Strong Geometry Consistency for Large Scale Partial-Duplicate Image Search

Junqiang Wang
School of Computer Science,
Nanjing University of Science and Technology
Nanjing, P.R. China
free.wjq@gmail.com

Jinhui Tang
School of Computer Science,
Nanjing University of Science and Technology
Nanjing, P.R. China
jinhuitang@mail.njust.edu.cn

Yu-Gang Jiang
School of Computer Science,
Fudan University
Shanghai, China
ygj@fudan.edu.cn

ABSTRACT

The state-of-the-art partial-duplicate image search systems reply heavily on the match of local features like SIFT. Independently matching local features across two images ignores the overall geometry structure and therefore may incur many false matches. To reduce such matches, several geometry verification methods have been proposed. This paper introduces a new geometry verification method named as Strong Geometry Consistency (SGC), which uses the orientation, scale and location information of the local feature points to accurately and quickly remove the false matches. We also propose a simple scale weighting (SW) strategy, which gives feature points with larger scales greater weights, based on the intuition that a larger-scale feature point tends to be more robust for image search as it occupies a larger area of an image. Extensive experiments performed on three popular datasets show that SGC significantly outperforms state-of-the-art geometry verification methods, and SW can further boost the performance with marginal additional computation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Image search, partial-duplicate images, geometry verification, scale weighting

1. INTRODUCTION

Given a query image and a large image dataset, the goal of this work is to accurately and efficiently identify partial-duplicate images, which may contain the same scenes or objects captured from different angles, or originally the same image altered manually in scale, contrast, etc. Figure 1 shows some example images. Automatic solutions for finding partial-duplicates have many potential applications like copyright violation detection, image annotation, web-scale image search, and so on.

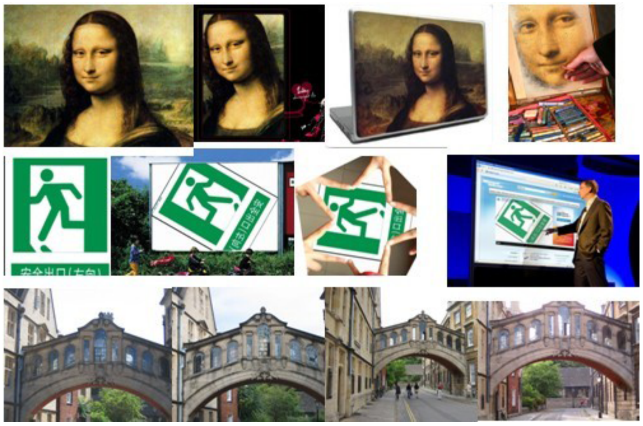


Figure 1: Examples of partial-duplicate images.

In the past decade, the performance of large scale partial-duplicate image search has been significantly boosted by using local invariant features [5, 4, 6, 7] and the well-known Bag-of-Features (a.k.a. Bag-of-Visual-Words) representation [11, 8]. See a subset of representative works in [2, 9, 10, 18, 17, 14, 19, 15, 3, 12, 16, 13]. A typical step in many of these approaches is to match similar local features between each pair of images. Instead of exhaustively evaluating pairwise similarity of feature points, a more efficient way is to assume that feature points falling into the same visual word are matched, which has been widely used. Through evaluating the number of matched patches, the similarity (and the likelihood of having duplicate image regions) between two images can be estimated. A major limitation of pure local feature matching is that it ignores the spatial locations of the local features, which have been proved to be very useful in improving search accuracy [9].

One conventional approach to alleviate the problem is to use RANSAC [7], which is able to perform careful geometry verification of the matched local features, but is computationally slow. To more efficiently verify the consistency of the geometry structures of the matched features, weak geometry consistency (WGC) was proposed [2]. One assumption in WGC is that the matched features should have similar scale and rotation changes. Therefore it uses the peak values of scale and orientation changes histograms to measure

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502166.

image similarity. An enhanced WGC (E-WGC) method was further proposed in [17], which holds that though the reverse transformation in scale and rotation, corresponding image regions will also share uniform or similar translation. The consistency of predicted translations is used as a measure of geometry consistency.

A different approach to tackle a similar problem, called Geometric Coding (GC), was proposed in [18]. GC encodes local feature points' spatial relationships, and iteratively removes matches that cause the largest inconsistency in geometry structure. Obviously, the computational complexity of GC is $O\left(n^2\right)$, which is much slower compared with WGC and E-WGC.

In this paper, we propose a new geometry verification measure called Strong Geometry Consistency (SGC), which is efficient to compute and robust to scale changes, rotation, slight 3D viewpoint changes, etc. The idea of SGC is motivated by WGC and E-WGC, with a key difference that SGC not only considers the magnitude of displacement but also fully uses translation in both x and y axes. In addition, SGC divides the matched features into several groups based on scale and orientation changes, which makes the evaluation of geometry structure consistency more strict and efficient. A simple scale weighting (SW) method is also proposed, which gives larger scale feature points greater weights, as features in larger scales cover larger image areas and are therefore expected to be more robust. In the following we first introduce our proposed SGC and SW and then verify their effectiveness using several benchmark datasets.

2. THE PROPOSED APPROACH

2.1 Strong Geometry Consistency

We start by briefly explaining the idea of WGC and E-WGC. Given two matched local feature points $p\left(x_p, y_p\right)$ and $q\left(x_q, y_q\right)$ where x and y represents the spatial coordinates of the two points, we can estimate the transformation from p to q as follows:

$$\begin{bmatrix} x_q \\ y_q \end{bmatrix} = s \times \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \times \begin{bmatrix} x_p \\ y_p \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \qquad (1)$$

where t_x, t_y indicate the translation in the horizontal and vertical directions respectively, s is a scaling parameter and θ is a rotation parameter. The two parameters can be approximated by [2]:

$$s = s_q/s_p, \theta = \theta_q - \theta_p \qquad (2)$$

where s_p and s_q are the scales of the local feature points p and q respectively, indicated by the local feature detectors like the Differences of Gaussian (DoG) [5]. Similarly θ_p and θ_q are the dominant orientations of the two points, which can also be found in the outputs of the local feature detectors. The authors of [9] have found that using the simple transformation model in Eq. (1) is just slightly worse than the much more expensive RANSAC approach.

In WGC, s and θ for all the matched feature pairs are collected, which form two histograms. The peak values of the two histograms are used to directly measure or slightly adjust the values of image similarities, because the peak values can indicate the consistency of scale changes and rotation between the two set of matched features. E-WGC is directly motivated by WGC. Based on Eq. (1), we can easily derive the translation (t_x, t_y):

$$\begin{bmatrix} t_x \\ t_y \end{bmatrix} = \begin{bmatrix} x_q \\ y_q \end{bmatrix} - s \times \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \times \begin{bmatrix} x_p \\ y_p \end{bmatrix} \qquad (3)$$

In E-WGC, a histogram of the L-2 norm of the translation is used instead of the histograms of scale and orientation differences, and the peak L-2 norm of all the matched pairs is used to adjust the similarity of two images.

Reducing a two dimensional translation vector to a single L-2 norm value in E-WGC makes it easier to compute the peak values, but apparently also loses important information. In SGC, we consider the original 2-d vector. Specifically, we first divide the matched feature pairs into k groups softly, based on their rotation angles. Within each group, we first find the pair which has the most similarly translated pairs, its translation is defined as the dominant translation of the group, and then remove feature pairs which have a significantly different translation with the dominant translation. After that, we use the number of remaining pairs in the largest group to represent the similarity of two images:

$$Sim(i, j) = N_{max} \qquad (4)$$

where N_{max} represents the cardinality of the largest group.

In addition to the benefit from directly using the 2-d translation vector instead of its L-2 norm, SGC is developed based on the assumption that there is a considerable portion of the image regions sharing similar scale changes, rotation and translation. In other words, the feature points with similar translation must have similar scale changes and rotation, while WGC and E-WGC do not explicitly consider this. This allows SGC to perform strict and accurate geometry consistency verification.

It is worth noting that in SGC the computational complexity of removing false feature pairs is $O\left(n^2\right)$, where n is the number of the matched feature pairs in each group. However, as feature pairs are divided into several much smaller groups by rotated angles, the speed is greatly improved since n becomes much smaller. We will experimentally evaluate the speed of SGC later.

Error Threshold of the Predicted Translation: While removing false matches in each feature pair group, we need a threshold of the predicted translation (t_x, t_y) to decide which pairs are false matches. Using a fixed threshold is a straightforward choice but is not very reliable as there can be many variations of duplicated image regions, which may lead to different levels of errors.

In this work, we propose a simple and efficient way to automatically and adaptively predict the most suitable threshold. To that end we need to make several error assumptions. First, the error of spatial locations of the detected feature points is (empirically) assumed to be within 5 pixels. Such an error exists because there may be a certain degree of offset in the detector outputs for the same interest region when image transformations exist. We also assume that the error of θ is within θ_{error} degrees. With these assumptions and Eq. (3), we can have the following error threshold of the translation:

$$\left\| \begin{bmatrix} t_{x_{error}} \\ t_{y_{error}} \end{bmatrix} \right\| \leq s \times \left\| \begin{bmatrix} cos\theta_{error} & -sin\theta_{error} \\ sin\theta_{error} & cos\theta_{error} \end{bmatrix} \times \begin{bmatrix} x_p + 5 \\ y_p + 5 \end{bmatrix} \right\| \qquad (5)$$

With Eq. (5), we have a better way of threshold estimation, which allows more accurate filtering of false matches, as will be validated in the experiments.

2.2 Scale Weighting

A key assumption of scale weighting is that a feature point (a region more rigorously), like SIFT, in a larger scale tends to be more robust for matching, since it occupies a larger image region and contains more information. One example is that a leg of a table may be matched to many images with similar objects like chairs, but a larger object like the head of a movie star can be more rarely wrongly matched. With this intuition in mind, we give the feature points in larger scales greater weights during the process of ejecting false feature pairs, where the weight is computed as :

$$SW = (s_1 \times s_2)^n \tag{6}$$

we experiment with different n from 0 to 1 on the 1M image dataset. In the experiments we observe that the optimal value of n is around 0.75. However, we simply use 1 for n as it is only marginally worse than the optimal choice, but is much more efficient to compute Eq. (6).

3. EXPERIMENTS

In this section, we experimentally validate the effectiveness of our proposed approach and compare with several existing geometric verification methods. All the experiments are performed on a regular desktop computer with a single 2.53GHz Xeon CPU and 16GB RAM. Throughout the experiments we use DoG as the local detector and SIFT as the descriptor [5]. To generate initial local feature matches for geometry-based filtering, we adopt a codebook of 200k visual words.

3.1 Datasets and Evaluation Measure

We adopt three popular benchmark datasets:

The first one is Oxford5K, which was introduced in [9]. It contains 11 different landmarks in Oxford and 55 query images. Together with a large number of background images that do not contain any of the landmarks, this dataset has totally 5,062 high resolution images.

The second dataset, called Holiday, was originally used in [2], which consists mainly of personal holiday photos. The dataset has 500 image groups, each of which depicts a different scene. In total there are 1,491 images. Following existing works, the first image of each scene group is used as the query image and the correct search results should be the remaining images within the same group.

The third dataset used in our experiments was collected by the authors of [18], called GCdup. It contains 1,104 partial-duplicate images in 33 groups, all downloaded from the Web. 100 images are selected as the query images in our experiments.

To make the task more challenging and realistic, we further adopt the MIRflickr 1 million dataset [1] as distracters. By gradually adding images from this dataset to the above 3 datasets, we can better understand the robustness and scalability of different approaches.

For performance metric, we compute the widely adopted Average Precision (AP) for each query and use mean AP (mAP) to measure the overall performance of multiple queries.

3.2 Alternative Methods for Comparison

To show the effectiveness of SGC and SW, we conduct comprehensive comparative studies with the following approaches: (1) Standard BOF [8], using a codebook of 1 million visual words; (2) BOF+WGC [2], where initial search results of BOF are refined using orientation and scale of matched feature pairs as described earlier in Section 2; (3) BOF+E-WGC [17]; (4) Re-ranking based on RANSAC [9], which facilitates good affine-transformation verification. Because the computational cost of RANSAC is very high, we only re-rank the top 300 candidate images. (5) GC [18].

Table 1: Average time cost of various geometric verification methods on the Holiday dataset.

Method	Without HE	With HE
WGC	1.16s	0.34s
E-WGC	1.63s	0.45s
RANSAC	5.58s	1.49s
GC	3.87s	0.97s
SGC	1.72s	0.49s
SGC+SW	1.74s	0.50s

All these approaches are based on our own implementations, which may not be exactly the same with the original papers. However, several of the approaches have been evaluated on one or more of the three datasets in corresponding references. We have checked into those works and found our obtained results very similar.

3.3 Results and Discussion

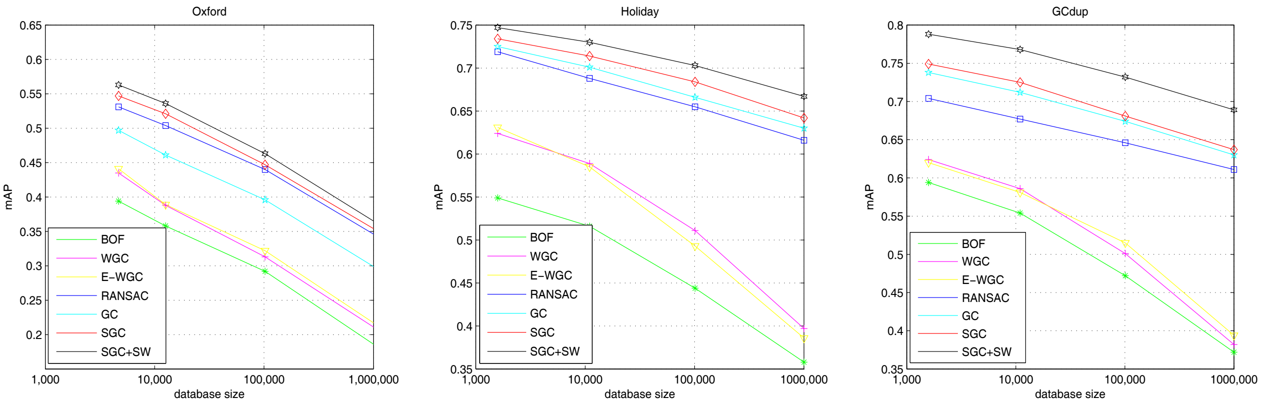
Figure 2 gives the performance of all the approaches. Overall we can see that the proposed SGC significantly outperforms all the existing methods including BoF, WGC, E-WGC, RANSAC, and GC. BoF is the worst since it does not have any geometry verification capability and therefore involves lots of false matches. WGC and E-WGC perform similar in our experiments, which are all not as good as stronger verification algorithms RANSAC and GC. Interestingly, we find that the similar scale weighting strategy SW is very useful, leading to significant improvements over all the three datasets. On the GCdup dataset, using SW (SGC+SW) improves SGC alone by as high as 7%. This confirms the fact that feature points with larger scales are much more important for partial-duplicate image search.

To validate the effect of our adaptive error threshold estimation method in SGC, we also experiment with several fixed error thresholds on the GCdup dataset, and found the best fixed error threshold is around 80, with which the performance is still around 2 percent lower than that using our adaptive error threshold. During the computation of the adaptive threshold, we set the error of θ as 7 degrees for Holiday and GCdup, and 15 degrees for Oxford5K where 3D viewpoint changes are more popular. Although this parameter can lead to different final performance, the difference is fairly minor and does not affect the general conclusion observed from Figure 2.

Table 1 further summarizes the average time cost of all the geometric verification methods for one image query on the Holiday dataset with 1 million image distracters. As can be seen, our SGC has similar speed to E-WGC and is much faster than RANSAC and GC. The fact that it is faster than GC is mainly because it divides feature points into several small groups. Dividing feature points, as one of the core ideas in SGC, also leads to performance gain. Further adding SW does not incur too much extra computation.

4. CONCLUSION

We have proposed a new geometry verification method



(a) Oxford5K

(b) Holiday

(c) GCdup

Figure 2: Performance (mAP) of various approaches on three datasets. The left end of each curve gives results on the original dataset without adding background images (distracters), while the other results are obtained on enlarged datasets with different numbers of distracters.

called SGC, which integrates the orientation, scale and location of the local feature points to quickly filter false matches from initial visual codebook based matching. On three popular datasets, we have shown that the proposed SGC significantly outperforms existing alternative solutions with minor additional computation. In addition, we also proposed a simple scale weighting strategy, which produces surprising good performance gains.

5. ACKNOWLEDGMENTS

This work was partially supported by the Natural Science Foundation of China (NSFC) under Grant 61103059, the National Basic Research Program of China (973 Program) under Grant 2012CB316304, the Program for New Century Excellent Talents in University under grant NCET-12-0632, The Natural Science Foundation of Jiangsu Province under Grant BK2012033 and BK2011700.

6. REFERENCES

[1] M. J. Huiskes, B. Thomee, , and M. Lew. New trends and ideas in visual concept detection. In *ACM MIR*, 2010.

[2] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[3] H. Jegou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *CVPR*, 2007.

[4] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998.

[5] D. Lowe. Distinctive image features from scale-invariant key points. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[7] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[8] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[11] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[12] J. Tang, R. Hong, S. Yan, G.-J. Qi and T.-S. Chua. Inferring semantic concepts from community contributed images and noisy tags. In *ACM Multimedia*, 2009.

[13] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi and R. Jain. Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. *ACM TIST*, 2011.

[14] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *CVPR*, 2009.

[15] W.Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian. Spatial coding for large scale partial-duplicate web image search. In *ACM Multimedia*, 2010.

[16] D.-Q. Zhang and S.-F. Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *ACM Multimedia*, 2004.

[17] W. Zhao, X. Wu, and C.-W. Ngo. On the annotation of web videos by efficient near-duplicate search. *IEEE TMM*, 12(5):448–461, 2010.

[18] W. Zhou, H. Li, Y. Lu, and Q. Tian. Large scale image search with geometric coding. In *ACM Multimedia*, 2011.

[19] W. Zhou, H. Li, Y. Lu, and Q. Tian. Sift match verification by geometric coding for large-scale partial-duplicate web image search. *ACM TOMCCAP*, 9(1), 2013.