“Where is the Interestingness?" Retrieving Appealing Video Scenes by Learning Flickr-based Graded Judgments

Miriam Redi
EURECOM, Sophia Antipolis
2229 route des crêtes
Sophia-Antipolis
redi@eurecom.fr

Bernard Merialdo
EURECOM, Sophia Antipolis
2229 route des crêtes
Sophia-Antipolis
merialdo@eurecom.fr

ABSTRACT

In this paper we describe a system that automatically extracts appealing scenes from a set of broadcasting videos. Unlike traditional computational aesthetic models that try to predict the hardly measurable degree of “beauty”, we chose to build a system that retrieves “interesting” scenes. We create a training database of Flickr images annotated with their corresponding Flickr “interestingness” degree. We then extract existing and novel aesthetic/semantic features from the training set. Based on such features, we build a graded-relevance “interestingness” model and we rank the test shots according to their predicted “interestingness”.

Categories and Subject Descriptors

I.4.8 [**Computing Methodologies**]: Scene Analysis

Keywords

Image aesthetics, interestingness, Semantic Indexing

1. INTRODUCTION

Automatic assessment of image beauty and appeal is becoming of crucial importance for the development of effective user centered visual applications. One of the tasks set for the ACM Multimedia Grand Challenge 2012 by the Japanese national public broadcasting channel NHK is indeed named “Where is the beauty?”. In this track, participants are provided with a set of broadcasting videos (from the NHK channel) describing Japanese famous landscapes and touristic sites. The task is to automatically extract the beautiful scenes in such corpus and rank them in terms of beauty.
Following the state of the art approaches, such as [1], in this paper we address this challenge using a Content Based Multimedia Retrieval Framework (CBMR), building a system able to rank the video scenes according to their appeal degree, based on learning techniques over discriminative visual features. Our *novel contributions* can be summarized in the flow of our proposed approach :
(1) We define a peculiar **Notion of Beauty**. In the NHK challenge, the notion of beauty is entrusted to the participants. According to our view, mere scenic beauty might be of limited importance when describing broadcasting data, that is typically edited to “attract” the passive TV user. We therefore choose to predict the *more informative property of “interestingness”*, namely an indicator representing how much the visual content is appealing for the audience, and how much curiosity it arouses.
(2) Since the NHK videos come without aesthetic labels, we build an **External Training Database** of Japan-related Flickr images to predict interestingness with our retrieval system. How do we build such dataset, judging pictures' interestingness without involving human subjectivity[1]? We exploit the *Flickr “interestingness” criteria*: a value representing the appeal of each photo in the Flickr collection, determined by number of views, comments, bookmarks, etc. We therefore build our training set by downloading a set of “interesting”, “average interesting” and “non interesting” Flickr images geotagged with the 20 most touristic Japanese landmarks.
(3) From both Flickr and NHK data we extract **New Compositional and Semantic Features**, including two new features we design for image *“symmetry” and “uniqueness”*.
(4) Using training features with their corresponding annotations, we train **Two Learning Frameworks**: (I) a traditional binary CBMR system (using just interesting/non interesting annotations), and (II) a *graded relevance framework* that can deal with multiple degree of annotations [4].
(5) Given the two models (I,II) computed in step (4) on Flickr images, we perform **Double Testing and Ranking** on the NHK data. We sample frames from the NHK videos, predict the interestingness score for each frame using both (I and II) systems, and we rank shots according to the highest score obtained by its frames. For the final submission, we *combine two lists of shots*, one based on the binary retrieval and the other based on graded annotations.

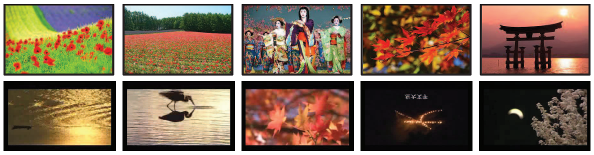

Figure 1: Flickr images (1^{st} row) and NHK “interesting” frames (2^{nd} row) according to our system

2. OUR SYSTEM

2.1 Feature extraction

From the Flickr-based and the NHK databases (i.e. the still frames extracted from the videos) we extract a set of aesthetic, affective and semantic features.

MM'12, October 29–November 2, 2012, Nara, Japan.
ACM 978-1-4503-1089-5/12/10.

[1] In order to have reliable judgments about the video beauty, we would need a large number of diverse human annotators.

Affective and Aesthetical Features

We compute a set of 43 features coming from emotion-based image recognition, computational aesthetics, and painting analysis, resulting in a feature vector $a = \{a(i)\}_{i=1}^{43}$, composed as follows: **(a) Color names [2]**, a(1-9). **(b) GLCM properties [2]**, a(10-19). **(c) HSV features [2]**, a(20-25). **(d) Level of detail [2]**, a(26). **(d) Rule of thirds [1]**, a(27-29). **(e) Low depth of field [1]**, a(30-38). **(f) Contrast [2]**, a(39). **(g) Symmetry**, a(40). We define our own measure of symmetry by extracting the Edge Histogram Descriptor [6] on both the left half and the right half of the image (here, major and minor diagonal bins are inverted) and retaining the difference. **(h) Image Order [5]**, a(41,42). **(i) Uniqueness**, a(43). How much is an image unique, e.g. it differs from the common image behavior? We measure the uniqueness by the Euclidean distance between the average spectrum of the images in the database and the spectrum of each image.

Semantic Features

The semantic content of an image plays an important role in determining its interestingness degree. In our system, we extract two semantic features, namely the MPEG7 **Edge Histogram Descriptor** $e = \{e(i)\}_{i=1}^{64}$ and the **Saliency Moments Descriptor** ($s = \{s(i)\}_{i=1}^{462}$) [3].

2.2 Learning and Ranking

We then use training features with their corresponding interesting/non interesting annotations to train a binary CBMR system, as shown in Sec 3.2. We then extend the training set to allow for non-binary annotations and feed a graded relevance interestingness-based retrieval system.

Binary Relevance System

The first system we build for ranking NHK video scenes is a traditional CBMR framework. We extract from the training Flickr images the feature vectors in Sec 3.1, we label each feature vector with a positive/negative label according to the interestingness of its corresponding image, and we use them as input to feed a set (one per feature) of Support Vector Machines (SVM) with RBF Kernel. We have now three feature-specific models able to distinguish between appealing/non appealing images.

On the NHK test set, we extract 16 frames per shot, for the 10 videos provided. We then classify the resulting frames with the feature-specific models: by doing so, we obtain, for each frame F an interestingness score $p_F(f)$, corresponding to the output probability of the f-specific SVM. Since we want to determine the interestingness of an entire shot S, and we have several frames per shot, we retain as interestingness score for a given shot the maximum of the scores of the frames belonging to that shot, namely $p_S(f) = max(p_{F \in S}(f))$. Finally, we compute the final interestingness score for each shot by linearly combining the output of the three feature-specific predictors.

Graded Relevance System

A partition of the training set based on binary annotations only can be too restrictive for the type of information we want to infer. An image can be appealing with different degrees for a given user, depending on the way the image is composed. As a matter of fact, in our second framework, similar to [4], we use 3 levels of labels and build a graded relevance retrieval system able to deal with multiple degrees of annotations.

First, we extend our training set by adding the images (and their corresponding features) that have been ranked by Flickr as being "average interesting". We then create two training subset: ($t1$) considers as positives the "interesting" images and as negative the "average interesting" plus the "non interesting" images; ($t2$) consider as positives the "interesting" and "average interesting" images, and as negative the "non interesting" ones. We then build a 2-layer model, similar to [4], and when we test on the NHK database, for each shot (processed as in the Binary Relevance System) we obtain two, complementary, feature-based interestingness score, namely $p1_S(f)$ and $p2_S(f)$, that we linearly combine into $p_S^{gr}(f)$, in order to retain the information coming from both levels of the multi-layer mode.

The Submitted Run

Based on the interestingness scores, we rank the test shots, and compose our final run by combining the top 5% of each list (5% of Binary System and 5% of Graded Relevance System). Moreover, we run a face detector on the final list, and eliminate those frames for which the detector identifies a face with size $\geq 1/4$ of the frame surface, since such frames are very likely not to contain sceneries or landscapes.

Evaluation

We present here some results on our development (Flickr) data evaluated with Mean Average Precision in the top 10 % images. Results show that our choices (e.g. semantic features, graded relevance retrieval) bring a substantial improvement in the final retrieval performances.

	(s)	**(e)**	**(a)**	**a+e+s**
Binary	0,16646227	0,11358656	0,17658801	0,18944645
Graded	0,17648103	0,12364157	0,20284673	0,21484038

3. CONCLUSIONS

We designed a framework for automatic extraction of interesting scenes from video shots. We train a binary and a graded relevance retrieval system based on interesting/not-interesting Flickr images through aesthetic and semantic features. Both systems concur in creating the final list of shots, ranked according to their interestingness degree.

4. REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. *Computer Vision–ECCV 2006*, pages 288–301, 2006.

[2] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, pages 83–92. ACM, 2010.

[3] M. Redi and B. Merialdo. Saliency moments for image categorization. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, 2011.

[4] M. Redi and B. Merialdo. A multimedia retrieval framework based on automatic graded relevance judgments. *Advances in Multimedia Modeling*, pages 300–311, 2012.

[5] J. Rigau, M. Feixas, and M. Sbert. Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity. *Computational Aesthetics in Graphics, Visualization, and Imaging*, 2007.

[6] C. Won, D. Park, and S. Park. Efficient use of mpeg-7 edge histogram descriptor. *Etri Journal*, 2002.