AUTOMATED FILM RHYTHM EXTRACTION FOR SCENE ANALYSIS

Brett Adams[†], *Chitra Dorai*[‡], *Svetha Venkatesh*[†]

Department of Computer Science[†]
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{adamsb, svetha}@cs.curtin.edu.au

IBM T. J. Watson Research Center[‡]
P.O. Box 704, Yorktown Heights
New York 10598, USA
dorai@watson.ibm.com

ABSTRACT

This paper examines film rhythm, an important expressive element in motion pictures, based on our ongoing study to exploit film grammar as a broad computational framework for the task of automated film and video understanding. Of the many, more or less elusive, narrative devices contributing to film rhythm, this paper discusses motion characteristics that form the basis of our analysis, and presents novel computational models for extracting rhythmic patterns induced through a perception of motion. In our rhythm model, motion behaviour is classified as being either *nonexistent*, *fluid* or *staccato* for a given shot. Shot neighbourhoods in movies are then grouped by proportional makeup of these motion behavioural classes to yield seven high-level rhythmic arrangements that prove to be adept at indicating likely scene content (e.g. dialogue or chase sequence) in our experiments. Underlying causes for this level of codification in our approach are postulated from film grammar, and are accompanied by detailed demonstration from real movies for the purposes of clarification.

1. INTRODUCTION

A large body of literature termed film grammar records rules and conventions commonly used in film. It gives insight into the crafting process that produces a film, and is essential to "thinking the filmmaker's thoughts after him," as it were. In our previous work, we have outlined how a systematic use of the knowledge of film production from all of its aspects can be exploited to build tools for the automatic understanding of films [3, 2, 1]. This level of inspection undergirded by tacitly and widely followed traditions and policies is vital to treating the raw data of film with integrity, as opposed to bringing a false set of interpretive rules to bear upon it.

In [3, 2, 1] we investigated the extraction of *tempo* or *pace* from film. Based on a thorough survey of film literature on the matter, we developed software tools with the ability to track and analyze the tempo of a film. Extracting certain features from this novel measure, we demonstrated that the dramatic development of a number of films can be reconstructed in their entirety for further use as semantic indexes for video summarization and search.

Another aspect of film that is present, albeit in a more elusive state, in film grammar, is that of film rhythm. Film rhythm has been referred to by some work in the field of automatic feature extraction from video and film (e.g. [5]), but attempts to extract it automatically are not forthcoming. Our work, is the first attempt at an algorithmic study of film rhythm — its constituents, types, and its linkages to scene content and story narration.

2. FILM RHYTHM

Movie watchers and makers alike attest to the existence of a rhythm of film. In its most generic definition, rhythm is an "organization of time" ([6, p. 90]. Film, being both a construction, and firmly tied to a timetable in at least one sense (i.e. running time, see [10, p. 246] for a list of *times* present in a film), naturally gives rise to a rhythm.

All sources, and indeed common sense, indicate that film rhythm can be very complex. Such a rich and versatile medium, film, having the ability to convey information both visually and aurally is bound to give rise to very intricate time relationships. [4] states that "the issue of rhythm in cinema is enormously complex and still not well understood," however, it goes on to say that it roughly involves "a beat or pulse, a pace, and a pattern of accents, or stronger and weaker beats". [4] lists "movement in the mise-en-scene, camera position and movement, the rhythm of sound" as well as *editing*, as many contributors to overall film rhythm.

Thus it would seem that, despite the fact that film rhythm in all its complexity is very hard to analyze, there remain some aspects which can be computed and will be of use in the process of automated film understanding. From the list above, the two elements that offer themselves as the most likely candidates for automatic extraction by simple means are editing patterns and motion characteristics. In this paper we restrict ourselves to the study of visual rhythm arising out of shot motion behaviour, and outline an approach for automatic rhythm extraction and classification. We derive three shot motion tendencies: *No Motion*, *fluid*, and *staccato*. Next we examine the persistence or lack of these fundamental types across shots, and categorize the neighbourhood behaviour into seven rhythmic types. Finally we show how each rhythmic type is typical of a particular scene content and its link to film grammar.

3. VISUAL RHYTHMIC ELEMENT – MOTION

Bordwell and Thompson [4] note that "frame mobility involves time as well as space, and filmmakers have realized that our sense of duration and *rhythm* is affected by the mobile frame". They later state that frame velocity can create expressive qualities, that "a camera movement can be fluid, staccato, hesitant, and so forth".

3.1. Classifying Shot Motion Behaviour

Taking a cue from such pointers, we have focused our work on classifying the motion behaviour of a shot as one of either *No Motion*, *Fluid Motion*, or *Staccato Motion*. No Motion class is

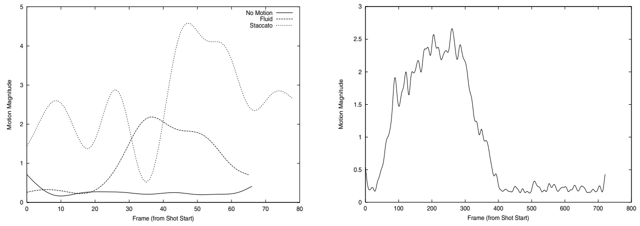
2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

self-explanatory, but the other two categories warrant some further examination.

Fluid and Staccato are terms that are well known in the realms of music appreciation. They are essentially opposites, and are discrete at the note level (i.e. a note transition can be indicated as either slurred or not), but form more of a spectrum at higher levels (e.g. phrase). A fluid transition implies a continuous change from one state to another. The distance between two states is achieved without jumping as it were. The change is not a pronounced one. Staccato transitions, on the other hand, draw attention to themselves by this property of pronounced change. Such transitions are of a much more discrete nature; they are abrupt.

When applied to motion within the frame, these definitions bear out in the following manner. The transition that is pertinent here is that of one motion velocity to another. We are interested in the characteristics of the change. Following from our above definitions, a fluid transition will be a smooth change of speed, while a staccato change will be abrupt (it is important to note here that we are not interested in the absolute level of the speed, that is a matter for tempo/pace [3]). The dominant feature in this analysis then is the first derivative of the motion speed. For a given transition, the higher the first derivative of motion speed, the more abrupt the change, and vice versa. Figure 1(a) shows samples of the three motion behaviour classes.



(a) (b)

Figure 1: (a) Examples of the three shot motion behaviour classes; (b) A long shot with well defined motion and no motion segments; both from "Lethal Weapon 2".

In summary, the motion categories may be defined as follows where $\bar{m}$ denotes average shot motion:

$$\begin{array}{ll} \text{No Motion:} & \bar{m} < T_1 \\ \text{Fluid:} & |\frac{\partial \bar{m}}{\partial t}| < T_2 \\ \text{Staccato:} & |\frac{\partial m}{\partial t}| > T_2 \end{array}$$

3.1.1. Method

Given a digital movie, an index of shot boundaries (specifically *cuts*) is created by means of the commercial software *WebFlix*. The raw shot motion data is computed by the *ptrz* algorithm used in [3], which produces a frame by frame estimate of camera movement. The raw pan and tilt computed are then filtered of anomalous values and smoothed with a sliding window. The smoothing artifact employed is a size 9 Savitzky-Golay window ([7]). This smoother was chosen because, in this case, faithful extraction of motion flow is of importance. Such a smoother is a good choice in the presence of non zero 2nd order motion.

Motion behaviour classification is essentially a two stage process. Shots are first labeled either *No Motion* or *Motion*, and then each Motion shot is classed as being either *Fluid* or *Staccato*.

Average motion (i.e. averaged sum of pan and tilt values for the shot) is used to determine whether the shot contains motion or not. A threshold is set, high enough to account for any small object contributions or slight framing corrections ([8, p. 223] on the cameraman's role: "Real camera movements are only made when authorized by the director, and further than that they are nearly always called for by him rather than anybody else"), but low enough to identify *real* camera movements (or close-up motion).

The Motion shots are then classified as either Fluid or Staccato by the following method. The first derivative of the motion is calculated for the duration of the Motion shot. The average of the first derivative magnitude is then taken for the shot, so as to normalize the value for shots of differing lengths. This value is then subject to a threshold to determine whether it is fluid (below the threshold) or staccato (above). Different policies could be applied for determining the threshold value and currently it is simply the average of the normalized value for all Motion shots in the given movie.

3.1.2. Experimental Results

In order to evaluate the effectiveness of our technique for categorizing shot motion behaviour, a ground truth list of shots for several movies was first constructed. Table 1 shows the varying levels of each type in some sections of film, taken from arbitrary sections of differing length from the three movies listed. Carrying out this task was useful in that it highlighted, at times, the difficult nature of the task. A shot can be short or long, it can have one dominant camera motion or many, it can have motion due to the environment or close-up movement, it can have one direction or many, and the classification can be influenced by elements such as the mood or tone of a sequence. However, that said, recognizing the somewhat spectral nature of the motion/no motion and fluid/staccato classification, the majority of shots can still be noted as one of no motion, fluid or staccato with a good level of confidence.

Table 1: Ground truth shot motion behaviour in film sections.

Film	No Motion	Fluid	Staccato	Total
Lethal Weapon 2	467	36	97	600
Colour Purple	192	36	49	277
Titanic	99	10	14	123

Table 1 indicates that no motion shots are generally the largest class of shot present in a given movie. The reader is refered to Table 2 for classification success on sections comprising about a half of both movies, The Mummy, and The Matrix. The motion threshold (T_1) is generally reliable in our experiments (85% plus correct classification), with the misclassified shots being of dubious class. The dubiousness lies in the question as to at what point an object becomes large enough, and hence the impact of its motion on the viewer, to cause the shot to be best designated as a motion shot. This is a difficult problem to solve as it has to do with two separate problems: The first being the inability to calculate scene depth (and hence the absolute motion of the camera/object) which feeds into the second problem, that being the difficulty in determining the psychological response to the identified movement. Does a close-up of slight motion have the same effect on a viewer as a long shot of physically larger motion? Such questions are beyond our scope, and found to be of only marginal impact on this analysis and our results.

With the figures from Table 1 we found that the maximum resolution of motion shots into either fluid or staccato by this measure using a threshold (T_2) for the movies analyzed is about 60 to 70%. On closer analysis, many of the shots that were being

Table 2: Classification results for half of The Mummy, and the Matrix.

Film	No Motion	Fluid	Staccato	Total
The Mummy	427/493	309/318	256/286	992/1097
The Matrix	680/765	314/333	281/316	1275/1414

wrongly classified were of long duration, with enough motion to be labelled as motion shots, but with significant periods of essentially no or small motion (see Figure 1(b)). Having classified a shot as containing motion (regardless of how long the shot is), the next step should be to analyze the characteristics of that motion. A better result should be obtained if sections of *No Motion* within the shot are ignored, with the analysis being focused on the sections of motion. For example, if a shot of long duration contains a pan at its beginning and contains no further purposeful camera motion, then this pan should be singled out and examined as to whether it is abrupt or fluid.

To that end, a further step was added to our automatic process: Sections of the Motion shot that are above a time threshold (reflecting the physical limitations of real camera movement and adding robustness to handle noisy fragments) and below a motion threshold are masked from further computation. The effect is that the 1st derivative for a freshly determined motion shot is normalized for the motion duration, not the shot length. The resulting maximum accuracy of classification by the original threshold applied to this improved process rises to 75-85% and above. The reader is again referred to Table 2 for an example of results of the fluid/staccato classification from The Mummy, and The Matrix, with the improved measure. For these results, the fluid/staccato threshold (T_2) was set automatically to the average of the 1st deriv. measure for all motion shots.

In future, fluid shots of short duration (of the order of 1 second) will be designated as staccato. Perceptually, the effect of such a shot is staccato, as it involves an abrupt jump to a different speed.

A final note on motion classification. There is another fundamental motion type that we have not attempted to classify here. It may be designated as either *Environmental* or *Subjective* motion. Typically produced by a hand-held, or roughly mounted camera, the effect is always to cause the viewer to feel immersed in the action [9]. While such shots are closest to staccato in nature, they often result in a fluid classification due to the averaging of the erratic pan and tilt components. We are currently working on methods of distinguishing this motion automatically.

3.2. Determining Rhythmic Classes from Shot Arrangements

Since rhythm is formed from the arrangement of a number of shots, the next step was to group and classify a *neighbourhood* of shots. We chose to classify the (sliding) neighbourhood centered on a shot by the percentage makeup of shot motion behaviour types. The following classes were defined:

1. Predominantly No Motion (NM)
2. Predominantly Fluid (F)
3. Predominantly Staccato (S)
4. Mixed, No Motion and Fluid (NM/F)
5. Mixed, No Motion and Staccato (NM/S)
6. Mixed, Fluid and Staccato (F/S)
7. Equally mixed, No Motion, Fluid, and Staccato (*All*)

The classes labelled "Predominantly" (No Motion, Fluid, Staccato) occur when 75% or more of the sliding window covers consistent type of shot behaviour. The "Mixed" classes occur where the two categories (e.g. no motion and fluid) are both between 25% and 75% of the window size. The "Equally Mixed" class label is given when a window satisfies none of the above criteria. The size of the sliding window is set to 41 shots and it was found to be about the order of the "clumps" of shots of interest, i.e. scenes. This value merely affects the level of detail for analysis.

From our initial experimental observation, this classification technique yielded useful information. Over 20 movies have had this process applied to them so far in their entirety, and the resulting rhythmic labels have been observed as being useful for finding sequence (scene) transitions, particularly if the neighbourhood size used is large (i.e. targeted at large trends).

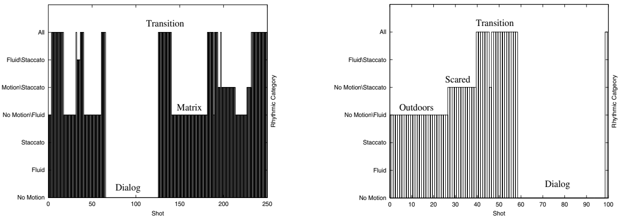
As a demonstration of the usefulness of the motion neighbourhood classifier, we undertook an experiment that locates persistent motion neighbourhood changes and compares them against a groundtruth of content change; see Table 3. Content is considered to have changed at scene and sequence boundaries (e.g. dialog to transition, or build-up to fight).

Table 3: Content change detection using motion neighbourhood.

Film	Content change	Correct det.	False pos.	False neg.
The Mummy	66	47	2	19
The Matrix	94	71	9	23

The reasonably high false negative rate is due to the fact that the groundtruth is fine grained in places, and consists of some categories that are semantically quite similar. This is due to the fact that the groundtruth was originally developed with a view to validating the observations of Section 3.3 (i.e. motion neighbourhood mapping to content, results forthcoming).

As a typical example of the type of content change that is detected, consider Figure 2(a), depicting an 18 minute segment from the film, "The Matrix", with a neighbourhood size of 41 shots. The graph shows a section of class *No Motion* centered approximately at shot 100. This section corresponds to the dialogue sequence between Morphius and Neo before Neo has entered the Matrix. At about shot 125 the classification shifts to *All* as Neo takes the pill and begins his descent into the real world Matrix; a time of transition. The graph plots the next section as *No Motion/Fluid*, ending at about shot 180, and corresponding to the sequence from Neo's realization of being in the physical Matrix to the point where he is rescued by Morphius and his crew. An examination of relationships between rhythmic class labels and motivating scene content is presented in Section 3.3.



(a) (b)

Figure 2: (a) An example of rhythmic category change (shot arrangements), from "The Matrix"; (b) Rhythmic class transitions from "The Truman Show".

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

Another example can be found in Figure 2(b), a section drawn from the movie, "The Truman Show". The category changes correspond to Truman suspecting that all is not right in his world (*No Motion/Fluid*), the testing of his power(*No Motion/Staccato*), a movement to find his friend(*All*), followed by a time of dialogue between the two (*No Motion*).

3.3. Film Rhythm and Scene Content Analysis

The most interesting results, however, lie in the degree of common scene content often found in the different rhythm classes. From examination we have discovered that a large proportion of the classes result from certain broad sequence types. A list of these and postulations for the causes of them from film grammar is presented in Table 4.

To gain an insight into this table, we will consider two shot arrangement categories in detail. First as an example of a *No Motion* rhythm neighbourhood, consider the section of dialog already depicted in Figure 2(a) between shots 60 and 125. The scene takes place exclusively within the confines of two rooms and consists almost exclusively of static alternating shots between Neo and Morphius. Dispersed among these shots there are small amounts of motion, used for dramatic effect. It is precisely the demands of the dialogue scene here, and in general, that dictate the lack of motion (*No Motion*) throughout.

Second, as an example of the *All* category consider the sequence from Truman in Figure 2(b), from shot 40 to 60. The sequence shows Truman go from outdoors, confused having tried his power to indoors in hushed conversation with his friend. The constituent shots of this piece include a static shot of Truman's current location (outdoors), a closer detail shot with staccato movement emphasizing his state, a subjective view of his friend entering the shop (where he wants to be), linking shots as he gets to the shop, and a number of subjective and objective shots of the new locale from different points of view that function as mini establishers. While this is specific to the Truman Show, the general principle should be apparent, i.e. for a considerable number of classes of locale/state transition it is necessary to migrate the current camera setup to a new stable position, whilst at the same time giving detail as to how the change is conducted. These filming requirements often result in use of shot motion types of all three categories, *No Motion, Fluid, and Staccato*, depending on the exact situation and creativity of the filmmaker, and hence, are classified in the *All* category.

4. CONCLUSION

Starting with the film literature concerning rhythm in film, we noted its complexity when taken as a whole and sought to focus our analysis on rhythmic elements expressed via motion. Taking a cue from music, but with analogous application to the domain of film, we defined the motion characteristics of a shot as being either No Motion, Fluid or Staccato. Tools for classifying a given shot index were developed and employed with a good level of accuracy in spite of the many difficulties inherent to the task. A natural extension to this work was pursued as the classification of a neighbourhood of shots, and may be considered analogous to the musical bar or phrase (though only in concept). The resulting shot rhythmic arrangement classes proved to be useful in that they are reflective of certain film content categories. Given the ability to find sections of likely scene content, the tools developed here could be applied to the task of identifying narrative structures and dramatic progression.

Table 4: Rhythm classes, sequence content and causes.

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)

5. REFERENCES

[1] B. Adams, C. Dorai, and S. Venkatesh. Role of shot length in characterizing tempo and dramatic story sections in motion pictures. In *IEEE Pacific Rim Conference on Multimedia 2000*, pages 54–57, Sydney, Australia, December 2000.

[2] B. Adams, C. Dorai, and S. Venkatesh. Study of shot length and motion as contributing factors to movie tempo. In *8th ACM International Conference on Multimedia*, pages 353–355, Los Angeles, California, November 2000.

[3] B. Adams, C. Dorai, and S. Venkatesh. Towards automatic extraction of expressive elements from motion pictures: Tempo. In *IEEE International Conference on Multimedia and Expo*, volume II, pages 641–645, New York City, USA, July 2000.

[4] D. Bordwell and K. Thompson. *Film Art, 5th Ed.* McGraw-Hill, 1997.

[5] W. Mahdi, L. Chen, and D. Fontaine. Improving the spatial-temporal clue based segmentation by the use of rhythm. In *Second European Conference, ECDL '98*, 1998.

[6] J. Mitry. *The Aesthetics and Psychology of the Cinema*. The Athlone Press, London, 1998.

[7] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1994.

[8] B. Salt. *Film Style and Technology: History and Analysis*. Starword, London, 1992.

[9] T. Sobchack and V. Sobchack. *An introduction to film*. Scot, Foresman and Company, 1987.

[10] H. Zettl. *Sight Sound Motion Applied Media Aesthetics*. Wadsworth Pub Co., 1998.

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE