RETRIEVAL OF VISUAL COMPOSITION IN FILM

Dalibor Mitrović, Matthias Zeppelzauer, Maia Zaharieva, Christian Breiteneder

Interactive Media Systems Group
Vienna University of Technology
Favoritenstrasse 9-11, Vienna, Austria
{lastname}@ims.tuwien.ac.at

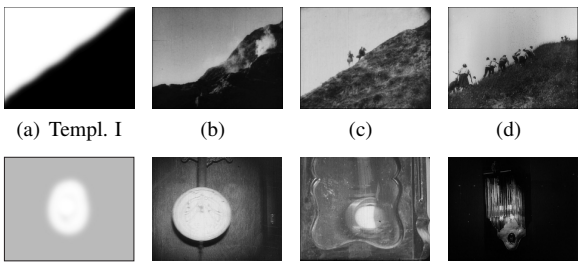
ABSTRACT

The spatial arrangement of visual elements of an image, i.e. the visual composition, is a research subject in the domain of visual arts which include painting, film, etc. Film experts face the problem of retrieval of visual compositions in film on a daily basis. Although, visual composition is a crucial element to consider in content-based video retrieval, little scientific effort has been invested into this problem so far. Actually, it is unclear if content-based retrieval of visual compositions is feasible. We present a user study conducted to investigate the feasibility of content-based retrieval of visual compositions as they are understood by film experts. For that reason, we create a data set derived from real world material and let the film experts evaluate the retrieval performance. The user study investigates the applicability of state-of-the-art visual features and shows differences in evaluations by film experts (test group) and computer scientists (reference group).

1. INTRODUCTION

Visual composition refers to the spatial arrangement of the visual elements of an image. In painting, the artist arranges the visual elements in a picture to evoke a certain impression. In film, the director arranges the elements in a scene and selects the camera's view.

Film experts want to identify recurring visual compositions (see Figure 1) because they want to analyze how compositions are used for conveying the message. Therefore, they closely inspect the films by hand which is a tedious and error-prone task. This fact motivates the use of automated methods. Currently, there is no accepted method for automated identification of visual compositions in film. Related work, such as [1], focuses on composition retrieval in news videos which follow much stricter composition rules than film.



(a) Templ. I (b) (c) (d)

(e) Templ. II (f) (g) (h)

Fig. 1. Two composition templates with three frames from different films that share the respective visual composition type.

This is the reason why related work is not applicable to films. It is still unclear whether or not visual compositions, *as understood by film experts*, can be represented and retrieved by low-level content-based features. In this paper we investigate the applicability of well-understood content-based retrieval methods in the novel domain of visual composition retrieval. For this purpose, we assemble a real world data set with the help of film experts in order to measure the retrieval performance. We need a novel data set in order to perform investigations under realistic conditions.

We see composition as the result of two concurrent processes. First, the adherence to certain *principles* and, second, the application of *formal elements*. Principles of composition include hard to grasp concepts like the dominant idea of the image as well as more tangible concepts like the gradation of lighting and the balance of the depicted elements. Formal elements among others include lines, shapes, textures and colors of depicted objects and surface areas. Formal elements are either purposely embedded into the image or they become apparent at a later time.

We design a system for retrieval of visual composition in film and perform a user study to test and answer the following hypothesis and research questions:

Hypothesis 1 *Low-level features are able to represent visual compositions.*

We pair combinations of features and single features with different proximity measures and let humans evaluate the retrieval results. These relevance judgments serve as a metric for a feature's ability to represent visual compositions. Additionally to the hypothesis, we investigate three research questions. RQ 1: Which content-based features perform best? RQ 2: Which proximity measure performs better? RQ 3: Do film experts judge the same retrieval results differently than computer scientists? We derive the third research question from the assumption that subjects with expertise in film studies better recognize the presence of compositions than subjects without this expertise.

2. TECHNIQUES

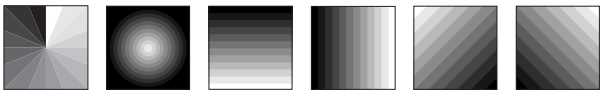
2.1. Content-Based Features

The formal elements and principles of composition can be divided into two groups, the tangible and the intangible ones. We focus on the *tangible* elements and principles. We expect that they can be captured with content-based features and thus are relevant for access to image databases.

First, we select Edge Histogram, Region Shape, and Homogeneous Texture which are defined in the MPEG-7 standard for multimedia content description [3].

Second, we employ the so called KANSEI features by Kobayashi et al. [4]. They propose the joint application of both a shape feature (KANSEI Shape) and a color feature.

Adaptations to the color feature become necessary because we employ frames from black and white films in this user study. First, we reduce the computation of the average color to one color channel, equaling the computation of the average intensity. Second, we discard the radiation-like mask 2(a) proposed in [4] in favor of two diagonal ones shown in Figures 2(e) and 2(f). This modification is based on recommendations of film experts. We name the modified feature KANSEI Intensity.



(a) (b) (c) (d) (e) (f)

Fig. 2. Masks defining the regions that are employed for the description of color and intensity distribution in the KANSEI color and intensity feature. Note that the shading only illustrates the spatial arrangement of the regions.

In addition to the *single* content-based features (see Table 1), we evaluate three feature *combinations* (summarized in Table 2) and a random feature. We obtain the feature combinations through concatenation of the single features' components. The random feature (RM) has 5 components with uniformly distributed pseudo-random values. The random feature defines a lower-bound of retrieval performance which we use to compare the other features with.

Name	Dim.	Type	Abbr.
MPEG-7 Edge Histogram	80	local	EH
MPEG-7 Homogeneous Texture	62	global	HT
MPEG-7 Region Shape	35	global	RS
KANSEI Intensity	60	local	KI
KANSEI Shape	64	local	KS
Random	5	-	RM

Table 1. Features used in the experiments.

Combination	Features	Abbr.
KANSEI features	<KI,KS>	KSI
MPEG-7 features	<EH,HT,RS>	MP7
KANSEI and MPEG-7	<EH,HT,RS,KI,KS>	ALL

Table 2. Feature combinations employed in the experiments.

2.2. Proximity Measures

We acquire retrieval results through similarity retrieval using Salton's Vector Space Model [6]. In order to preserve a certain objectivity we employ one similarity measure and one distance measure. We employ Cosine similarity and the Euclidean distance because they are two well-understood representatives of the respective groups of proximity measures.

2.3. Statistical Methods

We analyze the data quality of the content-based features using the weighted average loading indicator (WALDI), a measure for the information content based on Principal Component Analysis [5]. The WALDI summarizes the feature components' influence on the variability in the data. Feature components that describe much of the variability in the data obtain high scores while components that describe little variability obtain small WALDI scores. Evaluating the data quality of features helps in selecting compact (small number of components) and expressive (much variability explained) features.

Furthermore, we employ factorial analysis of variance [2] to identify significant differences in the means of the relevance judgments to test the hypothesis and to answer the research questions. Factorial analysis of variance (ANOVA) is a standard method employed in user studies. Significance tests with ANOVA allow for more objective statements than descriptive methods commonly used in information retrieval. ANOVA enables the evaluation of statistical properties of the investigated factors: content-based features, proximity measures, composition templates, and users' field of expertise.

3. USER STUDY

We conduct a user study to evaluate the applicability of low-level features for the retrieval of visual compositions in a real world scenario. We select 30 users for the study, 15 film experts (either film archivists or film scientists) as the test group and 15 computer scientists as a reference group. The reference group consists of computer scientists, because of two reasons. The first reason is that computer scientists frequently (mostly due to availability) serve as subjects in user studies concerned with information retrieval and we want to investigate the validity of this approach for the given task. The second reason is that the inclusion of computer scientists allows for a comparison of the two involved mindsets, on one side computer scientists as the creators of retrieval systems and novices regarding visual composition and on the other side film experts as specialists for visual composition and the real users of such a retrieval system.

The user study is performed with two sets of queries. The first set contains four pre-defined (common) query sketches which represent compositions typically sought after by film experts. These query sketches (see Figures 1(a), 1(e), 5(a), and 5(e)) were suggested by film experts prior to the study and later generated using a graphics tablet and a pressure sensitive brush. The common query sketches enable an objective comparison of two different user groups. The second set of query sketches is defined by the users themselves during the study. This set of query sketches enables the evaluation of the users' subjective satisfaction. The users first assess the retrieval performance regarding the four common query sketches and then draw and assess four individual query sketches.

We observe that the individual query sketches (see Figure 3) differ from the pre-defined ones in abstractness and the semantic content. Some individual query sketches are entirely abstract, e.g. a spiral, while others are strongly semantic, e.g. a heart. The performance of queries based on the semantics of sketches will probably suffer from the system's inability to process the semantics presented in the query. In the case of the abstract query images the retrieval performance depends on the frequency of such images in the data set. Note that the system supports user-generated query sketches as well as the use of existing images from known films, the web, etc. For the user study we employ sketches to reduce bias. For example, if existing images are used for the study, film experts could expect specific frames to be returned regardless of whether these frames are part of the data set or not.

Retrieval is performed on a data set that contains 6690 keyframes from six black and white archive films. The films are formalistic films which make frequent use of visual compositions. We select keyframes from all shots (including the ones without a distin-

	EH	HT	RS	KI	KS
WALDI	35%	25%	21%	50%	100%

Table 3. The information content represented by each feature measured with the WALDI technique relative to the best-scoring feature KANSEI Shape.

guishable composition) in order to enable an objective evaluation of the employed techniques creating a real-world scenario.

We implement a system that takes user-defined sketches of visual compositions as input and retrieves images similar to the sketch based on the features and proximity measures from Section 2. For each query sketch, we perform retrieval with the six content-based features listed in Table 1 and with the three feature combinations listed in Table 2. Each feature and feature combination is paired with both proximity measures (L2-norm and the cosine metric). This results in $(6$ *single features* $+ 3$ *feature combinations* $) * 2$ *metrics* $= 18$ different system configurations that are evaluated in the study. Each of the 18 result sets consists of the 16 best matches found in the data set and is assessed separately by each participant. We do not evaluate all possible system configurations to limit the duration of the study for each participant to an acceptable extent. Users spend 90 minutes to four hours to complete all assessments.

Prior to the assessment, we instructed the users to rate the visual similarity of the retrieved matches. All users were informed about the origin of the employed keyframes. Users not familiar with the term *visual composition* were briefed that the term refers to the spatial placement of visual elements inside an image.

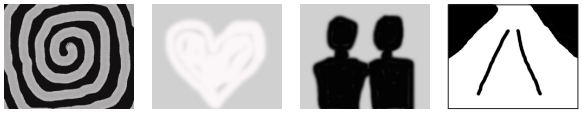
4. RESULTS

4.1. Data Quality of Features

An "ideal" feature has decorrelated components and a high score in regard of information content. Feature combinations should exhibit similar properties. Additionally, any two features in a combination should have low inter-feature correlations.

High information content is a necessary but not sufficient property of a good content-based feature. We analyze the features' expressiveness for the image data employed in this investigation. The analysis results are summarized in Table 3. We observe that KANSEI Shape scores highest followed by KANSEI Intensity. This means, they explain large amounts of variance contained in the feature data. The MPEG-7 features consistently have lower scores than the KANSEI features. Their expressiveness is limited in the context of the underlying image data.

In addition to the information content, we investigate intra-feature and the inter-feature correlations. Intra-feature correlations refer to the redundancies between the components of one single feature, while inter-feature correlations refer to the redundancies between components of two or more features. We compute Pearson's correlation coefficient between any two feature components and take its absolute value in order to obtain the correlation matrix depicted in Figure 4.



(a) (b) (c) (d)

Fig. 3. Individual query sketches generated by the users in the study.

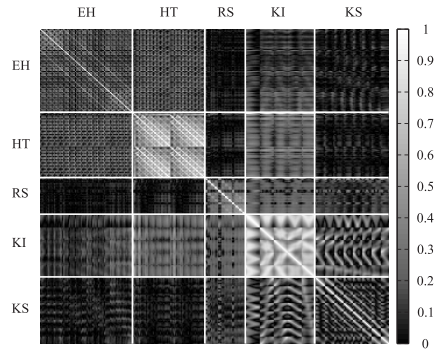


Fig. 4. The correlation matrix between all feature components. High values (light) indicate high correlations, low values (dark) indicate low correlations. The white lines mark the boundaries between features.

Ideally, the entire matrix would be dark (correlation of zero) except for the main diagonal which should be white (correlation of one). This would indicate that every feature component (and thus every feature) captures specific information that is not captured by any of the other components (and features).

On the intra-feature level, we observe strong correlations inside Homogeneous Texture and KANSEI Intensity. The correlations in Homogeneous Texture indicate that the energy and energy deviation of the captured frequency channels describe essentially the same information in the image data employed in this user study. The components of KANSEI Shape and Edge Histogram are moderately correlated. Both features base on neighboring image blocks which tend to have correlated content. Region Shape has the lowest correlations due to the independent basis functions of the ART.

On the inter-feature level, KANSEI Intensity is moderately correlated with all other features. The highest correlation is observed between KANSEI Intensity and KANSEI Shape. Region Shape has low correlations with other features, especially with the two MPEG-7 features. Homogeneous Texture correlates with some components of Edge Histogram. These correlations are expected since edges in an image introduce particular frequencies in the image's frequency domain representation.

4.2. Results of the user study

Hypothesis 1: Low-level features are able to represent visual compositions. We test this hypothesis by evaluating the Prec@16 obtained using the content-based features. Prec@16 is the proportion of relevant retrieval results in the result set of size 16. We choose Prec@16 in order to evaluate the complete result set our system retrieves. See Figure 5 for examples of composition sketches and relevant retrieval results. An evaluation of recall is not reasonable because there is no way to create a universally valid ground truth for the keyframes in the data set. A unique assignment of keyframes to composition types is not possible, since this assignment depends on the beholder's subjective assessment.

Table 4 lists the mean and standard deviation of Prec@16 for all features and combinations in the study. Note that a Prec@16 value of 1.00 can only be achieved if there are at least 16 relevant examples in the data set which is not the case for all tested sketches. Consequently, we are interested in the relative performance *differences* rather than in absolute precision values.

From the Prec@16 values, we observe that all single features and feature combinations outperform the random feature which cor-

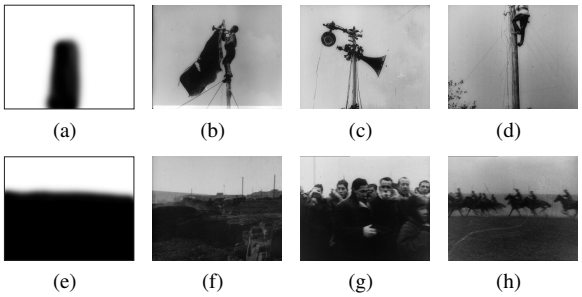


Fig. 5. Two of the pre-defined query sketches – 5(a), 5(e) – each with three relevant retrieval results.

	RM	EH	HT	RS	KI	KS	MP7	KSI	ALL
μ	0.07	0.22	0.30	0.34	0.42	0.54	0.37	0.49	0.53
σ	0.06	0.14	0.18	0.20	0.22	0.11	0.19	0.24	0.24

Table 4. Mean and standard deviation of Prec@16 for all features and combinations.

roborates the above hypothesis. The worst-performing real-world feature (Edge Histogram) yields an average Prec@16 of 0.22 while the random feature yields an average Prec@16 of 0.07.

RQ 1: Which content-based features perform best? Edge Histogram is outperformed by all other single features. The ANOVA confirms (using a level of significance of 5%) this and the following observations. The 0.04 difference between the mean Prec@16 of Homogeneous Texture and Region Shape is not significant. The performance differences of KANSEI Intensity and the other single features are significant. This makes KANSEI Intensity the second best single feature. The best performing single feature is KANSEI Shape. KANSEI Shape's performance supports the results of the statistical analysis based on WALDI. KANSEI Shape captures the variance in the data that is important for retrieval of visual compositions.

In the evaluation of feature combinations, both KSI and ALL outperform MP7. The performance difference between KSI and ALL is not statistically significant and, thus, there are two "best" feature combinations.

The performance differences between the single features and the combinations do not justify statements regarding a clear performance winner. It is nevertheless interesting that KANSEI Shape alone yields slightly higher precision than KSI and ALL. This means that a single feature achieves comparable performance to the feature combinations at lower computational costs.

RQ 2: Which proximity measure preforms better? In order to answer the second research question we analyze the performance differences between the two proximity measures. Cosine similarity yields an average Prec@16 of 0.41 with standard deviation 0.23. Euclidean distance yields an average Prec@16 of 0.39 with standard deviation of 0.22. Although, the Cosine similarity seems to be superior over the Euclidean distance, the factorial ANOVA reveals that there is no significant difference in the performance of the two proximity measures.

RQ 3: Do film experts judge the same retrieval results differently than computer scientists? We investigate the influence of the field of expertise by analyzing the differences in retrieval performance judgments between computer scientists and film experts. We ask both user groups offline to assess the retrieval system's general ability to represent visual compositions on a five-point scale (deficient - sufficient - satisfactory - good - excellent). Both user groups respond in the range from good to sufficient, with the median for both groups being satisfactory. The statistical analysis of the actual relevance judgments yields an average Prec@16 of 0.38 for computer scientists and of 0.43 for film experts with the same standard deviation of 0.22. These results indicate that film experts asses the relevance differently than the computer scientists. The ANOVA confirms the significance of this difference at a level of significance of 5%. We learn that given identical result sets, film experts rate the relevance of the presented images higher than computer scientists do. This observation is true for all four predefined query sketches employed in this study.

5. CONCLUSIONS

Visual composition is an important aspect of accessing visual arts and film. However, little effort has been invested into search and retrieval based on composition so far. We investigate the capability of low-level content-based features for the retrieval of visual compositions in a user study. Our findings suggest that low-level content-based features *are* capable of capturing composition as it is understood by film experts.

Additionally, we learn that film experts assess relevance of retrieval results to be higher than computer scientists which shows the influence of expertise for composition retrieval. This influence is linked to our finding that film experts, without being aware of it, perceive visual compositions only if there is a strong semantic connection between the query and the result image. Since the proposed technique focuses only on visual similarity film experts are presented with (for them) unexpected results which are semantically unrelated but visually similar. This allows the film experts to analyze visual compositions that they did not perceive before. One long-serving film expert even said: "*The computer sees more than man.*"

Acknowledgment

The authors wish to thank the anonymous referees whose insightful comments have significantly strengthened this paper. This investigation was partly supported by the Vienna Science and Technology Fund (WWTF) under grant no. CI06024: "Digital Formalism: The Vienna Vertov Collection."

6. REFERENCES

[1] J. Fauqueur and N. Boujemaa. Mental image search by boolean composition of region categories. *Multimedia Tools and Applications*, 31(1):95–117, 2006.

[2] S.R.A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, 1970.

[3] ISO-IEC. *Information Technology - Multimedia Content Description Interface*. Number 15938. ISO/IEC, Moving Pictures Expert Group, 1st edition, 2002.

[4] H. Kobayashi, Y. Okouchi, and S. Ota. Image retrieval system using kansei features. *PRICAI'98: Topics in Artificial Intelligence*, pages 626–635, 1998.

[5] D. Mitrovic, M. Zeppelzauer, and H. Eidenberger. Analysis of the data quality of audio features of environmental sounds. *Journal of Universal Knowledge Management*, 1(1):4–17, 2006.

[6] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.