Semi-automatic Video Assessment System*

Pedro Martins
NOVA-LINCS and DI, Faculdade de Ciências e Tecnologia,
Universidade NOVA de Lisboa, 2829-516 Caparica,
Portugal
pcp.martins@campus.fct.unl.pt

Nuno Correia
NOVA-LINCS and DI, Faculdade de Ciências e Tecnologia,
Universidade NOVA de Lisboa, 2829-516 Caparica,
Portugal
nmc@fct.unl.pt

ABSTRACT

This paper describes a system for semi-automatic quality assessment of user generated content (UGC) from large events. It uses image and video processing techniques[1] combined with a computational quality model that takes in account aesthetics and how human visual perception and attention mechanisms discriminate visual interest. We describe the approach and show that the developed system allows to sort and filter a large stream of UGC in an efficient and timely manner.

CCS CONCEPTS

• **Computing methodologies → Image processing;** • Computing methodologies → Visual content-based indexing and retrieval; • Computing methodologies → Graphics systems and interfaces; • Theory of computation → Support vector machines

KEYWORDS

UGC; large events; video quality assessment; aesthetics; interestingness; model of attention and perception.

ACM Reference format:

P. Martins, N. Correia. 2017. SIG Proceedings Paper in word Format. In *Proceedings of International Workshop on Content-based Multimedia Indexing, Florence Italy, June 2017 (CBMI'15)*, 6 pages.
DOI: 10.1145/3095713.3095748

1 INTRODUCTION

The availability of high-speed internet connections and the increasing rate of mobile phone usage combined with the advent of Social Media, created a growing stream of UGC (User Generated Content). Simultaneously the fast pace of technological development, originated new broadcast standards like Ultra High-Definition (UHD). From the convergence of these trends grows the urge to create a new visual interactive UHD experience, that includes UGC content, and to experiment in what way the introduction of UGC on Live Broadcast of Large Scale Events can enhance the quality of Experience (QOE) maintaining, at the same time, high standards for content quality.

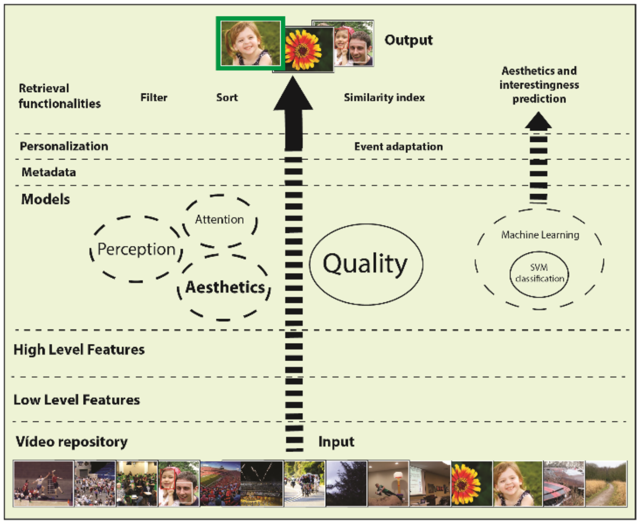


Figure 1: Concept overview

Fig. 1 shows an overview of our approach, where we attempt to combine in a video browsing interface, a balanced set of tools for efficient video discrimination. Simple filter and sort tools, based on a broad range of carefully selected human describable features, are used together with a similarity query tool and binary prediction values for aesthetics and interestingness. The next section presents related approaches. Section III provides a description of our system. Section IV and V describe the datasets and the results that were obtained so far. The paper ends with conclusions and directions for future work.

2 RELATED WORK

Some of the basic video features we use result from the mean or standard deviation of the values computed by applying image algorithms to video frames. We selected these image features, mainly, from previous approaches [1] where low level features, some based on photographic rules, are used to

[1] Mainly based on the OpenCV library.

CBMI, June 19-21, 2017, Florence, Italy
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5333-5/17/06...$15.00
http://dx.doi.org/10.1145/3095713.3095748

discriminate images based on aesthetic criteria. Or other projects [2] where a set of low level, describable features focused on color presentation and spatial composition were used to classify the aesthetics of images. Faces and facial features are also commonly accepted as being of great importance in terms of visual appeal, in [3] after face detection, is computed a rule of thirds score using a refined template technique. In [4] is proposed an approach to model spatiotemporal attention in video sequences where frame level saliency maps and motion contrast saliency maps are combined in a dynamic fashion. When motion contrast increases, the temporal model has a higher weight comparing to the spatial model. The proposed technique can detect the attended regions as well as attended actions in video sequences. From a content based image retrieval technique [5] came the inspiration for the indexing features based on a distance metric for color moments and Gabor filter texture measure. Studies on video aesthetics and interest are in an earlier stage than its image counterpart, nevertheless there are some experiments [6] where both image and video aesthetics are assessed using a combination of subject region detection with static and dynamic features. Also, where not only the temporal dimension is addressed [7] but also is taken in account the influence of objective quality towards aesthetic or interest of video content. Some common objective quality features found were frames-per-second, dimensions, bit depth, aspect-ratio, shakiness and blockiness. Other noteworthy features found were the motion-ratio and sharpness, difference measures between foreground and background. In [8] a video aesthetic assessment method is presented that combines a video representation integrating photographic and cinematographic rules, and a learning mechanism that takes video representation uncertainty into consideration. They compiled a dataset for the task of video aesthetic prediction, which we use as a way to test our overall method. A method for interest estimation in video [9] aims for a general prediction that most people would agree, not specifically tailored for a particular person or group. This project made two important contributions, they compiled two annotated datasets with interestingness score labels that we also use as ground truth for benchmarking our own interestingness binary classifier.

MPEG-7[2], is a multimedia content description standard, intended to provide additional functionality to previously released MPEG standards representing information about the content. Part 3 of this standard is dedicated to the specification of a large number of important visual descriptors like Color Layout or Edge Histogram. The specification also includes descriptor containers and basic supporting tools. The former consists of datatypes like the Grid Layout that provides representation of features on grids or the Visual Time Series used to represent temporal arrays of features. Overall these visual descriptors are very well suited for visual indexing and retrieval tasks.

In this work, we combine multiple features from the above body of work for visual quality assessment. There are several previous approaches targeting specific domains and applications that were integrated in this project into a common video evaluation framework.

3 SYSTEM DESCRIPTION

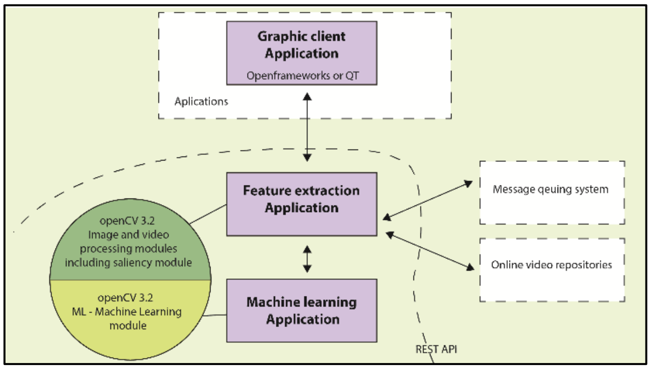

Graphic client Application
Openframeworks or QT
Aplications
Feature extraction Application
Message qeuing system
Online video repositories
openCV 3.2 Image and video processing modules including saliency module
openCV 3.2 ML - Machine Learning module
Machine learning Application
REST API

Figure 2: System architecture

In Fig. 2 we can see the system architecture. We use several algorithms, in a feature extraction application based on the OpenCV library, to compute features that describe meaningful visual properties as identified in the state-of-the-art literature. These features are then used for indexing based on color and texture, filtering, and sorting videos through the interface of a graphical application. Using a third application, we created several SVM (Support Vector Machines) classifiers based on existing ground truth annotated datasets. These classifiers are then used to predict aesthetics and interestingness on any new video sample.

3.1 Graphical Client Interface

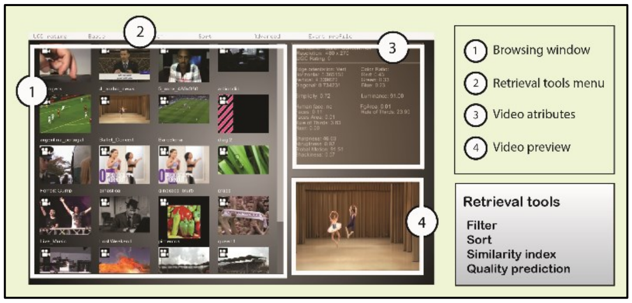


Figure 3: User interface

[2] http://mpeg.chiariglione.org/standards/mpeg-7

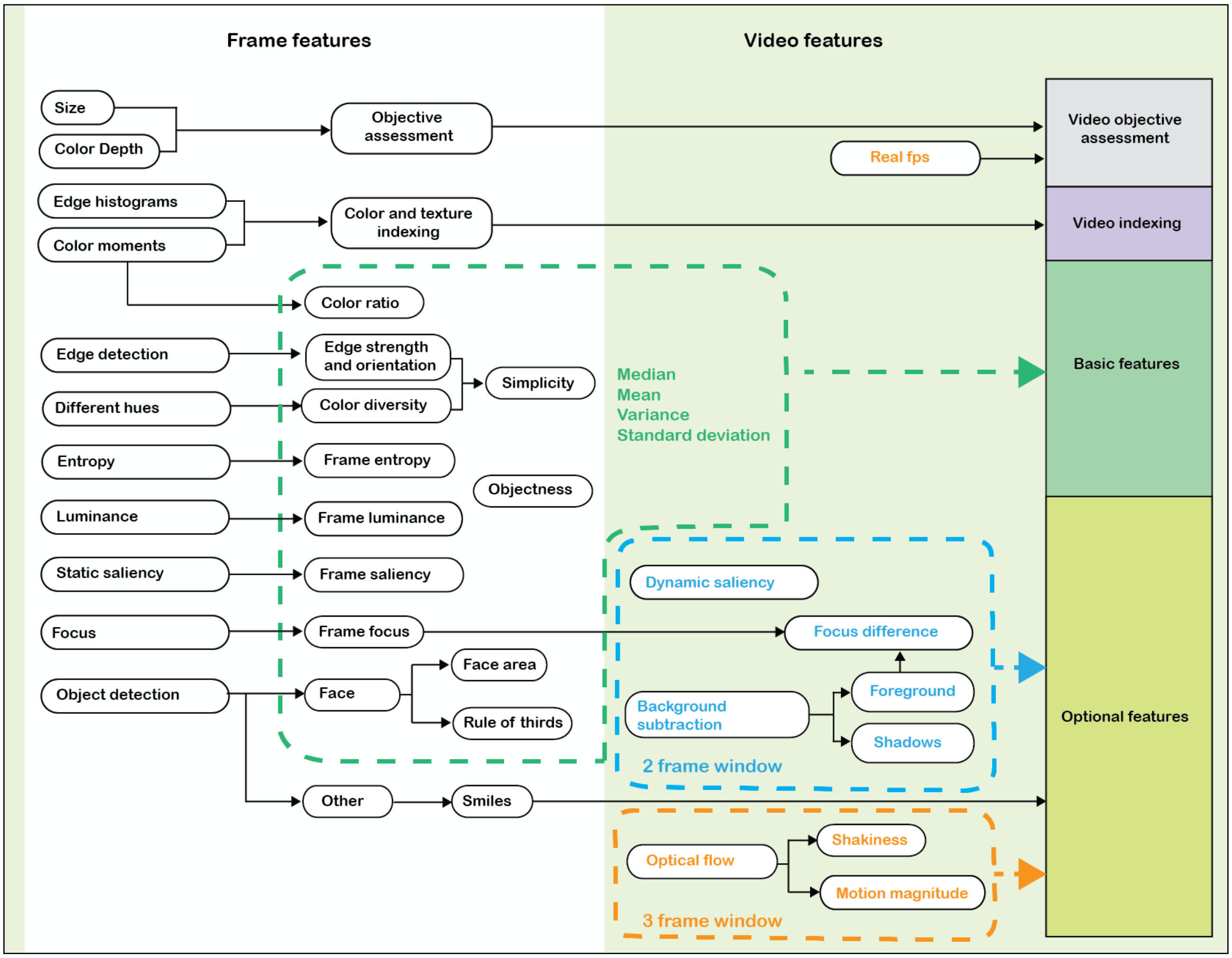
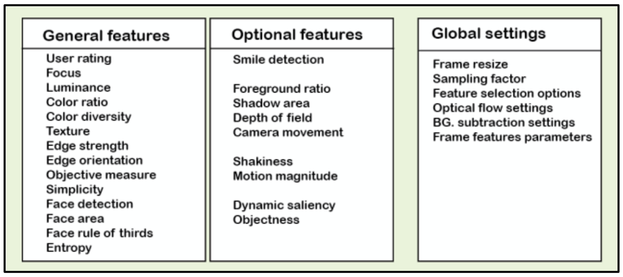


Figure 4: Frame and video level algorithms

Fig. 3 shows a view from our graphical application prototype that offers a simple way to browse a video repository. It starts by loading video metadata generated by the feature extractor application including binary prediction data generated using aesthetic and interestingness classifiers, built with the machine learning application. Afterwards, a window with the video thumbnails and a retrieval tool menu can be accessed and used to help the user discriminate videos by several criteria. There is also a video preview window and an information panel showing the metadata associated with the currently selected video.

3.2 Feature extraction application

This application is used to extract and compute visual features, and we can see in Fig. 4 a breakdown of this features. The values are saved as metadata to be used in the graphical application. After sampling, we use premade classifiers to generate binary predictions for the videos. Both the metadata and prediction values are saved in CSV format.


General features

User rating
Focus
Luminance
Color ratio
Color diversity
Texture
Edge strength
Edge orientation
Objective measure
Simplicity
Face detection
Face area
Face rule of thirds
Entropy

Optional features

Smile detection

Foreground ratio
Shadow area
Depth of field
Camera movement

Shakiness
Motion magnitude

Dynamic saliency
Objectness

Global settings

Frame resize
Sampling factor
Feature selection options
Optical flow settings
BG. subtraction settings
Frame features parameters

Figure 5: Describable features and configurations

By default, general purpose features are extracted, usually through fast to compute algorithms, that altogether contribute for a fast, rough assessment. Switching on optional features can help refine the quality model resulting in better performance on retrieval tasks but with the drawback that each additional feature adds to the overall computing time. These optional features can be switched on or off in configuration.xml, the same file that gives control over other important settings like the sampling rate or detection threshold values. Specific xml profiles can be built that override configuration.xml to tailor the quality criteria used on the assessment of some specific event (e.g., soccer match, live music).

We can see in Fig. 5 a list of the human describable features, arranged in general and optional features, together with the global settings of the feature extractor application. Fig. 6 depicts the influence of resizing and sampling factor in the overall computing time of the feature extraction stage. Changing the resize and sampling factors allows the adaptation to eventual time constraints at the expense of lower assessment performance. The sampling factor cannot be changed if any optional feature is used. In Fig. 7, we can see a time comparison using a combination of different feature options and resize. Both Fig. 6 and Fig. 7 computation times are based on a 62 seconds' video from YouTube with 1280 by 720 pixels and 30 fps. The test was carried out on a mid-range i7 laptop with 8Mb of RAM.

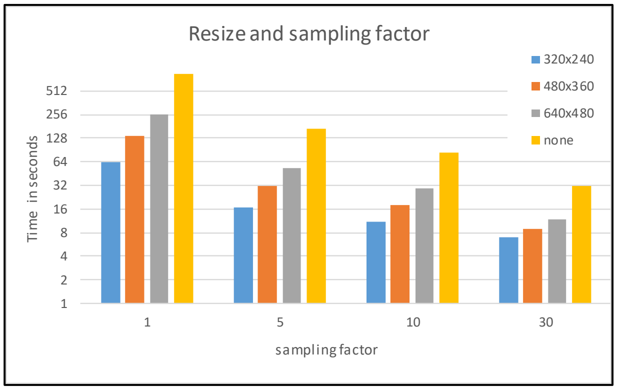


Figure 6: Resize and sampling factor

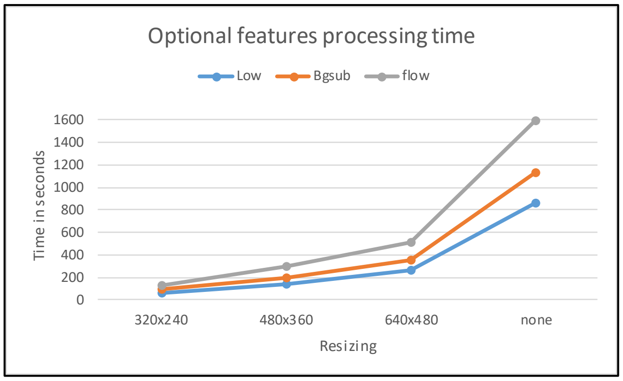


Figure 7: Optional features processing time

3.3 Machine learning application

SVMs are a useful technique for data classification. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value (i.e., the class labels) and several attributes (i.e., the features or observed variables). The goal of the SVMs is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. We then test many control samples with the classifier and compare the binary predictions obtained with our expectations (experimental ground truth binary value). The sample(s) to be evaluated by the classifier must match the format for the samples used to train the classifier. That is why we created classifiers for each combination of optional features. From the 36 video features used as metadata for the general retrieval tools of the graphical interface application, only a set of 27 were suited for SVM classification purposes.

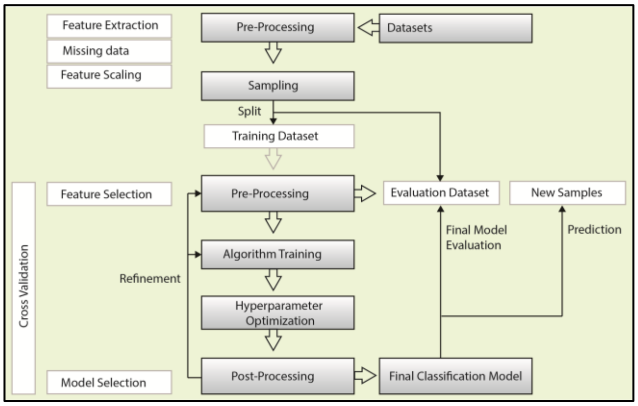


Figure 8: Supervised learning process

As we can see in Fig. 8 the supervised learning process involves many steps, and iterating through this process allows the refinement of the classification model. For this we use the machine learning application. With it, a learning algorithm can be trained, the hyper-parameters are optimized through grid-search, and cross validation tests are made. It also enables a statistical evaluation of the results were indicators like accuracy, precision, recall or false positive rate are taken in account. This eventually results in the selection of a final model used to generate a XML binary classifier. SVM uses the “kernel trick” to map a feature vector in a higher dimensionality space, using a kernel to build an optimal non-linear discriminative function (i.e., the hyper-plane). After some preliminary tests, we opted to use the Radial basis function (RBF) as it proved to be the best suited for our specific problem. We used our machine learning application to find the optimal hyper-parameters related to RBF. C is called the regularization constant or penalty and γ a parameter of the kernel. We performed linear and exponential grid-search to find both optimal values for each classifier.

4 DATASETS

The effectiveness of the machine learning component was refined and benchmarked using ground truth facilitated by image and video datasets where the subject is aesthetics and interestingness. All the datasets are compiled from UGC.

4.1 CERTH-ITI-VAQ700

A comprehensive video dataset for the problem of aesthetic quality assessment [8] with annotated scores for 700 (UGC) videos from YouTube, 350 videos are rated as being of high aesthetic quality and another 350 as being of low aesthetic quality.

4.2 Video Interestingness Database (VID)

Two benchmark datasets with ground-truth interestingness labels [9]. The first one (V.I.D. dataset A) consists in 1200 videos collected from Flickr which has a rank based on interestingness. The second (V.I.D. dataset B) consisted of 420 advertisement videos from YouTube. YouTube does not have an interestingness rank so to collect the interestingness scores, this dataset was subject to an experimental annotation procedure.

4.3 Photo.net

From the 20,278 total, we retrieved around 17,080 images all with annotated aesthetics scores compiled from images with more than 10 ratings from photography enthusiasts of this platform. This dataset was compiled during the study described in [10].

5 CLASSIFICATION RESULTS

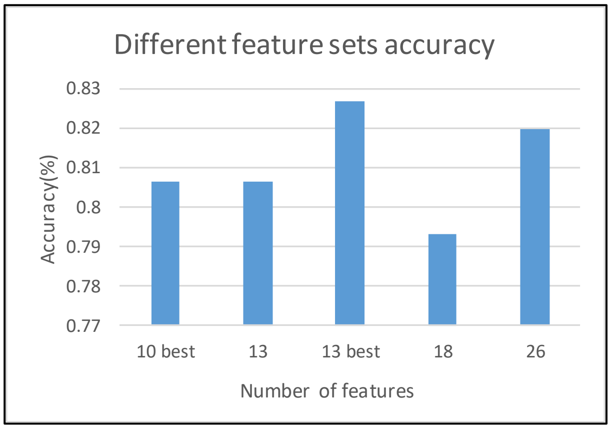


Figure 9: Feature sets accuracy

We computed features from the above datasets with our feature extractor application, and a subset of attributes was chosen. After normalization and proper formatting, we conducted a feature selection process where each feature was evaluated in groups (Fig. 9) and individually, we can see a breakdown in Fig. 10. Afterwards we made cross-validation tests for the individual feature classifiers. From those results, we selected the best performance set of features that in turn are used to extract a final feature vector from each dataset. Feeding this feature vector into a SVM is the initial step of the supervised learning process that eventually results in an adequate classifier.

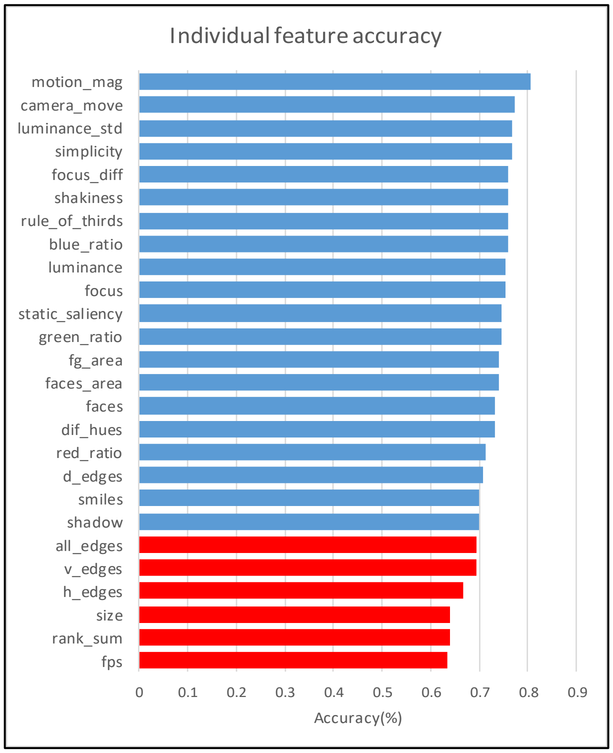


Figure 10: Individual features accuracy

We are still working on the Photo.net dataset, as we plan to provide aesthetic prediction, not only on video, but also on images.

5.1 CERTH-ITI-VAQ700

Using the CERT-ITI-VAQ700 dataset as ground truth we attained 82% accuracy for the task of aesthetic binary prediction with reasonable recall and false positive ratings. Over the precomputed feature vector, we used a sliding window of 50 evaluation samples against the remainder of 650 instances as training samples to train a classifier. We have done 14 tests, results are presented in Fig. 11, where the initial evaluation sample slides 50 positions until all the dataset instances have

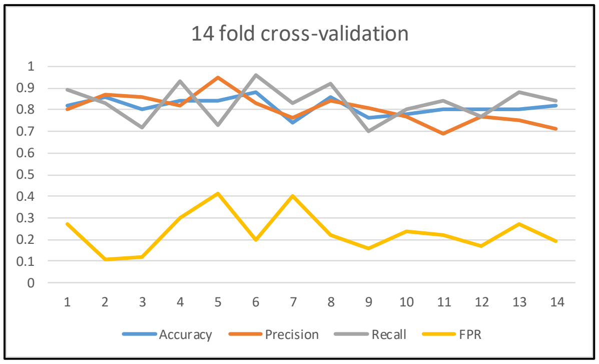


Figure 11: CERTH-ITI-VAQ700 cross-validation

been used both as evaluation and training samples. From the comparison of statistical indicators presented in Fig. 12 (being accuracy the most important) was assessed the classification performance.

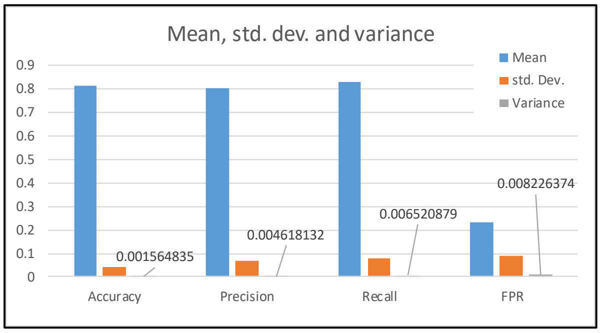


Figure 12: CERT-ITI-VAQ700 indicators

5.2 Video Interestingness Database (VID)

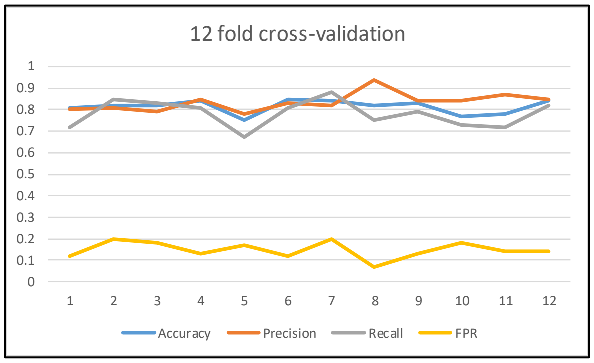


Figure 13: VID cross validation

From the first dataset with 1200 videos we extracted a feature vector and then conducted a 12-fold cross-validation test, we can see the results in Fig. 13. We followed the same cross-validation method, but in this case, we used an evaluation set of size 100 against the remainder of 1100 samples to train the classifier. From the statistical results shown in Fig. 14, we can observe that the classifier has a very good performance for the task of interestingness binary classification. It is also possible to see in Fig. 14 that this interestingness classifier has an accuracy above 81% with better precision and false positive rates then the aesthetic classifier.

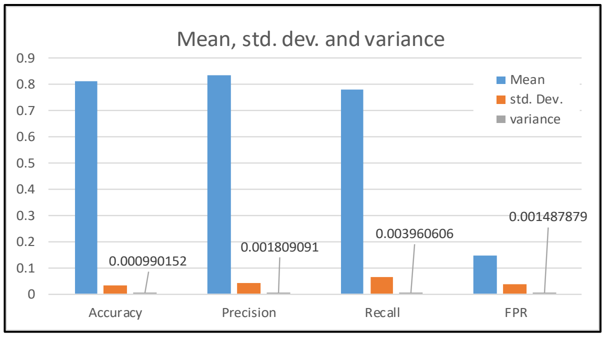


Figure 14: VID indicators

6 LIMITATIONS

So far, we disregard audio data or any type of semantic inference. The algorithms we use rely mostly on the color distribution of natural images and are not appropriate for synthetic images. We also assume that the video input has no posterior edition, and thus, no transitions are expected, leading to the absence of scene or shot detection.

7 CONCLUSIONS

In this paper, we present a system where a graphical video management application offers tools to discriminate videos based not only on aesthetics but also on human perception and attention mechanisms. We also train learning algorithms to predict video aesthetics and interestingness, and use color and texture data to compute a similarity index. Our visual discrimination concept is very flexible in terms of time constraints adaptation and context adaptation. It is also very compact and complete in terms of image properties representation. We have further plans to use our library of OpenCV based algorithms as a support for an infrastructure that recurs to parallelism and GPU computation, integrated with a messaging system, to provide VQA data to applications. We are also working on the classification process automation, to create classifiers on the fly in presence of different combinations of class attributes.

ACKNOWLEDGMENTS

This work was partially supported by Cognitus [11] an H2020 project.

REFERENCES

[1] R. Datta, D. Joshi, Jia Li and James Z. Wang. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. In Lecture Notes in Computer Science, vol. 3953, Proceedings of the European Conference on Computer Vision, Part III, pp. 288-301, Graz, Austria.

[2] Kuo-Yen Lo, K. Liu and C. Chen. 2012. Assessment of photo aesthetics with efficiency. International Conference on Pattern Recognition. Tsukuba.

[3] Shehroz S. Khan, D. Vogel. 2012. Evaluating visual aesthetics in photographic portraiture. CAe '12 Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging, (pp. 55-62). Annecy, France.

[4] Yun Zhai and M. Sha. 2006. Visual attention detection in video sequences using spatiotemporal cues. MM '06 Proceedings of the 14th ACM international conference on Multimedia, Pages 815-824, Santa Barbara, CA, USA.

[5] Z. Huang, P. P. K. Chan, Wing W. Y. Ng, and D. S. Yeung. 2010. Content-based image retrieval using color moment and Gabor texture feature. International Conference on Machine Learning and Cybernetics (ICMLC), Qingdao, China.

[6] Yiwen Luon and Xiaoou Tang. 2008. Photo and Video Quality Evaluation: Focusing on the Subject. ECCV '08 Proceedings of the 10th European Conference on Computer Vision: Part III Pages 386-399, Marseille, France.

[7] A. K. Moorthy. 2010. Towards Computational Models of the Visual Aesthetic Appeal of Consumer Videos. In Lecture Notes in Computer Science, European Conference on Computer Vision – ECCV 2010, vol. 6315, pages 1-14, Greece.

[8] C. Tzelepis, E. Mavridaki, V. Mezaris and I. Patras. 2016. Video aesthetic quality assessment using kernel Support Vector Machine with isotropic Gaussian sample uncertainty (KSVM-IGSU). In 2016 IEEE, International Conference on Image Processing (ICIP), Phoenix, AZ, USA.

[9] Y. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng and H. Yang. 2013. Understanding and predicting interestingness of videos. In AAAI'13 Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, pages 1113-1119, Washington, USA.

[10] R. Datta, Jia. Li and J. Z. Wang. 2008. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In 15th IEEE International Conference on Image Processing, ICIP 2008, San Diego, CA, USA.

[11] Project Cognitus. Accessed June 2017, at http://cognitus-h2020.eu