

Application of Computational Media Aesthetics Methodology to Extracting Color Semantics in Film

Ba Tu Truong
Dept. of Computer Science
Curtin University of
Technology
Perth, W. Australia
truongbt@cs.curtin.edu.au

Svetha Venkatesh
Dept. of Computer Science
Curtin University of
Technology
Perth, W. Australia
svetha@cs.curtin.edu.au

Chitra Dorai
IBM T. J. Watson Research
Center
Yorktown Heights
New York 10598, USA
dorai@watson.ibm.com

ABSTRACT

Using film grammar as the underpinning, we study the extraction of structures in video based on color using a wide configuration of clustering methods combined with existing and new similarity measures. We study the visualisation of these structures, which we call *Scene-Cluster Temporal Charts* and show how it can bring out the interweaving of different themes and settings in a film. We also extract color events that filmmakers use to draw/force a viewer's attention to a shot/scene. This is done by first extracting a set of colors used rarely in film, and then building a probabilistic model for color event detection. We demonstrate with experimental results from ten movies that our algorithms are effective in the extraction of both scene-cluster temporal charts and color events.

Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval—*Multimedia, Content Analysis and Indexing*

1. INTRODUCTION

There is still a large semantic gap between the rich meaning that users want when they query and browse media and the low-level nature of content descriptions that we can actually compute in current automatic content-annotation systems. Upon recognizing this problem, Dorai and Venkatesh [2] have proposed the *Computational Media Aesthetics* framework for high-level semantic analysis of media content. It is defined as the algorithmic study of a variety of image and aural elements in media with insights from media production [2]. It is also the computational analysis of principles that have arisen underlying their manipulation for clarifying, intensifying and interpreting an event for audience. This paper presents a study into the application of the Computational Media Aesthetics framework to the problem measuring and understanding the expressive function of color in film.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Multimedia '02, December 1-6, 2002, Juan-les-Pins, France.
Copyright 2002 ACM 1-58113-620-X/02/0012...\$5.00.

Film grammar, the body of knowledge and techniques for film production indicates that color is frequently manipulated in film to set up moods and thematic concepts. It details that color in film should be organized coherently or used to add excitement and drama to the film. Two interesting questions to examine using the Computational Media Aesthetics framework are: (a) Can we group scenes into different clusters so that elements of the same cluster share the same setting or mood (without explicitly labeling the setting/mood)?; (b) Can we detect 'color events', shots in which a filmmaker deliberately uses color to draw/force the attention of the viewer?

Answering these two questions is the *aim* of this study. Using film grammar as both motivational and guiding forces, we apply the framework to examine the feasibility of extracting high-level structures in film, which we term *scene-cluster temporal charts*. These charts essentially show the temporal progression and density of the clusters, thus clearly indicating the manner in which the filmmaker juxtaposes and interweaves different settings and moods. Using many configurations of clustering techniques and similarity measures, we demonstrate that the best performance is achieved using Ward's method combined with a novel similarity measure we propose. In addition, we note that directors use rare colors to draw/force the attention of a viewer to specific shots. We first automatically extract these colors and then construct an algorithm for the extraction of color events in a probabilistic framework

To our knowledge, the closest related work is discussed in [1] which tries to capture the expressive and emotional properties induced by colors and applies them to semantic retrieval of art paintings and commercial videos. [5] exploits color among other features to build a framework enabling high-level indexing of sample scenes such as rocks, sky, etc.

2. COLOR FUNCTIONS

According to Zettl [8], the primary expressive function of color is to make us feel a specific way and can be further divided into 3 somewhat overlapping sub-functions:

Expressing essential quality of an event: This function strongly requires knowledge about associations between color and objects in context, and it is difficult to obtain automatically.

Adding excitement and drama: Colors can both add additional drama or excitement to a scene/event or act as a principal event itself. This color function is referred in this work as ‘color event’. The term ‘event’ here is used to indicate something “unusual and noteworthy”. What constitutes a color event is the inclusion of certain colors which are unusual to gain the viewer’s attention or imply a certain aesthetic message. Examples of color events are “the colorful uniform of marching band, the brightly colored costume of the dancers, ...” [8, p.67]. A particular feature of color is the way in which certain hues attract our attention. Some colors have greater powers of attraction than others [8]. Zettl [8] and Neale [6] also point out that most of the film shots should not use colors or color arrangements that attract the viewer’s attention, as the main focus of attention in film is its drama. Therefore, we can deduce that there is a certain set of colors that cause the viewer attention, and these colors often occur rarely in film. In addition, the power of attention of a certain color composition is generally propositional to the ‘surprise’ level as perceived by the viewer. Therefore, the rarer a color composition is in film, the stronger it attracts the viewer’s attention.

Establishing mood: Another expressive function of color is for establishing or intensifying the mood of a scene. This application of color in film is often associated with the use of certain color palettes, tones or tints to ‘color’ the scene. The complexity, however, lies in the fact that though there is consistency in these observations, hard scientific evidence for such correlations do not exist, because as Zettl says “the perceptual effects are contextual, they rarely if ever occur in isolation; instead they usually operate in the context of other aesthetics variables” [8, p.57]. This is further complicated in film as [4, p.14] states: “In general, the ‘psychological’ interpretation of color is a very slippery business.” It is emphasized that “in art it is not the *absolute* relationship/associations that are decisive, but those *arbitrary* relationships within a system of images dictated by the particular work of art” [4, p.14]. Color is generally fitted to the scene to augment its dramatic value, be it mood (sad, happy), theme (indoor vs outdoor, warm vs cold) or a specific anchoring place (e.g., the office, the house).

3. FEATURE EXTRACTION

Given a digital stream of a movie in the form of MPEG-1-encoded data, shot detection is first carried out. Shot attributes such as average \mathcal{H} , \mathcal{L} , \mathcal{S} histograms are computed. The HLS space is quantized using 12 bins of hue, 5 bins of lightness, and 4 bins of saturation, resulting in a total of $113 = 1 + 1 + (5 - 2) + 12(4 - 1)(5 - 2)$ different colors in our final quantized color palette. [7] details the extraction of scene indices and how to compute overall shot/scene features from those of individual keyframes.

4. GROUNDTRUTH ESTABLISHMENT

In order to evaluate the effectiveness of our automated techniques for grouping scenes sharing the same setting and detecting color event shots, a groundtruth for 10 movies described in [7] was first constructed. The task is difficult at times due to its subjective nature. However, we can set up the groundtruth with a good level of confidence by following certain guidelines. Scenes are included into the same cluster according to the following criteria (applied in

the order of mentioning): (a) Consecutive segments of the same scene (mainly due to noisy scene indices; (b) scenes sharing the same locale at different times and under the same lighting conditions; (c) scenes belonging to the same theme (e.g., inside matrix vs outside matrix, day/outdoor vs night/outdoor, fictional world vs real world, etc). Individual shots are groundtruthed as containing a color event, if we can answer ‘yes’ to both the following questions: (a) Do certain colors cause instant attention/reaction from the viewer?; (b) Are there certain aesthetic implications by the filmmaker upon placing those colors in the shot?

5. DETECTING COLOR STRUCTURES

5.1 Clustering Method

We examine a number of traditional hierarchical clustering techniques including Single Linkage (**SL**), Complete Linkage (**CL**), Group-Average Linkage (**GAL**), Centroid (**CEN**), Median (**MED**), Ward’s Minimum Variance (**WARD**), Lance and Williams’s Beta Flexible (**BETALW**) and Belbin’s Alpha Beta Flexible (**BELBIN**) [3] for scene grouping.

Rand index (**R**) and its improved version namely, Adjusted Rand index (**R***) are two common measures of the agreement between the detected clusters and the groundtruth. We propose two new measures namely cluster recall (**CR**) and cluster precision (**CP**). Analogous to recall and precision in information retrieval, **CR** is defined as the ratio of correctly detected pairs in the same cluster, a to all possible pairs in the groundtruth clusters, U , while **CP** is defined as the ratio of correctly detected pairs to the all possible pairs in reported clusters, V :

$$\mathbf{CR} = \frac{a + n}{\sum_{i=1}^{n_U} \binom{u_i}{2} + n} \quad \mathbf{CP} = \frac{a + n}{\sum_{i=1}^{n_V} \binom{v_i}{2} + n}$$

Here, n is the number of elements to be clustered, and it is added, as we include n non-distinct pairs in the counting. Two partitions agree well when both **CP** and **CR** are high. When two partitions agree perfectly, both **CR** and **CP** are 1. High **CR** and low **CP** imply that small clusters in the groundtruth are grouped into bigger ones in detected clusters. On the other hand, low **CR** and high **CP** imply that clusters in the groundtruth are broken into smaller ones in detected clusters. Adjusted Rand index can serve as the overall measure of the performance while **CR** and **CP** provide more insight into the nature of clustering outputs.

5.2 Measuring Similarity

Measuring the visual similarity between image/shot/sequence, represented by the histogram, is the basis for any clustering technique. Let $\mathbf{F}_i[u]$ denotes bin u of frame \mathbf{F}_i . We propose a new metric, which mimics the process of measuring the similarity between two images by gradually excluding regions with highest similarity. We first form the color similarity matrix **I** based on the Euclidean distance between colors in \mathcal{H} , \mathcal{L} and \mathcal{S} space. Moreover, we set the distance to **INF** when the \mathcal{H} component of two colors are more than two hue levels apart. Other values are normalized to within range [0-1]. Let \mathbf{I}_{ut} denote the similarity between colors u and t . We define component similarity between two frames $\mathbf{F}_i, \mathbf{F}_j$ and two bins u, t as:

$$\mathbb{P}(\mathbf{F}_i, \mathbf{F}_j, u, t) = \mathbf{I}_{ut} \min(\mathbf{F}_i[u], \mathbf{F}_j[t])$$

The overall similarity of two frames can be calculated as the recursive sum of individual component similarities:

$$\mathbb{S}(\mathbf{F}_i, \mathbf{F}_j) = \begin{cases} \mathbb{S}(\mathbf{F}_i^*, \mathbf{F}_j^*) + \mathbb{P}(\mathbf{F}_i, \mathbf{F}_j, u_0, t_0) & \exists(u_0, t_0) \\ 0 & \text{otherwise.} \end{cases}$$

with

$$(u_0, t_0) = \{(u, t) \mid (\max_{1 \leq u, t \leq N} \mathbf{I}_{ut}), \mathbb{P}(\mathbf{F}_i, \mathbf{F}_j, u, t) > 0\}$$

$$\mathbf{F}_i^*[u] = \begin{cases} \mathbf{F}_i[u] - \min(\mathbf{F}_i[u_0], \mathbf{F}_j[t_0]) & \text{if } u = u_0 \\ \mathbf{F}_i[u] & \text{otherwise.} \end{cases}$$

$$\mathbf{F}_j^*[t] = \begin{cases} \mathbf{F}_j[t] - \min(\mathbf{F}_i[u_0], \mathbf{F}_j[t_0]) & \text{if } t = t_0 \\ \mathbf{F}_j[t] & \text{otherwise.} \end{cases}$$

This means that after taking the component similarity of the two most similar colors, the component similarities of the remaining part of the two histograms are recursively extracted until there does not exist any pair of colors from each histogram that has component similarity greater than zero, i.e., two colors are not similar at all or bin size of at least one of the colors is zero. The value of \mathbb{S} is then normalized by the total number of pixels of a frame. It should be noted that this measure is essentially the conventional bin-wise intersection metric when \mathbf{I} is the identity matrix.

5.3 Experimental Results

For each of the 10 movies in our data set, we first extract scene indices and compute their corresponding histograms. 24 different cluster hierarchies are then produced in our experiment using different clustering techniques. Only a single partition is extracted from each hierarchy by applying Mojena’s procedure [3]. Finally, the adjusted Rand index, cluster precision and cluster recall are computed by comparing these 24 partitions with the groundtruth. Our data shows that, except for the cluster recall of the Centroid and Median clustering techniques, the performance of our proposed similarity measure surpasses the two traditional measures. The Euclidian distance measure appears not to perform well. With respect to our similarity measure, Ward’s minimum variance method shows a slightly superior performance. Single Linkage is the worst performer, as it is affected most severely by the problem of grouping smaller clusters into the bigger ones, resulting in high values for cluster recall but very low values for cluster precision.

The best combination for clustering scenes in movies is therefore our proposed similarity measure and the Ward’s clustering method. Table 1 details the adjusted Rand index, cluster precision and cluster recall values for individual movies using this combination. Considering that the expected value of the adjusted Rand index is 0, this clustering method performs reasonably well. Higher cluster precision compared to cluster recall is quite desirable, as it is easy to merge clusters in post-processing. The best results are obtained from The Mummy and The Sleepy Hollow as colors in these movie are manipulated to great effect. Lower results are obtained for more ‘natural’ or ‘traveling’ movies such as The Siege, 12 Monkeys, and American Beauty. Groundtruthed clusters in 12 Monkeys tend to merge into bigger groups in the detected clusters as indicated by its low precision value. On the other hand, groundtruthed groups in American Beauty and The

Siege tends to be broken into many smaller groups. High precision is obtained for The Matrix, but its cluster recall is not very high as the ‘real world’ setting is broken into 3 different clusters.

Table 1: Color-Settings clustering results.

Movie	Clusters	R*	CP	CR
Star Wars I	16	0.37	0.53	0.53
The 13th Floor	14	0.49	0.83	0.52
The Matrix	7	0.54	0.91	0.55
Sleepy Hollow	14	0.63	0.76	0.73
Tall Tale	8	0.48	0.83	0.51
Chameleon	10	0.52	0.74	0.63
12 Monkeys	17	0.27	0.44	0.76
The Mummy	5	0.71	0.81	0.81
American Beauty	16	0.35	0.75	0.39
The Siege	8	0.25	0.62	0.34
Average	11.5	0.46	0.72	0.58

The clustered scenes can be presented as Scene-Cluster Temporal Graphs (**SCTC**). In this chart, scenes belonging to the same cluster are represented by horizontal bars at the same level. The duration of the bar is determined via its temporal length (e.g., time, shot, or scene). Clusters are sorted from the bottom according to the total length of associated bars. The top line contains a collection of clusters consisting of one single scene which may be empty. The vertical bars show the location of the scene boundary. The labels on the right are manually assigned based on the groundtruth, while the numbers on the left denote the order of these clusters in the cluster dendrogram (from left to right). The advantage of this chart is that it can show both the temporal progression and density of all clusters.

The **SCTC** of The Mummy is shown in Figure 1. There are three major settings where the film took place: indoor (hotels, library, etc), underground (inside the tomb) and outdoor day scenes (desert, city, etc). The chart clearly shows the interweaving of the three dominant setups and indicates the story progression as the treasure hunters enter and exit the tomb a number of times. The scenes on the boat are also set rather differently and are clustered together correctly. Further, though most of the scenes are in warm tones as this is typical for desert settings, The Mummy’s filmmakers occasionally use bluish tones to indicate darkness, night and mysterious atmospheres.

6. COLOR EVENTS

6.1 Measuring Color Event Likelihood

Color events can be detected by measuring the viewer ‘attention’. We can assume that this attention is generally proportional to the ‘surprise’ level of the viewer, which in turn can be characterized by the unlikelihood of the color compositions. Since we hypothesize that rare colors are used to create color events, a list \mathcal{M} of colors which have the average occurrence in multiple movies below a threshold is determined. Examining this list shows that these colors tend to cause color events. In addition, we build a measure of color distribution in film using a large collection of 23,500 shots from 13 different movies. Only colors in \mathcal{M} are taken into account.

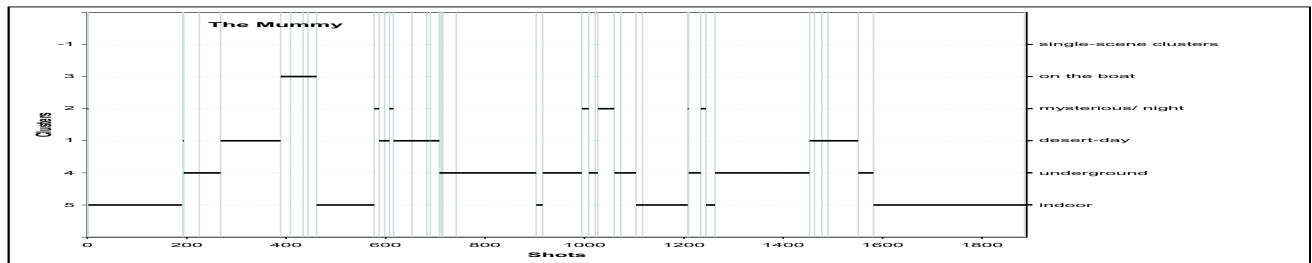


Figure 1: Scene-cluster Temporal Chart of The Mummy.

Assuming all colors are independent, we calculate the likelihood of obtaining a shot \mathbf{S} having histogram \mathbb{H} with respect to the color set \mathcal{M} as:

$$\mathcal{L}_{\mathcal{M}}(\mathbf{S}) = \prod_{c_i \in \mathcal{M}} \mathcal{P}(\mathbb{H}[C_i])$$

This can be converted into log-likelihood as:

$$\mathcal{L}_{\mathcal{M}}^*(\mathbf{S}) = \sum_{c_i \in \mathcal{M}} \log(\mathcal{P}_{C_i}(\mathbb{H}[C_i]))$$

The likelihood of a shot containing a color event can then be approximated by the following function:

$$\mathcal{E}_{\mathcal{M}}(\mathbf{S}) = -\mathcal{L}_{\mathcal{M}}^*(\mathbf{S})$$

$\mathcal{E}_{\mathcal{M}}(\mathbf{S})$ ranges from $-\infty$ to $\sum_{c_i \in \mathcal{M}} \log(\mathcal{P}_{C_i}(0))$ and can be normalized to a range $[0, 1]$.

Figure 2 depicts the color event likelihood of sample shot sequences from, The 13th Floor. The circles indicate that the shots are marked as containing color events in the groundtruth. It can be seen that we can threshold the color likelihood of each shot to identify color events.

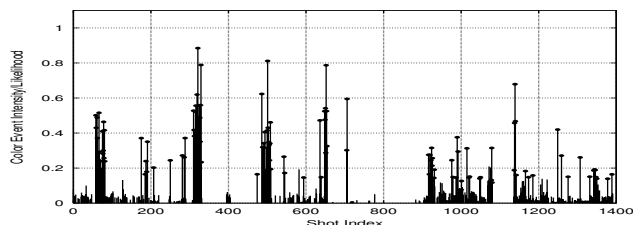


Figure 2: Color event likelihood of, The 13th Floor.

6.2 Experimental Results

Table 2 shows the results of our color event detection algorithm in terms of the number of events in the groundtruth and the percentage of recall (RECL) and precision (PREC). Statistics are collected at both shot and scene level. A scene is considered to contain a color event, if it contains at least one shot with a color event.

Overall the algorithm performs well, obtaining an average precision of 90.1 and recall of 84.9. The best result is obtained for The 13th Floor in which all color events such as laser light beams and computer graphics are high in intensity that quickly attract the viewer’s attention. The lowest result is obtained for The Matrix as the algorithm wrongly detects some greenish shots as color events (green is used to set the theme for inside-matrix scenes). Understandably, Star Wars I contains the largest number of color events, as light sabers fighting sequence as well as other computer generated graphics are inserted to excite the viewers.

Table 2: Color event detection results.

Movie	Scene Level			Shot Level		
	GT	PREC	RECL	GT	PREC	RECL
SWI	65	96.9	76.8	361	92.8	85.2
13F	27	96.3	96.3	124	96.8	99.2
MTX	23	87.0	71.4	49	81.6	71.4
SH	8	75.0	100	14	85.7	100
TT	17	94.1	64.0	35	91.4	80.0
CL	32	96.9	73.8	161	82.0	83.5
12M	29	96.6	82.4	84	90.5	88.4
MM	12	100	70.6	71	88.7	81.8
AB	40	90.0	75.0	133	91.7	85.9
SG	54	98.1	74.6	225	88.9	80.3
	307	94.8	76.6	1257	90.1	84.9

7. CONCLUSION

In this work, we have examined two important aspects of structuralising film content using color. We extract scene-cluster temporal charts, a visualization that brings out the progression and interweaving of themes and settings in movies. To extract scene clusters, we examine a variety of clustering algorithms using existing similarity measures as well as a new similarity measure that we propose. Further, we extract color events, shots and scenes that use color to force/draw the viewer’s attention. We first automatically detect a set of colors used rarely in film, and then build a probabilistic model for the extraction of color events. We have demonstrated the performance of our algorithms using a comprehensive set of 10 movies drawn across several genres.

8. REFERENCES

- [1] C. Colombo, A. D. Bimbo, and P. Pala. Semantics in information retrieval. *IEEE Multimedia*, 6(3):38–53, 1999.
- [2] C. Dorai and S. Venkatesh. Computational Media Aesthetics: Finding meaning beautiful. *IEEE Multimedia*, 8(4):10–12, October-December 2001.
- [3] B. S. Everitt. *Cluster Analysis*. Edward Arnold, 3rd edition, 1993.
- [4] W. Johnson. Coming to term with color. *Film Quarterly*, 20(1):2–22, 1996.
- [5] M. R. Naphade and T. S. Huang. A probabilistics framework for semantic video indexing, filtering and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, March 2001.
- [6] S. Neale. *Cinema and Technology: Image, Sound and Color*. MacMillan Education Ltd, 1985.
- [7] B. T. Truong, C. Dorai, and S. Venkatesh. Automatic scene extraction in motion pictures. Technical Report 1/2001, School of Computing, Curtin University of Technology, Perth, Western Austrlia, 2001.
- [8] H. Zettl. *Sight Sound Motion: Applied Media Aesthetics*. Wadsworth Publishing, 3rd edition, 1999.