3D RECONSTRUCTION FROM WEB HARVESTED IMAGES USING A FORENSIC QUALITY METRIC

Mattia Lecci, Simone Milani

Department of Information Engineering - University of Padova

ABSTRACT

Structure-from-Motion (SfM) algorithms have recently been employed to reconstruct 3D scenes or environments from large sets of unordered images which were harvested from the web. Unfortunately, the accuracy of the reconstruction is significantly affected by the quality and the amount of editing operated on the processed images. Indeed, 3D modelling can significantly benefit from including forensic analysis strategies that are able to reconstruct the processing history of the processed images and select the most reliable pieces of visual information.

The current paper presents an SfM strategy that orders the different views/images of the scene in the reconstruction process according to a *processing age* metric, i.e., a metric parameterizing the amount of processing stages operated on each image. Experimental results show that the proposed solution can improve the estimation accuracy of both 3D points and camera parameters with respect to state-of-the-art solutions.

Index Terms— Structure-from-Motion, processing age, quality metric, 3D reconstruction, multimedia forensics

1. INTRODUCTION

Every day billions of images are posted online by diverse users on heterogeneous data sharing platforms. This fact has recently suggested the development of big data image processing strategies that elaborate this huge amount of visual information to infer an enhanced understanding of the scene [3], build more effective storage systems [4], verify facts and much more. Some of the proposed solutions [5, 6] aim at reconstructing a 3D model of real scene from web-harvested images using generalized Structure-from-Motion (SfM) algorithms.

All these solutions are based on finding couples of conjugate points between couples of different images by matching the local feature descriptors. This coupling permits estimating the acquisition location and orientation for each camera [7] and, at the same time, computing the threedimensional coordinates of the points in the scene [8, 9]. Note that the processed pictures were taken at the same place (the scenario to be reconstructed) by different users with different light conditions, capture configurations, daytimes, occlusions, etc. All these elements make the finding of conjugate points more difficult, and therefore, robust local descriptors like SIFT [10] or SURF[11] are to be adopted.

In the 3D reconstruction, the processing order of the collected images, i.e., how to progressively-include them in the algorithm, plays a significant role. This order is called *track* [5], and experimental results have shown that the final accuracy improves by including first couples of pictures with a lot of conjugate points. This permits building more robust 3D models at the beginning of the reconstruction process. This ordering was very close to the acquisition order for traditional SfM strategies, which were designed for sequentially-captured image taken by a single moving camera [12, 13]. When processing unordered collections of images harvested from the web or from social platforms, this assumption is not valid any more. Moreover, image quality is important as well since degraded images present very few keypoints and noisy local descriptors (which could lead to wrong matches). Downloaded images could have been edited and re-compressed several times before their publication, and this alters keypoint locations and feature values of the associated descriptors [14]. Fig. 1 displays different version of image 3 from dataset `fountain` which have been compressed N_c times with different random QF values in the range $[65, 98]$. Graphs also display the keypoints found on each image. It is possible to notice that as the number of compression increases, the number of keypoints decreases and their locations and orientation could be significantly altered.

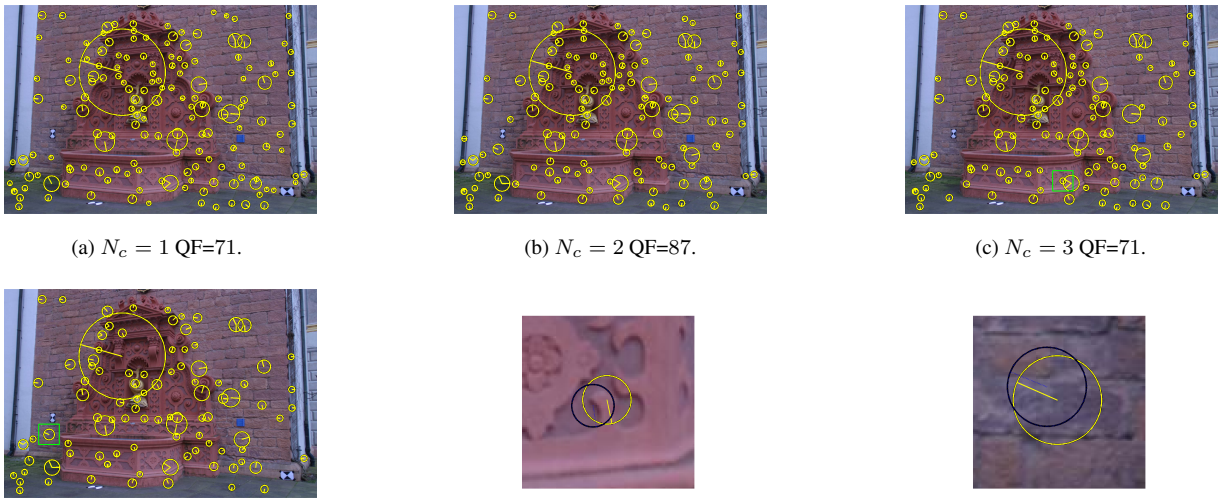
Most of the so-far-proposed approaches order images considering the number and the characteristics of matching conjugate points between couples of pictures [5, 15, 16]. At first , matching outliers are removed, a relational graph [17] between images is built where edge labels are computed from the number of matched points, and then the order is found by graph processing algorithms. Together with the number of matched keypoints, the approach presented in this paper takes into account the quality of the images in terms of number of editing steps. For every image in the dataset, the proposed solution estimates the "processing age", i.e., the number of the editing steps, using a no-reference forensic quality metrics based on the statistics of DCT coefficients [18]. Then, following the metric fusion strategy presented in other approaches [19], label values are changed in order to process images with a low processing age first. This change permits obtaining a higher accuracy both in the reconstructed point clouds, and in the estimated extrinsic camera parameters.

The remaining of the paper is organized as follows. Section 2 reviews some of the state-of-the-art techniques presented literature and highlights the main problems. Section 3 introduces the adopted processing age metric, and Section 4 presents how this can be used in the ordering. Section 5 reports the performance of the proposed solution in terms of accuracy, and Section 6 draws the final conclusions.

2. 3D MODELIZATION VIA STRUCTURE-FROM-MOTION: RELATED WORKS AND MAIN ISSUES

One of the first strategies to be presented is the *Bundler* algorithm [5], which starts estimating a 3D point cloud model of the scene by coupling pairs of images according to the matching local features

The work has been supported by the Robotic 3D [1] and by the CloudVision [2] projects, funded by the University of Padova, Italy.



(a) $N_c = 1$ QF=71. (b) $N_c = 2$ QF=87. (c) $N_c = 3$ QF=71.

(d) $N_c = 7$ QF=67. (e) Detail for image c. (f) Detail for image d.

Fig. 1. Effects of multiple compression stages on SIFT descriptors. QF values are referred to the last coding stage. Images (e), (f) reports a detail from images (c), and (d), respectively (green squares). Yellow circles are referred to the keypoint fount at the last coding stage , blue circles are referred to the corresponding keypoint found at the previous coding stage.

[10]. Similar to [5], the VisualSfM software [15] implements a multicore bundle adjustment using GPU-optimized SIFT.

For every image $I_i \in \mathcal{I}$ $(i = 0, \ldots, N-1)$, the approach in [5] compute a set of n_i SIFT keypoints $S_i = \{m_{i,k}, k = 0, \ldots, n_i\}$, where the pixel position $\mathbf{m}_{i,k}$ (in homogeneous coordinates) is associated to the k-th descriptor $\mathbf{s}_{i,k}$. Moreover, it is possible to assume that, without loss of generality, the pixel $\mathbf{m}_{i,k}$ is also associated to the a 3D point $\mathbf{P}_k$ via the pinhole camera model [9]

$$\mathbf{m}_{i,k} = K_i \left[R_i | \mathbf{t}_i\right] \mathbf{P}_k \tag{1}$$

where K_i is the intrinsic matrix, R_i and $\mathbf{t}_i$ define the orientation and the location of the camera that acquired image I_i.

Given an initial couple of images I_i, I_j, it is possible to generate the set $S_{i,j} = \{(\mathbf{m}_{i,k}, \mathbf{m}_{j,h}) |$ s.t. $\mathbf{s}_{i,k}$ matches $\mathbf{s}_{j,h}\}$. If no wrong matches are present, $k = h$ for all the couples in $S_{i,j}$ (i.e., they are associated to the same 3D point). At this point, both camera parameters and a sparse 3D model can be reconstructed via a resection-intersection strategy [9] and then refined by a bundle adjustment strategy [20, 21] implemented via a least-square minimization.

This initial reconstruction is then refined and enriched by additional points including the other images of the dataset following a specific track. It is necessary to find an ordering that avoids processing weakly-matched images before strongly-matched ones since the first couples that are included in the model have a stronger impact on the 3D point estimation [22].

To this purpose, SfM-based reconstruction strategies computes a similarity/dissimilarity metric for couples of images that permits inferring an optimal ordering. The solutions in [23, 24] computes a similarity value between every couple (I_i, I_j) by estimating the geometrical relation that links keypoint locations. These values are then used to label the edges of a weighted connectivity graph (which can be represented by a correspondence or a distance matrix), and an optimal sequence is found applying graph optimization algorithms. The approach in [5] estimates the fundamental matrix using the 8-points algorithm; this estimation permits removing outliers and generating the new set of matched points $S'_{i,j}$, whose cardinality is used to generate the correspondence matrix.

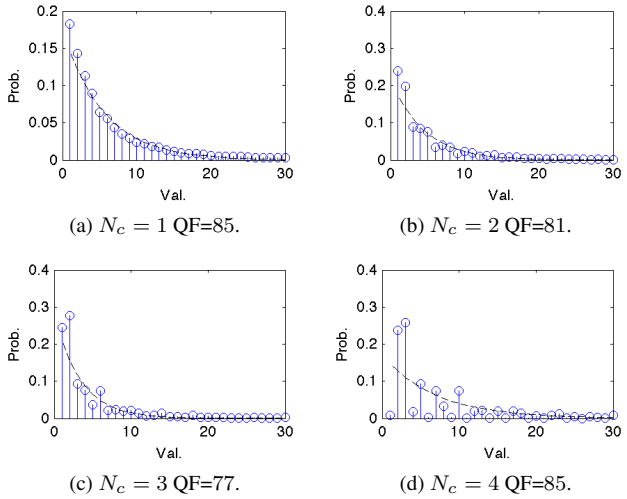
Differently from Bundler, the approach [16] divide the image clustering in two phases: a broad and a narrow phase. The broad phase operates an iterative Maximum Spanning Tree on a correspondence matrix where only a subset of all the SIFT keypoints (the ones with the largest scales) were considered. Then, a refinement is operated by the narrow phase using the MSAC algorithm leading to more accurate distance measurements.

A first attempt to include forensic analysis in SfM strategies was presented in [25], where an estimator of the interpolation parameters is used to identify the images that were acquired by the same camera model with similar configuration. According to this information, images are clustered in order to obtain a first coarse estimate of the intrisic camera parameters to be used in the general reconstruction algorithm.

Note that all these solutions do not take into consideration the quality of the processed images, which proves to be significant whenever the number of conjugate points decreases. In fact, as it is shown in Fig. 1, an image that has been edited and compressed several times presents a reduced number of matching keypoints/descriptors and the keypoint locations could be significantly altered (reducing the accuracy of estimation).

3. A FORENSIC PROCESSING AGE METRICS FOR STATIC IMAGES

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 [26], generalized Gaussian, laplacian+impulsive [27], and Cauchy [28]. In this work, we define the probability mass function (pmf) of the absolute values of quan-



(a) $N_c = 1$ QF=85. (b) $N_c = 2$ QF=81.

(c) $N_c = 3$ QF=77. (d) $N_c = 4$ QF=85.

Fig. 2. Probabilities $p_{0,1}(c)$ (blue values) and the fitted model $p^f_{0,1}(c)$ for the image 3 of dataset `fountain` compressed N_c times. Parameter QF is referred to the last quantization stage.

tized DCT coefficients c located at frequencies (u, v) as $p_{u,v}(c)$. According to the previous works on coefficient modellization, it is possible to approximate $p_{u,v}(c)$ as

$$p^f_{u,v}(c) = \Gamma e^{-\pi(c)}, \tag{2}$$

where $\pi(\cdot)$ is a polynomial of third 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.

Experimental results have showed that the fitting accuracy depends on the number of compression stage operated on the image. Fig. 2 shows that the deviation of the $p_{u,v}(c)$ from $p^f_{u,v}(c)$ increases with the number of coding operations. This fact has been highlighted in the multimedia forensic literature of the last years, and as a matter of fact, it suggests the possibility of parameterizing the number of operations applied on the image I_i via the fitting accuracy.

Since $p^f_{u,v}(c)$ and $p_{u,v}(c)$ are pmfs, their similarity (or the processing age a_i for the image I_i under analysis) can be measured using the Jensen-Shannon divergence, i.e.,

$$\begin{aligned} a_i \quad = D_{JS}(p||p^f) \quad &= \frac{1}{2}\sum_c p(c)\log_2\frac{p(c)}{p^f(c)} + \\ &\frac{1}{2}\sum_c p^f(c)\log_2\frac{p^f(c)}{p(c)} \end{aligned} \tag{3}$$

where indexes (u, v) have been omitted for the sake of clarity.

When referred to the data in Fig. 2, we have that the processing age values a_i are respectively 0.2314 (a), 0.2980 (b), 0.3848 (c), 1.6117 (d). Note that the corresponding PSNR values (in dB) with respect to the original uncompressed image (which is not available in a real case) are 40.61 (a), 39.44 (b), 38.02 (c), 37.52 (d). It is possible to conclude that the proposed no-reference metric permits ordering the processed image according to an objective quality evaluation. These values can be used to weight the similarity between images as it will described in the following section.

4. THE PROPOSED ALGORITHM

Frequently, SfM algorithms model the similarity $\sigma_{i,j}$ between couples of images (I_i, I_j) using the number of matching points, i.e., the cardinality $|S_{i,j}|$. The values $\sigma_{i,j}$ are then used to label the edges of a complete graph: by optimizing the graph, it is possible to find the optimal track. The proposed solution employs an iterative algorithm which operates on the matrices

$$M_{i,j} = \begin{cases} |S_{i,j}| & \forall\, j > i \\ 0 & otherwise \end{cases} \tag{4}$$

and

$$A_{i,j} = \frac{a_i + a_j}{2} \tag{5}$$

(which is normalized so that $\max_{i,j} A_{i,j} = 1$).
At first, the algorithm computes the matrix W as

$$W_{i,j} = \begin{cases} \frac{M_{i,j}}{A_{i,j}} & \text{if } M_{i,j} \geq .75\ M_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $M_{\max} = \max_{i,j} M_{i,j}$. At each iteration, the algorithm considers only the rows and columns of images already processed finding the couple of nodes (i, j) that maximizes $W_{i,j}$, includes these in the reconstruction (if not already present), removes the element (i, j) from matrix M, and then iterates recomputing W. In this way, the algorithm processes couples of images with many common points so that the first partial 3D models can be dense and accurately-aligned as much as possible; moreover, such accuracy contributes to minimizing estimation errors on both camera parameters and locations. Furthermore, by dividing matrix M with matrix A, images with the highest processing age are processed first (with respect to the same number of matches). The following section will show how this choice permits improving the reconstruction quality.

5. EXPERIMENTAL RESULTS

In our tests, we verified the performance of our similarity metric with some standard benchmark datasets with known ground-truth camera motion to quantitatively evaluate the reconstruction accuracy. To this purpose, we used the datasets related to the work [29]. We compared our results with values obtained by the standard Bundler's ordering method.

The processed dataset was generated as follows. Every image (which was initially available in lossless format) was compressed a random number of times (uniformly distributed between 1 and 10) using a JPEG coder. The adopted Quality Factor (QF) is obtained by computing different independent realizations of a Poisson distribution clamped between 50 and 100 and a mean of 80. This configuration was used since the generated QF values are very close to those frequently found on downloaded images. Lower QF values compromise the quality of images with modest bit rate saving.

In the analysis phase (generating a_i), every image is recompressed with $QF = 100$ (to enhance the effects of previous compressions), and then the proposed processing age metric is computed. In this computation we considered the coefficients located at frequency $(1, 0)$. For every dataset we performed this test 10 times and took the average results.

Table 1 reports the MSE values computed on the estimated camera parameters. The MSE of matrices R and t were almost always improved (improvement is more significant for datasets

Name	MSE(f)		MSE(R)		MSE(t)		RMS(PointCloud)	
	Bundler	Proposed	Bundler	Proposed	Bundler	Proposed	Bundler	Proposed
fountain-P11	283	277 (−2.3%)	0.278	0.355 (+28%)	11.1	4.05 (−63%)	0.147	0.112 (−24%)
Herz-Jesu-P8	846	1,376 (+63%)	0.0511	0.034 (−33%)	19.9	0.87 (−96%)	0.069	0.063 (+10%)
entry-P10	772	187 (−76%)	0.183	0.119 (−35%)	1.88	2.07 (+10%)		
castle-P19	1,775	1,378 (−22%)	0.328	0.241 (−27%)	5.32	3.69 (−31%)		
Herz-Jesu-P25	169	168 (−0.6%)	0.231	0.146 (−37%)	16.1	11.5 (−29%)	0.054	0.061 (+12%)
castle-P30	1.3e4	7.1e4 (+438%)	0.454	0.225 (−50%)	14.1	9.99 (−29%)		
Global mean	+67%		−26%		−40%		−0.7%	

Table 1. Accuracy of different parameters (focals f_i, orientations R_i, locations $\mathbf{t}_i$, and 3D points $\mathbf{P}_k$) for the proposed and the Bundler algorithms.

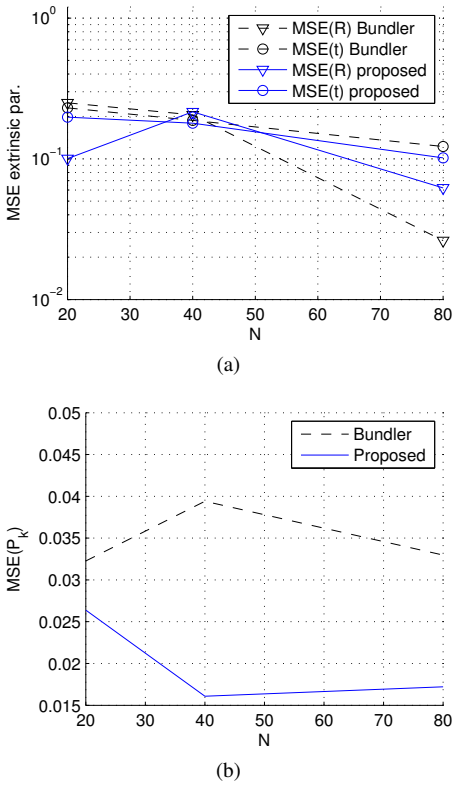


Fig. 3. Mean square error of the estimated orientation matrices $MSE(R)$, camera locations $MSE(t)$ (subplot a), and the point cloud $MSE(\mathbf{P}_k)$ (subplot b) for the proposed strategy and the Bundler approach.

`Herz-Jesu`). The MSE of the focal length, instead, is more variable (see the results for the dataset *castle-p30*, as an example). Other tests performed on this dataset seemed to have suffered from very high spikes of errors, too. Note that the fluctuations are not related to the dimension of the dataset; therefore, we are confident that the proposed algorithm can scale with no problems.

The last column shows the results obtained with the ICP (Iterative Closest Point) algorithm. For the analysis we used the ground truth high density meshes provided with some of the datasets. For faster calculations we subsampled 100.000 random points from them. It is possible to appreciate that in these cases the improvement is not so relevant.

Final tests were devoted to testing the proposed solution on real data (not synthetically generated). To this purpose, we applied the designed strategy to the `notredame` dataset [5] generating different subsets of images with different cardinalities. In this case, the reference model is provided by the dense point cloud generated by processing all the images in the dataset. Figure 3 reports the MSE values for the estimated orientation matrices R, the camera locations $\mathbf{t}$, and the point cloud $\mathbf{P}_k$ obtained with the proposed solution and with the standard Bundler approach on image subsets with different cardinalities. Data were generated averaging the results obtained from 10 different random subsamplings.

The results reported in Fig. 3(a) show that the proposed solution permits improving slightly the accuracy of camera parameters. The accuracy increment is more evident for the generated point cloud: the average MSE values obtained by the proposed solution are approximately 40 % lower than the values generated by the Bundler strategy. It is also possible to notice that the improvement in the accuracy of the extrinsic parameters increases as the number of images grows. Additional details are reported at [30].

6. CONCLUSIONS

The paper presented a new 3D reconstruction strategy for heterogeneous collections of images that employs a forensic quality metric to order images and build the 3D reconstruction track. The core strategy relies on processing first those images that were generated with a limited number of editings; this number is parameterized by a no-reference forensic processing age metric, which is computed from image pixels. The proposed solution permits improving the estimation accuracy of both the orientation and location coordinates in exchange for a little precision penalty in the focal length. Future work will be devoted in improving the reconstruction performance by including additional metrics in the ordering process.

7. REFERENCES

[1] S. Milani, "Robotic 3D Home Page," site: `http://www.dei.unipd.it/~sim1mil/ r3d`, 2014.

[2] S. Milani, "3D Cloud Vision Home Page," site: `http://www.dei.unipd.it/~sim1mil/ 3dcloudvision`, 2015.

[3] M. Valt, R. Salvatori, P. Plini, R. Salzano, M. Giusto, and M. Montagnoli, "Climate change: A new software to study the variations of snow images shot by web cam," in *Proc. of ISSW 2013*, oct 2013.

[4] Simone Milani, "Compression of multiple user photo galleries," *Image and Vision Computing, Special issue on*

Event-based Media Processing and Analysis, vol. 53, pp. 68 – 75, 2016, Available at: http://www.sciencedirect.com/science/article/pii/S0262885615001407 [Online].

[5] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *Int. J. Comput. Vision*, vol. 80, no. 2, pp. 189–210, Nov. 2008.

[6] R. Gherardi, M. Farenzena, and A. Fusiello, "Improving the efficiency of hierarchical structure-and-motion," in *Proc. of IEEE CVPR 2010*, June 2010, pp. 1594–1600.

[7] J.-M. Frahm, M. Pollefeys, B. Clipp, D. Gallup, R. Raguram, C. Wu, and C. Zach, "3D Reconstruction of Architectural Scenes from Uncalibrated Video Sequences," in *Proc. of ISPRS Workshop 3DARCH 2009*, Sept 2009.

[8] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H.-P. Seidel, "Markerless motion capture with unsynchronized moving cameras," in *Proc. of IEEE CVPR 2009*, June 2009, pp. 224–231.

[9] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[10] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, June 2008.

[12] R. A. Newcombe and A. J. Davison, "Live dense reconstruction with a single moving camera," in *Proc. of IEEE CVPR 2010*. June 2010, pp. 1498–1505, IEEE Computer Society.

[13] J. Stühmer, S. Gumhold, and D.Cremers, "Real-time dense geometry from a handheld camera," in *Proc. of 32nd DAGM Symposium*, Sept. 2010, pp. 11–20.

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

[15] C. Wu, "VisualSFM : A Visual Structure from Motion System," http://ccwu.me/vsfm/, Dec. 2014.

[16] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello, "Hierarchical structure-and-motion recovery from uncalibrated images," *Comput. Vis. Image Underst.*, vol. 140, no. C, pp. 127–143, Nov. 2015.

[17] B.Triggs, "Joint feature distributions for image correspondence," in *Proc. of ICCV 2001*, Vancouver, Canada, June 2001, vol. 2, pp. 201–208, IEEE.

[18] S. Milani, M. Fontana, P. Bestagini, and S. Tubaro, "Phylogenetic analysis of near-duplicate images using processing age metrics," in *Proc. of ICASSP 2016*, March 2016, pp. 2054–2058.

[19] A. Melloni, P. Bestagini, S. Milani, M. Tagliasacchi, A. Rocha, and S. Tubaro, "Image phylogeny through dissimilarity metrics fusion," in *Proc. of EUVIP 2014*, Dec 2014, pp. 1–6.

[20] B. Triggs, P. F. McLauchlan, R. Hartley, and A. W. Fitzgibbon, "Bundle adjustment ? a modern synthesis," in *Vision Algorithms: Theory and Practice*, vol. 1883 of *Lecture Notes in Computer Science*, pp. 298–372. Springer Berlin Heidelberg, 2000.

[21] S. Milani, "Three-dimensional reconstruction from heterogeneous video devices with camera-in-view information," in *Proc. of IEEE ICIP 2015*, Sept 2015, pp. 2050–2054.

[22] S. Gammeter, T. Quack, D. Tingdahl, and L. J. Van Gool, "Size Does Matter: Improving Object Recognition and 3D Reconstruction with Cross-Media Analysis of Image Clusters," in *Proc. of ECCV 2010*, Sept. 2010, pp. 734–747.

[23] D. Sarrut and S. Miguet, "Similarity measures for image registration," in *In First European Workshop on Content-Based Multimedia Indexing*, 1999, pp. 263–270.

[24] H. Kalinic, S. Loncaric, and B. Bijnens, "A novel image similarity measure for image registration," in *Proc. of ISPA 2011*, Sept 2011, pp. 195–199.

[25] Simone Milani and Enrico Tronca, "Improving three-dimensional reconstruction of buildings from web-harvested images using forensic clues," *Journal of Electronic Imaging*, vol. 26, no. 1, pp. 011009, 2016.

[26] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Process.*, vol. 9, no. 10, pp. 1661–1666, Oct. 2000.

[27] S. Milani, L. Celetto, and G.A. Mian, "An Accurate Low-Complexity Rate Control Algorithm Based on (ρ, E_q)-Domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 2, pp. 257–262, Feb. 2008.

[28] Y. Altunbasak and N. Kamaci, "An analysis of the DCT coefficient distribution with the H.264 video coder," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[29] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *Proc. of IEEE CVPR 2008*, June 2008, pp. 1–8.

[30] M. Lecci and S. Milani, "Supporting material for "3D Reconstruction from Web Harvested Images Using a Forensic Quality Metric" ," site: http://www.dei.unipd.it/∼sim1mil/3dcloudvision/parecon, 2017.