DA-IICT at MediaEval 2017: Objective prediction of media interestingness

Rashi Gupta, Manish Narwaria
Dhirubhai Ambani Institute of Information and Communication Technology
rashi.8496@gmail.com,manish_narwaria@daiict.ac.in

ABSTRACT

Interestingness is defined as the power of engaging and holding the curiosity. While humans can almost effortlessly rank and judge interestingness of a scene, automated prediction of interestingness for an arbitrary scene is a challenging problem. In this work, we attempt to develop a computational model for the said problem. Our approach is based on identifying and extracting context-specific features from video clips. These features are subsequently utilized in a predictor model to provide continuous scores that can be related to the interestingness of the scene in question. Such computational models can be useful in a automated analysis of videos (eg. movie, a CCTV footage or a clip from an advertisement).

1 INTRODUCTION

The aim of the task is to select content (image and video clips) which are considered to be the most interesting for a common viewer. This is a challenging task because interestingness of media is highly subjective, and can depend on multiple aspects including personal preferences, emotional state and the content itself. Therefore, as a first step, our goal in this task is to understand and extract signal related features which may, for instance, quantify visual appearance and audio information. Such features can then be mapped into an interestingness score via machine learning. Further details about the task and dataset can be retrieved from [1].

The said task falls under the broad areas of multimedia signal processing (image, video and audio) and machine learning. The former focuses on analysis and extraction of context-specific features from the signal. These may include color, contrast, complexity, audio characteristics etc. The primary goal of feature extraction is to obtain a more meaningful signal representation from the view point of capturing useful information pertaining to media interestingness. In the task, these features will be subsequently used as input to a regressor (eg. linear regression and multilayer perceptron). As the target value of such regression problem is known (equal to interestingness score given by a panel of human subjects), this is a supervised learning problem.

We note that a similar approach has been used in previous works such as [2], [3]. However, the key difference lies in the features used, and this is one of the contributions from the task. Also, the results shed light on some interesting aspects of interestingness that may not be fully captured by the current set of features. This can obviously be used to improve feature extraction, and in the process predict objective media interestingness scores that are closer to human judgments.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

2 APPROACH

2.1 Interestingness Features Computation

Following are the features extracted from the image for image subtask: colorfulness, contrast, complexity and visual attention. For video subtask, along with these features, audio feature Mel-frequency cepstral coefficients (MFCCs) is also computed to take audio of the clip in the account.

Colorfulness: We measure colorfulness as proposed by [6], Red-green and yellow-blue spaces are used where $\alpha = R - G$ and $\beta = 0.5(R + G) - B$ where σ_α^2, σ_β^2, μ_α, μ_β represent the variance and mean values along these two opponent color axes defined as:

$$\mu_\alpha = \frac{1}{N}\Sigma_{p=1}^{N}\alpha_P \text{ and } \sigma_\alpha^2 = \frac{1}{N}\Sigma_{p=1}^{N}(\alpha_P^2 - \mu_\alpha^2)$$

The equation formulates the ratio of the variance to the average chrominance in each of the opponent component:

$$colorfulness = 0.02 \times log(\frac{\sigma_\alpha^2}{|\mu_\alpha|^{0.2}}) \times log(\frac{\sigma_\beta^2}{|\mu_\beta|^{0.2}})$$

Contrast: We measure contrast as proposed by [5]. The main idea is to compute local contrast factors at various resolutions, and then to build a weighted average in order to get the global contrast factor. Let us denote the original pixel value with k, k = 0, 1,.. 254, 255. The first step is to apply gamma correction with $\gamma = 2.2$, and scale the input values to the [0,1] range. The corrected values linear luminance is $l = (\frac{k}{255})^\gamma$. The perceptual luminance L can be approximated with the square root of the linear luminance: $L = 100 \times \sqrt{l}$. Once the perceptual luminances are computed we have to compute local contrast. For each pixel we compute the average difference of L between the pixel and four neighboring pixels.

$$lc_i = \frac{|L_i - L_{i-1}| + |L_i - L_{i+1}| + |L_i - L_{i-w}| + |L_i - L_{i+w}|}{4}$$

The average local contrast for current resolution C_i is computed as the average local contrast lc_i over the whole image, where the image is w pixels wide and h pixels high.

$$C_i = \frac{1}{w \times h} \times \Sigma_{i=1}^{w \times h} lc_i$$

We have to compute the C_i for various resolutions. Once the C_i for original image is computed, we build a smaller resolution image, so that we combine 4 original pixels into one super pixel. The image width is half the original width and the image height is half the original height now. The C_i for various resolution can easily be computed and the process continues until we have only few huge superpixels in the image. Now that we have computed average local

contrasts C_i, we can compute the global contrast factor.

$GCF = \Sigma_{i=1}^{N} w_i \times C_i$

Complexity: We measure contrast by calculating Spatial Information as proposed by [7]. Let s_h and s_v denote gray-scale images filtered with horizontal and vertical Sobel kernels, respectively. $SI_r = \sqrt{s_h^2 + s_v^2}$ represents the magnitude of spatial information at every pixel. Mean and standard deviation of SI_r is used to calculate the complexity of an image. $SI_{mean} = \frac{1}{P}\Sigma SI_r$ and $SI_{stdev} = \sqrt{\frac{1}{P}\Sigma(SI_r^2 - SI_{mean}^2)}$
where P is the number of pixels in the image.

Visual attention: We propose a method to calculate attention of an image by computing saliency maps for the corresponding image. [4] implementation is used for saliency map computation. The mean of this saliency map at every pixel is the attention value.

Audio Extraction: For audio features, Mel-frequency cepstral coefficients (MFCCs) are computed and mean and standard deviation of a time frame is calculated. This is the feature vector for audio extraction.

Novelty: We propose a method to calculate novelty by firstly calculating saliency maps for the images. 8 X 8 average filter is convoluted and the mean is calculated for the consecutive saliency map images. If for both of the blocks, the average is less than the threshold (0.1), the block is avoided. Otherwise, for the two consecutive saliency map images mean squared error (MSE) is calculated. If MSE is less than the threshold, this block is also ignored, else it is considered. Hence, the mean of the MSE is calculated. Higher the value, more action has happened in the two consecutive frames.

2.2 Interestingness Prediction

For image subtask, we have used five features namely, colorfulness, contrast, complexity (mean and standard deviation) and attention. With the help of these features, we would learn our model for interestingness using linear regression.

For video subtask, along with these five features, we also added the feature vector for the audio. As there are more features in this case, we used multilayer perceptron. We used the mean image provided in the image subtask for computation of the feature vector of the video subtask as computation for all frames of the video was not feasible, given the time constraints.

3 EXPERIMENTAL RESULTS

3.1 Evaluation

For image subtask, using the five features with the linear regression as the learning algorithm the MAP@10 is coming up to be 0.0406 for the test set. For video subtask, along with these features and the audio feature vector, MAP@10 is 0.0636 for the test set where multilayer perceptron is the learning algorithm.

3.2 Analysis

For image subtask, the maximum value of average precision@10 is for those videos in which the top 10 images which are interesting in common view are more colorful, have high variation and contrast. It also includes the images which are eye catchy because of the visual attention feature. Example being the type of images that have many people gathering like in a rebellion or maybe in a meeting, or wearing clothes with vibrant setting say in a party.

Similarly, for the video subtask, the maximum value of average precision@10 is for those videos in which the top 10 clips which are interesting in common view are more colorful, have high complexity and traps attention. A clip which has high audio seems to attract more viewer. Example being a clip which shows a blast or people screaming are of more significance than a silent clip.

On the contrary, the lowest average precision@10 is for those kinds of videos in which the most interesting image are the ones which are less colorful and has very fewer variations. Example of such scenes includes the one in which say a dark wall with some strange symbols is painted. This may be because these symbols have some back story in the movie and hence are interesting in common view. Other being the scene where some explicit content is shown, it is usually shown in dark with very less variation and is arousing for humans. In such cases, the model tends to predict following kinds of images more interesting: a crowded place which may have no greater significance or complex buildings and road of no greater importance or just a lighted empty room.

For video subtask, the lowest average precision@10 is because, in these scenes, the audio is also negligible be it a moment of suspicion or any of the examples mentioned for the low value in image subtask and would instead predict those clips to be interesting which have higher audio along with the other features as in image subtask.

4 CONCLUSION

Interestingness of a scene is a subjective aspect and one that involves complex cognitive processes. However, certain features such as contrast, colorfulness and novelty of the scene can be assumed to play a part in the way humans quantify interestingness, irrespective of the type of scene. Therefore, in this work, we extracted and used such audio and visual features to develop a model for predicting interestingness. Such approach is, of course, an initial step towards building a more comprehensive model. The novelty feature proposed in this paper is not used for the current task due to time constraint and can be exploited in the future work.

ACKNOWLEDGMENTS

We thank Karan Thakkar for his fruitful help.

REFERENCES

[1] Demarty et al. 2017. Mediaeval 2017 predicting media interestingness task. (2017).

[2] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. 2013. Visual interestingness in image sequences. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 1017–1026.

[3] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1633–1640.

[4] Jonathan Harel, C Koch, and P Perona. 2006. A saliency implementation in matlab. *URL: http://www. klab. caltech. edu/˜ harel/share/gbvs. php* (2006).

[5] Kresimir Matkovic, László Neumann, Attila Neumann, Thomas Psik, and Werner Purgathofer. 2005. Global Contrast Factor-a New Approach to Image Contrast. *Computational Aesthetics* 2005 (2005), 159–168.

[6] Karen Panetta, Chen Gao, and Sos Agaian. 2013. No reference color image contrast and quality measures. *IEEE transactions on Consumer Electronics* 59, 3 (2013), 643–651.

[7] Honghai Yu and Stefan Winkler. 2013. Image complexity and spatial information. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*. IEEE, 12–17.