Content Description for Predicting Image Interestingness

Mihai Gabriel Constantin, Bogdan Ionescu
LAPI, University "Politehnica" of Bucharest, Romania
Email: {mgconstantin, bionescu}@imag.pub.ro

***Abstract*— In this article we analyze the prediction of image interestingness, a domain that is gaining importance in the fields such as recommendation systems, social media and advertising. We investigate the contribution of early and late fusion techniques, while using a set of image descriptors and analyze the best combinations that predict interestingness. Experimental validation is carried out on the MediaEval 2016 Predicting Media Interestingness image dataset. Results show the benefit of the introduction of late fusion approaches to solve the task, allowing to achieve better results than the state of the art.**

I. Introduction

Multimedia data interestingness is an area that is constantly gaining importance, considering the fact that image databases and collections are greatly growing in size and the need for a way to differentiate interesting images from the normal or uninteresting ones appeared. Concrete applications include multimedia recommendation systems, retrieval systems, video on demand scenarios, advertising, social media and education. This concept has been studied both from a psychological and a computer vision perspective.

Some psychological studies have associated and have shown the influence of certain factors like *"novelty", "uncertainty", "complexity"* and *"conflict"* [1], *"aesthetic", "pleasant", "exciting", "famous", "unusual"* etc. [6], while other researchers associated human interest with appraisal structures [14] with two components: *"novelty-complexity"* and *"coping potential"*.

From the perspective of computer vision techniques, many approaches have been developed to try to replicate the human factor in deciding the interestingness of media. For instance, authors in [6] studied several image descriptors including RGB values, SIFT histograms, GIST features, colorfulness, complexity, contrast, edge distribution, arousal features and composition of parts. The authors concluded that the most important features in determining interestingness are dependent on the type of images used in the learning process - for datasets that are composed of strong context dependent images the most important features were those that define the unusualness degree of samples, while for weaker context datasets general user preferences were more important.

The authors of [5] created three high level attributes: compositional attributes (based on concepts like rule of the thirds, depth of field, color theory and saliency), image content (detecting the presence of people or animals in the images) and sky-illumination attributes (cloudy, clear or sunset skies).

An interesting perspective is to investigate the link between *social* and *visual* interestingness [7]. Social interestingness was determined by analyzing social media scores provided by websites and defined by likes and shares, while visual interestingness was determined by pure visual features. The final results shown that there is a low correlation between these two concepts, however visual interestingness was shown to have a high correlation with the aesthetic concept. The authors created an interestingness predictor system using HSV colorspace, Local Binary Patterns, saliency and Histogram of Oriented Gradients as descriptors.

In the previously described works the authors usually used different databases and validation scenarios, thus a comparison between these approaches is hard to make, as well as to experiment other combination of methods or advanced description schemes. Although a literature exists on computer vision techniques, the advancement of the state of the art in this area is still at its earliest stages. The first and so far only common evaluation dataset regarding image interestingness is MediaEval 2016 Predicting Media Interestingness [4] Task, a benchmarking competition that bases its validation protocol on a Video-on-Demand (VoD) scenario and requires the participants to rank the image and video samples according to an interestingness score.

In this context, this paper proposes an in-depth analysis of the performance of various consacrated image description techniques for image interestingness and investigates the potential of data fusion for boosting even more the classification performance. In particular several experimental tests with early and late fusion techniques prove the benefits of fusion, allowing to achieve state of the art results. The contributions of this paper can be summarized with: (i) we provide an in-depth evaluation of most of the dedicated image descriptors for predicting image interestingness in a real-world scenario; (ii) we demonstrate the potential of appropriate late fusion to boost the performance; (iii) we setup a new baseline for the Predicting Media Interestingness Task by outperforming the performance of the other participants; (iv) evaluation is carried out on a standard data set [4] making the results both relevant and reproducible.

The rest of this paper will present the content descriptors that we investigate in Section II, the proposed early and late data fusion techniques in Section III, the experimental results in Section IV and finally Section V concludes the paper.

II. Content Description

In this section we describe some of the most effective description approaches used to predict image interestingness. Some of the approaches are inspired from connected fields such as aesthetics, style, image composition and color theory.

- *Hue, Saturation, Value from HSV space (HSV, 3 values)* and *Hue, Saturation, Lightness from HSL space (HSL, 3 values)*. These two color spaces were used in [2] and [10] and are each implemented as average therefore generating 3 values for each of these features.

- *Colorfulness (3 values)*. Three different methods of colorfulness calculation were used to generate this feature. As described in [2] and [8] this measure was obtained by separating the RGB space in 64 equal cubes, and calculating the Earth Mover's Distance [12] (v_1) and Quadratic-form distance (v_2) between two distributions: D_1 - the color occurrence frequency in each cube and D_0 - a reference distribution that generated 1/64 frequency for each sampling point. Also a standard deviation of colorfulness (v_3) was calculated for each image as presented in [8].

- *Hue Descriptors (HueDesc, 7 values)*. According to [10] when calculating hue pixel statistics, it is enough to only take into account pixels with saturation $I_S > 0.2$ and lightness $0.15 < I_L < 0.95$, hues becoming irrelevant to human perception outside these parameters. Therefore we calculated the most frequent hue for each image accordingly (v_1). A number of hues present (v_2) and missing (v_3) was calculated by dividing the hue interval into 20 equal subdivisions, and calculating a histogram of the image a hue is considered present or absent, according to [10] and [8], if it's histogram values are: $h > 0.1 \times Q$ or $h < 0.01 \times Q$ where Q is the maximum theoretical value on the histogram. Also for the present and missing values we calculated a maximum contrast (v_4 and v_5), defined by [10] as the arc-length distance on the hue wheel between the present or missing hue values. Another two values (v_6 and v_7) calculate the percentage of pixels inside the most present hue [8] and the number of insignificant hues defined by [9] as $v_7 = 20 - count(h > 0.05 \times Q)$.

- *Hue Models (HueModel, 10 values)* [10] proposed the idea that some hue models are more pleasant than others from a human perception point of view. The distance to 9 such hue models (v_1,...,v_9) and the hue model that fit the image best (v_{10}) were calculated according to the arc-length methods proposed by [10] and [8].

- *Brightness (4 values)*. Arithmetic (v_1) and logarithmic (v_2) average of brightness across the image was calculated according to the methods defined by [10]. Also two measures of contrast calculated on the brightness histogram (v_3 and v_4) were calculated by dividing the histogram into 100 [10] and 255 [9] bins.

- *Edge (3 values)*. Edge energy was used by [10] and [9] as a measure of edge distribution, by calculating the smallest bounding box that encapsulates 81% of the edge energy (v_1) and 96.04% (v_2) respectively. Authors in [8] calculated a sum of edges (v_3) as average of the Sobel images for red, green and blue channels.

- *Texture (2 values)*. The authors of [8] defined the range of texture as the average value of the sum of the maximum difference (v_1) and standard deviation (v_2) for the hue, saturation and value channels of a 3-by-3 bounding box around each pixel.

- *RGB Entropy (RGBEntropy, 3 values)*. Authors in [8] used the entropy of the red (v_1), green (v_2) and blue (v_3) channels as a descriptor of randomness and image texture.

- *HSV wavelet (HSVWavelet, 9 values)*. A three level Daubechies wavelet transform [3] was implemented in [2] as a measure of smoothness and therefore image quality on the hue, saturation and value channels, therefore generating 9 values (v_1,...,v_9). Another feature was calculating by averaging these nine values, generating *HSV wavelet average (aHSVWavelet, 3 values)*.

- *Average HSV - Rule of Thirds (aHSVRot, 3 values)*. The authors of [2] calculated average hue, saturation and value for the central portion of an image divided in 9 equal regions. This approach focused on the "Rule of the thirds" concept from the image composition domain.

- *Average HSL - Focus (aHSLFocus, 3 values)*. The central portion of the image was extended by a small factor in [10], and average hue, saturation and lightness were calculated for the resulted region. The factor used by [8] for this approach was 0.1.

- *Largest 5 segments (LargSegm, 7 values)*. Authors in [2] proposed using segmentation for determining objects' importance and using the largest objects for determining certain features and used the ratio of the 5 largest segments compared to the size of the image (v_3,...,v_7) and the number of these segments that are larger than 1% of the image (v_2) as features. The final value (v_1) was determined by the number of color based clusters obtained through K-Means in the LUV space.

- *Centroids (5 values)*. The authors of [2] calculated the geometric position of the 5 large segments, by placing their centroid (v_1,...,v_5) in one of the 3×3 portions of the image, according to $v_\alpha = (10r + c), where(r, c) \in \{(1, 1), ..., (3, 3)\}$.

- *Hue 5 segments (HueSegm, 5 values)*, *Saturation 5 segments (SatSegm, 5 values)*, *Value 5 segments (ValSegm, 5 values)* and *Brightness 3 segments (BrightSegm, 3 values)*. The average hue, saturation and value was calculated for each of the 5 largest segments, as proposed by [2]. Also, according to [10] the average brightness for the largest 3 segment was calculated.

- *Color model 5 segments (ColorSegm, 2 values)*. Authors in [2] calculates average color spread (v_1) and average complementary colors (v_2) for these 5 segments, citing the fact that opposed colors could be aesthetically pleasing when used together.

- *Coordinates 3 segments (CoordSegm, 6 values)*. Authors in [10] calculated the mass center for each of the 3 largest segments and used the horizontal and vertical positions as features indicating spatial arrangement. *Mass variance 3 segments*

(MassVarSegm, 3 values and *Skewness 3 segments (SkewSegm, 3 values)* were also calculates in the same paper, being used as shape defining features.

- Contrasts between segments (ContrastSegm, 4 values). Contrast values between the average hue (v_1), saturation (v_2), brightness (v_3) and blur (v_4) were calculated for the largest 5 segments, as maximum differences between any two segments, as indicated by [10] and [8].

- Low Depth of Field (DoF, 3 values). According to the method described by [2], the image was divided into 16 equal portions denoted by M_1,...,M_{16} and w_3 the set of wavelet coefficients in the high-frequency of the hue (v_1), saturation (v_2) and value (v_3) components, therefore generating the following equations:

$$v_\alpha = \frac{\sum_{(x,y)\in M_6 \cup M_7 \cup M_{10} \cup M_{11}} w_3(x,y)}{\sum_{i=1}^{16} \sum_{(x,y)\in M_i}} \quad (1)$$

III. Data Fusion

As a testing framework, the prediction of interestingness is carried out by employing a Support Vector Machine (SVM) classifier with linear, polynomial and RBF kernels and with different coefficients, considering that SVM are well known to provide state of the art results. Apart from employing individual features, we investigate the use of fusion techniques.

Early fusion consists of combining, before the classification, several individual features, and using the resulting concatenated feature as an input for the classification algorithm.

Late fusion, on the other hand, performs the fusion, late in the system, typically after the classification is carried out. The aim is to fuse the results obtained with different systems exploiting their complementary effectiveness. Basically, the late fusion consists of designing a functional that combines the out confidence level of several, weaker, systems. We experimented 4 strategies for our late fusion approach, namely: *CombSum, CombMean, CombMax* and *CombMin* which are described in the following.

Considering that, for each given image denoted Img, we dispose of a set of N classifiers, *CombMean* will have the following formula:

$$CombMean(Img) = \sum_{i=1}^{N} w_i c_i. \quad (2)$$

where c_i represents the confidence values generated by each of the N classifiers for that image and w_i represents weights assigned to each classifier. In our approach we chose values for the w_i weights according to the final reverse rank of each classifier involved, $w_i = 1/(2^{rrank(i)})$, where the $rrank$ function returns 0 for the best classifier, 1 for the second best and so on. *CombSum* is a case where the weights have the same value $w_1 =, ..., = w_N$.

The final two late fusion strategies use only the minimum (*CombMin*) or maximum (*CombMax*) of the confidence values c_i for each image.

IV. Experimental results

Experimentation is carried out on the MediaEval 2016 Predicting Media Interestingness Task, and particularly on the image dataset. The dataset is composed of a development data, containing 5,054 images (out of which 473 are labeled as interesting), intended for training the models, and a testing dataset consisting of 2,342 images (out of which 241 images as interesting) for the actual validation. The official evaluation metric is Mean Average Precision (MAP).

A. Experimenting with individual group of features

Our first experiment consisted of evaluating all the features presented in Section II. Generally the individual feature approach had lower MAP results than either of the fusion approaches, as we will show in the following subsection. Each of these descriptors was trained with a SVM system, with linear kernel, polynomial kernel and RBF kernel, each having the degree and gamma parameters in 2^k, where $k \in [-6, ..., 6]$.

The best 5 performing features are as follows: aHSVwavelet (MAP = 0.2057), SatSegm (MAP = 0.2057), HSV (MAP = 0.2051), RGBEntropy (MAP = 0.2023) and CoordSegm (MAP = 0.2008). For the learning systems, RBF was the best performing kernel. However individual features do not seem to generate a trend regarding the types of features that will perform best, having in the top 5 texture, image objects and color descriptors.

B. Experimenting with early fusion

Regarding the early fusion we chose to run every possible combination of 2 features, generating a total of 300 new combinations, and we took the best 10 of them and again combined them with other features generating an additional 230 combinations. Further combinations were performed with 4 or more features, each time taking the best combinations from earlier runs and concatenating features. For classification, we vary the SVM parameters as in the previous experiment.

The best 5 performing combinations are: SatSegm + MassVarSegm + SkewSegm (MAP = 0.2363), aHSVWavelet + HueSegm + SatSegm (MAP = 0.2261), HSL + LargSegm + BrightSegm (MAP = 0.2232), ColorSegm + SkewSegm (MAP = 0.2231) and aHSVWavelet + HueSegm + SatSegm + MassVarSegm (MAP = 0.219).

Again the best results were achieved with RBF kernel, however this time we can see some patterns emerging, and some logically related features are starting to give better results - for example MassVarSegm and SkewSegm representing the mass variance and skewness for the largest 3 segmentation objects, HueSegm and SatSegm representing the saturation and value for the largest 5 segmentation objects. Another important observation is that many of the features present in this top5 deal with descriptors concerning the largest segments in the images. A final observation is that the concatenation of all the features produces a MAP of 0.1936, lower than other results in the early fusion approach. A reason for that may be the redundancy in the merged information, early fusion being unable to exploit the effectiveness of each data source.

TABLE I

COMPARISON WITH MEDIAEVAL 2016 PREDICTING MEDIA INTERESTINGNESS TASK STATE OF THE ART RESULTS

Approach	MAP	Description
Late fusion	0.2485	CombMax (SVM-RBF with aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)
Late fusion	0.2451	CombMean (SVM-RBF with aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm and HSL + LargSegm + BrightSegm)
Late fusion	0.2448	CombMean (SVM-RBF with aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)
Late fusion	0.2408	CombSum (SVM-RBF with aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm)
Late fusion	0.2403	CombMax (SVM-RBF with aHSVWavelet + HueSegm + SatSegm and and SatSegm + MassVarSegm + SkewSegm and HSL + LargSegm + BrightSegm)
Early fusion	0.2363	SVM-RBF (aHSVWavelet + HueSegm + SatSegm)
TUD-MMC [11]	0.2336	Normalized confidence score of color histogram and face area detection
Tehnicolor run 1 [13]	0.2336	AlexNet fc7 + SVM
Tehnicolor run 2 [13]	0.2315	AlexNet fc7 + MLP

C. Experimenting with late fusion

Several top performing classifiers from the individual features and early fusion approach experiments were then combined in the late fusion approach, giving us the best results as we will show in the following.

Our results show that the best performing strategy is the CombMax (MAP = 0.2485), followed by CombMean (MAP = 0.2451), CombSum (MAP = 0.2408) and CombMin (MAP = 0.2069). While CombMax, CombMean and CombSum definitely improved our final results, CombMin had lower results than the best early fusion approach. As expected, the best performing strategies are the ones that include the best combinations of features, and the best achieved score overall was a MAP of 0.2485 for a CombMax strategy combining the early fusion features aHSVWavelet + HueSegm + SatSegm and SatSegm + MassVarSegm + SkewSegm.

D. Comparison with state-of-the-art

Our final experiment involved comparing our results with the MediaEval 2016 Predicting Media Interestingness image subtask teams results [4]. The results are presented in Table I.

No individual feature performed above the state-of-the-art MAP 0.2336 [11], [13], however we have shown that there is at least an early fusion scheme that can out-perform that score, and that many late fusion combinations perform above this target value. The 5 best MAP scores were all the results of late fusion techniques, with a maximum score of 0.2485, which means an improvement of 0.0149 percentage points compared to the state of the art.

V. CONCLUSION

In our paper we approached the problem of predicting image interestingness. We experimented with several features and early and late fusion combination strategies. We experimented on the MediaEval2016 Predicting Media Interestingness image subtask dataset. Our results show that at least one early fusion combination performs above the best result reported at MediaEval2016, while the late fusion approach improves our scores even more, many strategies surpassing this value. Our best MAP score was 0.2485, while the best score reported at MediaEval was 0.2336, thus proving the benefits of data fusion techniques for the descriptors we studied. Another interesting result is the fact that the insertion of object or segment analysis in the feature set and in the fusion techniques has the greatest impact on the overall score. Future work should address the addition of more features to the learning systems, including but not limited to CNN features and the study of our system's results on a video dataset.

ACKNOWLEDGMENT

Part of this work was funded under research grant PN-III-P2-2.1-PED-2016-1065, agreement 30PED/2017, project SPOTTER.

REFERENCES

[1] D. E. Berlyne. *Conflict, arousal and curiosity.* Mc-Graw-Hill, 1960.

[2] R. Datta, D. Joshi, J. Li and J. Z. Wang, *Studying aesthetics in photographic images using a computational approach.* In European Conference on Computer Vision (pp. 288-301). Springer Berlin Heidelberg, 2006.

[3] I. Daubechies, *Ten lectures on wavelets.* Society for industrial and applied mathematics, 1992.

[4] C. H. Demarty, M. Sjberg, B. Ionescu, T. T. Do, H. Wang, N. Q. Duong and F. Lefebvre. *Mediaeval 2016 predicting media interestingness task.* In Proc. of the MediaEval 2016 Workshop, Hilversum, Netherlands, 2016.

[5] S. Dhar, V. Ordonez, and T. L. Berg, *High level describable attributes for predicting aesthetics and interestingness.* In IEEE International Conference on Computer Vision and Pattern Recognition, 2011.

[6] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater and L. van Gool, *The interestingness of images.* In ICCV International Conference on Computer Vision, 2013.

[7] L.-C. Hsieh, W. H. Hsu and H.-C.Wang, *Investigating and predicting social and visual image interestingness on social media by crowdsourcing.* In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 43094313. IEEE, 2014.

[8] A.F. Haas, M. Guibert, A. Foerschner, T. Co, S. Calhoun, E. George, M. Hatay, E. Dinsdale, S.A. Sandin, J.E. Smith, M.J.A. Vermeij, B. Felts, P. Dustan, P. Salamon and F. Rohwer, *Can we measure beauty? Computational evaluation of coral reef aesthetics.* PeerJ 3:e1390, 2015.

[9] Y. Ke, X. Tang and F. Jing, *The design of high-level features for photo quality assessment.* In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on (Vol. 1, pp. 419-426). IEEE, 2006.

[10] C. Li and T. Chen, *Aesthetic visual quality assessment of paintings.* IEEE Journal of Selected Topics in Signal Processing, 3(2), 236-252, 2009.

[11] C. Liem, TUD-MMC at MediaEval 2016 Predicting Media Interestingness Task. In Proceed- ings of the MediaEval Workshop, Hilversum, Netherlands, October 2016.

[12] Y. Rubner, C. Tomasi and L. J. Guibas, *The earth mover's distance as a metric for image retrieval.* International journal of computer vision, 40(2), 99-121, 2000.

[13] Y. Shen, C.-H. Demarty, and N. Q. K. Duong. Technicolor@MediaEval 2016 Predicting Media Interestingness Task. In Proceedings of the MediaEval Workshop, Hilversum, Netherlands, October 2016.

[14] P. J. Silvia, *What is interesting? exploring the appraisal structure of interest.* Emotion, 5(1):89, 2005.