# LOCAL GEOMETRIC CONSISTENCY CONSTRAINT FOR IMAGE RETRIEVAL

*Hongtao Xie[1,2], Ke Gao[1], Yongdong Zhang[1], Jintao Li[1]*

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[2]Graduate University of Chinese Academy of Sciences, Beijing, 100049, China

## ABSTRACT

In state-of-the-art image retrieval systems, an image is represented by bag-of-features (BOF). As BOF representation discards geometric relationships among local features, exploiting geometric constraints as post-processing procedure has been shown to greatly improve retrieval precision. However, full geometric constraints are computationally expensive and weak geometric constraints have limited range of applications. To efficiently handle common transformations and deformations, we present a novel local geometric consistency constraint (LGC) method. It utilizes the local similarity characteristic of deformations, and measures the pairwise geometric similarity of matches between two sets of local features. Besides, we propose a new method to accurately calculate the transformation matrix between two matched features, with the information provided by their local neighbors. Experiments performed on famous datasets show the excellent performance of our method.

***Index Terms***— Image retrieval, Geometric Consistency Constraints

## 1. INTRODUCTION

Given a query image, the goal of image retrieval is to find out its similar images which contain the same scene or object in a large corpus of images. It plays an important role in many applications, such as copyright protection, image registration and redundant image filtering [1][2][3] .

State-of-the-art image retrieval systems [2][3][4][5][6] build on the bag-of-features (BOF) [7]representation. The local descriptors [8] are quantized into visual words. An image is represented by the frequency histogram of visual words obtained by assigning each descriptor of the image to the closest visual word. Measuring the similarity between two images is then performed by bin-to-bin matching of their histograms. By adopting an inverted file index of visual words, these systems can match two images efficiently [7]. While favorable for simplicity and scalability, BOF representation ignores the geometric relationships among local features [5][6]. So it brings about false feature matches and reduces the accuracy in image retrieval.

To address the issue caused by BOF quantization, geometric consistency constraints are applied to eliminate false matches. In [2][3][9], full geometric consistency constraints (FGC) have been proposed by adding a re-ranking step that computes an affine transformation between the query image and a short list of candidate images. The affine transformation is usually estimated by Hough scheme [8] and RANSAC–based method [10]. FGC can verify geometric consistency effectively, but it is computationally expensive and not appropriate when a large amount of images need to be verified. To exploit the geometrical information without explicitly estimating the affine transformation, weak geometric consistency constraints (WGC) [6] and enhancement of WGC (E-WGC) [1] are proposed. WGC and E-WGC assume that similar images undergo uniform transformations (global scale, rotation and translation changes), so the characteristic scales, the dominant orientations and positions are consistently modified for all the features. Thus, WGC and E-WGC construct three histograms, representing the scale, orientation and translation consistency between the matches in two images respectively. The peaks of these histograms are used to refine search accuracy [1][6]. WGC and E-WGC are intuitive and simple to implement. However, they have strong assumptions and can only work under uniform transformations between the query image and candidate images. If viewpoint changes and nonrigid deformations take place, they are useless.

In this article we present a novel local geometric consistency constraint (LGC) method. It is based on the local similarity characteristic of deformations [11], and measures the pairwise geometric similarity of matches between two sets of features. To accurately calculate the transformation matrix between two matched features, we propose a new method which utilizes the information provided by their local neighbors. Compared to the state-of-the-art methods, LGC can not only deal with uniform transforms, but also handle viewpoint changes and nonrigid deformations. Besides, it is simple and runs efficiently. Experiments performed on three annotated datasets show the effectiveness of LGC, as well as its efficiency.

This paper is organized as follows. Section 2 presents the strategy applied for LGC. The experiment results are presented in section 3 and section 4 concludes this paper.

## 2. LOCAL GEOMETRIC CONSISTENCY CONSTRAINT

Geometric consistency constraints are useful post-processing procedures of image retrieval to re-rank the candidate images. As full geometric constraints [2][3][9] are costly and weak geometric constraints [1][6] have limited range of applications, we need a new geometric consistency verification method which can not only handle common transformations and deformations, but also run efficiently.

### 2.1. LGC

Full geometric consistency constraints and weak geometric consistency constraints are all based on the global geometric information of the query image and the candidate images. But the global information cannot reflect different local changes. Thus, we use the local geometric information to exploit geometric consistency constraint.

By observing the images under affine changes and non-rigid deformations, we obtain an important phenomenon. For an image and its deformed counterpart, all feature correspondences do not form global compactness in their geometric similarity owing to deformations, but deformed parts are locally connected by some mediating parts. That is all the matches do not form uniform transformations in geometric similarity (such as scale, orientation and translation). But locally, for a pair of matches, the deformations of them are nearly similar. Thus, a connectedness criterion exists and the deformations have local similarity [11].

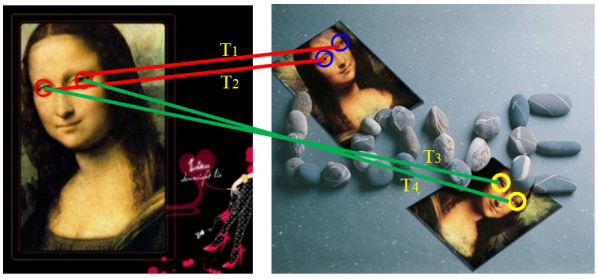



**Fig.1.** The illustration of local similarity of deformations

As illustrated in Fig.1, the red and green lines represent feature correspondences between two similar images. For the transformations $T_1$, $T_2$, $T_3$ and $T_4$, they have the following properties: $T_1$ is approximately equal to $T_2$, and $T_3$ is approximately equal to $T_4$. But $T_1$ and $T_2$ are very different from $T_3$ and $T_4$. So the similar transformations are locally adjacent, reflecting the same visual objects. Based on this characteristic, we propose the LGC method to measure the similarity of these images.

We first define the geometric consistency measure function. For two points $p(x_p, y_p)$ and $q(x_q, y_q)$ and a match $c = ((x_p, y_p),(x_q, y_q), T)$, the transformation matrix $T$ from $p$ to $q$ can be calculated by the method proposed in section 2.2. Then for two matches $c_i = ((x_i, y_i),(x_i', y_i'), T_i)$ and $c_j = ((x_j, y_j),(x_j', y_j'), T_j)$, the pairwise geometric dissimilarity between $c_i$ and $c_j$ is defined by:

$$d_{pgc}(c_i, c_j) = \frac{1}{2}\left(\left|X_j' - T_i X_j\right| + \left|X_i' - T_j X_i\right|\right), \quad (1)$$

where $X_k = [x_k, y_k]^t, X_k' = [x_k', y_k']^t, k = i, j$, and $|\cdot|$ denotes the Euclidean distance. $d_{pgc}(c_i, c_j)$ will be small if $T_i$ and $T_j$ are similar to each other. Exploiting the homographies of two matches, this pairwise geometric dissimilarity measure provides a discriminative basis for our LGC method.

LGC executes as follows:

1) For all the matches returned by BOF, $d_{pgc}(c_i, c_j)$ between $c_i = ((x_i, y_i),(x_i', y_i'), T_i)$ and $c_j = ((x_j, y_j),(x_j', y_j'), T_j)$ is calculated, only if they satisfy $d(c_i, c_j) < \delta_D$.
$$d(c_i, c_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad (2)$$
where $(x_i, y_i)$ and $(x_j, y_j)$ are the coordinates of the two query points. $\delta_D$ is a predefined threshold, which reflects the local similarity of deformations. In experiment, $\delta_D$ is set to 40.
2) A histogram $h^{pgc}$, referring to the pairwise geometric consistency, is constructed on the values of $d_{pgc}(c_i, c_j)$. For robustness, linear interpolation is adopted in building histogram. Then its peak is used to re-rank the candidate images similarity as WGC [6].

Intuitively, for a pair of similar images, $h^{pgc}$ has an obvious peak and the peak is located at or near the first bin of the histogram. For a pair of dissimilar images, as the matches are distributed in disorder, $h^{pgc}$ does not have an obvious peak.

### 2.2. Transformation Matrix Calculation

For LGC algorithm, the calculation of transformation matrix $T$ is of significant importance, as it provides the basic data. So we need to calculate $T$ accurately.

Given two matched points $p(x_p, y_p)$ and $q(x_q, y_q)$, WGC [6] and E-WGC [1]estimate the transformation $T$ from $p$ to $q$ as:

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

In (3), there are four parameters: the scale factor $s$, the rotation angle $\theta$ and the translation components $t_x$ and $t_y$. In WGC and E-WGC, $s$ and $\theta$ are simply derived from the information of local features:

$$s = 2^{(s_q - s_p)} \text{ and } \theta = \theta_q - \theta_p, \quad (4)$$

where $s_p$ and $s_q$ are the characteristic scales of points $p$ and $q$, $\theta_p$ and $\theta_q$ are the dominant orientations of these two points. In this way, $T$ can be estimated easily. But the calculation is not accurate, for the following two reasons:

- The characteristic scales and dominant orientations of local features cannot be precisely confirmed by the local feature detectors. In fact, these values are quantized into constants in local feature detection [8].

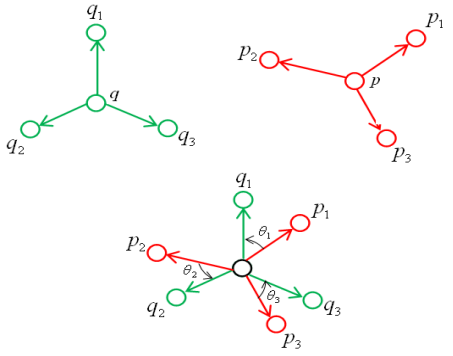
- Some local features have more than one dominant orientation [8], which causes confusion in computing $\theta$.

To accurately estimate the values of $s$ and $\theta$, we propose a new method which utilizes the information of the local neighbors around $p$ and $q$. We call it transformation estimation with local information (TELI).

TELI executes as follows:

1) For points $p$ and $q$, we find their $k$ (=10) nearest neighbors $P_N = (p_1, p_2, ..., p_k)$ and $Q_N = (q_1, q_2, ..., q_k)$ in the query image and candidate image respectively. Then, we get the matched neighbors of these two points:
$$M_N(p,q) = \left\{ (p_i, q_j) \mid p_i \in P_N \wedge q_j \in Q_N \wedge V_{p_i} = V_{q_j} \right\}.$$
$V_m$ is the visual word of feature $m$.
2) For the candidate match $(p,q)$, it will be eliminated if $|M_N(p,q)| < t$. $|M_N(p,q)|$ is the size of $M_N(p,q)$ and $t$ is set to 5.
3) If $|M_N(p,q)| \geq t$, we can obtain two support sets for $s$ and $\theta$, respectively:
$$s_N = \left\{ (s_{q_j} - s_{p_i}) \mid (p_i, q_j) \in M_N(p,q) \right\},$$
$$\theta_N = \left\{ (\theta_{q_j} - \theta_{p_i}) \mid (p_i, q_j) \in M_N(p,q) \right\}.$$
Then $s$ and $\theta$ in (3) are calculated as:
$$\log s = \lambda \times (s_q - s_p) + (1-\lambda) \times ave(s_N),$$
$$\theta = \lambda \times (\theta_q - \theta_p) + (1-\lambda) \times ave(\theta_N).$$

Where $ave(E)$ is the average value of the elements in set $E$, and $\lambda$ controls the relative weight and is set to 0.6.When $p$ or $q$ has several dominant orientations, we will take the largest one. Fig.2 shows the calculation of rotation angle $\theta$ from $p$ to $q$.




$$\theta = \lambda \times (\theta_q - \theta_p) + (1-\lambda) \times ave(\theta_1, \theta_2, \theta_3)$$

**Fig.2.** The calculation of rotation angle $\theta$ from $p$ to $q$.

With the information of nearest neighbors, we can estimate the transformation matrix between two matched points much more accurately than relying on them alone.

## 3. EXPERIMENTAL RESULTS

In this section, we evaluate LGC method for image retrieval on three annotated datasets: Ukbench dataset [4], Oxford5k dataset [2] and Toys dataset [13]. These sets are often used to evaluate image retrieval systems. Some typical examples of these sets are shown in Fig.3. Note that significant viewpoint changes and nonrigid deformations take place in Ukbench set and Toys set respectively, which makes the task of image search much more challenging. Our experiment environment is: Intel Xeon quad-core E5506 2.13GHz with 24GB memory.

In our experiment, local features are obtained by the Hessian-Affine [12] detector and the SIFT descriptor [8]. We build an image search system upon the scheme proposed by [6]. Hamming embedding (HE, 24-bit hamming code) is applied to filter out false matches. Visual words are learned from independent dataset and the size of visual vocabulary is 20000. We experiment with different sizes of visual vocabulary, and find the 20000 vocabulary to give the best overall performance. The performance metric applied in our experiment is mean average precision (mAP), as used in [2][6]. For each query image we calculate its precision-recall curve, from which we obtain its average precision and then take the mean value over all queries.

For the effectiveness evaluation of LGC, we compare the performance of the following approaches: 1) BOF [7], the standard BOF method; 2) HE+WGC [6], the search result of BOF is refined by HE and WGC; 3) HE+FGC[2], the search result of BOF is optimized by HE and FGC; 4) HE+s-LGC: the WGC in 2) is replaced by simplified LGC (without TELI); 5) HE+LGC: the WGC in 2) is replaced by LGC (with TELI). As FGC is time-consuming, we apply it to a short list of the top 150 candidate images.

**Table 1:** Search accuracy for five methods on Ukbench, Oxford5k, and Toys datasets.

| | Ukbench | Oxford5k | Toys |
|---|---|---|---|
| BOF | 0.744 | 0.343 | 0.442 |
| HE+WGC | 0.842 | 0.607 | 0.445 |
| HE+FGC | 0.887 | 0.69 | 0.507 |
| HE+s-LGC | 0.873 | 0.631 | 0.653 |
| HE+LGC | 0.881 | 0.684 | 0.71 |

Table 1 compares the above five approaches, leading to four major observations:

(1) Compared to BOF, HE+WGC can improve the values of mAP. This is because that BOF quantization ignores geometric relationships among visual words and induces many false matches. It indicates the importance of post-processing after the search based on BOF.
(2) LGC significantly improves the search precision, even without TELI. Compared to BOF, HE+s-LGC achieves 17.3%, 83.9% and 47.7% improvements in mAP on the three sets respectively, and HE+s-LGC is better than HE+WGC on these sets.
(3) With TELI, the search accuracy gets further improvement. The reason is that TELI can not only eliminate false matches, but also calculate the transformation matrix much more accurately.

(4) HE+LGC achieves almost the same performance as HE+FGC on the Ukbench and Oxford5k sets. On the Toys set, HE+LGC far exceeds HE+FGC. This because nonrigid deformations take place in this set. But FGC is based on the global geometric changes between the query image and the candidate images. As LGC measures the geometric consistency locally, it can effectively deal with these challenges.

From these comparisons, we can conclude that LGC can deal with uniform transforms (Oxford5k set), viewpoint changes (Ukbench set) and nonrigid deformations (Toys dataset). Fig.3 shows the matching results of the images in these sets. We can see that many false matches are removed by LGC. It demonstrates the effectiveness of LGC.

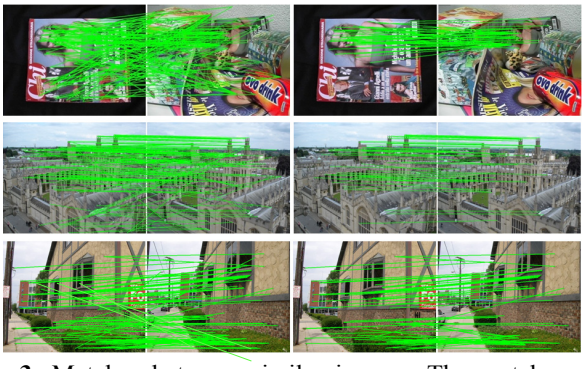


**Fig.3.** Matches between similar images. The matches are highlighted as green lines. Left: matches obtained by BOF. Right: matches obtained by HE+LGC. The images are selected from Toys, Oxford 5k, and Ukbench sets respectively.

Fig.4 shows the average query time per image for these methods. As the time cost for FGC is too large (more than 200s), we do not draw it in Fig.4. We can see that the time performance of LGC is comparable to WGC. Compared to WGC, LGC has to calculate the local pairwise geometric similarity additionally. But due to the introduction of TELI, many false matches are discarded, which reduces the computations of LGC. Besides, only one histogram is generated in LGC. So LGC run efficiently.

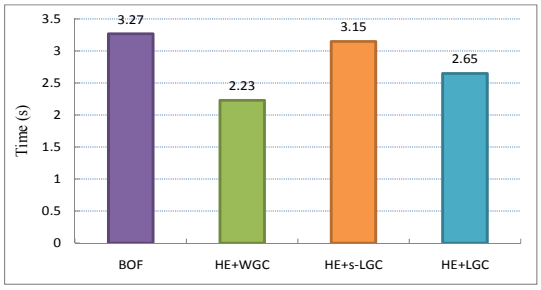



**Fig.4.** Query time per image for four methods.

## 4. CONCLUSIONS

We introduce LGC to improve the accuracy of image search. It exploits the local similarity characteristic of deformations, and measures the pairwise geometric similarity of matches between two sets of features. Compared to the state-of-the-art methods, LGC can not only deal with common transformations and nonrigid deformations, but also has excellent time performance. Besides, we propose TELI to accurately calculate the transformation matrix between two matched features. TELI can further improve the accuracy of image retrieval. Experiments performed on famous datasets illustrate the effectiveness and efficiency of our method.

## 5. ACKNOWLEDGEMENTS

This work is supported by the National Basic Research Program of China (973 Program, 2007CB311100); National Nature Science Foundation of China (61003163); National High Technology and Research Development Program of China (863 Program, 2009AA01A403); Co-building Program of Beijing Municipal Education Commission.

## 6. REFERENCES

[1] W.-L. Zhao, X. Wu and C.-W. Ngo, On the Annotation of Web Videos by Efficient Near-duplicate Search, IEEE. Trans. on Multimedia, 12(5), pp. 448–461, 2010.

[2] J. Philbin, O. Chum, et al, Object retrieval with large vocabularies and fast spatial matching, CVPR, 2007.

[3] M. Perdoch, O. Chum and J. Matas, Efficient Representation of Local Geometry for Large Scale Object Retrieval. CVPR, 2009.

[4] D. Nister and H. Stewenius, Scalable recognition with a vocabulary tree, CVPR, 2006.

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

[6] H. Jegou, M. Douze and C. Schmid, Improving bag-of-features for large scale image search, IJCV, vol. 87, no. 3, pp. 316–336, 2010.

[7] J. Sivic and A. Zisserman, Efficient visual search of videos cast as text retrieval, PAMI, vol. 31, no. 4, pp. 591– 606, Feb. 2009.

[8] D. Lowe, Distinctive image features from scale-invariant keypoints, IJCV, vol. 60, pp. 91–110, 2004

[9] O. Chum, M. Perdoch, et al, Geometric min-hashing finding a (thick) needle in a haystack, CVPR, 2009.

[10] O. Chum, J. Matas, et al, Enhancing RANSAC by generalized model optimization. ACCV, 2004.

[11] B. Fischer and J. M. Buhmann, Path-based clustering for grouping of smooth curves and texture segmentation. PAMI, 25(4):513–518, 2003.

[12] K. Mikolajczyk, Binaries for affine covariant region descriptors, http://www.robots.ox.ac.uk/vgg/research/affine/, 2007.

[13] V. Ferrari, T. Tuytelaars and L.V. Gool, Simultaneous Object Recognition and Segmentation by Image Exploration, ECCV, 2004.