Generating an Album with the Best Media Using Computer Vision

Tancredo Souza[1], João Paulo Lima[1,2(✉)], Veronica Teichrieb[1], Carla Nascimento[3], Fabio Q. B. da Silva[4], Andre L. M. Santos[4], and Helder Pinho[5]

[1] Voxar Labs, Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
{tantan,jpsml,vt}@cin.ufpe.br
[2] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, Recife, Brazil
joao.mlima@ufrpe.br
[3] Projeto de Pesquisa e Desenvolvimento CIn/Samsung, Universidade Federal de Pernambuco, Recife, Brazil
cmpn@cin.ufpe.br
[4] Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
{fabio,almss}@cin.ufpe.br
[5] Samsung Instituto de Desenvolvimento para a Informática, Campinas, Brazil
helder.p@sidi.org.br

Abstract. Due to the increase in smartphone usage, it became easier to register memorable moments with a more accessible camera. To ensure a nice capture was made, users often take multiple shots from a scene, later filtering them based on some quality criteria. However, sometimes this may be unfeasible to do manually. To address this issue, this work initially defines relevant characteristics present in a good personal picture or video. We then show how to automatically search for these aspects using computer vision algorithms, successfully assessing personal media based on these aspects. Moreover, we show that it was possible to use this proposed solution in a real-world application, improving the generation of a personal album containing the best pictures and videos.

Keywords: Heuristics · Image · Video · Content analysis
Media description · Album generation

1 Introduction

In recent years, smartphones enabled users to easily register memorable events, being possible to even dismiss the need of a dedicated camera. Pictures and videos from vacation trips, birthday parties or friend meetings are worth keeping to yourself or sharing with others in social network. It is important, therefore, to assure that these records were made properly.

A. Marcus and W. Wang (Eds.): DUXU 2018, LNCS 10919, pp. 338–352, 2018.
https://doi.org/10.1007/978-3-319-91803-7_25

In this sense, users take similar shots from a scene, and afterwards choose the best ones to keep. This evaluation is done by filtering the media based on some personal quality criteria. For example, blurred images, videos with noticeable shakiness or photographs where not everyone appeared in their best pose are often discarded. However, sometimes there is just too much content to analyze, and doing this manually may take too much effort or accidentally cause the loss of media with decent quality.

Personal galleries in smartphones help the users to manage their captures by offering some filters to decrease the quantity of pictures and videos displayed, such as the Apple's Photos[1] app, which filters the content based on specific time periods or locations. While this may help reducing the amount of media to analyze, it is still needed to manually evaluate them. We propose an automatic evaluation method using computer vision that dismisses this manual step, and is able to create an album containing the best pictures and videos.

But how to distinguish a good capture from a bad one? To differentiate high quality shots from bad quality shots is more natural to humans. This means that users without knowledge in photography concepts are able to often determine their preferable picture or video. Figure 1 shows two pictures of a scene as an example.

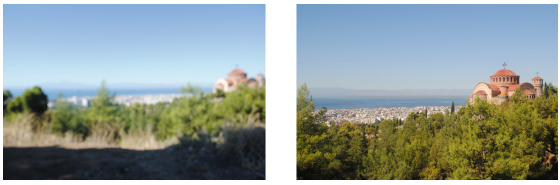


Fig. 1. Two picture examples. It is clear that the left picture is less pleasant than the right one.

On the other hand, this is more difficult for computers. As stated in [4], challenges in this task include modeling the photographic rules computationally, knowing the aesthetic differences between images from different genres (e.g. night scenes, close-shot object, scenery) and having a large human-annotated dataset for robust testing.

Thus, we organize this work as follows. In Sect. 2 we initially discuss related works that aim to automatically assess the quality of the user's media. Afterwards, we describe in Sect. 3 our developed automatic personal media evaluation, based on searching for a set of important characteristics using heuristics-based and adaptive computer vision algorithms. Section 4 presents the results obtained in datasets of personal pictures and videos. We also emphasize this work's contribution by analyzing its integration to a real-world application, improving the generation of an album containing the user's media with the best quality. Finally, Sect. 5 presents our final considerations to this work and to the prospects of the discussed problem scenario.

[1] https://www.apple.com/ios/photos/.

2 Related Work

Over the past years, researchers proposed different ways to automatically assess personal media. We can broadly organize this evaluation into two different approaches [4]: those based on heuristics and those based on adaptive machine learning.

The use of heuristics evaluates a picture or a video based on principles of photography – such as the *rule of thirds* (Fig. 2) – and evaluating high-level semantics like lighting composition, blurring and color.

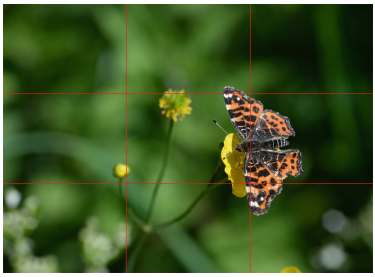
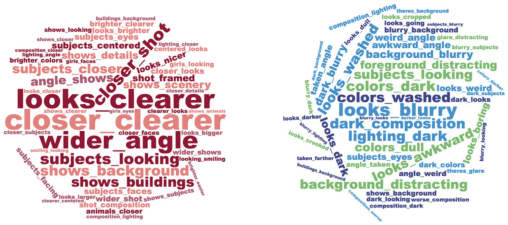


Fig. 2. The rule of thirds states that the subject of interest should be placed along the indicated red lines. (Color figure online)

Although this achieves good results when tested in a very diverse picture database [6], it only considers the media's visual aesthetics. However, the presence of other people in a personal picture or video matters to the user [3]. As an example, it is often desirable that the people framed are looking at the camera. This means that, in this context, assessing personal media quality does not necessarily imply on strictly following professional photograph rules, because users in some cases may ignore these aspects. In this sense, our work combines a subset of these aspects with the people and scene descriptions when evaluating the quality of personal media.

The machine learning approaches train a model using a dataset of pictures or videos with a human-annotated ground-truth. This model is able to extract information just from the pixels of the input media, learning from that data how to adaptively assess the media's quality. In order to perform a decent training and evaluation of the trained model, the size and diversity of the dataset must be sufficiently large. This, on the other hand, demands significant effort.

To workaround this issue, the work developed in [3] described the construction of a crowdsourced dataset with more than $10,000$ images. They annotated the dataset by asking for participants to choose only one picture from random pairs, justifying their choice. With this information, shown in Fig. 3, they proposed a variety of methods for learning human preferences when evaluating photo series. Although having more than $10,000$ images collected and annotated, it was still insufficient to successfully train a machine learning model for this task.



(a) (b)

Fig. 3. Word clouds visualizing common photo preferences of the participants for (a) preferred photos and (b) rejected photos. Image obtained from [3].

In our proposed work, it is considered not only visual aspects, but also the scene context description, using the presence of the people and objects framed to adaptively perform such media evaluation. Also, we can extend these analyses to personal video quality assessment [8,9,17]. For videos, on the other hand, we can also analyze the scene's *spatiotemporal information*: the information regarding the movement of the framed objects.

3 Photograph and Video Quality Assessment Method

This work lies between these two approaches: the use of heuristics applied to visual features and the use of machine learning algorithms. Searching for visual characteristics is done by applying adaptive computer vision algorithms, which will be detailed in this section.

Our analysis is segmented into four main categories:

- *People Detection*: extracts characteristics of the people framed.
- *Image Properties Computation*: describes the visual aesthetics of the image.
- *Image Content Analysis*: analyzes the context of the picture.
- *Video Processing*: describes the media's spatiotemporal information.

This section will then analyze in more depth each one of these groups, showing which characteristics they analyze and the methods used for searching and describing these aspects.

3.1 People Detection

As previously discussed, it is important to consider the presence of a person in a capture, as it may influence the user's preference. It was observed that, in personal pictures, it is often desirable that the people framed are smiling or with their eyes open. Initially, it is applied a face detector in order to know whether there are faces in a picture. Then, we proceed to analyze these detected faces, describing if their eyes were open and if a smile was visible. We proceed to detail how this work describes these characteristics in an image.

Face Detection. Face detection is done using the SeetaFace [15] method. An overview of SeetaFace Detection is shown in Fig. 4. It feeds its input to a cascade of classifiers, which analyzes a set of characteristics that identify the presence of faces in a picture. Due to its cascade format, the face detector uses the result collected from a given classifier as additional information for the next in the cascade, improving its prediction confidence. As a final output, it estimates the detected faces' bounding box (Fig. 5).

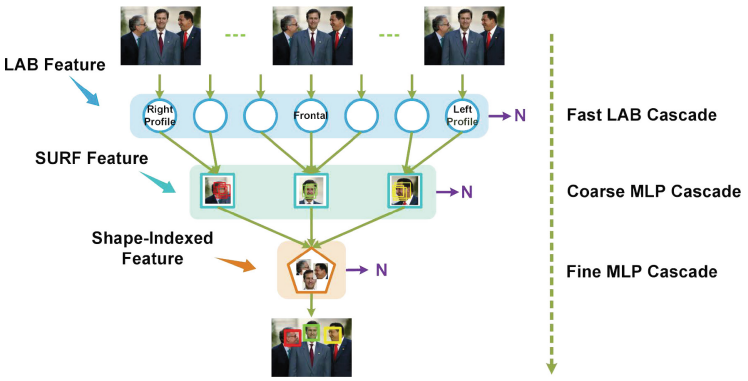


Fig. 4. SeetaFace Detection method overview. Image obtained from [15].

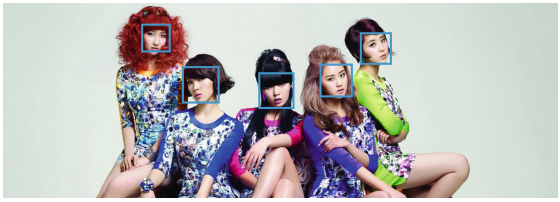


Fig. 5. Faces detected with the SeetaFace method.

Open Eyes Detection. Our developed method for open eyes detection consists of four steps:

- *Face Landmark Detection*: Given the bounding rectangle of a detected face by SeetaFace, we use the method proposed in [18] to obtain the position of the eyes (Fig. 6a).
- *Eye Region Estimation*: For each eye location retrieved by the alignment step, regions centered on these positions are estimated, as illustrated in Fig. 6b.
- *Eye Classification*: This step consists of classifying each region as containing an open eye or not. This is done using an open eyes detector, based on a trained Haar cascade classifier [14]. This makes the regions closer to the detected eyes, as shown in Fig. 6c.

– *Open Eyes Region Estimation*: If any of the face's eyes is classified as open, the detection was considered successful. It is thus estimated the position of the eyes as a final result (Fig. 6d).

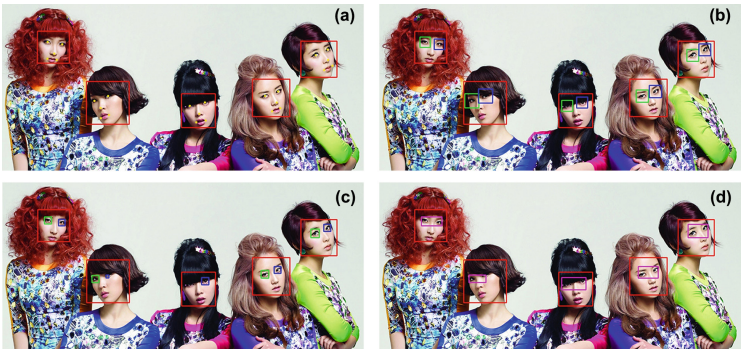


Fig. 6. Open Eyes Detection steps: (a) Input faces bounding rectangles (red) and face landmarks estimated by SeetaFace Alignment (yellow). (b) Estimated regions of the left (blue) and right (green) eyes. (c) Refined regions of the left (blue) and right (green) eyes, detected by the Haar classifier. (d) Estimated area of the open eyes (magenta). (Color figure online)

Smile Detection. Given the detected bounding box of a face, we only consider its lower half rectangle for the smile detection, because it is where one expects to find a mouth, as illustrated in Fig. 7a. Then, it is checked if there is a smile inside each region. For this task, another trained Haar cascade classifier is employed for detecting the smiles, returning a rectangle containing the detected smile's location (Fig. 7b).

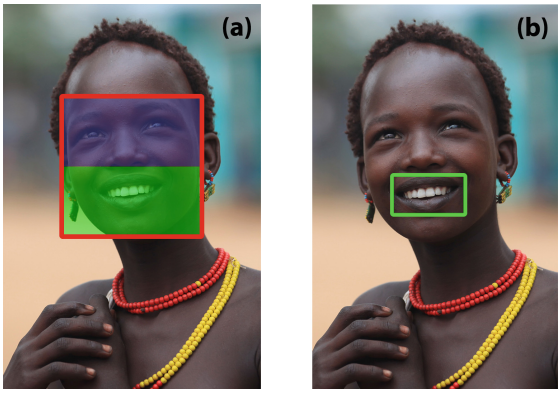


Fig. 7. (a) Only the lower half of the detected face is considered. (b) Refined bounding box for the smile detected by the Haar classifier.

3.2 Image Properties Computation

In images, noticeable motion blur and bad exposure make it difficult to recognize the framed elements or objects (Fig. 8). Thus, we perform the detection of these aspects in a picture.

Fig. 8. Examples of undesirable conditions, such as bad lighting or blurring.

It is worth mentioning that the presence of aspects such as blur or brightness does not imply that the media has bad quality. For example, aspects like blur are not necessarily associated to a bad capture. As shown in Fig. 9, this can be done for artistic purposes. We disconsidered subjective cases like these when performing our media quality evaluation, analyzing these cases just like the others.

Fig. 9. An example of an artistically blurred capture.

Also, since a personal gallery may contain various pictures of the same scene, it is possible that more than one picture has good quality. In consequence, this could cause similar photographs ending up in the generated album. To avoid this, our work proposes the detection of similar images.

Blur Detection. The approach employed for blur detection is based on the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [10]. This technique uses mean subtracted contrast normalized (MSCN) coefficients, which can provide a quantitative description of the image just using its pixels. It was

noted that an image exhibits a histogram of MSCN coefficients with a Gaussian like appearance, as shown in Fig. 10. Thus, [10] proposed a statistical relationship between these coefficients, being able to describe the visual aspects of the image using this Gaussian curve, such as blur. A blurred image will cause a characteristic distortion in these values, causing the Gaussian curve to change.

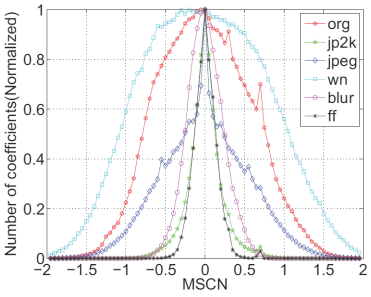


Fig. 10. Histogram of MSCN coefficients as a Gaussian curve. Image obtained from [10].

These curve aspects are used to train a support vector machine (SVM) model by using a set of blurred and unblurred training images. This enables the model to identify if an image is blurred or not, by only considering these mentioned aspects.

Exposure Detection. Exposure detection is inspired by the Zone System [1], which assigns numbers from 0 through 10 to different brightness values for an image - where 0 represents black, 5 middle gray, and 10 pure white. These values are known as zones, and are illustrated in Fig. 11.

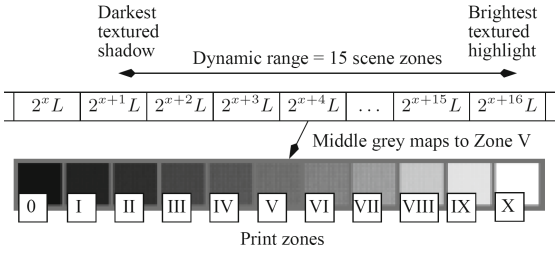


Fig. 11. The Zone System. Image obtained from [12].

This information is used to estimate a value related to the overall illumination of the scene, called the scene's *key value* [12]. This key value is then used to classify an image as underexposed (very dark), illuminated neutrally or overexposed (very bright). Figure 12 shows that underexposed images have a lower

key value, images with neutral illumination are associated with intermediate key values and overexposed images present a higher key value.

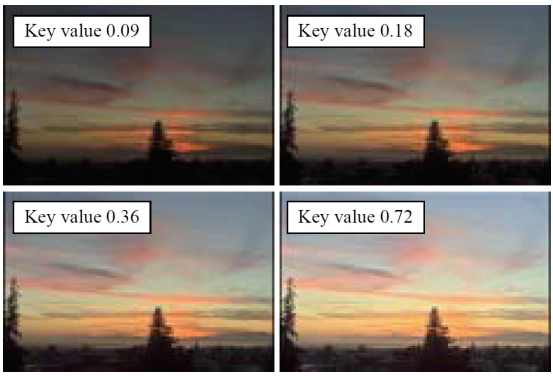


Fig. 12. Different image exposures and their corresponding key values. Image obtained from [12].

Similar Images Detection. Our similar images detection algorithm is based on perceptual hashing [16], which only uses the information of the pixels. An overview of the similarity detection is shown in Fig. 13.

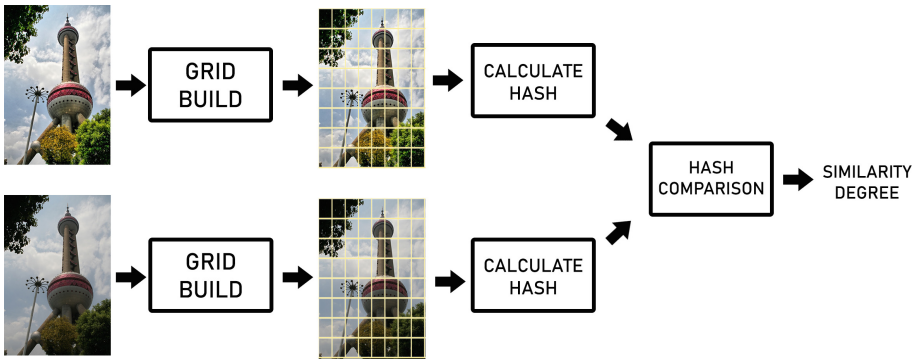


Fig. 13. Blockhash perceptual hashing method overview.

The two input images are, initially, segmented into grid blocks. By segmenting the images using a grid, it is able to make a more specific comparison of the two inputs, increasing the robustness of the technique. Afterwards, we use a hash to assign an individual binary code for each block, creating a representation of that image as an array containing only 0s and 1s.

Finally, given the two binary codes of the input images, we calculate their similarity degree by estimating how much they differ, using hamming distance. Thus, it is expected that similar images will have a high similarity degree.

3.3 Image Content Analysis

For humans it is possible, as an example, to identify that a photograph is related to a party by recognizing balloons, a cake, party hats, etc. (Fig. 14). In this sense, this work enables understanding the scene's context based on the objects captured. Using object detection, it is possible to use the classes of the detected objects to obtain this context information.

Fig. 14. Party pictures or videos often presents distinctive objects.

Object Detection. The approach adopted for object detection is based on the YOLO system [11]. The process performed by YOLO is described in Fig. 15.

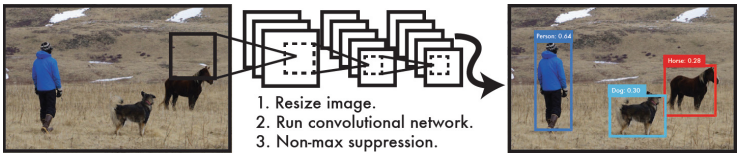


Fig. 15. YOLO (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence. Image obtained from [11].

Using a single convolutional neural network, it segments the input image into various regions. Each region is associated with a probability of containing an object of some class (e.g. dog, car or bicycle), as shown in Fig. 16.

After increasing its confidence in its predictions, this method finally considers the object classes with the highest probabilities and outputs the corresponding bounding boxes for each object class detected.

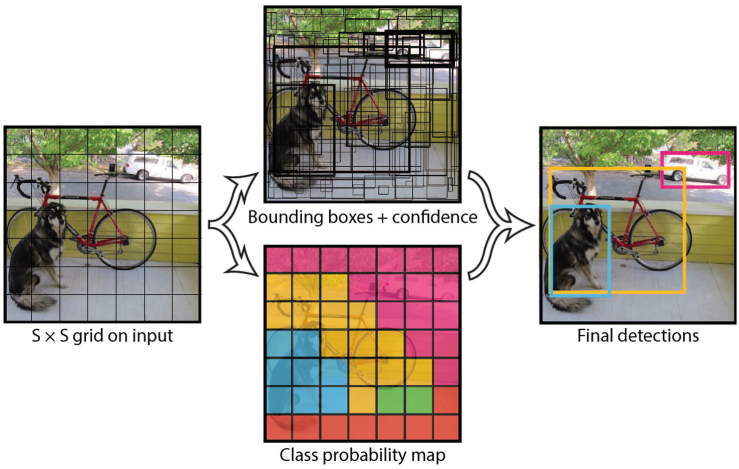


Fig. 16. YOLO divides the image into an $S \times S$ grid and for each grid cell predicts bounding boxes, confidence for those boxes, and class probabilities. Image obtained from [11].

3.4 Video Processing

All of the previous analyses for images can be extended to video frames. This is done by analyzing each individual video frame and averaging these characteristics for the whole video. We, therefore, combine an individual frame analysis along with the media's spatiotemporal information. By describing the stableness of the video, we can estimate if a video is shaky or steady. In this section, we detail the detection of what is called hand shakiness.

Hand Shakiness Detection. The approach employed for hand shakiness detection is based on the method proposed by [17]. This technique uses the frame's *optical flow* information, which is the structure of the movement between two consecutive frames. Figure 17 shows a visualization of this structure.

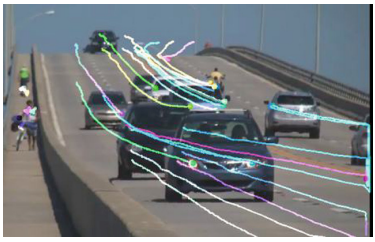


Fig. 17. The optical flow of a scene.

Since only the motion of the camera describes the video's movement stableness, it is expected that framed objects moving in the scene should not be considered. It is then necessary to distinguish the movement of the objects from the movement of the camera itself. As proposed by [17], extracting the frame's optical flow information from its border area addresses this issue, increasing the robustness of this method (Fig. 18).

Fig. 18. Extracting the information from the (a) entire area may cause false alarms, whereas reducing to the (b) border area may provide a more robust analysis. Image obtained from [17].

Finally, we describe the video's shakiness using its optical flow information. In unstable videos, the movement direction changes very frequently [8]. Also, if the amount of movement is very significant, this results in a perceptual shaky recording. Using the optical flow of the video, it is possible to perceive the changes in the camera's movement direction.

Thus, we intuitively define the hand shakiness degree calculation as: *If the direction of the movement changed between two frames, then how much did it change?* For a video, we accumulate the hand shakiness degree and average it to obtain an estimation for each frame. A shaky video will have a higher average value of hand shakiness, whereas a more steady one will have a lower average value.

4 Evaluation

4.1 Dataset

Since this work aims to assess the quality of personal media, our own dataset is composed of pictures and videos that would belong to a personal gallery. As discussed before, these captures register, for example, vacation trips, birthday parties, etc. The images and videos from our evaluation dataset were manually annotated. This annotation described the presence or absence of some particular characteristics (e.g. blur, smile, underexposed, shaky video, etc.). Our dataset was then segmented into groups regarding their visual aspects.

Training Datasets. The adaptive machine learning models (blur, face and object detectors) require a dedicated dataset to train their respective neural networks. The BRISQUE model for blur detection was trained using the CERTH Image Blur dataset [7]. The SeetaFace and YOLO models for face and object detection were made available by the authors as a previously trained model, which were used in our work.

Evaluation Datasets. For video quality assessment, the public availability of evaluation datasets containing personal recordings is scarce. To address this issue, a dataset called CERTH VAQ-700 [13] was made available, containing hundreds of personal videos useful for evaluating automatic video quality assessment techniques. However, it was not very clear in the dataset's annotations if a video was considered shaky or steady. Thus, we manually selected and annotated a subset of these videos, marking for each one the presence or absence of hand shakiness. Finally, the developed exposure detection method was evaluated using the ImageCLEF 2011 dataset [2] and the YOLO object detector was tested in the MS-COCO 2015 dataset [5].

4.2 Implementation

This proposed solution was implemented in C++ using the computer vision library OpenCV. The testing execution times were obtained using a computer with an Intel Core i7-5500U @ 2.40 GHz processor and 16 GB RAM.

4.3 Results

It was calculated a score for each technique, determining how effective the algorithm was in its task when tested on a dataset segment (0 means complete inaccuracy and 1 is perfect accuracy). Table 1 shows the scores obtained when executing our techniques in the described set of images and videos, along with the average execution times per image/video frame for the developed methods.

Table 1. Final results in the evaluation dataset.

Media description	Evaluation dataset	Data amount	Score	Execution time
Face detection	Ours	180	0.9013	1.40 s
Open eyes detection	Ours	180	0.8425	1.93 s
Smile detection	Ours	180	0.7724	1.42 s
Blur detection	Ours	69	0.6818	1.21 s
Exposure detection	ImageCLEF 2011 [2]	8000	0.1934	0.17 s
Similar images detection	Ours	51	0.7246	0.37 s
Object detection	MS-COCO [5]	20288	0.4400	4.62 s
Hand shakiness detection	CERTH VAQ-700 [13]	50	0.9340	0.30 s

4.4 Real-World Application

All the proposed methods, except for object and hand shakiness detection, are currently integrated to a Windows application available worldwide. It was perceived an improvement over existing functionalities, generating a more agreeable album containing the user's best pictures and videos. While this emphasizes this work's contribution, it also shows the efficient optimization of the developed solution, dismissing the need of significant computational power to use it.

5 Conclusion

This work proposed an automatic approach for personal media evaluation, using computer vision. We conducted an initial discussion about important media characteristics that are often more perceivable to the layman users. This set of visual aspects was then searched in the users' pictures and videos by adaptive computer vision algorithms. It was possible to improve the application's album generation functionality of a real-world application, only including the agreeable best quality media, emphasizing this work's contribution.

However, as experimentation shows, there is still room for improvement. One can argue that there are subjective cases, where it is not that easy to decide, for example, if a picture was artistically blurred or if a video had a shaky movement. In our scenario, we decided to dismiss these more complex cases, because, as discussed before, quality assessment is already harder to computers than to humans. We showed that, instead of using deep features only recognizable computationally, it is possible to provide to computers a more natural understanding of the scene. Giving a sense of scene context to a machine makes its media evaluation more similar to the analysis made naturally by humans. Thus, we believe that it is possible to make much more future progress in this direction.

Acknowledgements. The results presented in this paper have been developed as part of a collaborative project between Samsung Institute for Development of Informatics (Samsung/SIDI) and the Centre of Informatics at the Federal University of Pernambuco (CIn/UFPE), financed by Samsung Eletronica da Amazonia Ltda., under the auspices of the Brazilian Federal Law of Informatics no. 8248/91. The authors would like to thank the support received from the Samsung/SIDI team. Professor Fabio Q. B. da Silva holds a research grant from the Brazilian National Research Council (CNPq), process #314523/2009-0.

References

1. Adams, A.: The Negative: Exposure and Development Basic Photo 2, vol. 98. Morgan and Lester, New York (1948)
2. Bosch, M.: Imageclef experimental evaluation in visual information retrieval. Inf. Retr. **1**, 4 (2016)
3. Chang, H., Yu, F., Wang, J., Ashley, D., Finkelstein, A.: Automatic triage for a photo series. ACM Trans. Graph. (TOG) **35**(4), 148 (2016)

4. Deng, Y., Loy, C.C., Tang, X.: Image aesthetic assessment: an experimental survey. IEEE Sig. Process. Mag. **34**(4), 80–106 (2017)
5. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
6. Luo, Y., Tang, X.: Photo and video quality evaluation: focusing on the subject. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 386–399. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88690-7_29
7. Mavridaki, E., Mezaris, V.: No-reference blur assessment in natural images using fourier transform and spatial pyramids. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 566–570. IEEE (2014)
8. Mei, T., Hua, X.S., Zhu, C.Z., Zhou, H.Q., Li, S.: Home video visual quality assessment with spatiotemporal factors. IEEE Trans. Circuits Syst. Video Technol. **17**(6), 699–706 (2007)
9. Mei, T., Zhu, C.Z., Zhou, H.Q., Hua, X.S.: Spatio-temporal quality assessment for home videos. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 439–442. ACM (2005)
10. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), pp. 723–727, November 2011
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
12. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. ACM Trans. Graph. (TOG) **21**(3), 267–276 (2002)
13. Tzelepis, C., Mavridaki, E., Mezaris, V., Patras, I.: 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). pp. 2410–2414. IEEE (2016)
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, p. I. IEEE (2001)
15. Wu, S., Kan, M., He, Z., Shan, S., Chen, X.: Funnel-structured cascade for multi-view face detection with alignment-awareness. Neurocomputing **221**, 138–145 (2017)
16. Yang, B., Gu, F., Niu, X.: Block mean value based image perceptual hashing. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2006, pp. 167–172. IEEE (2006)
17. Yang, C.Y., Yeh, H.H., Chen, C.S.: Video aesthetic quality assessment by combining semantically independent and dependent features. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1165–1168. IEEE (2011)
18. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 1–16. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_1