PHYLOGENETIC ANALYSIS OF NEAR-DUPLICATE IMAGES USING PROCESSING AGE METRICS

S. Milani[1], *M. Fontana*[1], *P. Bestagini*[2], *S. Tubaro*[2]

1 :Department of Information Engineering, University of Padova, Italy
e-mail: simone.milani@dei.unipd.it, marco.fontana.5@studenti.unipd.it
2 :Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
e-mail: paolo.bestagini@polimi.it, stefano.tubaro@polimi.it

ABSTRACT

Recent researches on image forensics have led to the design of algorithms to study the phylogenetic relationship between near-duplicate (ND) images. The proposed solutions aim at reconstructing the image phylogeny tree (IPT), and they have immediate applications in security, law and copyright enforcement, and news tracking services. Anyway, the effectiveness of such strategies strictly depends on the accuracy in characterizing image similarities. In this paper, we show that it is possible to take into account additional information to better reconstruct the IPT. More specifically, we propose a set of features that blindly model the processing age of an image, i.e., how much an image has been edited in its lifetime. By exploiting these features, it is possible to improve the performance of IPT reconstruction by increasing the accuracy and reducing the computational complexity.

***Index Terms*—** image phylogeny, processing age, multimedia forensics, near-duplicate images

1. INTRODUCTION

Thanks to the availability of accessible and usable authoring and sharing tools, the publication and distribution of digital images online is a relatively-easy task. Unfortunately, this has brought several new issues and problems from legal and social point-of-views. Images and videos can "mutate" as they spread out and some of the modified versions are not always authorized [1]. After posting an image online, other users can copy, resize and/or re-encode it and then repost different versions, thus generating similar but not identical copies, often referred to as near-duplicates (NDs).

In the last decade, several research groups have successfully focused on the design and deployment of algorithms for the detection and recognition of the near-duplicates of a document. A far more challenging task that has been vastly overlooked until recently, arises when we want to identify which document is the original within a set of NDs, and the structure of their generation [2, 3]. These relations can be well described by means of a structure called image phylogeny tree (IPT) (see an example in Fig. 1).The term was mutuated from biology given the analogy with the analysis of the mutation process that occur to living organisms in biology.

In order to reconstruct the IPT given a set of ND images, state-of-the-art algorithms compute similarities/dissimilarities metrics between every pair of images [4, 5]. Then, images are associated to the nodes of a weighted directed graph (where weights correspond to the dissimilarity values), and a minimum spanning tree algorithm is run.

Unfortunately, these solutions are less robust in identifying the correct relations whenever dissimilarity values are noisy [6] or some nodes are missing in the graphs (since some images, which are present in the original IPT, are not available to the analyst). In this paper we show that it is possible to overcome this problem by introducing a set of features that are strictly related to the processing age (PA) of the image, i.e., to the number of processing steps that have been applied to it. PA metrics are blindly computed on each single image and permits ordering near-duplicate images from the earliest processing stages to the latest ones. The reconstruction of the IPT benefits from these features since it is possible to avoid computing dissimilarities between pair of images which presents different PA values. This fact permits reducing significantly the computational complexity of the approach since the calculation of dissimilarities between image is the most demanding task in IPT reconstruction algorithms. Moreover, whenever some nodes are still missing, it is possible to place every image at the correct depth of the original IPT.

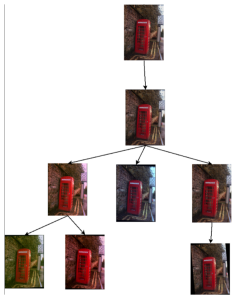

Fig. 1. Image phylogeny tree between ND images.

In the following, in Section 2 the IPT reconstruction problem is formulated. Section 3 describes the proposed features, and Section 4 describes how to use them in estimating the IPT. Experimental results are reported in Section 5. Finally, Section 6 concludes the paper.

2. RECONSTRUCTION OF IMAGE PHYLOGENY TREES

State-of-the-art algorithms for IPT reconstruction approximately follow a common pipeline [4, 6]. In this section we briefly outline this pipeline that serves as a background for the rest of the paper. For more details on each step, please refer to [4].

Given a set of K near-duplicate images I_k, $k = 1, \ldots, K$, first the dissimilarity matrix $D = [d_{h,k}]$ is computed. Each element $d_{h,k}$ corresponds to the dissimilarity between images I_h and I_k. This value is related to the likelihood of I_k to be parent of I_h and it is

computed as

$$d_{h,k} = \mathcal{L}(I_h, I_{k\to h}), \tag{1}$$

where $\mathcal{L}$ is any dissimilarity metric (e.g., mean squared error) and $I_{k\to h}$ is a transformed version of I_k obtained via a set of processing steps that maximize the similarity with I_h. In order to estimate $I_{k\to h}$, we adopted the method reported in [6].

Once dissimilarity matrix D has been computed, it can be interpreted as a complete directed graph, where each node is an image, and each directional branch is the dissimilarity $d_{h,k}$. Computing a minimum spanning tree (MST) on the graph permits estimating the IPT that generated the set of images. In our approach, we adopted a solution based on optimum branching algorithm (OB) [7].

Notice that dissimilarity computation is typically the most demanding task in IPT reconstruction algorithms due to the estimation of $I_{k\to h}$ for each image pair (i.e., $K(K-1)$ cases). For this reason, in the following, we propose an approach that helps reducing the number of dissimilarity computations.

3. A PROCESSING AGE METRIC FOR IMAGES

Several works in multimedia forensics have shown that for every image it is possible to compute a set of statistics following a pre-defined model. Every modification operated on the image alters these statistics so that they deviate from the original model [8]: these deviations can be considered as traces (footprints) of the alterations and depends on both the type and the number of operations [9, 10]. As a matter of fact, they can be used to estimate the processing age (PA) of an image.

As an example, previous works have shown that the statistics of DCT coefficients c of natural images can be well modelled by parametric probability density function (pdf) such as Laplacian [11], generalized Gaussian [12], laplacian+impulsive [13], and Cauchy [14]. Alternatively, if DCT coefficients are quantized (e.g., due to JPEG compression), further studies have shown that the pdf of their first digits (FDs) with base M, i.e.,

$$m = \mathrm{FD}_M(c) = \left\lfloor \frac{|c|}{M^{\lfloor \log_M |c| \rfloor}} \right\rfloor. \tag{2}$$

can be well modelled by some parametric functions. If $M = 10$ is used, FD distribution should follow a logarithmic curve defined by Benford's law [10, 15].

In this work, we define the probability mass function (pmf) of the absolute values of quantized DCT coefficients c located at frequencies (i, j) as $p_{i,j}(c)$. Similarly, we define the pmf of the FDs m computed on the quantized coefficients c located at frequencies (i, j) as $P_{i,j}(m)$. The fitting model that we use for DCT coefficients and FDs is defined as

$$\begin{aligned} p^f_{i,j}(c) &= \Gamma e^{-\pi(c)}, \\ P^f_{i,j}(m) &= \Gamma e^{-\pi(m)}, \end{aligned} \tag{3}$$

where the first or second equation is adopted depending on the use of DCT coefficients or FDs, $\pi(\cdot)$ is a polynomial of second degree and Γ is a normalizing constant. In this way, it is possible to include both a Laplacian and a Gaussian model for the absolute value of quantized coefficients avoiding the fitting problems related to the generalized Gaussian.

Given an image I, it is possible to compute the coefficients c (or the FDs m) at different frequencies (i, j), and find the best fitting model $p^f_{i,j}(c)$ (or $P^f_{i,j}(m)$) according to (3). Fig. 2 shows that the pmf of both DCT coefficients and FDs deviates from the model (3) when we apply a series of operations to the first image of the UCID dataset [16]. More specifically, the plots in Fig. 2 report the statistics $p_{i,j}(c)$ and $P_{i,j}(m)$ with the corresponding fitted models when one (a,d), two (b,e) and three (c,f) operations randomly selected among rotation, rescaling and cropping are applied, followed by a JPEG compression. Parameter values will be reported in Section 5. Figs. 2 d, e and f report the plots of $P_{i,j}(m)$ and the corresponding fitted model, considering $M = 100$ (in order to have a more reliable statistic). It is possible to notice that the fitting error $\|p_{i,j}(c) - p^f_{i,j}(c)\|$ (or $\|P_{i,j}(m) - P^f_{i,j}(m)\|$) increases as the number of transformation increases.

From these premises, it is possible to employ the fitting error for the coefficients at some locations as a processing age measurements that helps building the IPT. The more the statistics fit the model, the less an image has been processed and altered. Therefore, PA metrics should focus on the divergence between $p_{i,j}(c)$ and $p^f_{i,j}(c)$ (or, equivalently, between $P_{i,j}(m)$ and $P^f_{i,j}(m)$). To this purpose, we tested different divergence measures $D_X(P||P^f)$ to find out the one that performs better (the index X will be described in the following).

Since divergence is not a symmetric function, we usually evaluate the sum of divergencies

$$D_X(P, P^f) = D_X(P||P^f) + D_X(P^f||P). \tag{4}$$

In our analysis, we considered only coefficients at frequencies $(0, 1)$ and $(1, 0)$ since they are more stable and their statistics are mildly affected by transformations like small rotations and rescalings. Moreover, their statistics can be well modelled by a simple Laplacian or Gaussian random variable for non-edited images and progressively deviate from the proposed model as the number of alterations increases.

From these premises, the processing age metrics associated to divergence D_X can be defined as

$$\text{PA-Xc} = \frac{D_X(p_{0,1}(c), p^f_{0,1}(c)) + D_X(p_{1,0}(c), p^f_{1,0}(c))}{2} \tag{5}$$

whenever referred to coefficients, and

$$\text{PA-Xfd} = \frac{D_X(P_{0,1}(m), P^f_{0,1}(m)) + D_X(P_{1,0}(m), P^f_{1,0}(m))}{2} \tag{6}$$

whenever referred to FD statistics.

3.1. Processing age from Kullback-Leibler divergence

In a first set of tests, we considered the well-known Kullback-Leibler divergence D_{KL} which can be expressed via the equation

$$D_{KL}(P||P^f) = \sum_m P(m) \log_2 \frac{P(m)}{P^f(m)}. \tag{7}$$

Including equation (7) into equations (4), (5), and (6), we obtains the processing age metrics PA-KLDc and PA-KLDfd (where reference X =KLD is associated to the Kullback-Leibler divergence)

3.2. Processing age from Renyi divergence

Experimental results showed that the values of PA metrics based on the Kullback-Leibler divergence become very dense as the number of processing steps increases. Moreover, Kullback-Leibler divergence values change significantly for low-entropy statistics, but for medium and high entropy pmf the divergence values vary quite

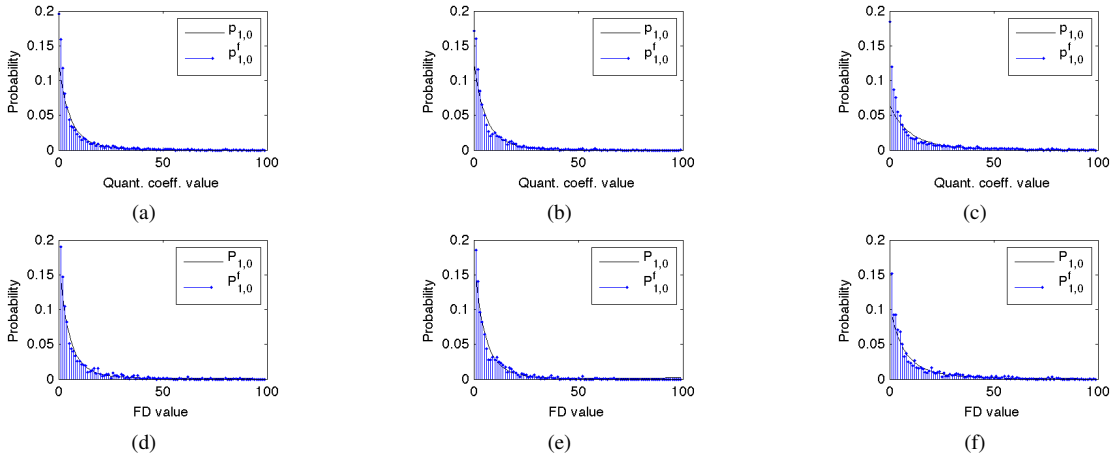

Fig. 2. Probability mass function for coefficients ($p_{1,0}(c)$) and FDs ($P_{1,0}(c)$) and the corresponding fitted model ($p^f_{1,0}(c)$ and $P^f_{1,0}(c)$, respectively) for image 0 of UCID dataset after different number of transformations (indexes are omitted). Upper graphs: $p^f_{1,0}(c)$; lower graphs: $P_{1,0}(c)$. (a,d): one transformation; (b,e): two transformations; (c,f): three transformations. The fitting error values are 0.0010 (a), 0.0011 (b), 0.0017 (c), 0.025 (d), 0.026 (e), and 0.027 (f).

slowly. In order to obtain a better separation of PA values, we considered the Renyi divergence D^{α}_{R} [17] with parameter $\alpha \neq 1$, which can be written as

$$D^{\alpha}_{R}(P||P^f) = \frac{1}{\alpha - 1} \log_2 \left(\sum_m P(m)^{\alpha} P^f(m)^{1-\alpha} \right) \quad (8)$$

Similarly, the associated processing age metrics will be called PA-RENc and PA-RENfd, where the parameter α is set to 0.5. Different values of α were tested, but only $\alpha < 1$ permitted obtaining a good performance since metric values are distributed on a wider range whenever the statistics have medium or high entropy (i.e., images are non-trivial or excessively compressed).

3.3. Processing age from Tsallis divergence

In the end, we tested the possibility of achieving a better characterization of the processing age using Tsallis entropy [18]. Performing the same generalization from entropy to divergence used for Shannon and Renyi entropy, it is possible to write the Tsallis divergence D_T as

$$D^{\alpha}_{T}(P||P^f) = \frac{1}{1 - \alpha} \left(1 - \sum_m (P(m)^{\alpha} P^f(m)^{1-\alpha}) \right). \quad (9)$$

Similarly, the associated processing age metrics, generated via equations (5) and (6), are PA-TSAc and PA-TSAfd, where the parameter α is set to 0.5. after a set of tests.

In the following, we will explain how PA metrics can be included in the phylogenetic analysis

4. RECONSTRUCTING IMAGE PROCESSING TREES VIA PROCESSING AGE METRICS

Given a set of K ND images I_k, $k = 1, \ldots, K$, it is possible to reconstruct the IPT via the following strategy.

At first, the algorithm computes the processing age PA(I_k) for every image (using one of the proposed techniques). Notice that PA computations increase linearly (i.e., K) with the number of images rather than quadratically as the dissimilarities (i.e., $K(K-1)$).

Then, the age of images are compared in order to exclude unlikely parent-child relations. As an example, if PA(I_h) $<$ PA(I_k), then it is less probable that the h-th image is the parent of the k-th image. As a matter of fact, there is no need to compute the dissimilarity $d_{k,h}$.

Unfortunately, processing age metrics are affected by errors, and therefore, if the values PA(I_h) and PA(I_k) are too close they could not be reliable enough. In order to solve this, we introduced a threshold δ which avoid removing the parent-child dependencies when the PA values are too close, i.e.,

procedure REMOVEDEPENDENCY(k, h)
  if $|\text{PA}(I_h) - \text{PA}(I_k)| \geq \delta$ **then**
    if PA(I_h) $<$ PA(I_k) **then**
      remove dependency $I_h \rightarrow I_k$
    else
      remove dependency $I_k \rightarrow I_h$
  else
    do not remove anything.

According to this procedure it is possible to compute only some values of the D matrix (i.e., those for which the dependency has not been removed). Once D has been populated, the standard Oriented Kruskal [4] or optimum branching [19] algorithms can be used to reconstruct the IPT.

Note that the parameter δ can control both the accuracy of reconstruction and the computational complexity. As δ increases, the percentage of parent-child dependencies (and, consequently, of computed dissimilarities) decreases leading to reduced computational complexity.

In the following section, we will present how this parameter also affect the reconstruction performance.

5. EXPERIMENTAL RESULTS

In order to validate our method, we built a dataset starting from images of the UCID database [16] as suggested in [4]. More specifically, we built 50 trees of 10 and 30 nodes for a total number of $50 \times (10 + 30) = 2000$ images. The root of each tree is a dif-

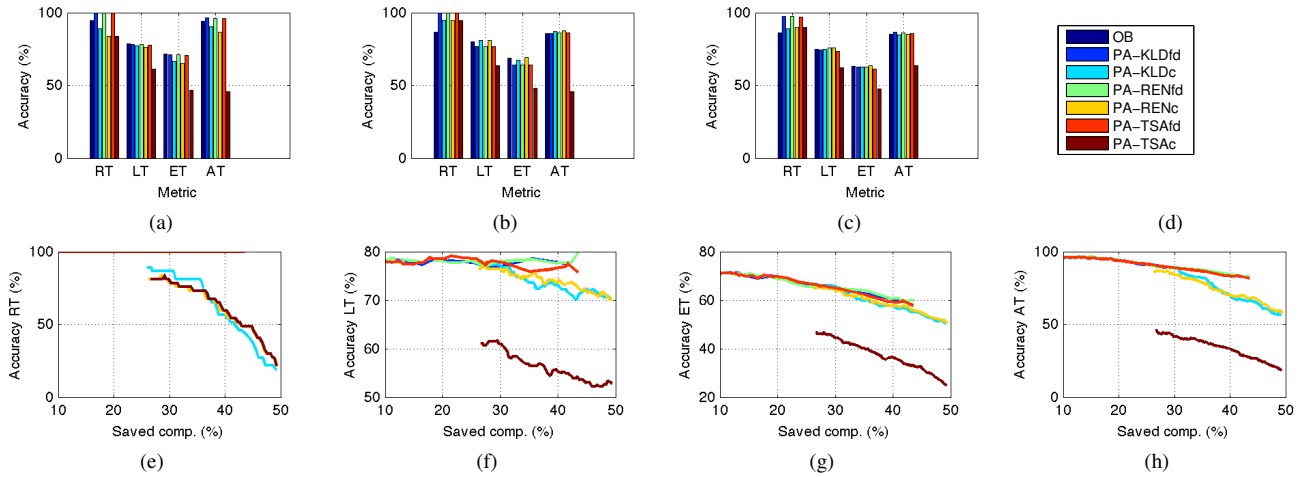

Fig. 3. Performance for different PA metrics in IPT reconstruction. Vertical bars in the first row plots the accuracy for each metric under different test conditions. a) 10 nodes ; b) 30 nodes; c) 30 nodes with removal probability 20%; d) legend. The second row reports the accuracy of parameters RT, LT, ET, AT vs. computational saving for different PA metrics. e) RT, f) LT, g) ET, h) AT.

ferent image compressed in JPEG format. The other nodes of each tree are obtained applying a set of possible transformations (up to 4) randomly-chosen among resampling (factor $\in [.7, 1.2]$), cropping ($[1, 5]$ %), rotating (angle $\in [-5, 5]$ deg), and JPEG compression (QF $\in [85, 87, 90, 92, 95]$). The parameters for each operations are chosen randomly as well. All operations are terminated by a JPEG compression. In our tests, we reconstructed the trees using both the reference algorithm [7] and our proposed method (i.e., testing separately each one of the processing age metrics before D estimation and using OB for the IPT reconstruction).

The reconstruction performance was evaluated using the same metrics reported in [4, 7] , which aim at evaluating the precision in identifying the roots, the leaves, the ancestors and the edges of the original IPT. Naming $\hat{T}$ the estimated IPT and T the correct one, the metric RT reports the percentage of correctly identified roots, i.e., the percentage of $\hat{T}$ where $\mathrm{root}(T) \in \mathrm{root}(\hat{T})$. The metric ET reports the percentage of correctly-identified edges, i.e. $|\mathrm{edges}(\hat{T}) \cap \mathrm{edges}(T)|/|\mathrm{edges}(T)|$, while LT denotes the percentage of correct leaves ($|\mathrm{leaves}(\hat{T}) \cap \mathrm{leaves}(T)|/|\mathrm{leaves}(T)|$). The metric AT evaluates the effectiveness of the algorithm estimating the pairs node-ancestor in the chains. More precisely, given that $A(T)$ denotes the set of pairs of nodes (I_h, I_k) such that I_h is a direct ancestor of I_k, AT evaluates $|A(\hat{T}) \cap A(T)|/|A(T)|$.

Figures 3 a and b report the performance for different PA metrics on trees with different number of nodes/images. In general, it is possible to notice that accuracy obtained using PA metrics improves with respect to standard reference algorithm. More precisely, the accuracy increment obtained by `PA-KLDfd`, `PA-RENfd`, `PA-TSAfd` is the highest for all the metrics. It is possible to notice that this increment is more evident as the number of nodes in the tree increases. In fact, they permit obtaining 100% accuracy for the RT metric while the reference algorithm has a lower performance for large trees (see Fig. 3 c). It is also worth considering that computing the divergences on the coefficient statistics lead to a poorer performance since the accuracy decreases for all the metrics. This fact is more evident when using Tsallis divergence (`PA-TSAfd` vs. `PA-TSAc`), while Renyi and Kullback-Leibler divergence approximately present the same accuracies when applied to coefficient and FD statistics (made exception for the RT metric).

Additional tests were done on incomplete sets, i.e., removing images from the set assuming that the pictures that generated the tree are not fully available. In this case, we performed a random removal of images/nodes with varying removal probability. Fig. 3 c report the accuracies of PA-based IPT reconstruction for a tree of 30 nodes with removal probability equal to 20 %. It is possible to notice that the benefits of using PA metrics are more evident in this case since the accuracy of the reference method decreases. The bars show an increment of 10% in the accuracy of metric RT.

On average, PA metrics permit saving about 30% of dissimilarity computations between images (with respect to OB). Anyway, further investigations were performed in order to characterize the dependency between accuracy and computational effort. To this purpose, we varied the parameter δ (see Section 4) for all the PA metrics computing the obtained RT, LT, ET, AT accuracy and the percentage of skipped dissimilarity computations. The second row in Figure 3 reports the accuracy vs. the computational saving.

Although many PA metrics present the same accuracy, complexity analysis permits noting that for IPTs of 10 images (Fig. 3 e, f, g and h) the metric `PA-TSAfd` provides the best performance since the accuracy decrement as the computational saving increases is minimum for all the considered metrics. It is also worth of consideration using different metrics according to the type of information about the underlying IPT we need to extract. If the analyst needs to identify the root of the tree, he/she simply looks for the image whose coefficient statistics is highly conforming with respect to the fitted model. In case other metrics are targeted (LT, ET and AT), characterizing different fitness levels could be extremely useful to understand the interrelations between images.

6. CONCLUSIONS

In this paper we presented the possibility of including a processing age metric in the reconstruction of IPTs. The age of a processed digital image can be approximated by measuring the fitting of DCT coefficients and FDs statistics with respect to a parametric model. The proposed feature permits both reducing the computational complexity and improve the final accuracy. Future work will be devoted to test the proposed approach on a large scale scenario and to other media format such as audio tracks and video sequences.

7. REFERENCES

[1] L. Gaborini, P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Multi-clue image tampering localization," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2014.

[2] Z. Dias, S. Goldenstein, and A. Rocha, "Large-scale image phylogeny: Tracing image ancestral relationships," *IEEE MultiMedia (MM)*, vol. 20, pp. 58–70, 2013.

[3] L. Kennedy and S.-F. Chang, "Internet image archaeology: Automatically tracing the manipulation history of photographs on the web," in *ACM International Conference on Multimedia (ACM-MM)*, 2008.

[4] Z. Dias, A. Rocha, and S. Goldenstein, "Image phylogeny by minimal spanning trees," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 7, pp. 774–788, 2012.

[5] A. De Rosa, F. Uccheddu, A. Costanzo, A. Piva, and M. Barni, "Exploring image dependencies: a new challenge in image forensics," in *SPIE Conference on Media Forensics and Security (MFS)*, 2010.

[6] A. Melloni, P. Bestagini, S. Milani, M. Tagliasacchi, A. Rocha, and S. Tubaro, "Image phylogeny through dissimilarity metrics fusion," in *European Workshop on Visual Information Processing (EUVIP)*, 2014.

[7] Zanoni Dias, Siome Goldenstein, and Anderson Rocha, "Exploring heuristic and optimum branching algorithms for image phylogeny," *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 1124 – 1134, 2013.

[8] A. Piva, "An overview on image forensics," *ISRN Signal Processing*, vol. 2013, pp. 22, 2013.

[9] T. Bianchi, A. De Rosa, and A. Piva, "Improved DCT coefficient analysis for forgery localization in JPEG images," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[10] S. Milani, M. Tagliasacchi, and S. Tubaro, "Discriminating multiple JPEG compressions using first digit features," *APSIPA Transactions on Signal and Information Processing*, vol. 3, pp. e19, 2014.

[11] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Transactions on Image Processing (TIP)*, vol. 9, pp. 1661–1666, 2000.

[12] G. Calvagno, C. Ghirardi, G.A. Mian, and R. Rinaldo, "Modeling of subband data for buffer control," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 402–408, Apr. 1997.

[13] S. Milani, L. Celetto, and G. A. Mian, "An accurate low-complexity rate control algorithm based on (ρ, e_q)-domain," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 18, pp. 257–262, 2008.

[14] Y. Altunbasak and N. Kamaci, "An analysis of the dct coefficient distribution with the h.264 video coder," in *Proc. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, May 2004, vol. 3, pp. iii–177–80 vol.3.

[15] S. Milani, P. Bestagini, M. Tagliasacchi, and S. Tubaro, "Multiple compression detection for video sequences," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2012.

[16] G. Schaefer and M. Stich, "UCID - An uncompressed colour image database," in *SPIE Conference Storage and Retrieval Methods and Applications for Multimedia*, 2004.

[17] A. Renyi, "On measures of information and entropy," in *Berkeley Symposium on Mathematics, Statistics and Probability*, 1961.

[18] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, pp. 479–487, 1988.

[19] J. Edmonds, "Optimum branchings," *Journal of Research of National Institute of Standards and Technology*, vol. 71B, pp. 233–240, 1967.