Semantic Indexing and Computational Aesthetics: Interactions, Bridges And Boundaries

[Extended Abstract] *

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

ABSTRACT

Semantic Indexing and Computational Aesthetics are two closely related fields. For some aspects they are similar, complementary for others, and sometimes completely disjoint. Semantic Indexing is about automatically identifying content in natural images, namely recognizing objects and scenes. Computational Aesthetics provides a set of techniques to automatically assign a beauty degree to a given image. In our work, we enrich both types of visual analysis by exploring the synergy of those two fields. We investigate the role of Semantic Indexing techniques for Computational Aesthetics Frameworks, and, vice versa, the importance of Aesthetic features for Semantic Indexing prediction. We show the benefits and the limits of this synergy, and propose some improvements in this direction.

Categories and Subject Descriptors

I.4.7 [**Image Processing and Computer Vision**]: Feature Extraction

1. INTRODUCTION

With the increasing amount of visual content surrounding us, automatic image analysis systems become more and more important for the development of effective and efficient user centered visual applications. Research on Semantic Indexing (SI) [1, 8, 13] has already accomplished great advances in the field of automatic scene and object categorization. However, in the recent years, a new field for automatic image analysis has attracted the attention of multimedia researchers: Computational Aesthetics (CA), namely a set of techniques to automatically assess the image beauty and appeal. While SI systems predict the presence of given semantics in an image, Computational Aesthetics frameworks predict the aesthetic degree of its visual content.

Figure 1: Similar images, similar aesthetics.

Semantic Indexing techniques are generally more focused on the analysis of the *content* of the image: they learn models based on *semantic* features, namely low-dimensional description of the image content. Semantic features are extracted either locally [5], by studying the local shape of the edges, or by globally analyzing [8] the behavior of the image. On the other hand, Computational Aesthetic frameworks [2, 7] learn models able to predict image beauty based on *compositional* features, that describe how much an image is following given photographic rules [2], and what is the general arrangement and layout of the image. While SI features give information about the content, CA features collect the attributes related to the shooting process and the image composition.

Despite their different applications and underlying features, Semantic Indexing and Computational Aesthetics systems are closely related fields, as they both address image analysis issues. From a technical point of view, they share the same learning framework, adopted by CA systems from SI. In both cases, a model is learnt on annotated (with content or aesthetic degree) training data (namely semantic or compositional features) through machine learning techniques, and then used to label (with object/scenes categories or beauty degree) a test image. But analogies between SI and CA are not limited to their implementations. Content and Aesthetics are closely related in natural images also from a perceptive point of view. For example, as prove in [3] the type of object depicted in an image can influence the aesthetic judgment (e.g. people, animals, faces). Moreover, it is well known in photographic theory [4] that the image shooting process and its composition technique, as well as the emotion vehiculated and the degree of visual appeal, change according to the content to be depicted. Content is therefore important to determine the image composition,

*A full version of this paper is available as *Author's Guide to Preparing ACM SIG Proceedings Using LaTeX2ε and BibTeX* at `www.acm.org/eaddress.htm`

WOODSTOCK '97 El Paso, Texas USA

and, subsequently, its aesthetic degree. Similarly, given this relation, groups of semantically similar images must share the same compositional attributes, making compositional-aesthetic information an additional cue for SI.

These observations regarding the junctions between these two fields suggest us that the synergy between Semantic Indexing and Computational Aesthetics can help both image category and aesthetic degree prediction. We indeed investigate here the borders, boundaries and intersections between these two fields with the aim of improving the global effectiveness of CA and SI systems. Since the general frameworks for CA and SI have the same structure, it is practically straight forward to combine the knowledge of these two conjoint fields. Content and aesthetics are complementary source of information regarding the image depicted, and we can exploit their combination to enrich both the aesthetic and semantic learning.

In our work, we therefore address two questions: (1) How is Semantic Analysis influencing Aesthetic prediction? First, having a strong background on Semantic Indexing, we use for CA analysis the rules and features that we originally created for SI problems. We indeed explore the role of graded relevance learning systems and of our holistic semantic features [8] for image appeal prediction. (2) How is Computational aesthetics information affecting in Semantic prediction? To answer this issue, we look at the prediction improvements obtained by adding compositional features to classic SI frameworks for scene recognition. In both cases, we show that the combination of SI and CA information brings substantial improvements to both types of systems. We also observe that that there are some limits beyond which the interactions between this two fields is not bringing any new information, e.g. the quality of images that we consider.

2. THE SYNERGY OF AESTHETIC AND SEMANTIC ANALYSIS

In our work, we test the benefits that Semantic Indexing and Computational Aesthetic fields achieve through their interactions and we show the limits of this approach. We apply semantic indexing frameworks to aesthetic prediction and we borrow holistic features from both fields to improve the effectiveness of their prediction systems. In order to explore this combination, we perform two classes of experiments: (1) we show the role of holistic semantic features for image interestingness prediction through graded relevance-based learning, and (2) we build a scene recognition system that embeds several holistic aesthetic, affective and artistic features. In both cases, we show that the prediction (Aesthetic, for (1) and Semantic, for (2)) is improved through the combination of the two conjoint fields.

2.1 Retrieving Appealing Images by Learning Flickr-based Graded Judgments

Our first work that investigates the combination aesthetics-semantics is a learning framework that predicts the image "interestingness", typically related to the image "beauty" (see [11] for details). Our aim here is to build a system that, given an image (or a video sequence), can output a value corresponding to the appeal of its visual content.

We chose to model and predict the image interestingness using a SI framework, based on learning techniques over discriminative visual (semantic and aesthetic) features.

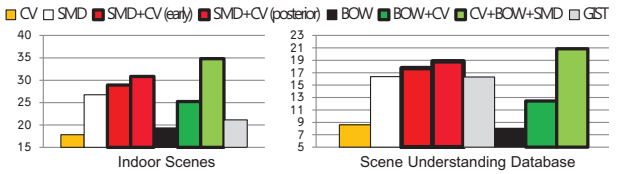

Figure 2: CA at the service of SI, results.

We therefore first create a training/test database of Flickr images annotated with their corresponding Flickr "interestingness" degree (Non Interesting, Average Interesting, Very Interesting). We then compute a set of "compositional" features from emotion-based image recognition, computational aesthetics, and painting analysis[1]. We additionally create two new features for image appeal analysis, namely an edge-histogram[16] based measure of **Symmetry** and a **Uniqueness** feature based on spectral analysis representing how much a given image differs from the standard image behavior. We combine the resulting features in a 43-dimensional "compositional" feature vector (CV). Since the aesthetic appreciation and the image composition can change according to the image content, we also extract two *semantic* features, namely the MPEG7 **Edge Histogram Descriptor** (EHD) [16] and our **Saliency Moments Descriptor** (SMD) [8]. Why did we choose to embed such features into our aesthetic predictor? Both are semantic features related to aesthetics. EHD represents a holistic summarization of the image composition, which is typically very important to define the aesthetic degree of an image, and SMD summarizes the image content with visual saliency information, which has been proved [17] to be closely related to image aesthetics. We will combine the contributions of compositional and semantic features using posterior fusion.

We then model the feature space for interestingness prediction using a non-binary learning framework. We re-use our graded relevance learning framework for video retrieval [10], namely a semantic indexing system that can deal with multiple degrees of annotations. Such system can suit well the variety of the interestingness scores that we have in our training set. We therefore use training features with their corresponding very/average/non interesting annotations to train the graded relevance system. The resulting model will be able to predict the interestingness degree of a test image given the values of its semantic-compositional features.

In our experiments, we split the database into a train and a test subset, we learn a graded-relevance-based model for each type of feature (CV, EHD, SMD) and combine them with posterior fusion. We evaluate the results on the test subset using Mean Average Precision (MAP) over the list of the images that have been ranked in the top 10 %.

	SMD	EHD	CV	SMD+CV	ALL
Binary	0,16646	0,11358	0,17658	0,18918	0,18944
Graded	0,17648	0,12364	0,20284	0,21361	0,21484

In order to show the effectiveness of our approach, we also compare the results with a traditional binary-relevance system that uses the same setup. Results in Table 1 show the performances of the three features (CV, EHD, SMD) used as stand-alone descriptors and the prediction improvement

[1]The existing features we compute are: (a) Color names [6],(b) GLCM properties [6]. (c) HSV features [6]. (d) Level of detail [6]. (e) Rule of thirds [2]. (f) Low depth of field [2]. (g) Contrast [6]. (h) Image Order [12].

after posterior fusion (+ 7% over CV-based classification only, which show the importance of semantic features in the interestingness-based retrieval) in the binary-relevance system. We can also observe the improvement (+ 6% for SMD-based retrieval, +8% for EHD-based retrieval, +15% for CV-based retrieval, + 14% for all features) obtained using graded relevance retrieval.

2.2 Enhancing semantic features with compositional analysis for scene recognition

The second work we propose to investigate the relations between aesthetics and semantics focuses on the semantic scene categorization task, namely the automatic prediction of the image scene category (where was the image taken?) based on a pre-defined set of scene classes.

While traditional scene recognition systems are based on features that represent the image semantics, i.e. their content, here we go beyond the mere content representation, exploiting another cue for information: the image composition, which summarizes its aesthetic and affective properties, its layout and artistic traits. Why using this type of information for semantic analysis? It is well known in photography theory [4] that the photographic techniques and intent change according to the content, and it has been verified in [14] that groups of semantically similar images can share the same compositional attributes.

We therefore build a scene categorization system (see [9] for more details) that embeds and combine both aesthetic-compositional and semantic features. We extract from popular scene categorization datasets traditional semantic features such as the SMD [8] and the Bag of Words (BOW) [1]. Moreover, we analyze the image composition by storing the values of affective, aesthetic and artistic features in the compositional vector, computed as in Sec. 2.1.

We then use Support Vector Machines to model both sources of information (compositional and semantic), and predict the scene category. By doing so, we can use our compositional feature vector for scene classification and verify its discriminative power for scene categorization. We then combine the semantic and compositional information using different fusion methods (early, namely the feature combination, and posterior, namely the prediction combination).

We test the effectiveness of our approach for scene classification on a variety of challenging datasets for scene recognition, including the SUN [18] dataset, that contains around 400 categories of very diverse scenes. For each database, we first compute the classification accuracy given the model built using each feature (BOW, SMD, CV in Fig. 2). We then look at the classification performances resulting from using our CV as a stand-alone descriptor: in all databases, CV classification performances are much better than a random classifier, which tells us that CV (aesthetic information) carries some discriminative power for scene recognition. We then combine CV with SMD in a single early fused descriptor, showing that the early combination of semantic and aesthetic analysis brings substantial classification improvement (up to +8% compared to semantic analysis only). Finally, we combine the predictions of the semantic-only and compositional-only models with posterior linear fusion. Due to the complementarity of compositional and semantic features, the categorization system benefits from this late fusion (+ 13-15% over semantic-only categorization).

2.3 The Boundaries: Where is The Limit?

Our experiments on the interactions between CA and SI pointed out not only the benefits of this synergy, but also the limitations and boundaries that this approach implies.

One first observation is that the boundary between semantic and aesthetic features is not always well established. For example, we can consider the EHD. The Edge Histogram [16] was originally built for image and video similarity assessment and concept detection, and it summarizes the local distribution of the relevant edges in the image. However, edge distribution can be seen as both semantic (it gives information about the shape of the objects) and compositional information (it can be seen as a general description of the image layout). Therefore, when we combine this type of semantic feature with the compositional analysis we perform with our compositional feature vector, we obtain a limited improvement (+0,57%) of the general MAP.

Another important deduction from our experiments is the limit of the effectiveness of compositional features for image categorization. The shooting process does not always follow compositional rules, and not always artistic and affective traits are defined. These attributes can be typically found in professional pictures, while there is a lack of attention regarding composition when dealing with user-generated or amateur pictures. For this reason, compositional features may not be clearly discriminative for the semantic classification in some type of datasets. As an example, we tried to use compositional features for the Semantic Indexing Task of the TrecVID [13] edition of 2010. The CV performances for retrieval were close to zero, and the weight assigned by the posterior fusion to the CV descriptor was null. The reason of this failure is the nature of the TrecVID videos: they are user-generated videos randomly taken by the internet. The frames therefore do not follow any particular photographic rule, and no particular attention is given on the artistic/affective factors.

3. RELATED WORK

We show here the novelty introduced by our work, one of the first attempt to improve Semantic Indexing and Computational Aesthetics through their interaction.

Semantic Indexing works generally by building frameworks for scene categorization using holistic features [8, 18], for object recognition using local features [1], or for concept detection for video retrieval [13]. Generally, such systems use local or global visual features that represent the pure image content, without considering all the information coming from the image composition, layout and shooting style. However, are compositional features useful for semantic analysis? In our work, we address this question by creating a scene categorization system that embeds some compositional features. To our knowledge, the closest work is the one presented by Van Gemert [15], that incorporates into the spatial pyramid descriptor some style attributes for object recognition. Our work differs from [15] first because of the final application (scene vs object recognition), and second because we directly apply the compositional features for semantic analysis rather than using composition to extend an existing algorithm.

On the other hand, existing aesthetic image analysis frameworks automatically define the beauty degree of an image, generally by using learning systems trained on compositional features. Datta et al. in their pioneer work [2] learn features

that model photography rules, and Wong et Al improve it in [17] by adding saliency information in the prediction framework. Here, we go beyond the pure compositional analysis by extending the pool of features used for aesthetic prediction, and embedding semantic features in the CA framework. The use of semantic features for aesthetic prediction has been explored in [3], where semantic concepts such as animals, scenes, people, are detected and the probability of their presence is used as an attribute to predict image aesthetics. Our work differs from the one in [3] because we do not train any concept model (in order to avoid complexity and prediction noise generated by the low precision of semantic indexing systems), but we instead use the semantic features in an unsupervised way, and predict the aesthetics of an image given its semantic content without explicitly labeling it. Moreover, we also improve the CA learning framework by using a graded relevance semantic indexing system, previously used for video retrieval [10].

4. FUTURE WORK

Can we further investigate the synergy between those two fields? Two main tracks can be followed for our future work.

Improving CA with SI. As said, content plays an important role for aesthetic prediction, and different contents will generally show different compositional arrangements. We therefore aim to build a content-aware aesthetic framework with multiple aesthetic models, each one built according to the characteristics of a group of visually similar images. Some work has been done in this direction by Obrador et Al. [7], that build different aesthetic models for different image categories, using pre-defined manually labeled image categories. However, the relevance of an image to one category is not always binary, as shown in [10], thus changing the compositional rules and the aesthetic appreciation. Moreover, even if extended with automatic classification, such work would be strongly dependent on the classifier performances. Our idea is to perform an unsupervised pregrouping of the training images, by automatically defining a set of appearance-based clusters based on semantic features. We could then infer an aesthetic model for each "semantic" cluster, and then predict the aesthetic degree of the image according to its group and to its aesthetic features.

Improving SI with CA. On the other hand, we can improve the scene categorization system by looking at the compositional features that are more useful to distinguish each class from the others. For example, symmetry might be more useful to identify a skyscraper scene, rather than contrast. For each classifier, we could design a set of category-specific compositional vector, which can be constructed based on the discriminative ability of each feature for the class.

5. REFERENCES

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22. Citeseer, 2004.

[2] 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.

[3] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011.

[4] B. Krages. *Photography: the art of composition*. Allworth Press, 2005.

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[6] 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.

[7] P. Obrador, M. Saad, P. Suryanarayan, and N. Oliver. Towards category-based aesthetic models of photographs. *Advances in Multimedia Modeling*, pages 63–76, 2012.

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

[9] M. Redi and B. Merialdo. Enhancing semantic features with compositional analysis for scene recognition. *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 446–455, 2012.

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

[11] M. Redi and B. Merialdo. Where is the interestingness?: retrieving appealing videoscenes by learning flickr-based graded judgments. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1363–1364. ACM, 2012.

[12] 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.

[13] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06*, New York, NY, USA, 2006. ACM Press.

[14] J. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 14. ACM, 2011.

[15] J. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 14. ACM, 2011.

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

[17] L. Wong and K. Low. Saliency-enhanced image aesthetics class prediction. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 997–1000. Ieee, 2009.

[18] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.