Mining Semantic Structures in Movies

Yuya Matsuo[1], Kimiaki Shirahama[1], and Kuniaki Uehara[1]

Graduate School of Science and Technology, Kobe University,
Nada, Kobe, 657-8501, Japan
{yuya, kimi, uehara}@ai.cs.scitec.kobe-u.ac.jp

Abstract. Video editing is the process of selecting and joining various fragments of video material (called shots) to create a final video sequence. In an editing process, there are many possible ways to make a transition from one shot to another. So, the quality of a created video depends on the editor's skills, that is, the quality of the video created by professional editors is much higher than that of amateurs. Importantly, professional video editors carry out video editing based on their own editing patterns in order to successfully convey their intention to viewers.

In this paper, we concentrate on extraction of useful editing patterns from movies by applying data mining technique. The patterns extracted from movies are called 'semantic structure'. We propose two methods to extract two types of semantic structure about the connections between consecutive shots, and the relation between character's appearance and what is happening to him/her. Finally, based on the extracted semantic structure, our video editing support system [2] suggests hints to amateurs how to produce a new, more attractive video.

1 Introduction

Ever since we have developed the video editing support system [2], we have been engaged in discovering useful editing patterns from movies by applying data mining techniques, called 'video data mining'. Movies are communicative media, and there are a lot of editing patterns which have been employed for a long time and known to successfully convey editor's intention to viewers. In fact, a famous and popular movie persuasively conveys the idea of the movie editors. Furthermore, a movie generally consists of a large number of fragments of video material (called shots), which are joined together during the editing process. It is possible to mine the editing patterns specific to the movie. For the reasons mentioned above, we extract useful editing patterns from movies, called 'semantic structure'.

A movie is one of the multimedia which is the integration of video and audio media, both known as 'continuous media'. One continuous media consists of a sequence of media quanta (i.e. video frames or audio samples), and it conveys meaning only when media quanta are continuously presented in the appropriate order and timing. Moreover, these video and audio streams need to be synchronized so that a desired audio is heard when the corresponding video frame is presented. We organize such a movie into a multi-stream, where each single stream is a sequence of metadata derived from visual and audio features. Then, we apply data mining technique to the movie to extract the useful editing patterns as semantic structure.

2 Semantic Structure

We plan to mine the following three types of semantic structure from movies.

Syntactical Association Rules: These are called ‘video grammar’ to define the syntactical connections between consecutive shots. That is, we don’t consider the semantic content but only technical features of each shot. Specifically, a shot is indexed by the following ‘raw-level’ metadata: *shotsize*, *camerawork*, and *duration*. An example of a syntactical association rule is that, two shots whose *shotsize*s are extremely different cannot be connected. In that case, viewers are confused by the visual leap. In [1], we have already extracted such syntactical association rules from movies.

Semantic Association Rules: These are the detailed syntactical association rules, which depend not only on the syntactical connections between consecutive shots, but also on the semantic content connections of consecutive shots. A shot is indexed by the above raw-level metadata and the following ‘semantic-level’ metadata: *Character*, *BGM (BackGround Music)* and *Sound*.

Topic Continuities: A ‘topic’ is an interval, where one meaningful episode of a target character is presented, that is, his/her topic is continuous. For example, in a topic, a character talks to someone, or he/she makes love with a partner. Such topics are marked by character’s appearance and disappearance. After detecting topics, we extract ‘topic aspects’ which are character’s appearance and disappearance patterns used by the movie editor to present certain types of semantic concepts.

3 Mining Semantic Association Rules from Multi-Stream Data

In this section, we focus also on semantic-level metadata in order to extract the semantic association rules. For example, in the event where the leading character and the secondary one are standing face to face and fighting, the editor uses the leading character’s close shot (TS) with long duration many times, effectively using background music as well. This is one of the semantic association rules revealing the intention of an editor to emphasize on the leading character.

The following definitions show the attributes of the metadata:

Shotsize: *Shotsize* is selected according to the distance from the camera to the objects. *Shotsize* is classified into loose shot (LS), medium shot (MS) and tight shot (TS).

Camerawork: *Camerawork* means camera movement including its direction, such as Fix, RightPan, UpPan, ZoomIn, and so on.

Duration (*sec*): $Duration$ (= $(EndFrameNo - StartFrameNo)$ / 30) means the duration of a cut. The value of *Duration* is classified according to the distribution of each cut's length. The cut duration plays an important role to convey the meaning of the cut.

Character: The attribute *Character* represents the character who is displayed on the screen. In movies, editors may change the editing pattern according to each character. So in each scene, the most important character should appear in many cuts, and the character's cut duration must be relatively long.

BGM: The attribute *BGM* represents the background music in the cut. The editors employ BGM with a certain intention, so the BGM influences the rhythm of the scene. For example, the scene during the same BGM is assumed to have a constant rhythm. Therefore, even in the case of the scenes of the same type, the editing rhythm changes whether the same BGM is present or not in the scene.

Sound: The attribute *Sound* represents the sound in the cut. Here, it means the sound that adds some effects to the scene, such as a gunshot, an explosion, a strike and so on. The *Sound* also influences the rhythm of the scene. For example, the scenes that have many sounds are assumed to be the battle scenes, and therefore they should be edited in a fast rhythm.

Cut No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**		
Shotsize	LS	MS	TS	MS	LS	MS	TS	TS	LS	MS	MS	LS	MS	TS	.	.
Camerawork	fix	fix	fix	pan	fix	pan	fix	fix	fix	fix	pan	fix	fix	fix	·	·
Duration[sec]	1.5	2	3	4	2	3	2	3	6	3	10	2	4	5	·	·
Character	A	B	A	-	A	B	A	B	A	-	A	A	B	A	·	·
BGM	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	-	-	-	-	·	·
Sound	Y	-	-	-	Y	Y	-	-	Y	Y	Y	Y	Y	-	·	·

time

Fig. 1. Video stream indexed by the metadata.

As illustrated in Fig. 1, we formulate the video stream that is indexed by the metadata. Cut *No. 1* represents the content that "*Shotsize* is *LS*, *Camerawork* is fix, *Duration* of Cut *No. 1* is 1.5 *seconds*, the *Character* name displayed on this cut is *A*, and *BGM* and *Sound* are present in this cut". Since the pattern extraction algorithm from the multi-stream data needs to consider vast amounts of candidate patterns in the data, the work of extracting significant patterns from multi-stream data takes huge amount of time. Therefore, it is essential to develop an effective algorithm which reduces the amount of redundant calculation.

We present a new mining method to extract the significant patterns from multi-stream data. Our method accepts a set of multi-stream time-series data as input. Each element of the pattern whose length is 1, $x_i = (v_1, v_2, \cdots, v_6)$ represents the metadata that can be identified by using our mining method. In order to express the pattern over multiple cuts, it is required to express it as a combination of x_i chronological patterns. That is, we denote the pattern over multiple cuts whose length is n as the form $y_i = \{x_1, \cdots, x_n\}(n > 1)$. The following shows the procedure of our searching algorithm.

1. **Generate the candidate pattern whose length is 1.**
   First of all, our method generates the group of the possible candidate patterns whose length is 1, after scanning the symbols in multi-stream data.
2. **For each candidate pattern, determine the searching position in the data using Boyer-Moore approach.**
   Boyer-Moore approach [3] is known as one of the fastest string matching algorithms. For pattern $P\ (= p_1p_2 \cdots p_m)$ and text $T\ (= t_1t_2 \cdots t_n)$ of length m and n respectively, it can complete the string matching in average time order complexity $O(\frac{n}{m} \log_k m)$, where k is the number of symbols. For each candidate pattern generated in procedure 1, determine the searching position where the current searching pattern may occur, using Boyer-Moore approach focused on a single stream data.
3. **At all the positions determined in procedure 2, perform matching between each candidate pattern and the original data, using dynamic programming algorithm.**
   The dynamic programming algorithm (DP) for approximate string matching to compute the edit distance between two strings A and B of length m_1 and m_2 respectively computes a matrix $C_{0 \dots m_1, 0 \dots m_2}$ that consists of m_1 columns and m_2 rows. The value $C_{i,j}$ holds the edit distance between $A_{1 \dots i}$ and $B_{1 \dots j}$. A_i is the symbol of A at i-th position. $A_{i \dots j}$ represents a substring of A enclosed between i-th and j-th symbol.

$$C_{i,0} \leftarrow i, \quad C_{0,j} \leftarrow j \tag{1}$$

$$\begin{aligned} C_{i,j} \leftarrow \mathit{if} \quad & A_i = B_j \quad \mathit{then} \quad C_{i-1,j-1}, \\ & \mathit{else} \quad 1 + \min(C_{i-1,j-1}, C_{i-1,j}, C_{i,j-1}) \end{aligned} \tag{2}$$

   The edit distance $ed(A, B)$ is the final value of C_{m_1,m_2}. The rationale of the formula is that if $A_i = B_j$ then the cost to convert $A_{1 \dots i}$ into $B_{1 \dots j}$ is the cost of converting $A_{1 \dots i-1}$ into $B_{1 \dots j-1}$. Otherwise, we have to select one among three choices: (a) convert $A_{1 \dots i-1}$ into $B_{1 \dots j-1}$ and replace A_i by B_j. (b) convert $A_{1 \dots i-1}$ into $B_{1 \dots j}$ and delete A_i, or (c) convert $A_{1 \dots i}$ into $B_{1 \dots j-1}$ and insert B_j. At all the positions determined in procedure 2, we perform the same matching process.

In Fig. 2, (1) search the symbols “ABC” that are the candidate pattern of stream 1, by applying Boyer-Moore approach focused on stream 1. (2) perform matching from stream 2 to 5 between the candidate pattern and the original data by the dynamic programming algorithm at the five places obtained in (1). During matching, count the places in the original data whose edit distance to the candidate pattern is at most the number of allowed errors. These are used to measure the significance of each pattern in procedure 4. From Fig. 2, we can say that the most appropriate alignments in each stream by DP are $\{C, D, E\}$, $\{H, I, J\}$, $\{P, Q, R\}$ and $\{X, Y, Z\}$, respectively.

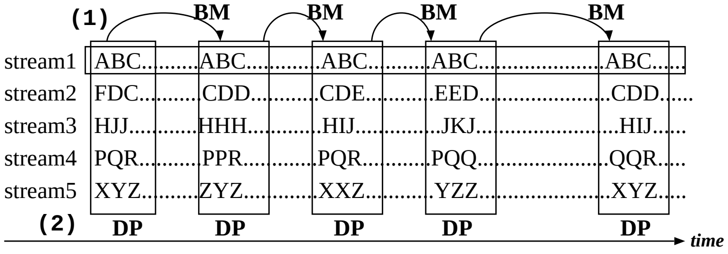

(1)
BM
BM
BM
BM
stream1 ABC.........ABC..........ABC..........ABC.....................ABC.....
stream2 FDC.........CDD..........CDE..........EED.....................CDD.....
stream3 HJJ..........HHH..........HIJ...........JKJ......................HIJ.....
stream4 PQR.........PPR..........PQR..........PQQ.....................QQR.....
stream5 XYZ.........ZYZ..........XXZ..........YZZ.....................XYZ.....
(2)
DP
DP
DP
DP
DP
time

Fig. 2. The mining method using Boyer-Moore approach (focusing on stream 1) and Dynamic Programming.

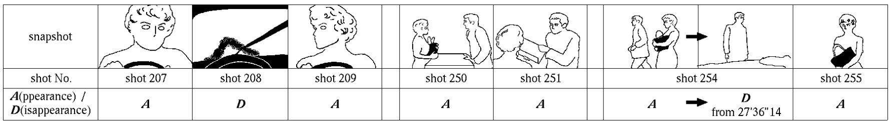
4. **Using *J-measure* that measures the significance of the pattern, and remove the non-significant patterns from the group of the candidate patterns.**
   In this procedure, by adopting *J-measure* shown in the following equation (3), we evaluate the importance of each pattern.
   $$J(B;A) = P(B|A)*\left\{P(B|A)\log_2\frac{P(B|A)}{P(B)} + (1-P(B|A))\log_2\frac{1-P(B|A)}{1-P(B)}\right\} \quad (3)$$
   In the above equation, $P(A)$ and $P(B)$ represent the appearance probability of A, B in time-series data. $P(B|A)$ represents the value of conditional probability that shows the appearance B when A appears. The first term of this equation is the weight of the adaptive flexibility in the rule, the second term is the variation of the probability distributions of the consequential part by the appearance of the conditional part. It shows the intensity of the relevance between the conditional part and the consequential one. By using *J-measure*, patterns that don't satisfy the preset threshold value are regarded as non-significant, and therefore can be removed from the group of the candidate patterns.

5. **Generate the candidate patterns whose length is incremented by** 1**, and repeat procedure** 2 **until candidate patterns exist.**
   Generate the new candidate patterns whose lengths are incremented by 1, using the remaining candidate patterns. Go back to procedure 2.

We proposed a method for mining the technical semantic structure based on shots to extract patterns. By adopting the above mining method, we can cut about 40% of calculation time compared with the method described in [1].

4 Extracting Topics based on Character's Rhythm

In this section, we describe a technique for extracting the semantic structure as perceived by the viewer. Character's appearance and disappearance patterns are essential for conveying the editor's intention to the viewer. Fig. 3 represents the several shots from Alfred Hitchcock's movie "PSYCHO". Here, there are two characters, *Marion* and *Norman*. In the bottom row of the table in Fig. 3, *A* and *D* indicate *Marion*'s appearance and disappearance in a shot, respectively.



snapshot									
shot No.	shot 207	shot 208	shot 209		shot 250	shot 251		shot 254	shot 255
A(ppearance) / ***D***(isappearance)	***A***	***D***	***A***		***A***	***A***		***A*** → ***D*** from 27'36"14	***A***

Fig. 3. Several shots from Alfred Hitchcock's movie PSYCHO

In a movie, each character definitely has his/her own 'action flow'. We will describe *Marion*'s action flow by using Fig. 3. In *shot 207* of *Marion*'s appearance, her 'action' is driving her car. In *shot 208* of *Marion*'s disappearance, looming light is to be seen outside her car as the 'surrounding response' to her previous action. A surrounding response means the response by other characters (*Norman* at the end of *shot 254*), or changes in the surrounding setting (*shot 208*), when she disappears from the screen. Thus, character's action flow conveys to a viewer what action he/she performs during his/her appearance, and what surrounding response occurs during his/her disappearance.

Furthermore, each action or surrounding response must be consistent with the previous one. In *Marion*'s appearance in *shot 209*, she glances outside while driving her car as a 'reaction' to the surrounding response in *shot 208*. Another form of character's reaction can be found in *Marion*'s appearance in *shot 251* (*Marion* is listening to *Norman*) as a reaction to her action in *shot 250* (*Marion* is talking to *Norman*). While maintaining these consistent connections of character's actions and surrounding responses, his/her action flow continues until he/she no longer appears in the movie. Through character's action flow, a viewer comprehends what is happened to that character.

Meanwhile, movie editors intertwine characters' action flows to build the story of the movie. At this stage, they pay special attention to how long a character's action is presented. The duration of a character's action (i.e. his/her appearance) is determined based on the context of his/her situation and action. Similarly, the duration of a surrounding response (i.e. character's disappearance) is determined by the context of other character's situation and action in regard to main character's previous action. Consequently, we can assume that, as long as a character performs a particular action in a similar context of his/her situation, the durations of his/her appearance don't vary so much and the same is true for the durations of his/her disappearance. Therefore, in a topic, character's 'rhythm' which consists of durations of his/her appearance and disappearance, stays roughly constant.

To realize the above concept, we index metadata for each shot, representing the intervals where a character appears in the shot and the intervals where he/she disappears. Some metadata indexed for *Marion*, are shown in Fig. 3, where except *shot 254*, she keeps appearing or disappearing during a shot (indexed by only A or D). Then, *Marion*'s rhythm from *shot 194* to *267* can be drawn as shown in Fig. 4, by scanning durations of her appearance and disappearance one by one. Starting from the left in Fig. 4, the circles correspond to *Marion*'s appearance or disappearance in *shot 207*, *208*, *209*, *250*, *251*, *start of 254*, *end of 254*, and *255*, respectively (Fig. 3). A positive value indicates the duration of *Marion*'s appearance, and a negative value indicates the duration of her disappearance. Note that, if some durations of *Marion*'s disappearance are continuous, we join those durations into one total duration. Because *Marion* reacts to the surrounding response which has occurred in the total duration.

The best example reflecting our assumption of character's rhythm is the interval from *Marion*'s *198th* to *226th* disappearance (*topic 2*). In *topic 2*, the durations of *Marion*'s appearance and disappearance are shown as alternating positive and negative values of the same amplitude, that is, her rhythm is relatively constant. Moreover, this interval is considered as the topic associated with the semantic concept (*Marion* drives her car at night). But, character's rhythm may be rugged in a topic, because one semantic concept may consist of more precise sub-concepts. For example, in *topic 3* (*Marion* calls *Norman* after reaching his motel), *Marion* walks around looking for *Norman* and she waits impatiently beeping the horn. So, we need to divide a movie into topics where character's rhythm is relatively constant with some degree of error.

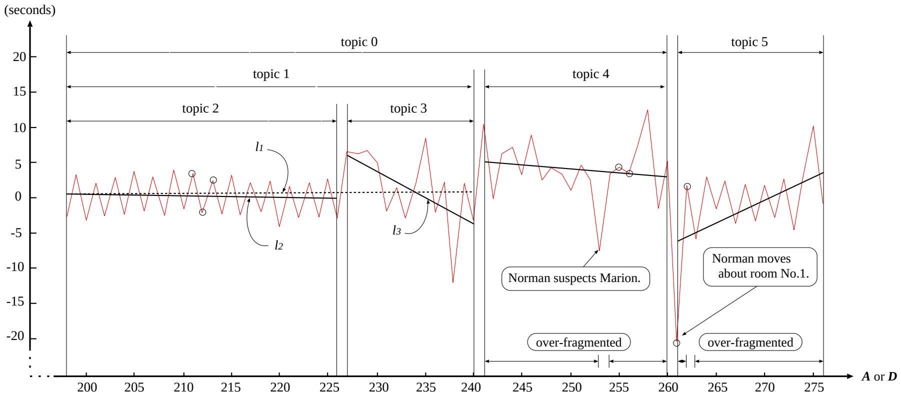


Fig. 4. Our rhythmic model of character's actions and surrounding responses

For character's rhythm, we figure out topics where his/her rhythm keeps roughly constant, by employing time series segmentation method [4]. This method recursively divides a segment of character's rhythm into two sub-segments in a top-down approach, beginning with the whole range of his/her rhythm. In each recurrence, our method considers every possible division of a segment into two sub-segments, while approximating character's rhythm in each sub-segment by the least-square method. Then, our method finds the optimal location within the segment where the sum of the approximation errors in the two sub-segments is minimum. If dividing the segment at the optimal location results in a significant approximation (i.e. the reduction of approximation error is more than the pre-specified threshold value), our method divides this segment at this location. Otherwise, our method no more divides the segment.

We concretely show the above process of our method with *Marion*'s rhythm in Fig. 4. Now our method is considering *Marion*'s rhythm in *topic 1* approximated by l_1. Clearly, *Marion*'s rhythm is difficult to be regarded as constant. In order to approximate *Marion*'s rhythm more precisely, our method divide *topic 1* into *topic 2* and *topic 3* where her rhythm is approximated by l_2 and l_3, respectively. In consequence, the approximation error can be reduced significantly. Next, our method considers *Marion*'s rhythm in *topic 2*. In *topic 2*, the positive and negative values of *Marion*'s rhythm alternate with similar deviations from l_2, in other words, her rhythm is constant enough to eliminate the need for further division. That is, even if our method further divides *topic 2*, a significant reduction of approximation error can't be expected.

Note that, in *topic 1*, *Marion*'s rhythm remains relatively constant, although from a broader perspective. By minimizing the sum of deviations of *Marion*'s rhythm from the approximated lines, l_1 and l_4, the pair *(topic 1, topic 4)* is chosen as the optimal pair within *topic 0*. This means that *Marion*'s rhythm in *topic 1* and *topic 4* is as constant as possible within *topic 0*. Since our method selects such an optimal pair of sub-topics in each recurrence, we can obtain the binary tree consisting of topics where character's rhythm is constant at a multiple-abstraction level.

5 Experiments

5.1 Semantic Asscociation Rules

We implemented the method stated in section 3 to extract the semantic association rules. The movie used for this purpose is "Star Wars Episode V" directed by George Lucas. This movie is one of the representative works in a modern movie, and has the aspect of a speedy deployment. The scenes indexed by the metadata are the battle scenes in the movie. The following shows the examples of the semantic association rules that are extracted by our mining algorithm.

Presence of *BGM* and *Sound* in the scene: The editing patterns are influenced by whether BGM or sound is present or not in the scene. That is, even in the case of the scenes of the same type, BGM scenes are edited in a fast rhythm compared with non-BGM scenes. Non-BGM scenes totally contain a lot of long conversation scenes and should be edited in a slow rhythm. On the other hand, the scenes with BGM or sound contain a lot of quick camera motions combined with quick motions of the character, so the scene is obviously edited in a fast rhythm.

Character's appearance: The editing patterns are influenced by the type of the character. Using rapid transition of the shotsize ($LS \Rightarrow TS, TS \Rightarrow LS$) frequently, the editor shows the leading character in the close shot (TS) with long duration to emphasize him in the scene. We can find that the editor intentionally uses rapid transition of the shotsize, and accents on the scene in order to emphasize a certain character. This kind of editing technique is prohibited in classic video grammar [2].

Intelligible connection: Especially in a fast rhythm scene, the editor connects cuts, avoiding the rapid change of direction between adjacent cuts. For example, if the camera pans right combined with a right motion of the character, it is inappropriate to follow up the character using left-pan in the subsequent cut. That is, the connections such as $RightPan \Rightarrow LeftPan$ and $UpPan \Rightarrow DownPan$ are not allowed in a fast rhythm scene, because such rapid change confuses viewers.

5.2 Topic Continuities

In classical Hollywood movies, movie editors faithfully observe the principles of video editing to concentrate viewer's attention on the story. We analyze two Alfred Hitchcock's movies, "PSYCHO" and "North by Northwest", in order to extract topic continuities of *Marion* and *Roger*, respectively.

We first describe whether topics detected by our method described in section 4, are really associated with a semantic concept or not. It is actually inevitable that our method excessively divides some topics into smaller fragments, which can't be considered to form a semantic concept individually. This is caused by the temporary change of character's rhythm to emphasize the situation, where character's action (or surrounding response) is presented in a long duration. Otherwise, the topic forms its semantic concept by combining some fragments where the character performs actions of different nature. For example, *topic 4* in Fig. 4 is over-divided at *Marion*'s *253th* disappearance, where *Norman*'s suspicion of *Marion*'s previous action is emphasized. In the case of *topic 5* (*Marion* is talking to *Norman* in the room), *Norman* moves around showing the room to *Marion* (*261th* disappearance) before their conversation starts. Even in such a topic, our method estimates that character's rhythm is relatively constant, because this topic is the parent of the fragments in the resulting tree. In this way, character's rhythm must not be drastically changed with no special reason to present a new semantic concept in a movie.

We realize the following association between character's action and its duration: if a character actively moves, the action is presented in a long duration, and on the other hand, a short duration yields if he/she hardly moves. The meaning of this association is that presenting an uneventful action in a long duration bores viewers. Furthermore, we extracted the following interesting topic aspects which are specific ways to present certain types of semantic concepts:

Uneventful topic aspect: When a character performs the uneventful actions during a topic, such uneventful actions are presented with a fast and constant rhythm. For example, when a character talks to someone else while standing still, the speech is very short in duration. Otherwise, the reaction of the listener is presented while someone else keeps on speaking to him/her.

Thrilling topic aspect: In spite of the active movements of a character and the surroundings (e.g. he/she is desperately running away from the enemy), such active actions are presented in short durations, which makes a viewer more excited.

Romantic topic aspect: In spite of the inactive movements of a character and his/her partner (they are mainly hugging and kissing), these actions are presented in very long durations to create a romantic mood. However, when a character talks to his/her partner, their actions are often presented in much shorter durations than those characteristic for the love moments. So, character's rhythm is very distorted in a romantic topic.

6 Conclusion and Future Work

In this paper, we proposed the methods of pattern extraction by using data mining techniques from movies. In Section 3, we proposed the mining method using both raw-level and semantic-level metadata in order to extract the semantic association rules. About the metadata BGM, we extracted the semantic structures by determining whether BGM is present or not. But it may be assumed that the rhythms or the patterns of editing change according to the types or the volume of BGM. Since editors add such semantic-level metadata with a certain intention, it is necessary to consider how to index the metadata more carefully.

In Section 4, we proposed a method for extracting topics based on character's rhythm. We uncovered some factors affecting character's rhythm, such as character's movement and the type of topic. These factors act on character's rhythm in different ways. While character's movements determine the durations of his/her appearance, the type of topic is related to his/her overall rhythm in a topic. Taking into account these different factors, we aim to classify the characters' rhythms from various topics in a unified way. Once achieved, this enables the use of the action type and the topic type as search keys for rhythms. By editing video material based on the rhythms close to our desired rhythms, we can produce a new enhanced video.

References

1. Y. Matsuo, K. Shirahama and K. Uehara: Video Data Mining: Extracting Cinematic Rules from Movie. In Proc. of 4th International Workshop on Multimedia Data Mining MDM/KDD. (2003) 18-27
2. M. Kumano, Y. Ariki, M. Amano, K. Uehara, K. Shunto and K. Tsukada: Video Editing Support System Based on Video Grammar and Content Analysis. In Proc. of 16th International Conference on Pattern Recognition. (2002) 346-354
3. R. Boyer and S. Moore: A Fast String Searching Algorithm. Communications of the ACM. volume 20 (1977) 762-772
4. V. Guralnik and J. Srivastava: Event Detection from Time Series Data. In Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (1999) 33-42