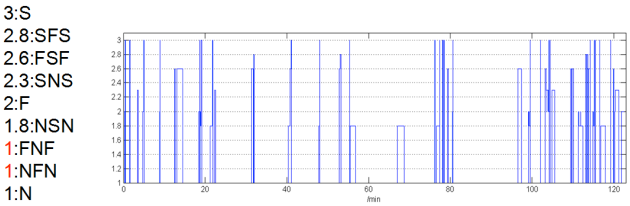


Figure 9: Showing the rhythm result of The Matrix, which arises from different weight strategy. In this picture, the weight of class FNF, NFN and class N are equal.

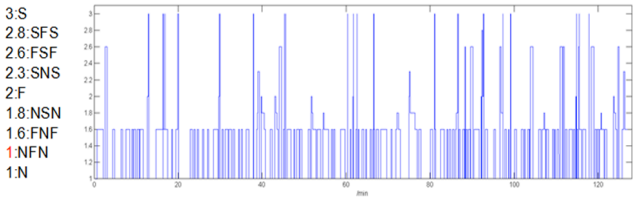


Figure 10: Showing the rhythm result of The Lord of Rings:The Return of the King with the same weight bewteen class NFN and class N.

Conclusion

We implemented previously proposed method of motion tracking in video segments to classify shot types – and subsequently narrative events. We make use of motion tracking method to detect camera movement and achieve a good level accuracy for detecting cuts and classifying the shots. Further, we provide middle-scale classification theory and introduce a new algorithm and as set of features to analyze film rhythm. Results show that mid-scale features perform well in detecting cuts and identifying shot types. Further, mid-scale features improve performance of narrative event identification proposed in prior work.

There are several limitations in our method. Here we do not discuss evaluation of the method with respect to

precision and recall (F-Scores) for detecting fundamental classes. Also, the camera motion pattern of movies are recorded and judged by ourselves. While it is easy to define No Motion class, we don't have a gold standard to compare against. Most narrative analysis is done manually for this paper. One direction of future work may be to have a user study to help us to decide shot types.

Another area of work is in broadening this analysis to different types of movies in terms of length, genre, platform, etc. There might be characteristic rhythm features that can be learned automatically from a corpus of annotated videos across genres. With limited extension of this work it could be applied to indexing and summarization of videos through identification of interesting segments.

References

Andrey, Vavilin., and Kang-Hyun, Jo. 2012. *Camera Motion Estimation Based on Edge Structure Analysis.* New Frontiers in Graph Theory.

Xingquan, Zhu., and Xiangyan, Yue. eds. 2002. *Qualitative Camera Motion Classification for Content-Based Video Indexing.* Advances in Multimedia Information Processing.

Srinivasan, M., and Venkatesh, S, and Hosie, R. 1997. *Qualitative extraction of camera parameters*. Pattern Recognition.

Renxiang, Li., and Bing, Zeng. 1994. *A new three-step search algorithm for block motion estimation.* Circuits and Systems for Video Technology, IEEE Transactions.

Brett, Adams., and Chitra, Dorai., and Svetha, Venkatesh. *Automated Film Rhythm Extraction For Scene Analysis.* 2001.ICME.