DETECTING CARTOONS: A CASE STUDY IN AUTOMATIC VIDEO-GENRE CLASSIFICATION

*Tzvetanka I. Ianeva**

Departamento de Informática,
Universidad de Valencia, Valencia, Spain
tzveta.ianeva@uv.es

Arjen P. de Vries, Hein Röhrig

Centrum voor Wiskunde en Informatica,
Amsterdam, The Netherlands
{arjen,hein}@acm.org

ABSTRACT

This paper presents a new approach for classifying individual video frames as being a 'cartoon' or a 'photographic image'. The task arose from experiments performed at the TREC-2002 video retrieval benchmark: 'cartoons' are returned unexpectedly at high ranks even if the query gave only 'photographic' image examples. Distinguishing between the two genres has proved difficult because of their large intra-class variation. In addition to image descriptors used in prior cartoon-classification work, we introduce novel descriptors like ones based on the pattern spectrum of parabolic size distributions derived from parabolic granulometries and the complexity of the image signal approximated by its compression ratio. We evaluate the effectiveness of the proposed feature set for classification (using Support Vector Machines) on a large set of keyframes from the TREC-2002 video track collection and a set of web images. The paper reports the identification error rates against the number of images used as training set. The system is compared with one that classifies Web images as photographs or graphics and its superior performance is evident.

1. INTRODUCTION

TREC is a series of workshops for large scale evaluation of information retrieval technology (e.g., see [1]). The goal is to test retrieval technology on realistic test collections. TREC-2001 has introduced a video retrieval task, on a collection of (copyright free) videos produced between the 1930s and the 1970s (including advertising, educational, industrial, and amateur films). The videos vary in their age, productional style, and quality [2].

Our experiments for the search task at the video track studied a generic probabilistic retrieval model that ranks shots based on the content of their keyframe image and speech transcript [3]. Evaluating the results, we noticed that the model does not distinguish sufficiently between 'cartoons' and other keyframe images.[1] Of course, one generally does not expect a 'cartoon' as query result unless explicitly asked for; consequently, returning these 'cartoons' by mistake results in a lower precision of our system.

The objective of this study is to implement a classifier that distinguishes 'cartoon' keyframes from 'non-cartoons'. The problem can be viewed as a case study of automatic genre classification. The paper describes an approach which employs both grayscale and color image features. The output from various feature extractions is combined in a Support Vector Machine (SVM) training process to produce a classification model. The results demonstrate a small error rate on both the TREC-2002 video corpus and a collection of images gathered from the WWW.

The main contributions of this research are a rigorous analysis of the classification results on a large corpus, the use of image morphology in the feature set, as well as the good results achieved in spite of difficult, inhomogeneous data.

1.1. Related work

Roach et al. published an approach for the classification of video fragments as cartoons using motion only [4]. Yet, their database consisted of only 8 cartoon and 20 non-cartoon sequences, so it is difficult to predict how it would perform on the TREC corpus, and their data set is not publicly available. Another recent effort addressed the classification of video into seven categories (including cartoons) [5]. Two of our features are similar to theirs, but our approach is different and the experiments are incomparable.

A closely related problem is the automatic classification of WWW images as photographs or graphics; examples are the WebSeek search engine [6] and the systems described in [7, 8]. Unfortunately, the most discriminative features used in these works take advantage of some characteristics of web images that do not exist in video keyframes, notably the aspect ratio and the word occurrence in the image URLs. We applied the farthest neighbor histogram descriptor suggested by [8] to our data collection, but this characteristic is expensive to compute without resulting in improved error rates.

The photo/graphics classifier of [8] had been previously implemented in our group as part of the Acoi system [9]. A decision rule classifier (C4.5, [10]) has been trained on the features given in [8] on varying quantities of training data. However, as the results in Section 3 show, the features do not provide enough (or even none) discriminating power in the case of photo/cartoon classification on the TREC video collection. Conversely, the same implementation has a classification score of 0.9 on a data set of 14,040 photos and 9,512 graphics harvested from the WWW.

*This work has been done while at CWI, supported by grants GV FPIB01_362, CTESPR/2002/12 and the ICES-KIS MIA project.

[1]The class 'cartoon' is defined more precisely in Section 1.2.

Image Descriptor	Dim.	$\mathcal{E}(p)$	$\mathcal{E}(c)$	$\mathcal{E}(t)$
average saturation	1	0.26	0.45	0.27
threshold brightness	1	0.27	0.48	0.29
color histogram	45	0.19	0.28	0.19
edge-direction histogram	40	0.48	0.24	0.46
compression ratio	1	0.36	0.42	0.36
multi-scale pat.spectrum	30	0.36	0.44	0.36

Table 1. Overview of Image Descriptors. Training set 500 photos and 500 cartoons; test set 12,526 photos and 1,120 cartoons; all sets disjoint random samples from TREC-2002 keyframes.

1.2. What is a Cartoon?

We call images or parts of images *photographic material* or *photos* if they have been obtained by means of a photographic camera (movie or still). We call images *cartoons* if they do not contain any photographic material. Some distinguishing features of cartoons are:[2]

Few, simple, and strong colors: The abstraction in transforming a real-world scene into the cartoon world leads to a reduction of colors and exaggeration in saturation.

Patches of uniform color: Textures are often simplified to uniform color.

Strong black edges: The large patches of uniform color are often surrounded by strong black edges.

Text: Educational cartoons, charts, etc. often contain large text that is typically horizontal and not distorted by a perspective transformation. Moreover, the fonts are chosen to be readable and the colors to give good contrast.

Given this list, it may seem easy to separate the cartoons from the other keyframes. But in practice the problem is not as simple as expected. Part of the problem lies in the low quality of the video streams in the collection. More problematically, keyframes may contain an artificial building, or, people in an theatrical scene. Clearly, by the definition given before, both are in the 'photograph' class, but such 'mixed' images cause quite a challenge for the machine. We invite the reader to look at the examples shown in Figure 1 to see some keyframes that sparked some heated discussion in our group while making the ground truth.

2. IMAGE DESCRIPTORS

We first outline the image descriptors used previously and then explain our innovations. Refer to Table 1 for an overview of the image descriptors. Their individual 'naive' usefulness is given by the error on photos $\mathcal{E}(p)$, error on cartoons $\mathcal{E}(c)$, and total error $\mathcal{E}(t)$, when performing classification with machine learning (as outlined in Section 3) using the given image descriptor alone.

2.1. Saturation, Brightness, and Color Histogram

We compute average color saturation (the S channel of the image in the HSV color space) and the ratio of pixels with brightness greater than 0.4. Similar descriptors were used in [5], and we observed a good correlation with class membership alike. We compute a $3\times3\times5$ histogram of the image in the HSV color space and use the 45 numbers (normalized by the number of pixels) as 45 image descriptors.

2.2. Edge Detection

Cartoons are expected to have strong black edges. We measure the distribution of edges in an image by a histogram of the angles and absolute values of the gradient in each point: we approximate the horizontal and vertical derivatives by separately filtering the image with the horizontal and vertical Sobel filters. Let $\partial/\partial xI(x,y)$ and $\partial/\partial yI(x,y)$ denote the horizontal and vertical derivatives of the image I at point (x,y). Then the angle $\theta(x,y)$ of the gradient at (x,y) satisfies $\tan\theta(x,y) = (\partial/\partial yI)/(\partial/\partial xI)$ and for the absolute value $m(x,y)$ holds $m(x,y) = ((\partial/\partial xI(x,y))^2 + (\partial/\partial yI(x,y))^2)^{1/2}$. From our approximations of $\theta(x,y)$ and $m(x,y)$ we compute a two-dimensional histogram with eight bins for angles and five bins for absolute value; this yields 40 image descriptors normalized to values between 0 and 1. This approach is inspired by [11] where a similar technique is outlined.

2.3. Compression

Cartoons are expected to have a more simple composition than photographic images: they typically have few colors, simple geometric shapes, etc. As an approximation to the generic minimum description-length (Kolmogorov complexity [12]) of the image, we considered compression ratios with common lossy and lossless image-compression techniques.

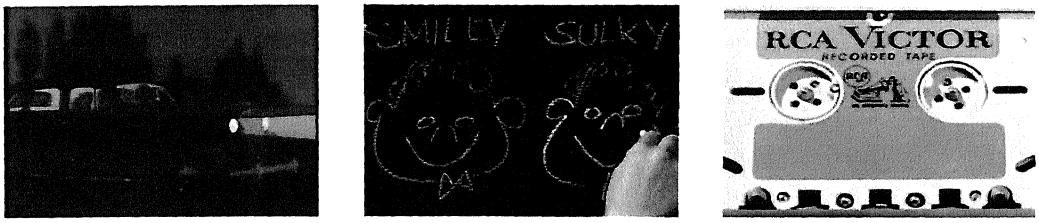
Experiments with lossy compression with JPEG at various quality settings gave little correlation. This is probably due to our source material, which was captured from TV signals (blurring sharp edges and adding noise to uniform colors) and stored in the JPEG format. On the other hand, quantizing to 256 colors and lossless compression using the PNG format gave compression ratios with high predictive power.

2.4. Pattern Spectrum

Granulometry is a widely used tool in mathematical morphology for determining the size distribution of objects in an image without explicitly segmenting each object first. Intuitively, granulometry treats image objects as particles whose sizes can be established by sifting them through sieves of increasing mesh width, and collecting what remains in the sieve after each pass. One pass of sifting and retaining the residue is analogous to the morphological 'opening' of an image using a 'structuring element' of a certain size. Morphological opening of the image performs an erosion followed by a dilation using the same structuring element.

In mathematical terms, a granulometry is defined by a transformation Φ_λ with size parameter λ that satisfies the *Anti-extensivity*, *Increasingness* and *Absorption* axioms [13]. Of particular interest are granulometries generated by openings by scaled versions of a single convex structuring element B, i.e., $\Phi_\lambda(I) = I \circ \lambda B$. There exist many applications that use flat structuring elements; e.g., rectangular

[2] Since in this work we were only concerned with single frames, we did not consider time-dependent properties of cartoons.



(a) (b) (c)

Fig. 1. Sample keyframes from TREC-2002. (a) a 'clear' cartoon image, but the exaggerated light beams are the only distinguishing feature. (b) illustrates the difficulty of mixed images. For (c), only the sparkle of the metallic parts and the shadows justify the classification as 'photograph.'

size distributions are used to characterize visual similarity of document images [14]. However, in our case the objects do not have a particular geometric shape and therefore we need a generic and nonflat structuring element. For nonflat structuring elements B, the erosion $\epsilon_B(I)$ of the grayscale image I is defined as $[\epsilon_B(I)](x,y) = \min_{(a,b)\in \mathrm{dom}\ B}\{I(x+a, y+b) - B(a,b)\}$, and dilation $\delta_B(I)$ analogously, i.e., the nonflat "shape" is subtracted or added before taking the minimum or maximum, respectively [13].

Van den Boomgaard and Smeulders [15] have shown that the Gaussian function used in linear convolutions has as morphological analogue the parabola. We use a parabola as structuring element because it is the unique structuring function that is both rotational symmetric as well as dimensional decomposable [16]. Erosion (and analogously dilation) with a n-dimensional parabola can be decomposed into the erosion (dilation) with n one-dimensional parabolae, reducing the complexity by a factor n [17].

Formally, the two-dimensional parabola of scale λ is $B_\lambda(x,y) = -(x^2+y^2)/\lambda$. The decomposition property allows us to compute erosion with B_λ efficiently using one-dimensional structuring elements: $\epsilon_{B_\lambda}(I) = \epsilon_{V_\lambda}(\epsilon_{H_\lambda}(I))$ for H_λ and V_λ the one-dimensional horizontal and vertical scale-λ parabolae, respectively: $H_\lambda(x,y) = -\infty$ for $y \neq 0$, $H_\lambda(x,0) = -x^2/\lambda$, and $V_\lambda(x,y) = H_\lambda(y,x)$. Of course, the same optimizations also apply to dilation $\delta_{B_\lambda}(I)$.

Let $\Phi_\lambda(I)$ denote the opening of the grayscale image I with parabola B_λ. For values $\lambda_1 < \cdots < \lambda_k$, the normalized size distribution induced by the granulometry Φ_λ is $s(i) = 1 - \sum_{x,y}[\Phi_{\lambda_i}(I)](x,y) / \sum_{x,y} I(x,y)$. The corresponding pattern spectrum is $p(i) = s(i+1) - s(i)$. Based on the distinguishing features of cartoons like large patches of uniform color we expect differences between cartoons and photographs in the pattern spectrum: a peak in the pattern spectrum at a given size indicates that there are many objects of that size in the image. Hence, we store as image descriptors a 'small scale' pattern spectrum with $\lambda_i = 2.55i$, $i = 1,\ldots,20$ and a 'large scale' pattern spectrum with $\lambda_i = 1275i$, $i = 1,\ldots 10$.

3. EXPERIMENTAL RESULTS

As mentioned in the introduction, we used keyframes extracted by Westerveld et al. [3] from the TREC-2002 video track [2]. These keyframes consist of some 24,000 JPEG images of dimension 340×252 and average file size 46 kBytes. 14,000 of these images were classified manually into the categories 'photograph' (13,026), 'cartoon' (1,620), and 'borderline' (354). From this data we randomly selected subsets of equal numbers of photographs and cartoons for training and cross-validation.

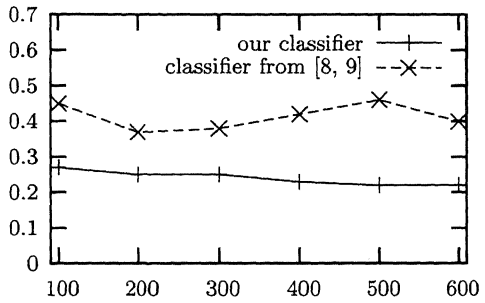


Fig. 2. Size of learning set versus error. Test set constant size 1,000 photos and 1,000 cartoons.

For some of our one-dimensional image descriptors, we have a clear intuition how they distinguish cartoons from photographs and 'learning' to use such a descriptor reduces to determining a good threshold value. For others (especially the various histograms) it is more convenient to classify automatically the patterns that are typical for cartoons or photographs. For such generic classification tasks, a popular and often successful technique is Support Vector Machine (SVM) learning [18], which has built-in guards against overfitting and can be tailored to known structure in the data by the choice of the kernel.

We used the OSU SVM Classifier Toolbox for Matlab [19]; after initial experiments with both polynomial and RBF kernels, we opted for a Gaussian RBF kernel with variance $\sigma^2 = 0.25$. The smaller σ^2, the more complex will the resulting model be, hence the more training data is needed and the greater is the danger of overfitting. Figure 2 shows that the classification error obtained does not depend much on the size of the training set.

For comparison, we also trained a classifier on the previously mentioned Web data (23,552 images). Even without

Image Descriptors	$\mathcal{E}(p)$	$\mathcal{E}(c)$	$\mathcal{E}(t)$
all	0.17	0.25	0.18
all w/o brightness	0.16	0.27	0.17
all w/o saturation	0.16	0.27	0.17
all w/o color histogram	0.25	0.27	0.25
all w/o edge direction	0.19	0.24	0.19
all w/o compression	0.18	0.24	0.18
all w/o multi-scale pat.spect.	0.18	0.25	0.18

Table 2. Relative power of descriptors. Sets as in Table 1

using image descriptors depending on file type, dimension ratio, or smallest dimension (very good indicators of logos, banners, etc.), we obtain a 92% correct classification (slightly better than the decision tree classifier mentioned before [8] that uses file type and image size). When adding these properties to the feature vectors, our classification rate improves marginally (to 94%); unlike [8], we have not noticed a significant difference in accuracy depending on the image file type (JPEG or GIF). Conversely, when training and testing the decision tree on our data, the classification accuracy suffers severely from the fact that all keyframes are of the same size; the error rate goes up to over 45%. We believe that this comparison demonstrates that the approach outlined in the paper is generic and can be applied equally well to distinguish graphics from photos — without using derived properties like image size: *only the visual content is modeled in our characteristics!*

4. CONCLUSIONS AND OUTLOOK

We have shown that a generic image classifier based on well-chosen visual features can distinguish cartoons from photos on a difficult video corpus, and identify the graphics in a collection of Web images. The results can most likely be improved further only using higher level, semantic descriptions. Even for an anticipated easy problem like this one, our experience shows that people use a significant amount of world knowledge (like shining objects, shadows, human body parts, and so on). Low-level characteristics that we have not used (like the temporal structure of the shot) may help a little bit further, but it is unlikely that this can bridge the semantic gap (again, see Figure 1).

Recall our original goal: filter out cartoons from the returned results in the TREC-2002 results. Although we reduced the error significantly, only one out of ten keyframes in the corpus is a cartoon; hence, always deciding the image is photographic is still better than using the classifier. If we wanted to never miss a cartoon, the classifier is quite useful; unfortunately, most queries prefer not to miss photographic keyframes more than to view some cartoons by mistake. This problem is not unique to our current scenario. The 'feature detection task' at TREC has shown similar accuracy of classifiers for other events (like 'people', 'outdoors', and 'landscape'). The example query 'give me outdoors images without people' will thus throw out too many outdoors images. We conclude that current systems have inherent problems when facing infrequent events and therefore cannot answer requests with negative conditions.

Acknowledgments: We thank A. Bagdanov for his advice concerning parabolic granulometries, M. Windhouwer for the photo/graphics classifier comparison, and R. Cilibrasi for help with image compression.

5. REFERENCES

[1] E. M. Voorhees and D. K. Harman, Eds., *The Eleventh Text Retrieval Conference (TREC-2002)*, 2002, To appear.
[2] A.F. Smeaton and P. Over, "The TREC 2002 video track report," In Voorhees and Harman [1], pp. 171–181, To appear.
[3] T. Westerveld, A.P. de Vries, A. van Ballegooij, F.M. G. de Jong, and D. Hiemstra, "A probabilistic multimedia retrieval model and its evaluation," *EURASIP Journal on Applied Signal Processing*, 2003.
[4] M. Roach, J. S. Mason, and M. Pawlewski, "Motion-based classification of cartoons," in *Int. Symposium on Intelligent Multimedia*, 2001.
[5] B. T. Truong, C. Dorai, and S. Venkatesh, "Automatic genre identification for content-based video categorization," in *Proc. 15th International Conference on Pattern Recognition*, Barcelona, Spain, September 2000, vol. II, pp. 230–233.
[6] J. R. Smith and S.-F. Chang, "Searching for images and videos on the world wide web," Tech. Rep. 459-96-25, Center for Communications Research, Columbia University, 1996.
[7] N. C. Rowe and B. Frew, "Automatic caption localization for photographs on word wide web pages," Tech. Rep., Department of Computer Science, Naval Postgraduate School, 1997.
[8] V. Athitsos, M. J. Swain, and C. Frankel, "Distinguishing photographs and graphics on the world wide web," in *Workshop on Content-Based Access of Image and Video Libraries*, Puerto Rico, June 1997.
[9] M. A. Windhouwer, A. R. Schmidt, and M. L. Kersten, "Acoi: A System for Indexing Multimedia Objects." in *International Workshop on Information Integration and Web-based Applications & Services*, Yogyakarta, Indonesia, November 1999.
[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
[11] B. Adams et al., "IBM research TREC 2002 video retrieval system," In Voorhees and Harman [1], pp. 182–198, To appear.
[12] M. Li and P. M. B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer, Berlin, second edition, 1997.
[13] P. Soille, *Morphological Image Analysis Principles and Applications*, Springer-Verlag, Berlin,Heidelberg, 1999.
[14] A. D. Bagdanov and M. Worring, "Multi-scale document description using rectangular granulometries," in *Document Analysis Systems V*, D. Lopresti, J. Hu, and R. Kashi, Eds., Princeton, NJ, August 2002, vol. 2423 of *LNCS*, pp. 445–456, Springer.
[15] R. v.d. Boomgaard and A. W. M. Smeulders, "The morphological structure of images, the differential equations of morphological scale-space," *IEEE transactions on pattern analysis and machine intelligence*, vol. 16, no. 4, pp. 1101–1113, 1994.
[16] R. v. d. Boomgaard, "The morphological equivalent of the Gauss convolution," *Nieuw Archief voor Wiskunde*, vol. 38, no. 3, pp. 219–236, 1992.
[17] E. A. Engbers, R. v. d. Boomgaard, and A. W. M. Smeulders, "Decomposition of separable concave structuring functions," *Journal of Mathematical Imaging and Vision*, vol. 15, pp. 181–195, 2001.
[18] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2001.
[19] J. Ma, Y. Zhao, and S. Ahalt, "OSU SVM classifier Matlab toolbox," `http://eewww.eng.ohio-state.edu/~maj/osu_svm/`.