Walter Scheirer

Department of Computer Science and Engineering

College of Engineering

University of Notre Dame

Fitzpatrick Hall of Engineering,

Notre Dame, Indiana, 46556

P +1 574 631 2436

walter.scheirer@nd.edu

June 14th, 2018

**Dear T.IP Editors,**

We would like to submit our revised **manuscript** “Image Provenance Analysis at Scale” for consideration to *IEEE Transactions on Image Processing (T*.*IP)*.

**Paper title:** Image Provenance Analysis at Scale

**Name of the submitted file:** provenance\_v2.pdf

**Version:** 2

**Authors:** Daniel Moreira, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin Bowyer, Patrick Flynn, Anderson Rocha, and Walter J. Scheirer

**Contact information for corresponding author:**

Walter Scheirer

Phone +1 574 631 2436

walter.scheirer@nd.edu

We thank the Reviewers and the Editor for their important contributions towards improving this work as well as for their patience with this submission. The three reviews containing comments were detailed and insightful, and this has helped us tremendously in our revision process (the fourth reviewer was satisfied with the article as-is). Overall, we are encouraged by the positive response this work received during the first round of reviewing. In the current version, we have addressed all of the points raised by the reviewers. The comments have allowed us to produce a better manuscript, which we believe is now ready for publication in *IEEE Transactions on Image Processing.*

Major changes in this submission include:

1. Report of the latencies of the experimented solutions through their average query times (i.e., the average time spent for processing a single query over 1,000,000 images), in an effort to elucidate the efficiency and scalability of the proposed method.
2. Clarification that the present manuscript constitutes, to the best of our knowledge, the first effort to publicize the current state of the art in provenance analysis to the scientific community. We are now explicitly linking the proposed method to our participation in the NIST Nimble Challenge as the NDPURDUE team.
3. Improvements in the flow and presentation of the manuscript text, as suggested by the reviews.

In the following pages, we present explanations for each of the comments from the reviews of the first version of the manuscript. Note that we also plan on releasing our experimental code upon publication to facilitate the use of this new image provenance framework across existing and new problems in visual recognition. The newly introduced Reddit Photoshop Battles dataset will also be released at that time. If there is any question regarding this submission, please let us know.

Thanks in advance.

Sincerely yours,

The authors

|  |
| --- |
| **Associate Editor** |

**General Comment.** Based on the enclosed set of reviews, your manuscript requires a MAJOR REVISION (RQ). While the reviewers commend your work, a few common points of criticism need to be addressed including comparison to other state-of-the-art methods and clarification of efficiency and scalability of the proposed method.

**Answer**

*We thank the editor and the reviewers for diligently dealing with our submission and for the very constructive feedback. Besides addressing all the other reviewers’ minor observations, we tackled both points of major criticism (comparison to the state of the art and clarification of the performance of the proposed method). To make their identification easier, all the consequent modifications in the manuscript are currently highlighted in red, except the very minor ones (writing-related) to avoid cluttering.*

|  |
| --- |
| **Review #1** |

**General Comment.** This work proposed a provenance analysis pipeline that consists of the 10 image processing stages based on SURF features. This work presents a system for addressing the provenance analysis and demonstrates outperforming results in the experiments. There are some major issues for this paper that are listed as below.

**Answer**

*We thank the reviewer for recognizing the contributions of the paper. We address each of the comments raised in the review in what follows.*

**\*\*\* \*\*\* \*\*\***

**Comment 1.** This work comprises a large system by the 10 image processing stages which obstacles the readability of this paper and leaves the room for numerous handcraft parameters tuning. On the other hand, the convolution neural networks (CNN) based approaches conduct an end-to-end optimization that eliminates a huge amount of parameter tuning efforts in each stage. The work claims that the proposed method is a faster alternative than deep learning. This property should be clarified or demonstrated since there are emerging efficient CNN architectures (e.g. MobileNet, ShuffleNet, Distillation method).

**Answer**

*We understand that the 10-stage system might affect the readability of the paper, when considered as a whole. The division of the solution into two major steps (Provenance Image Filtering and Provenance Graph Construction), as presented in Section III, aims at facilitating a clear understanding. Indeed, although the second step may depend on the results of the first one, they conceptually consist of very distinct tasks, each one containing 5 of the 10 image processing stages. Such a division is also expressed through Figures 1 and 2.*

*We agree that a single CNN-based architecture for all 10 stages would reduce the need for handcrafted parameter tuning. However, the formulation of a joint training optimization over all of the tasks of the pipeline is not straightforward, and at present, very difficult to solve. The reason for such difficulty is mostly related to the Provenance Graph Construction stage. As a part of our current research efforts, we are investigating ways of engineering CNNs that are able to describe the pairwise relations between images that might share content, including the direction of the relation, but preliminary results are still incipient and need further investigation.*

*Nonetheless, we can still consider a CNN-based solution for the feature extraction task of the Image Filtering step. The table below summarizes experiments that we have conducted to verify the runtimes of ShuffleNet [1] (for which functional code was available) and SURF, including the usage of both GPU- and CPU-based implementations. In all experiments, we report timing numbers for the exact same set of 10,000 images from the Tiny ImageNet [2] dataset.*

|  |  |  |
| --- | --- | --- |
|  | *ShuffleNet (feedforward time)* | *SURF (detection and description time)* |
| *GPU* | *3 min 24 sec (using 2 GPUs)* | *39 sec (using 1 GPU)* |
| *CPU* | *6 min 44 sec (using 24 cores)* | *1 min 32 sec (using 5 cores)* |

*As one might observe, SURF is significantly faster than ShuffleNet for both configurations (either GPU or CPU), despite using fewer hardware resources. Such findings agree with the observations provided by Zagoruyko and Komodakis [3], where they report that SIFT descriptors on CPU – reportedly slower than SURF – are twice as fast as their CNN-based image patch matching solution running on GPU. To justify our feature extraction approach, we are now mentioning their findings in Section II (Related Work), which one may find in the beginning of the second column of page 3, in the current revised version of the manuscript.*

*[1] X. Zhang, X. Zhou, M. Lin, and J. Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In: Proceedings of IEEE CVPR. 2018*

*[2] Tiny ImageNet Visual Recognition Challenge. Available at:* [*https://tiny-imagenet.herokuapp.com/*](https://tiny-imagenet.herokuapp.com/)*. Accessed on June 10, 2018.*

*[3] S. Zagoruyko and N. Komodakis. Learning to Compare Image Patches via Convolutional Neural Networks. In: Proceedings of IEEE CVPR. 2015.*

**\*\*\* \*\*\* \*\*\***

**Comment 2.** This work claims the proposed method can be applied at scale. This property should be addressed or clarified explicitly in the experiments by the throughput or latency metrics.

**Answer**

*In response to this comment, we are now reporting the average time spent per query in the Provenance Image Filtering step, which expresses the latency of the system, as suggested by this review. Table I was updated accordingly, as well as some discussion added to Session V (Results).*

*As one can now observe, the best solution (IVFADC-DSURF-IF) spends, on average, 2.2 minutes to query one image among 1,000,000 samples and to retrieve 200 directly and transitively related images, when running on a 24-core CPU system operating at 2.4 GHz and eight TITAN Xp GPUs in a distributed file system environment. In addition, if the system is supposed to only retrieve the directly related images (as happens in a typical image retrieval solution), the IVFADC-DSURF alternative performs the same task in around 32 seconds (0.54 minutes) on the same hardware. All this information has been added to the manuscript.*

**\*\*\* \*\*\* \*\*\***

**Comment 3.** The ground truth of the Reddit Photoshop battle is automatically inferenced from the underlying comments. The quality of these inferenced results must be clarified or verified since the inferred result is the foundation of the whole analysis for this dataset.

**Answer** *As we now explain in Section IV.3, we verified all of the inferenced graphs by hand, taking care of removing cases with missing images linked to the comments, or with disconnected nodes. The result was 184 provenance graphs, which are assured to be connected, directed acyclic, and highly novel. Nonetheless, it is important to mention that, in some cases, linked images might share semantic content rather than visual content, posing a challenge to the currently proposed solution. Regardless of that, we decided to keep those cases as an effort to make the dataset as realistic and novel as possible. This has been mentioned in the last paragraph of Section V.B.*

**\*\*\* \*\*\* \*\*\***

**Comment 4.** The experimental result in NIST Nimble Challenge only compares the author’s previous work. The performance comparison of NIST Nimble Challenge (or Media Forensics Challenge as stated in CVPR workshop 2017) should compare the results from other research works under the same evaluation conditions.

**Answer**

*This is a very good suggestion. We now explain in the beginning of Section IV.A.1 that we were actually participants of the challenge mentioned by the reviewer, where we were enrolled as the NDPURDUE team and had the opportunity to submit the herein proposed solution. As reported by Fiscus et al. [1], in that challenge, we obtained the best results for end-to-end provenance image filtering and graph construction, as one may observe in slide 28. To the best of our knowledge, no other participants have made their solutions available to the scientific community as of yet.*

*With respect to the evaluation conditions, NC2017-Eval-Ver1 is the latest NIST evaluation provenance dataset, which was used to generate the results reported in [1]. Nevertheless, no complete provenance ground-truth is currently made available for this dataset, preventing us from reproducing the results reported in [1] (i.e., the evaluation was sequestered). As we have explained in the paper, through Section IV.A.1, that led us to adopt a protocol where NC2017-Dev1-Beta4 (whose ground-truth is available) is complemented with images from NC2017-Eval-Ver1 up to the point of having one million distractors. This is our effort to provide a reproducible protocol with respect to dataset availability.  
  
[1] J. Fiscus, H. Guan, Y. Lee, A. Yates, A. Delgado, D. Zhou, D. Joy, A. Pereira. The 2017 Nimble Challenge Evaluation: Results and Future Directions. In: CVPR Workshop on Media Forensics. Available at* [*https://www.nist.gov/sites/default/files/documents/2017/07/31/nist2017mediaforensicsworkshop\_20170726.pdf*](https://www.nist.gov/sites/default/files/documents/2017/07/31/nist2017mediaforensicsworkshop_20170726.pdf)*. Accessed on May 9, 2018.*

|  |
| --- |
| **Review #3** |

**Comment 1.** In Sec III-B 1) the author introduces two similarity metrics: GCM-based metric and MI-based metrics, each later paired with a graph construction algorithm. This indicates some kind of incompletion in my opinion. Theoretically these two metrics have their own pros and cons: GCM-based metric measures the geometric alignment and structure similarity of the matching object, while the MI-based metric measures the similarity of pixel distribution. From the experiment result it seems that the GCM metric does much better in the Edge overlap while MI metric does better in the image retrieval and is able to measure edge direction. Why not consider a similarity metric that combines the benefits of both metrics instead of evaluating them separately which yields suboptimal result in both? Even by the current evaluation method why can’t we look at all possible combination of graph construction algorithm and similarity metrics: Krustal-GCM, Krustal-MI, Cluster-GCM, Cluster-MI.

**Answer**

*This review comment is right when it says that the two metrics – number of geometrically-sound matched interest points, in GCM, and mutual information between homographically paired image patches, in MI – have their own pros and cons. However, as is depicted in Figure 2 and explained in Section III.B.1, the MI-based solution is built upon the results of the GCM-based one. While the GCM-based approach presents the drawback of delivering only symmetric image-pairwise adjacency matrices (this is explained in the bottom of page 6, highlighted in red for the reader’s convenience), the MI-based approach works by receiving those symmetric matrices as input and generating asymmetric adjacency matrices (this is explained in the beginning of page 7, highlighted in red for convenience). As a consequence, the GCM output gives no clues for guessing the directions of the image relationships, being only amenable to the application of Kruskal’s algorithm, which was designed to find the optimum undirected spanning tree within a graph. This is the reason why the we do not report the Cluster-GCM configuration, as questioned in this review.*

*In addition to the MI-based solution being obtained from the GCM one, according to what is explained in Section III.B.2, the Cluster-based solution explicitly relies on both the GCM- and MI-based adjacency matrices (we are conveniently highlighting the related text in red), meaning that, in fact, it benefits from both metrics (GCM and MI), exactly as questioned in the review. Therefore, it is not necessary to have configurations such as Cluster-GCM or as Cluster-MI, since Cluster uses both strategies.*

*Nonetheless, thanks to this review, we realized that the previous labels given to the experimental graph construction configurations were confusing, justifying the review’s doubts. For that reason, we changed them from GCM-SURF and GCM-MSER to Kruskal-SURF and Kruskal-MSER, respectively, as well as from MI-SURF and MI-MSER to Cluster-SURF and Cluster-MSER. With these modifications and the proper updates to the text of Sections IV and V (respectively Experimental Setup and Results), we believe the presentation of results is more clear and correct.*

**\*\*\* \*\*\* \*\*\***

**Comment 2.** The image search algorithm described in Sec. III-A 3) reads a bit confusing due to abuse of notation. Also some equation is questionable. For example, Equation (5) seems to mean sequentially finding the best match images based on the best matched features found by KNN. However, the iterative expression 5) only removes V\_i in the argmax expression. If you mean what I stated above shouldn’t you remove V0 – V\_(i-1) from argmax as well?

**Answer**

*The review is correct in the understanding of the image search step and, indeed, the iterative process expressed in equation 5 was confusing. For that reason, Sections III.A.2 and III.A.3 are now rewritten, with improvements to the mathematical notation, as well as the replacement of equation 5 by text, which we believe now expresses better the idea of K-NN feature-wise image voting.*

**\*\*\* \*\*\* \*\*\***

**Comment 3.** The iterative filtering is a quite expensive operation, which is usually avoided by retrieval algorithm considering the huge number of images considered. Even though the retrieval result looks better with the iterative filtering, we still need to carefully evaluate the gain in the context of additional cost of the iterative filtering, especially since you claim that you avoid the CNN based methods due to efficiency reason.

**Answer**

*The review is correct when they say that iterative filtering is an expensive operation. We are now reporting, in Table I, the average times spent by each solution for processing a single query over 1,000,000 images. According to the results, the best solution without iterative filtering (namely IVFADC-DSURF) takes around 32 seconds (0.54 minute) to retrieve 200 directly related images, with a recall at the top-50 most related images (R@50) equal to 0.882. If iterative filtering is added, though, (expressed by the IVFADC-DSURF-IF option), the system takes around 2.2 minutes to perform the same task, with an improvement in R@50 towards 0.907. As we are now discussing in the end of Section V.A, if time is not a strong constraint, the increase of 3% in recall and the assurance that not only the images directly related to the query will be retrieved, but also the transitively related ones, may justify the use of iterative filtering.*

*Last but not least, with respect to the performance of CNN-based solutions, we are now mentioning in Section II (Related Work) the work of Zagoruyko and Komodakis [1], where they report that SIFT descriptors on CPU – reportedly slower than SURF – are twice as fast as their CNN-based image patch matching solution running on GPU. That justifies our avoidance of CNN-based methods, together with the observation that CNN-based solutions directly apply to low-level image description, but do not straightforwardly cope with the need for indexing (and searching over) 1,000,000 images.*

*[1] S. Zagoruyko and N. Komodakis. Learning to Compare Image Patches via Convolutional Neural Networks. In: Proceedings of IEEE CVPR. 2015.*

**\*\*\* \*\*\* \*\*\***

**Comment 4.** Since you compare the proposed algorithm with Krustal algorithm in the experiment section, it’s better to include a paragraph in the method section to discuss the advantage of the cluster-based graph construction algorithm over traditional Krustal algorithm to justify the complexity.

**Answer**

*As might already be known, the main advantage of the proposed clustered solution over Kruskal’s algorithm is that the former provides directed provenance graphs, in opposition to the undirected optimum spanning trees returned by the latter. In order to make this aspect clear upfront, we added that information to the first paragraph of Section III.B.2 (inside Section III), as suggested by this review.*

**\*\*\* \*\*\* \*\*\***

**Comment 5.** The Introduction section gives a pretty good overview of related work in this field and I appreciate it. I think it’s better to move the contribution overview after the related work overview and clearly state the key problems addressed in this work that are different or not properly addressed in previous work.

**Answer**

*We have moved the contribution overview to the end of the related work section. We agree that this change helped to make the differences and contributions of this work more clearly stated.*

**\*\*\* \*\*\* \*\*\***

**Comment 6.** The large-scale infra proposed in Sec.III A (5) doesn’t read like a good fit in the algorithm context. Maybe consider moving it to the implementation details in the experiment section?

**Answer**

*We moved the details of the large-scale infrastructure to the experimental setup section and it indeed improved the readability of the paper.*

**\*\*\* \*\*\* \*\*\***

**Comment 7.** As stated above, retrieval algorithm needs to be evaluated in the context of efficiency. Can you add the result for comparison of retrieval time of different approaches being compared in the table? A more thorough evaluation should look at the recall constraint by a certain CPU budget. e.g. How good the algorithm can perform if we must return the result within a certain execution time. This might mean choosing a smaller K to make time for the iterative filtering.

**Answer**

*As mentioned above, we are now reporting, in the last column of Table I, the average time for processing a single query over 1,000,000 images and retrieving 200 related images. This gives the reader an idea of the latency of the experimental solutions, as well as to help put them into perspective.*

*In addition, some pairwise comparisons are useful for revealing how one solution builds upon the other. That is the case, for example, of IVFADC-SURF2k and IVFADC-SURF5k, whose only difference is the number of extracted SURF interest points (the k value mentioned in this review comment). As we discuss in the third paragraph of Section V.A, a reduction in the value of k from 5,000 to 2,000 interest points accelerates the retrieval from 33 to approximately 10 seconds, at the expense of reducing the recall at the top-50 most related images (R@50) from 0.876 to 0.713. Using a k value of 2,000 would certainly leave more time for iterative filtering (IF), as mentioned by the review, but such a significant reduction in recall impairs IF by design, since IF departs from the first set of results to refine the retrieved images and find the transitively related examples. For that reason, we opt, in the paper, to apply IF over the best one-tier solution, namely IVFADC-DSURF, which leads to IVFADC-DSURF-IF, the best configuration tested thus far.*

**\*\*\* \*\*\* \*\*\***

**Comment 8.** A more fundamental question about the evaluation metric EO and VEO: how stable is the ground truth label for the edge of the provenance graph? It seems to me that how to connect the graph from the relevant images is quite subjective and might not have unique answer if you ask different human labeler to draw the graph. e.g. in Fig. 6, it also makes sense if we directly connect the bottom left image to the root. If the graph construction is very flexible and there’s no unique solution, how important is it to predict the EO exactly is questionable.

**Answer**

*The observation that “graph construction is very flexible and there’s no unique solution” is correct in some cases. However, we are operating in a more constrained evaluation scenario with ground-truth, as described by NIST. By definition and in accordance to NIST [1], the ground-truth label is a consequence of the actual composition history of the images. For that reason there is only one correct provenance graph for each case, which is used as the reference for computing the VO, EO and VEO metrics. As we explain in the second paragraph of Section IV.B, VO, EO and VEO measure the overlap of a candidate graph with respect to the ground-truth graph. Given that the groundtruth graph is always unique (including its sets of nodes and of edges), the assessment of these metrics is stable.*

*In the mentioned case of Figure 4 (formerly Figure 6), the unique ground-truth graph is the one depicted there, which was inferred from the Reddit users’ comments. A solution that connects the root to the bottom left image, as suggested by the reviewer, would be penalized for (1) providing an extra edge that is not present in the ground-truth (from the root to the leaf at hand), and (2) missing the correct edge between the leaf and its parent.*

*[1] National Institute of Standards and Technology. Nimble Challenge 2017 Evaluation. https://www.nist.gov/itl/iad/mig/*

*nimble-challenge-2017-evaluation*

|  |
| --- |
| **Review #4** |

**General Comment.** The authors present an end-to-end processing pipeline for image provenance analysis, which works at real-world scale. Here are some points that I would like to see in the paper:

**Answer**

*We thank the reviewer for the comments. Each one is addressed in what follows.*

**\*\*\* \*\*\* \*\*\***

**Comment 1.** Comparison with Image Modification Detection.

**Answer**

*This recommendation is appropriate when image forensics is considered in a broader way. However, in this paper, we concentrate on the more specific problem of provenance analysis. Different from image manipulation detection, which is focused on the analysis of a single image and aims at deciding if an image was manipulated in some way or not (including variations of localizing the modified image regions – our team has explored that idea in prior work [1]), provenance analysis is focused on discovering the relations between images, in terms of shared content and probable history of generation. Thinking in this direction, we do not see a way of comparing the solutions, since they refer to different problems, with distinct metrics and datasets. However, provenance analysis might certainly benefit from image modification detection in the future, in the sense of using manipulation detectors to improve the definition of the direction of provenance (from pristine to manipulated images, for instance). We will certainly keep track of this idea for future work.*

*[1] J. Brogan, P. Bestagini, A. Bharati, A. Pinto, D. Moreira, K. Bowyer, P. Flynn, A. Rocha, and W. Scheirer, “Spotting the difference: Context retrieval and analysis for improved forgery detection and localization,” in IEEE ICIP, 2017.*

**\*\*\* \*\*\* \*\*\***

**Comment 2.** Have better comparison with state-of-the-art approaches. So far, the authors compare with [16, 38] (the same group research).

**Answer**

*As is explained in the related work section, the problem of provenance analysis was only recently posed by the American National Institute of Standards and Technology (NIST). Indeed, in 2017, NIST brought the issue into light for the scientific community through the Media Forensics Nimble Challenge, which was presented at a workshop held at IEEE CVPR [1]. We apologize for not mentioning that fact before. As we are now explaining in the beginning of Section IV.A.1, we were participants of that challenge, enrolled as the NDPURDUE team. According to the results reported in [1], with the herein proposed solution, we obtained state-of-the-art results for end-to-end provenance image filtering and graph construction, as one may observe in slide 28. In addition, to the best of our knowledge, no other participants have made their solutions available to the scientific community, justifying the apparent lack of comparison to other approaches.  
  
[1] J. Fiscus, H. Guan, Y. Lee, A. Yates, A. Delgado, D. Zhou, D. Joy, A. Pereira. The 2017 Nimble Challenge Evaluation: Results and Future Directions. In: CVPR Workshop on Media Forensics. Available at* [*https://www.nist.gov/sites/default/files/documents/2017/07/31/nist2017mediaforensicsworkshop\_20170726.pdf*](https://www.nist.gov/sites/default/files/documents/2017/07/31/nist2017mediaforensicsworkshop_20170726.pdf)*. Accessed on May 9, 2018.*

**\*\*\* \*\*\* \*\*\***

**Comment 3.** The authors should report the experiment on different image modification including transformation and lighting, illumination.

**Answer**

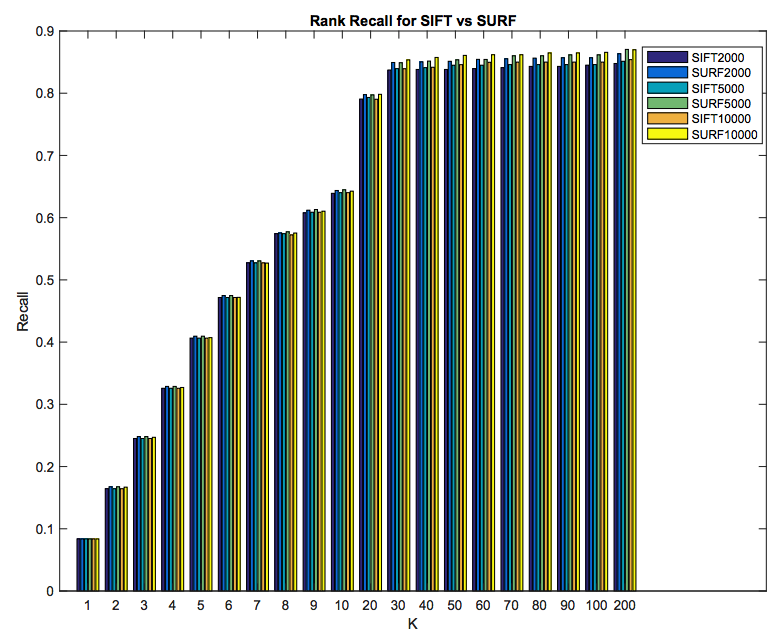
*This comment is related to the issue tackled in the answer of the first comment. The currently available datasets for image manipulation are mostly focused on the analysis of single images, which makes them inappropriate for evaluating provenance analysis. The available datasets of provenance analysis, in turn, do contain various image modifications (including transformations in lighting and illumination), but the manipulations have not been explicitly categorized, preventing us from reporting results for specific manipulations. The organization of the datasets are this way because they were created with the initial aim of benchmarking the relations between images, in terms of generation history, rather than manipulation detection. For that reason, ground-truth and metrics for image manipulation detection are out of scope.*

**\*\*\* \*\*\* \*\*\***

**Comment 4.** Besides SURF, I would suggest the authors evaluate the system on SIFT, Dense SIFT.

**Answer**

*In preliminary experiments, we conducted tests with different interest point descriptors, including SIFT. Nonetheless, in accordance with the findings of previous work from the literature [1], the usage of one technique or the other did not significantly affect the image representation capacity. Figure 1 shows a comparison of the impact of different detectors and descriptors over the recall of images from the NIST dataset.*

**

*As one might observe, the application of either SIFT or SURF, with different interest point densities (from 2,000 to 10,000 interest points), does not result in significant improvements of the recall, for different rank sizes (defined by K). Due to our aim of providing a solution that is fast and works at scale, we opted for reporting only the experiments with 5,000 SURF interest points (which is also faster than SIFT). In addition, the employment of 5,000 interest points constituted one of the attempted improvements over our previous solution in [2], which, besides a simpler image indexing technique, had used less SURF interest points (from 500 to 2,000 points).*

*For such reasons, we are keeping the manuscript as it is, with respect to the usage of other interest point detectors and descriptors. We thank the reviewer, though, for bringing the issue into discussion.*

*[1] T. Tuytelaars, K. Mikolajczyk, Local invariant feature detectors: a survey, Found. Trends Comput. Graph. Vis. 3 (3) (2008) 177–280.*

*[2] A. Pinto, D. Moreira, A. Bharati, J. Brogan, K. Bowyer, P. Flynn, W. Scheirer, and A. Rocha. Provenance filtering for multimedia phylogeny, in IEEE ICIP, 2017*