Real-time Classification of Gorilla Video Segments in Affective Categories Using Crowd-Sourced Annotations

J. Schavemaker, E. Thomas and A. Havekes

TNO Media & Network Services, Technical Sciences, Brassersplein, Delft, The Netherlands

Abstract

In this contribution we present a method to classify segments of gorilla videos[1] in different affective categories. The classification method is trained by crowd sourcing affective annotation. The trained classification then uses video features (computed from the video segments) to classify a new video segment into one of different affective categories: exciting, boring, scary, and moving. As video features we propose to use features based on optical flow. As classification method we propose to use a k-NN classifier for quick relearning possibilities. We validate our method with an experiment with multiple recordings of gorillas from different video cameras and an annotation crowd from within our company.

Key words: affective computing, video annotation, crowd sourcing, classification, optical flow.

Introduction

Live streaming videos are an important part of multimedia that company websites can offer. When broadcasting 'raw' live video the chances are high that the content does not meet the expectations of the viewing audience. This is mainly because the video is not filtered or produced: the video offers no content (e.g. nothing is happening) or the content is in such a format that it is hard to recognize (e.g. overview instead of a close-up of the interesting action). For example, for a zoo that has a digital camera infrastructure that captures the activities of their inhabitants with video cameras and wants to broadcast that live, it is essential that the videos show the activities that will evoke emotions in the audience, preferably content in close-up because that is attractive to the general viewing public.

A real-time automatic way of classifying videos into affective categories denoting its relevance would be an attractive asset. This contribution describes a method that learns the classification from labeled examples from crowd sourcing. The classification learns to link measurable image features with affective viewing categories.

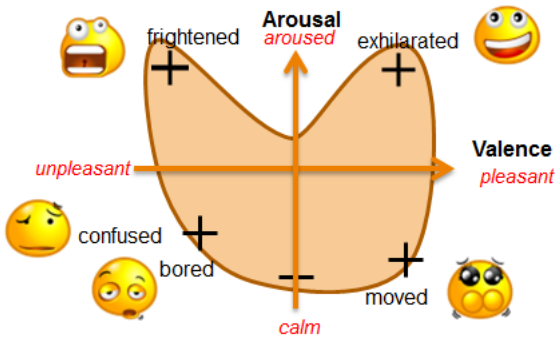
Related Work

Image annotation is the process by which a computer system or person assigns metadata in the form of captioning or keywords to a digital image. Affective annotation records the viewer's emotion; in general it is a subjective annotation. A common model of describing emotions is the two-dimensional arousal-valence diagram of Russel [7], see Figure 1. Common approaches along this model are to define one emotion for each quadrant of the arousal-valence diagram or to use negative, neutral and positive valence values [1].

Hanjalic et al [4] present an affective video content representation and modeling. Their approach is content-based: they compute affective annotations from the video itself. In their representation viewer emotions are described in a 2D arousal space. They note that motion in video increases arousal but valence is not modified. Motion, audio and shot lengths are used as low-level features in this paper. Hu [5] gives a broader overview of content-based features, see Figure 1. Soleymani et al [3] describe the development of a viewer-reported boredom corpus by crowd sourcing affective annotation of video. Using Mturk, crowdsourcing provides an effective means of collecting the

[1] We would like to thank Apenheul for use of their video material and additional support.

viewer affective response annotations needed to create a corpus to be used in the development of automatic prediction of viewer reported boredom.



Domain	Features	
Video Structure	Shot boundary detection	Threshold-Based Approach: Statistical Learning-Based Approach:
	Key frame extraction (frames that best reflect the shot contents)	Sequential Comparison Between Frames Global Comparison Between Frames Reference Frame Clustering Curve Simplification Objects/Events
	Scene Segmentation	Key Frame-Based Approach Audio and Vision Integration-Based Approach Background-Based Approach
Picture	Color	color histograms, color moments, color correlograms, a mixture of Gaussian models, etc
	Texture	Tamura features, simultaneous autoregressive models, orientation features, wavelet transformation-based texture features, co-occurrence matrices, etc.
	Shape	detect edges then describe the distribution of the edges using a histogram
Object	dominant color, texture, size, etc., corresponding to the objects	
Motion	Statistics	model the distribution of global or local motions of points in the video
	Trajectory	extracted by modeling the motion trajectories of objects in videos
	Objects' Relationship	relationships between objects using a symbolic representation scheme

Figure 1. (left) Arousal-valence diagram. (right) Different types of content-based features by Hu [5].

Method description

In this section we describe our classification method. The method is part of the following scheme:

1. phase 1 : affective annotation of video segments by crowd sourcing;
2. phase 2 : learning video classification from annotations and features;
3. phase 3 : real-time classification of new video segments.

Annotation of video segments by crowd sourcing

The first phase consists of collecting videos and annotating videos by crowd sourcing. For the experiments we have collected videos of the Gorilla Island of the Apenheul[2]. The Gorilla Island is recorded with eleven high-definition cameras capturing different views of the Gorilla Island; see Figure 2 for some example views. Our video collection consists of videos that have been captured for a number of days. The videos have been segmented and transcoded (using FFmpeg[3]) into fragments of 10 seconds, from which a selection is made.

Figure 2. Different camera views of the Gorilla Island in the Apenheul (images courtesy Apenheul).

The annotation of the videos is performed with the help of an annotation crowd consisting of more than 100 participants within our research company and a web-based annotation tool. Every participant views a number of randomly selected video segments from the video pool. The participant selects the appropriate affective category that best fits her/his emotional state after viewing the video segment, see Figure 4. This annotation process is

[2] http://www.apenheul.nl/

[3] http://www.ffmpeg.org/

implemented as a web-based video player with additional functionality to record the annotation. Every participant is also asked to enter age, gender and nationality. This allows future analysis for tailoring classification to different viewing groups.

Learning video classification from annotations and features

To train a classifier that classifies a video segment into an affective category the relationship between annotation and video content must be made. In order to do that we construct a content-based feature vector for every video segment that captures the video content. In our method we have chosen for optical-flow features: the hypothesis is that optical flow correlates with arousal (amount of motion) and valence (type of motion) of the viewer. Optical flow is commonly used to classify (human) activities using a histogram approach like histograms of oriented optical flow (HOOF) [2]. This approach bears similarities with the popular histogram of oriented gradients (HOG) approach by Dalal et al.

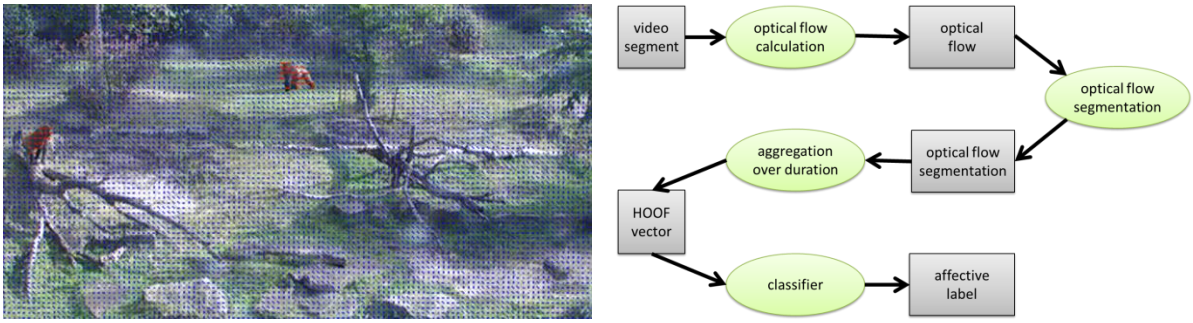


Figure 3. (left) Optical flow segmentation, blue: background, red: foreground. (right) Real-time classification diagram.

For our experiments optical flow is computed for every video frame using the Dual TV L1 optical-flow algorithm described in [7] and as implemented in the OpenCV[4] library. After calculation, the optical flow is segmented to extract the foreground optical flow actions and not let noise and clutter distort the histogram. To do that, we compute per camera the mean and standard deviation of optical flow magnitude per image block for a set of videos without gorillas. When an optical flow vector is larger in magnitude than corresponding mean magnitude plus 3 times standard deviation the optical-flow vector contributes to the feature computation. The segmented optical flow is aggregated over the duration of the complete video segment to create a histogram of oriented optical flow (HOOF). The HOOF vectors with the associated labels are used to train the classifier.

Real-time classification of new video segments

Real-time classification of new video segments (without annotation) proceeds much in the same way as learning. For a new video segment, optical flow is computed, segmented and aggregated into a HOOF feature vector. The HOOF feature vector is fed into the classifier to obtain a classification of the video in one of the categories.

Experiments

For the experiments we use a dataset with video segments of 3 cameras. The total number of video segments for every camera equals: 505, 461, and 167. The crowd annotation consists of 1623, 644, and 532 ratings (per video segment multiple ratings exist, ranging from 1 to 10 raters). The inter-rater reliability has been computed from the segments having 3 ratings. The Fleiss' kappa [6] equals 0.26 in this case, corresponding with a 'fair agreement' according to Landis and Koch[9]. The dominant rating is 'boring', followed by 'exciting', 'moving' and 'scary'. Ratings differ per camera, for example camera 1 (the feeding area) has about 65% boring, 24% exciting, 11% moving and less than 1% scary. Videos for which there is no majority vote of the rates are removed from the set

[4] http://opencv.org/

because of the fair rater agreement. Next, the dataset is divided in a train and test set of video segments for each camera and per camera a classifier is trained. The train set per camera consists of the video segments (and computed feature vectors) with the corresponding labels from the raters. So, one video segment in the train set can occur multiple times with all the labels from the different raters. For the test set, every video segment has one ground truth label that equals the majority vote of its ratings. As a classifier we use a k-NN classifier with a major voting scheme on the nearest neighbors of the sample (k = 11). Figure 4 presents the classification accuracies for the different cameras and the confusion matrix (actual versus predicted classification percentages) for camera 1.

camera	1	3	7
accuracy	84,80%	94,04%	87,90%
#train videos	171	218	58
#test videos	171	219	58

	moving	boring	exciting
moving	1%	3%	1%
boring	0%	77%	5%
exciting	0%	6%	8%

Figure 4. (left) Web-based annotation tool. (right top) Classification accuracies. (right bottom) Confusion matrix camera 1.

Figure 4 shows high accuracies but that is mainly because the dominant class is ‘boring’. From the confusion matrix we may deduce that there is some accuracy for ‘moving’ and ‘exciting’ albeit marginal. This is largely because ‘moving’ and ‘boring’ will overlap in features space as well as to some extent ‘exciting’ and ‘boring’.

Conclusions

In this contribution we have presented a method to classify segments of gorilla videos in different affective categories. The classification method uses crowd sourcing annotation which is combined with features computed from the video segments to classify a new video segment into different categories: exciting, boring, scary, and moving. We validated our method with a small experiment with gorilla videos and a small crowd. The computed optical-flow features show promising classification results. Future work will focus on a larger experiment.

References

1. E. van den Broek, “Affective signal processing (asp): Unraveling the mystery of emotions,” Ph.D. dissertation, University of Twente, 2011.

2. G. H. Rizwan Chaudhry, Avinash Ravichandran and R. Vidal, “Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions,” CVPR, 2009.

3. Mohammad Soleymani; Martha Larson; "Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus"; Workshop on Crowdsourcing for Search Evaluation SIGIR’10, July 19–23, 2010, Geneva, Switzerland.

4. A. L.-q. X. Hanjalic, “Affective video content representation and modelling,” Multimedia, IEEE Transactions on, vol. 7, no. 1, pp. 143–154, February 2005.

5. S. Weiming Hu; Nianhua Xie; Li Li; Xianglin Zeng; Maybank, “A survey on visual content-based video indexing and retrieval,” Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 41, pp. 791–819, 2011.

6. Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." Psychological Bulletin, Vol. 76, No. 5 pp. 378–382

7. J. A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, vol. 39, pp. 1161–1178, 1980.

8. C. Zach, T. Pock and H. Bischof. “A Duality Based Approach for Realtime TV-L1 Optical Flow”, In Proceedings of Pattern Recognition (DAGM), Heidelberg, Germany, pp. 214-223, 2007.

9. Landis, J. R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in Biometrics. Vol. 33, pp. 159–174