2017 16th IEEE International Conference on Machine Learning and Applications

Exposing Computer Generated Images by Eye's Region Classification via Transfer Learning of VGG19 CNN

Tiago Carvalho*, Edmar R. S. de Rezende†, Matheus T. P. Alves*, Fernanda K. C. Balieiro* and Ricardo B. Sovat*
*Federal Institute of São Paulo (IFSP), Campinas-SP, Brazil 13069-901
†CTI Renato Archer, Campinas-SP, Brazil 13069-901

***Abstract*—The advance of computer graphics techniques comes revolutionizing games and movie's industries. Creating very realistic characters totally from computer graphics models is, nowadays, a reality. However, this advance comes with a big price: the realism of images is so big that it is difficult to realize when we are facing a computer generated image or a real photo. In this paper we propose a new approach for highly realistic computer generated images detection by exploring inconsistencies into the region of the eyes. Such inconsistencies are captured exploring the expression power of features extracted via transfer learning approach with VGG19 Deep Neural Network model. Unlike the state-of-the-art approaches, which looks to evaluate the entire image, proposed method focuses in specific regions (eyes) where computer graphics modeling still needs improvements. Experiments conducted over two different datasets containing extremely realistic images achieved an accuracy of 0.80 and an AUC of 0.88.**

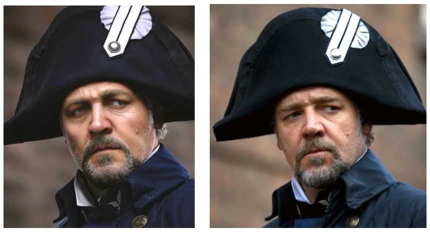
I. Introduction

People that grew up reading comic books are probably familiar with the famous quote: "Parker, get me pictures of Spider-Man", yelled by J. J. Jameson[1] on different occasions. This fictional situation expresses the power that images yielded some years ago, when newspapers were one of the main information channels, and images were strongly respected and considered "a piece of truth". Unfortunately, nowadays, we can not put as much trust on images as before. Mainly because, there are many techniques and tools to create and manipulate images.

To create an image or a movie containing computer generated (CG) people with a high degree of realism is more than a possibility. Nowadays, it is a reality. One example of such realism has been depicted in the last Star Wars movie[2] where the actress Carrie Fisher has been digitally reproduced with the same appearance of the beginning of her carrier, in the 70's. And this realism, many times, deceives even the most skeptical expert.

However, as pointed by *Holmes et.al.* [1], once the perfect CG image generation goal is achieved, it brings with itself challenges for other science areas like, for example, the challenge of discerning between a photo generated (PG) — the one generated by a digital camera — and a CG image. Figure 1 illustrates this hard task.

But, even in a very realistic image of a person, there are details, many times not noticeable by human eyes, that contains inconsistencies when compared with real human body parts. Regions of eyes and hair, for example, are some of the regions more difficult to reproduce in CG images. And it is in these kinds of inconsistencies that Digital Forensics community support their methods. The work proposed by *Conotter et.al.* [2], for example, use information associated with blood flow captured from videos involving people. On the other hand, *Tokuda et.al.* [3] use machine learning techniques to hand-craft different features and combine them into a detection method.



(a) CG (b) PG

Fig. 1. CG image of the actor Russel Crowe and a PG image of the actor in the same pose. Realism degree presented in CG image is extremely high.

Going through machine learning path, this work proposes a new method for CG image detection which uses VGG19 architecture [4], without last fully-connected layers, with weights of the model pre-trained on ImageNet dataset, to extract features from eyes regions from images containing a person. In the place of last fully-connected layers we place a trained classifier able to classify eyes from PG and CG images with 0.80 accuracy. It is straightforward to realize that an eye correctly classified as CG leads to a CG image.

Among the main contributions of this paper we can highlight: (1) the use of regions not explored before (eyes) in CG image detection problem; (2) the proposition of a new approach for CG images detection based on a combination of inconsistencies present in eyes regions with a transfer learning approach for inconsistencies encoding as features; (3) an accuracy around 0.80 in tests performed over two different datasets containing highly realistic images.

The rest of the paper is organized as follows: Section II describes some works related with CG images detection in Digital Forensics literature. Section III describes in details the proposed methodology. Section IV presents the main experiments performed for methodology validation, exposing the obtained results and comparing them with literature methods. Finally, Section V presents the main conclusions and future research directions.

II. Related Work

Several works in the literature have studied the importance of distinguishing CG and PG images. *Holmes et.al.* [1], for example, highlight user's difficult in CG image detection task. Farid [5] by other hand, analyses the impact of CG images when analyzing child pornography images.

In the direction of methods for detecting CG images, *Tokuda et.al.* [3] propose to use a combination of a big number of feature extraction algorithms and different classifiers fusion techniques. Despite evaluating a big number of CG and PG images, the authors did not focused in highly realistic images.

Recently *Tan et.al.* [6], based upon the statement that texture features has a strong ability to distinguish CG and PG images, have used Local Ternary Patterns (LTP) for feature

[1] John Jonah Jameson and Peter Parker are fictional characters of Spider-Man stories in the Marvel Comics Universe.

[2] http://www.imdb.com/title/tt3748528/

0-7695-6321-X/17/31.00 ©2017 IEEE
DOI 10.1109/ICMLA.2017.00-47

posterior classification using a Support Vector Machine (SVM) [7] classifier. The authors report good results but, again, highly realistic images have not been evaluated.

Despite present good results, most of the methods for CG image detection does not explore specific details into the images as other forensics problems explore. In image splicing detection, for example, Johnson and Farid [8] explore direction of light estimated from eye's specular highlights to identify fake images. The same idea has been applied by *Saboia et.al.* [9] but now using a machine learning improvements.

In the next section we introduce an approach that also explores details in the eye's region to accomplish CG image detection.

III. Proposed Method

To distinguish between PG and CG images, we propose a new method constructed under the assumption that eye's region keeps important information for CG image detection. Specially, when removing eye's specular highlight from previously detected and cropped eye region, resulting images (without reflection) of CG images present much more artifacts than PG images (also without reflection). Since human eyes are a hard part to be produced in CG images, we believe these artifacts result from imperfections in computer graphics techniques applied to generate eyes.

Once each eye from an image is located and it has its specular highlight removed, as detailed in Section III-A, we then take advantage of transfer learning process to use VGG19 architecture as a feature extractor, producing a set of bottleneck features. Finally, features extracted from each eye are fed into a Machine Learning (ML) classifier to detect if an eye, and consequently an image, is or not, produced by CG.

A. Eye's Specular Highlights Removal

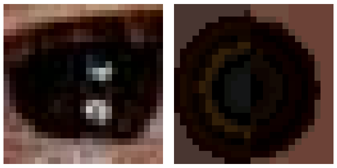
The first step of proposed method is to detect the eye region for, next, separate specular and diffuse components for each eye of a person image. Detection of eye region is an straightforward task and can be performed manually or using some automatic approach. We use Viola and Jones [10] algorithm.

Next, we need to perform eye's specular highlights removal. To accomplish this task, we have employed an approach based on Nishino and Nayar [11] work.

Given an image containing a human eye, the method proposed by Nishino and Nayar [11] uses a geometric model of the cornea to estimate 3D coordinates and orientation of the cornea in the cameras coordinate frame from a single image.

According to the authors, under weak-perspective projection, the major axis of the ellipse detected in the image corresponds to the diameter of the circular limbus, which delimits cornea region in 3D space. Once estimated the image parameters of the limbus, Nishino and Nayar's [11] method is able to compute the 3D orientation of the cornea and the direction of light ray for each cornea point.

Finally, given each light ray value, we are able to separate diffuse and specular components in cornea region. The result of this separation is depicted in Figure 2.



(a) Cornea Region (b) Diffuse Image

Fig. 2. Eye's specular highlight removal using Nishino and Nayar's [11] approach. For each pixel in the eye, we calculate light source properties (direction and intensity), removing specular highlights in the end of the process.

B. Bottleneck Features Extraction via VGG19 Architecture

Once eye region has been extracted and eye's specular highlights have been removed, the next step of proposed method is to encode this region information in useful features which can be fed into a machine learning classifier.

Instead of carrying out a laborious hand-crafted process, we take advantage of transfer learning process, which consists in transferring the parameters of a neural network trained with one dataset for a specific task to another problem with a different dataset and task [12]. This way, we can use part of a trained model (without top layers) and feed it with our eye images, in a way to produce a feature vector for each eye image. To accomplish this task, we choose the VGG19 model [4], producing feature vectors with 25,088 dimensions.

The VGG network architecture was initially proposed by Simonyan and Zisserman [4]. The VGG architecture with 16 layers (VGG16) and with 19 layers (VGG19) were the basis of their ImageNet Challenge 2014 submission, where the Visual Geometry Group (VGG) team secured the first and the second places in the localization and classification tracks respectively.

The VGG19 architecture is structured starting with five blocks of convolutional layers followed by three fully-connected layers. Convolutional layers use 3×3 kernels with a stride of 1 and padding of 1 to ensure that each activation map retains the same spatial dimensions as the previous layer. A rectified linear unit (ReLU) activation [13] is performed right after each convolution and a max pooling operation is occasionally used to reduce the spatial dimension. Max pooling layers use 2×2 kernels with a stride of 2 and no padding to ensure that each spatial dimension of the activation map from the previous layer is halved. Two fully-connected layers with 4,096 ReLU activated units are then used before the final 1,000 fully-connected softmax layer.

The convolutional blocks can be seen as feature extraction layers. The activation maps generated by these layers are called bottleneck features.

C. Top Classifier

As presented in Section III-B, VGG19 architecture [4] uses in the end of architecture three fully-connected layers to perform classification task. However, the proposed method takes advantage of VGG19 architecture just to generate specialized features based on data, without the necessity of hand-crafted extraction. Instead of using fully-connected layers to perform the classification, we replace them by a Support Vector Machine (SVM) [7] with a linear kernel.

The SVM classifier has been chosen based on two main factors: first, for many years, SVM was the most used Machine Learning classifier in classification problems; second, given small size of evaluated datasets (as detailed in Section IV-A), the use of the full VGG19 model would not be recommended given the need for a larger number of training samples.

Thus, the last step of the proposed method is to use the bottleneck features of each eye, previously extracted, as input for an SVM. Once SVM model has been trained, given a new test image, we classify both eyes from image.

Figure 3 summarizes the entire pipeline of proposed method.

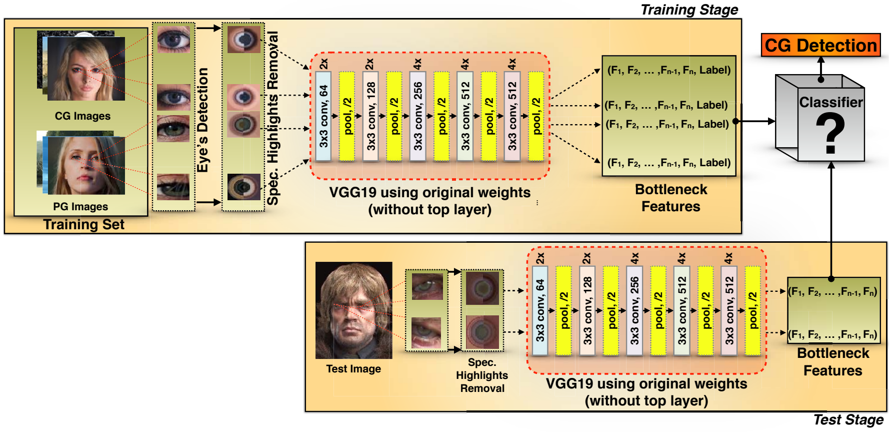


Fig. 3. Overview of proposed method. Given a new image, the method detects eye's region and remove eye's specular highlights. Then, this eye part is used as input into the convolutional blocks of VGG19 model, which is responsible for bottleneck features extraction. Finally, a trained classifier is used at the end of the pipeline to classify each eye as CG or not.

IV. Experiments and Results

To validate the proposed method, we have performed different rounds of experiments on two datasets containing highly realistic images, one proposed by *Holmes et.al.* [1] and another one collected by ourselves on the Internet.

A. Datasets

1) DSCG1: The first dataset were proposed in *Holmes et.al.* [1] work. It is composed of 30 extremely realistic CG images of faces collected from different websites and 30 PG images. The 30 CG images depicts 15 male and 15 female faces with image resolution varying from 389×600 to 557×600 pixels. For each CG image, the dataset also contains a very similar PG image (sometimes depicting the same person). PG images have resolutions varying from 387×600 to 594×600 pixels.

2) DSCG2: The second dataset has been collected by ourselves from different websites. Trying to follow the *Holmes et.al.* [1] path, for each CG image collected, we also collected a PG image as similar as possible. Since we are not interested in evaluating gender detection, we did not concern about the number of males and females in the images. Dataset is composed of 160 images (80 CG and 80 PG) with resolution from 236×308 to 1569×2000 for CG images and from 194×259 to 5616×3744 for PG images.

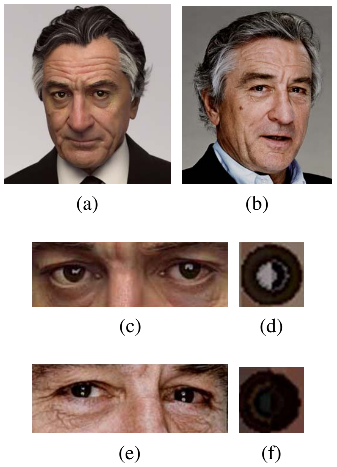
B. Implementation Details

We implemented the proposed method using Python 3.5 with Keras 2.0.3[3] and ThensorFlow 1.0.1[4]. All performed tests have been executed in a machine with an Intel(R) Xeon(R) CPU E5-2620 2.00GHz with 96GB of RAM and a Titan XP GPU.

C. Qualitative Analysis of Eye's Specular Highlights Removal

Our first round of experiments focuses on showing some qualitative results of eye's specular highlight removal. To perform this, we use Viola and Jones approach [10] to detect eye location. Then, we apply the Nishino and Nayar [11] algorithm to detect the light ray for each pixel in cornea region, removing specular components for each eye in an image.

Figure 4 depicts an example of qualitative results for eye's specular highlights removal. Note that the image produced after the reflection removal from CG have a larger number of artifacts (white dots). Furthermore, along all tests that we have conducted, it has been observed that the method performs better in dark eyes than in clear eyes. This is due to the fact that the contrast is higher between the background color and the color of reflected ray.


(a) (b)

(c) (d)

(e) (f)

Fig. 4. Eye's specular highlights removal. Images (a) and (b) present, respectively, a CG image and a PG image from *Holmes et.al.* [1] dataset. Image (c) represents a zoom view of (a) and image (d) represents left eye from (c) after specular reflection removal. In the same way, image (e) represents a zoom view of (b) and image (f) represents left eye from (e) after specular highlights removal. It is not difficult to realize that (d) image presents much more artifacts, mainly in central part, than (f) image.

D. Validation Protocol

For the next rounds of experiments, to avoid bias in performed tests, we applied the 10-fold cross-validation protocol, reporting besides average accuracy, ROC curve and area under the curve (AUC) for each round of experiments.

E. DSCG1 classification using Support Vector Machines

To measure the performance of the proposed method, at this round of experiments, we have used a Support Vector Machine (SVM) [7]

[3] https://keras.io

[4] https://www.tensorflow.org

classifier to perform classification of DSCG1 dataset.

We use an SVM with linear kernel where the parameter C has been obtained through a gridsearch process with $C \in \left[10^{-2}, 10^{-1}, ..., 10^{2}\right]$. The best C obtained was 10^{-2}.

Figure 5 depicts ROC curve for this round of experiments. True positive rate reports the amount of eye's regions from CG images correctly classified while false positive rate reports the amount of eye's regions from PG images misclassified as CG. The area under the curve (AUC) is 0.73 with an average accuracy of 0.68.

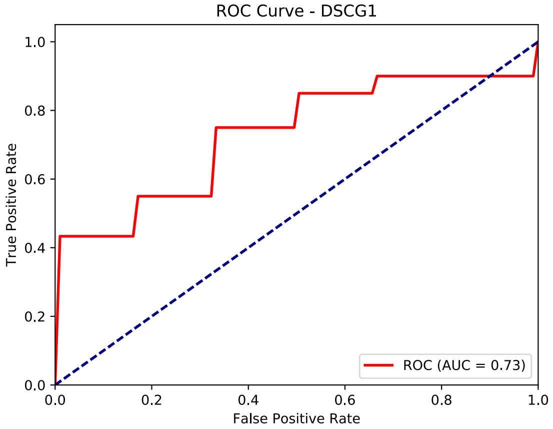


Fig. 5. ROC curve provided by classification using SVM for 10-fold cross-validation in DSCG1.

F. DSCG2 classification using Support Vector Machines

As performed in Section IV-E, at this round of experiments, we have used a Support Vector Machine (SVM) [7] classifier to perform classification of DSCG2 dataset.

Reproducing the same scenario of previous section, at this round of experiments we use an SVM with linear kernel where the parameter C has been obtained through a gridsearch process with $C \in \left[10^{-2}, 10^{-1}, ..., 10^{2}\right]$. The best C obtained was 10^{-2}. Figure 6 depicts ROC curve of this round. The AUC is 0.84 with an average accuracy of 0.76.

G. Full Image vs. Eye's Region vs. Eye's Region with Specular Highlight Removal

In Section **??** we showed how transfer learning process and bottleneck features extraction help to improve separability of samples in the CG image detection problem.

Now, we show the contribution of eye's specular highlights removal for features extraction in CG detection process. Most of the literature methods that use machine learning approaches for CG detection, works under the entire image domain, using features extracted from the entire image. So, at this round of experiments, we use DSCG1 and DSCG2 datasets to evaluate effectiveness of extracting bottleneck features from different scenarios:

- **DIF:** bottleneck features are extracted from eyes region after eye's specular removal, as detailed in Section III-A;
- **NoDIF:** bottleneck features are extracted from eyes region without eye's specular removal. In other words, at this scenario, we just crop a Region of Interest (RoI) in eye region an use this RoI as input for bottleneck features extraction;
- **IMG:** bottleneck features are extracted from the entire image.

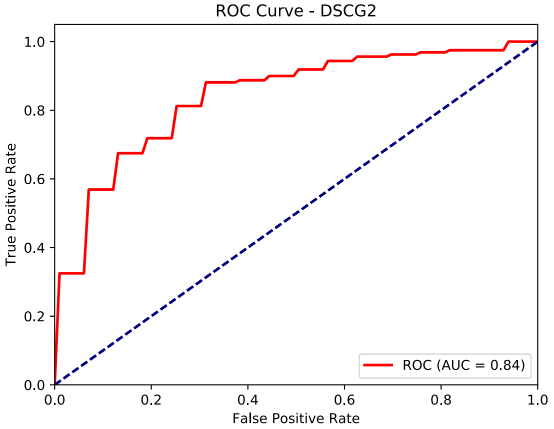


Fig. 6. ROC curve provided by classification using SVM for 10-fold cross-validation in DSCG2.

We use an SVM with linear kernel where the parameter C has been obtained through a gridsearch process with $C \in \left[10^{-2}, 10^{-1}, ..., 10^{2}\right]$. The best C obtained was 10^{-2}. Furthermore, we applied the same validation protocol as described in Section IV-D.

In DSCG1 dataset, NoDIF scenario presents the better result with an accuracy of 0.77, followed by DIF and IMG scenarios with an accuracy of 0.67 and 0.53, respectively.

In DSCG2 dataset, DIF scenario presents better result with an accuracy of 0.76, followed by NoDIF and IMG scenarios with an accuracy of 0.72 and 0.53, respectively.

This round of experiments shows that eye region provides most effective information (features) in CG detection when compared with the entire image. Also, this experiment shows that DIF scenario presents better results in DSCG2 dataset than in DSCG1, where NoDIF scenario performed better in DSCG1.

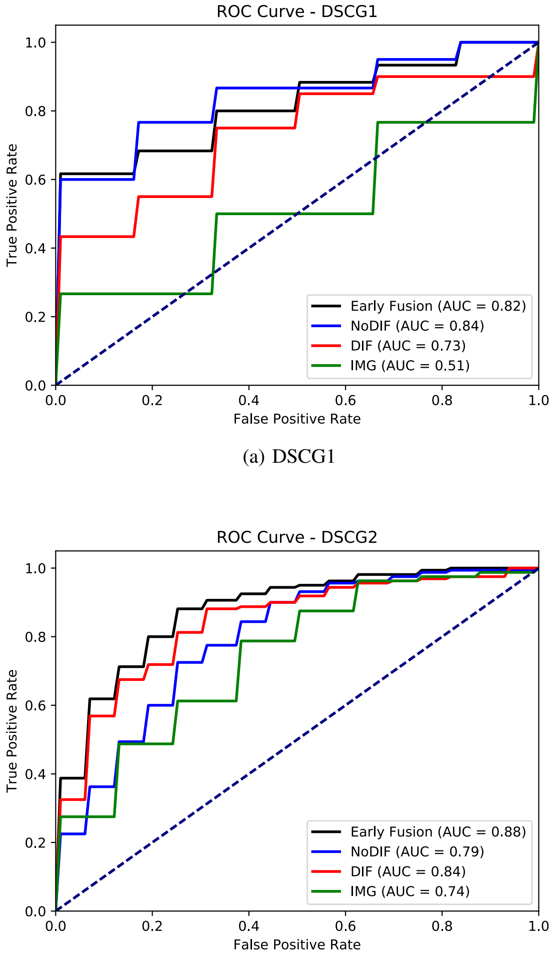
H. Early Fusion of Eyes Features

In machine learning field, different combinations of features × classifiers can produce different results when evaluating the same samples [14], [15]. This can be seen in Section IV-G, where we presented classification results when using three different approaches (NoDIF, DIF, IMG) as input for proposed method. Many times, these different results can provide complementary information which, when combined, can lead to an improved result.

Since NoDIF and DIF approaches have presented the best results, and given that each one of these approaches performed better in one of tested datasets, at this round of experiments we evaluate the performance in an early fusion [16] process. In early fusion, given a sample, we first extract bottleneck features from NoDIF and DIF scenarios. Then, we concatenate these features into a single feature vector, which is used to feed our classifier.

Here we also use an SVM with linear kernel where the parameter C has been obtained through a gridsearch process with $C \in \left[10^{-2}, 10^{-1}, ..., 10^{2}\right]$. The best C obtained was 10^{-2}. We applied the same validation protocol as described in Section IV-D.

Figure 7 depicts previously results compared with early fusion in DSCG1 and DSCG2. In DSCG1, we obtained an accuracy of 0.78 with an AUC 0.82 while in DSCG2 achieved accuracy was 0.80 with an AUC of 0.88.



(a) DSCG1

(b) DSCG2

Fig. 7. ROC curves using early fusion. Achieved accuracy in DSCG1 was 0.78 while in DSCG2 accuracy was 0.80.

As depicted in Figure 7(a), despite the better accuracy rate achieved, early fusion did not improve ROC curve in DSCG1. On the other hand, in DSCG2, as depicted in Figure 7(b), early fusion provides more effective features, resulting in a better accuracy and ROC.

V. Conclusions and Research Directions

This work proposed a new CG image detection method which takes advantage of the inconsistencies occurred in eye's region of CG images containing people. Given an image of a person, we extracted features from eye's region, with and without remove specular highlights, using a deep convolutional neural network model based on VGG19 and transfer learning approach. Locating eye's region using Viola and Jones algorithm [10], we remove reflection using the process described in Section III. This process highlights artifacts present, specially, in eye's region of CG images.

Each eye image is fed into our deep CNN model and, as result, we obtain a 25,088 dimension feature vector, here called bottleneck features, which is used to train a machine learning classifier in PG and CG eye classification.

To validate proposed method, we perform different rounds of experiments where we measure important characteristics as the quality of eye's specular highlight removal (and consequently the presence and absence of artifacts), the contribution of bottleneck features in classes separation, and the contribution of eye's specular highlights removal for CG detection.

Furthermore, after realizing that eye's region with and without specular highlights removal provides complementary information, we propose the use of an early fusion process to construct more distinctive features, which improves our best results in 0.04, achieving an accuracy of 0.8 with an AUC of 0.88.

Finally, as future research directions, we intend to explore the fusion of bottleneck features extracted from different deep CNNs models.

Acknowledgments

The authors would like to thank the financial support of IFSP-Campinas, FAPESP (grant 2016/13011-9) and CNPq (423797/2016-6). We also would like to thank the authors *Tokuda et.al.* [3] who helped us with dataset acquirement.

References

[1] O. Holmes, M. S. Banks, and H. Farid, "Assessing and improving the identification of computer-generated portraits," *ACM TAP*, vol. 13, no. 2, p. 12, 2016.

[2] V. Conotter, E. Bodnari, G. Boato, and H. Farid, "Physiologically-based detection of computer generated faces in video," in *IEEE ICIP*, 2014, pp. 248–252.

[3] E. Tokuda, H. Pedrini, and A. Rocha, "Computer generated images vs. digital photographs: A synergetic feature and classifier combination approach," *Elsevier JVCI*, vol. 24, no. 8, pp. 1276 – 1292, 2013.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] H. Farid, "Creating and detecting doctored and virtual images: Implications to the child pornography prevention act," Tech. Rep. TR2004-518, 2004.

[6] D. Q. Tan, X. J. Shen, and H. P. Qin, J.and Chen, "Detecting computer generated images based on local ternary count," *Springer PRIA*, vol. 26, no. 4, pp. 720–725, 2016.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[8] M. Johnson and H. Farid, "Exposing digital forgeries through specular highlights on the eye," in *IHW*, 2007.

[9] P. Saboia, T. Carvalho, and A. Rocha, "Eye specular highlights telltales for digital forensics: A machine learning approach," in *ICIP*, 2011.

[10] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, May 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb

[11] K. NISHINO and S. K. NAYAR, "Eyes for relighting," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 704–711, 2004.

[12] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Adv Neural Inf Process Syst*, 2014, pp. 3320–3328.

[13] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[14] T. Carvalho, F. A. Faria, H. Pedrini, R. da S. Torres, and A. Rocha, "Illuminant-based transformed spaces for image forensics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 4, pp. 720–733, 2016.

[15] F. A. Faria, J. A. dos Santos, A. Rocha, and R. da S. Torres, "A framework for selection and fusion of pattern classifiers in multimedia recognition," *Pattern Recognition Letters*, vol. 39, pp. 52 – 64, 2014, advances in Pattern Recognition and Computer Vision. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865513002870

[16] A. Ferreira, L. Bondi, L. Baroffio, P. Bestagini, J. Huang, J. A. dos Santos, S. Tubaro, and A. Rocha, "Data-driven feature characterization techniques for laser printer attribution," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 8, pp. 1860–1873, 2017.