Image Retrieval Based on Anisotropic Scaling and Shearing Invariant Geometric Coherence

Xiaomeng Wu
NTT Communication Science Laboratories
Kanagawa, Japan
wu.xiaomeng@lab.ntt.co.jp

Kunio Kashino
NTT Communication Science Laboratories
Kanagawa, Japan
kashino.kunio@lab.ntt.co.jp

***Abstract*—Imposing a spatial coherence constraint on image matching is becoming a necessity for local feature based object retrieval. We tackle the affine invariance problem of the prior spatial coherence model and propose a novel approach for geometrically stable image retrieval. Compared with related studies focusing simply on translation, rotation, and isotropic scaling, our approach can deal with more significant transformations including anisotropic scaling and shearing. Our contribution consists of revisiting the first-order affine adaptation approach and extending its application to represent the geometric coherence of a second-order local feature structure. We comprehensively evaluated our approach using Flickr Logos 32, Holiday, and Oxford Buildings benchmarks. Extensive experimentation and comparisons with state-of-the-art spatial coherence models demonstrate the superiority of our approach in image retrieval tasks.**

***Keywords*-feature extraction; geometry; image retrieval**

I. Introduction

The Bag-Of-Words (BOW) model based on local features has been shown to be successful in object retrieval. Many approaches have been proposed for imposing a spatial coherence constraint on images to enhance the discriminative power of BOW. For example, the Locally Optimized RANdom SAmple Consensus (LO-RANSAC) [1] repeatedly and randomly selects the inter-image correspondences of local features and computes the parameters of a geometric model fitting the sample. It subsequently finds all inliers to the model and evaluates the quality of the parameters as a barometer of spatial coherence. LO-RANSAC achieves state-of-the-art discriminative power, but is computationally expensive due to the iterative affine adaptation.

Among non-iterative solutions, Kalantidis et al. [2] and Yang et al. [3] explored the higher-order intra-image co-occurrence of local features. Both approaches are faster but less discriminative than LO-RANSAC due to the exclusion of a geometric coherence constraint. Zhang et al. [4] describe the long-range spatial layout of local features by computing a Hough transform in the Euclidean space. It is invariant to translation but achieves limited robustness as regards rotation and scaling. Wu et al. [5] measured the spatial conherence by projecting the local features inside each maximally stable extremal region along Cartesian coordinate axes. The approach achieves scale invariance but remains sensitive to rotation. Liu et al. [6] explored the second-order spatial structure of local features by using the k-nearest neighbor and embedded the relative distance and the relative principal angle between them into an image representation to obey a larger variety of affine invariance. The same level of invariance was also achieved in earlier studies of image recognition [7], [8].

Liu's approach [6] and its related studies [7], [8] are invariant to translation, rotation, and isotropic scaling. The images constituting the input to a computer vision system are however subject to perspective distortions. Of the above geometry-based approaches, LO-RANSAC is the only one that is invariant to the full 6-degree of freedom (DOF) affine transformations. In this paper, we propose a novel approach for geometrically stable image retrieval. The main contributions include: 1. the proposal of a geometric coherence constraint based on Hessian-based affine adaptation that is fully invariant to the 6-DOF transformations; 2. the design of a prior spatial coherence model that adapts the proposed geometric coherence constraint, achieves a higher efficiency than RANSAC, and achieves a higher retrieval discernment than other related studies. The rest of the paper is organized as follows. We revisit Hessian-based affine adaptation and develop our proposal of a geometric coherence constraint in Sect. II. Section III describes the adaptation of our approach to content-based image indexing and matching. We then present our experiments in Sect. IV and discuss future work in Sect. V.

II. Geometric Coherence Constraint

A. Hessian-Based Affine Adaptation

Lindeberg [9] has proposed a methodology for iteratively adapting the shape of a smoothing kernel to the region near an image point. Provided that this iteration converges, the consequent fixed point will be invariant affine transformations. For an image I_L, let $\mathbf{x}_L = (x_L, y_L)^T$ be an image point in I_L. Introducing an affine transformation

$$\xi_R = A\xi_L \tag{1}$$

where A is a 2×2 matrix, we can define a transformed image I_R as

$$I_L(\mathbf{x}_L + \xi_L) = I_R(\mathbf{x}_R + \xi_R) \tag{2}$$

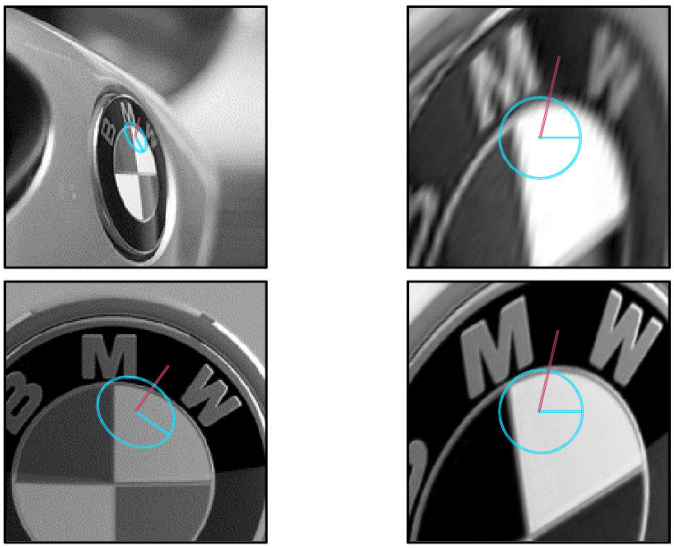
Equation 2 indicates the correspondence between each pixel $\mathbf{x}_L + \xi_L$ in I_L and each pixel $\mathbf{x}_R + \xi_R$ in I_R. Given two arbitrary images that are related as regards an unknown transformation

A due to a certain viewpoint change, if we can estimate A, then the two images can be successfully matched regardless of the viewpoint change. Lindeberg [9] demonstrated that given a correspondence between two image points $\mathbf{x}_L$ and $\mathbf{x}_R$, A can be estimated from measurements of the affine-adapted second-moment matrices M_L and M_R. Mikolajczyk and Schmid [10] continued with this formulation and showed that $M^{1/2}$ can transform the original anisotropic regions around $\mathbf{x}_L$ and $\mathbf{x}_R$ into two isotropic regions that are related through a rotation matrix $\mathcal{R}$. The eigenvalues and eigenvectors of M characterize the curvature and shape of the ellipsoid known as an affine region (AR). The isotropic regions can be thought of as a normalized reference (NR). A can then be estimated by:

$$A = A_R^{-1} A_L \tag{3}$$

$$A_{(\cdot)} = \mathcal{R}_{(\cdot)} M_{(\cdot)}^{1/2} \tag{4}$$

Mikolajczyk et al. [10] also proposed an iterative algorithm for estimating M and subsequently A. $\mathcal{R}$ can be recovered using gradient methods.



(a) Initial images $I(\xi)$. (b) Normalized images.

Figure 1. Affine normalization.

Figure 1 shows an example of affine normalization. Figure 1a corresponds to the initial images I_L and I_R and Fig. 1b corresponds to the normalized images $I_\varphi^{(L)}$ and $I_\varphi^{(R)}$. The ARs and their corresponding NRs are highlighted in blue. Although the viewpoints of the initial images vary significantly, the NRs are almost identical. Any feature computed from the NRs, usually known as a local feature, becomes invariant to the transformation A. Given $u(\mathbf{x})$ as such a feature descriptor and $\mathbf{x}$ as the image point, we denote the correspondence as

$$u(\mathbf{x}_L) = u(\mathbf{x}_R) \tag{5}$$

Motivated by the analogy to the 1-gram model, we may call this equality a first-order local feature coherence. In Sect. II-B, we extend the application of this first-order coherence to model the images' second-order geometric coherence.

B. Second-Order Geometric Coherence

Given a local feature coherence between the image points $\mathbf{x}_L$ and $\mathbf{x}_R$ detected from two images satisfying Eq. 5, Eq. 2 holds provided that the position vectors ξ_L and ξ_R are related according to Eq. 1. We subsequently consider whether another local feature coherence exists, given that Eq. 5 and Eq. 1 are satisfied, between the image points $\mathbf{y}_L = \mathbf{x}_L + \xi_L$ and $\mathbf{y}_R = \mathbf{x}_R + \xi_R$ such that

$$u(\mathbf{y}_L) = u(\mathbf{y}_R) \tag{6}$$

Equation 6 naturally holds if the ARs of $\mathbf{y}_L$ and $\mathbf{y}_R$ are inside those of $\mathbf{x}_L$ and $\mathbf{x}_R$. Otherwise, the equality depends on certain circumstances. Provided that 1. we can temporarily ignore non-affine transformations, e.g. illumination changes, 2. both $\mathbf{x}$ and $\mathbf{y}$ are inside the same object, and 3. we can confine our attention to near-planar and rigid objects, then Eq. 6 should hold and there should be coherence between $\mathbf{y}_L$ and $\mathbf{y}_R$. Needless to say, we assume that I_L and I_R include the same object.

We now invert the problem and assume that we have two pairs of image points $(\mathbf{x}_L, \mathbf{y}_L)$ and $(\mathbf{x}_R, \mathbf{y}_R)$ that satisfy Eqs. 5 and 6. Let $\xi_L = \mathbf{y}_L - \mathbf{x}_L$ and $\xi_R = \mathbf{y}_R - \mathbf{x}_R$ be the position vectors heading from $\mathbf{x}$ to $\mathbf{y}$. The normalized position vectors ξ_φ can be computed by projecting ξ_L and ξ_R onto the normalized space:

$$\xi_\varphi^{(\cdot)} = \mathcal{R}_{(\cdot)} M_{(\cdot)}^{1/2} \xi_{(\cdot)} \tag{7}$$

where $(\cdot)$ denotes L or R. If it is observed that

$$\xi_\varphi^L = \xi_\varphi^R \tag{8}$$

then $(\mathbf{x}_L, \mathbf{y}_L)$ and $(\mathbf{x}_R, \mathbf{y}_R)$ provide evidence for the belief that I_L and I_R include the same object under a certain transformation. Given the collection of all such evidence detected across I_L and I_R, a similarity can be formulated by measuring the global geometric coherence between the two images. This similarity is naturally more discriminative than the local feature coherence alone because a co-occurrence constraint is imposed by Eqs. 5 and 6 and a geometric constraint is imposed by Eq. 8. It is also highly robust as regards viewpoint changes because the constraint is invariant to the transformation A as discussed above.

Figure 1 shows an example where the central point $\mathbf{x}$ is highlighed in blue and the satellite point $\mathbf{y}$ in pink. The position vectors heading from $\mathbf{x}$ to $\mathbf{y}$ show a noticable declination in the initial images, but become almost identical after being projected onto the normalized space. We adapt this constraint to image representation and matching in Sect. III.

III. Image Retrieval

A. Indexing

Given an image I, a 2-tuple $\mathbf{t} = (\mathbf{x}, \mathbf{y})$ is a pair of points $\mathbf{x} \in I$ and $\mathbf{y} \in I$. Given a position vector heading from $\mathbf{x}$ to

$\mathbf{y}$ of $\xi(\mathbf{t}) = \mathbf{y} - \mathbf{x}$, a normalized position vector $\xi_\varphi(\mathbf{t})$ can be computed by projecting $\xi(\mathbf{t})$:

$$\xi_\varphi(\mathbf{t}) = \mathcal{R}(\mathbf{x}) M(\mathbf{x})^{1/2} \xi(\mathbf{t}) \tag{9}$$

The appearance and geometric characteristics of $\mathbf{t}$ can be represented by

$$h(\mathbf{t}) = \langle u(\mathbf{x}), u(\mathbf{y}), \xi_\varphi(\mathbf{t}) \rangle \tag{10}$$

where $u(\cdot)$ is a feature descriptor. Given two images I_L and I_R, every geometrically coherent correspondence can be found by thresholding the similarity between each $h(\mathbf{t}_L)$ and $h(\mathbf{t}_R)$. Inspired by the success of descriptor quantization, we consider the simplest visual vocabulary and Hough transform to assign certain visual and geometric terms to a tuple. We quantize $u(\cdot)$ into a visual term $\hat{u}(\cdot)$ using a visual vocabulary constructed with an approximated k-means. We also transform ξ_φ from Cartesian coordinates to log-polar coordinates $(\rho_\varphi, \alpha_\varphi)$ with ρ_φ being the log radius and α_φ being the log-polar angle. ρ_φ is further transformed into two bins by a threshold $\epsilon_\rho = \log \sqrt{\det M(\mathbf{x})}$, which is the AR scale of $\mathbf{x}$. α_φ is transformed into four bins by an equal division of $[0, 2\pi)$. We thus have an asymmetric visual phrase

$$\hat{f}(\mathbf{t}) = \langle \hat{u}(\mathbf{x}), \hat{u}(\mathbf{y}), \hat{\rho}_\varphi(\mathbf{t}), \hat{\alpha}_\varphi(\mathbf{t}) \rangle \tag{11}$$

describing the co-occurrence and geometry of the tuple. Depending on the parameterization, quantization usually incurs significant information loss [11]. Since ξ_φ has only two dimensions, we tap into its descriptor space by preserving ξ_φ in the tuple's representation:

$$\hat{h}(\mathbf{t}) = \langle \hat{f}(\mathbf{t}), \xi_\varphi(\mathbf{t}) \rangle \tag{12}$$

Given two tuples $\mathbf{t}_L$ and $\mathbf{t}_R$, the correspondence can be determined by imposing $\hat{f}(\mathbf{t}_L) = \hat{f}(\mathbf{t}_R)$ on them and thresholding the similarity between $\xi_\varphi(\mathbf{t}_L)$ and $\xi_\varphi(\mathbf{t}_R)$. This enables an efficient search with an inverted index in which each key corresponds to a visual phrase $\hat{f}(\mathbf{t})$ and each mapped value corresponds to a pair consisting of image ID I and $\xi_\varphi(\mathbf{t})$.

An image is regarded as a set of tuples. In theory, the total number of tuples is quadratic. In practice, we can restrict the set of tuples to that of the nearest neighbors of each local feature in the image space [6], [3]. Given a local feature $\mathbf{x}$, let its k-nearest neighbors be $\mathcal{N}_k(\mathbf{x})$. The set of tuples in the image I thus contains all pairs of $\mathbf{x} \in I$ with their neighbors:

$$N = \{\mathbf{t} \in I^2, \mathbf{t} = (\mathbf{x}, \mathbf{y}) : \mathbf{y} \in \mathcal{N}_k(\mathbf{x})\} \tag{13}$$

B. Matching

Given two images I_L and I_R, a set of candidate correspondences is obtained based on the visual phrase:

$$C_{L,R} = \{(\mathbf{t}_L, \mathbf{t}_R) \in N_L \times N_R : \hat{f}(\mathbf{t}_L) = \hat{f}(\mathbf{t}_R)\} \tag{14}$$

It follows that each correspondence should contribute to the similarity score according to how far apart $\xi_\varphi(\mathbf{t}_L)$ and $\xi_\varphi(\mathbf{t}_R)$ are. We define this contribution using a kernel function $\kappa(\xi_\varphi(\mathbf{t}_L), \xi_\varphi(\mathbf{t}_R))$, which gives rise to a similarity between I_L and I_R:

$$S(I_L, I_R) = \frac{\sum_{(\mathbf{t}_L, \mathbf{t}_R) \in C_{L,R}} \kappa(\xi_\varphi(\mathbf{t}_L), \xi_\varphi(\mathbf{t}_R))}{\Pi} \tag{15}$$

where $\Pi = \parallel N_L \parallel + \parallel N_R \parallel$ is a penalty function. It is more reasonable to choose a kernel related to Euclidean distance since ξ_φ is basically a position vector. We define κ as a radial basis function (RBF) kernel with a free parameter σ. Π penalizes the images with a very large number of features that cause confusion between unrelated images.

Note that Eq. 15 does not take the inverse document frequency (IDF) into account. This indirectly avoids the computation and storage cost, but the actual motivation lies in the fact that IDF is less helpful for geometry-based matching. IDF was designed to reduce the negative effect of confusing local features, e.g. those deriving from finely-textured patterns. In our approach, the visual phrase describes both the co-occurrence and the geometry of semi-local regions and so is highly discriminative and rarely creates confusion. We show evidence of this insight in Sect. IV-C.

IV. Experimentation

A. Setting

For our evaluation, we use three datasets: Flickr Logos 32 (FL32) [12], Holiday (HD) [13], and Oxford Buildings (OB) [14], which are compared in Table I. We employ the same Hessian-based region detector [10] for all datasets to extract local features. We compare our approach in an image retrieval scenario with the BOW and other spatial coherence models including multi-scale Delaunay triangulation (MSDT) [2], spatial co-occurrence kernel (SCK) [3], and Liu's approach [6]. Table II qualitatively compares these approaches. In this table, $\surd$ indicates calling into account or being invariant, and $\times$ the reverse. Order indicates the number of elements in each tuple. Isotropic and anisotropic indicate the corresponding types of scaling. Note that MSDT uses a graph model instead of k-nearest neighbor (k-NN) for neighborhood extraction, which differs from the other spatial coherence models. The other approaches, e.g. query expansion [15], Hamming embedding [13], and soft assignment [16], are not tested but are compatible with our approach. We measure the performance using mean average precision (MAP) and mean precision at top-4 (MP@4) [17]. For MSDT, the partition size is varied from 0.1 to 1 and the overlap ratio is varied from 0 to 0.9; for SCK, Liu's approach, and our approach, the parameter k used in k-NN is varied from 10 to 100.

B. Parameter Examination

In our approach, there are three parameters that influence the retrieval performance. They are the k used in k-NN, the free parameter σ of the RBF kernel, and the penalty function Π. We tune k in Sect. IV-C. We tested the MAP with various σ values, where a larger value corresponds to greater robustness as regards noises but less discriminative power and vice versa. The best MAP stabilized within $\sigma \in [4, 6]$ for all datasets.

Table I
DATASET COMPARISON.

Dataset	**FL32**	**HD**	**OB**
Category	Logo	Scenery	Building
Num. of Queries	960	500	55
Num. of Images	4.3K	1.5K	5.1K
Num. of Clusters	1M	0.2M	1M
Descriptor	Root SIFT	SIFT	Root SIFT

Table II
APPROACH COMPARISON.

Approach	**BOW**	**MSDT**	**SCK**	**Liu**	**Ours**
Co-occur.	×	√	√	√	√
Order		3	2	2	2
Geometry	×	×	×	√	√
Translation				√	√
Rotation				√	√
Isotropic				√	√
Anisotropic				×	√
Shearing				×	√

We chose $\sigma = 5$ for all subsequent experiments. We also compared five functions handling Π including a constant 1, two linear functions, and two quadratic functions related to $\| N_L \|$ and $\| N_R \|$. These choices were inspired by the similarity functions defined in information retrieval [18]. Since our approach is based on a second-order structure, it is superficially more reasonable to choose quadratic functions in theory. However, in practice the best MAP was stabilized with linear functions. Quadratic functions tend to be overly punitive because of the high discriminative power of the visual phrase defined in Eq. 11. We therefore chose $\| N_L \| + \| N_R \|$ for all subsequent experiments.

C. Comparison

Table III compares the best accuracies of various approaches. In general, spatial coherence models are superior to BOW with the exception being MSDT. The poor performance of MSDT may be because the graph model has a lower capacity for neighborhood association than k-NN and the third-order co-occurrence constraint is too sensitive to feature description errors. In contrast, SCK, Liu's approach, and our approach exhibited greatly improved performance gain compared with BOW. Figure 2 compares the relationship between the MAP and the k used in k-NN. Our approach obtains a higher MAP than SCK and Liu's approach for the same k in all cases. A larger number k usually leads to a higher MAP for all these approaches. A larger k allows the model to capture the spatial characteristics of larger objects, but may also cause more confusion between unrelated objects. The degradation of robustness has a negative impact on retrieval when the visual vocabulary is small, as reported by Liu [6]. In contrast, we used a 1M-cluster vocabulary in our approach and so avoided the impact of false responses. Hence the curves in Fig. 2 became monotonic.

Table III
BEST ACCURACY COMPARISON (%).

	FL32		**HD**	**OB**	
	MAP	**MP@4**	**MAP**	**MAP**	**MP@4**
BOW	54.3	79.8	54.7	70.9	94.1
MSDT [2]	54.8	81.1	49.5	70.1	94.1
SCK [3]	63.4	87.5	63.0	72.8	95.5
Liu [6]	65.3	89.5	66.2	73.6	**95.9**
Ours	**67.5**	**90.9**	**67.4**	**76.3**	**95.9**

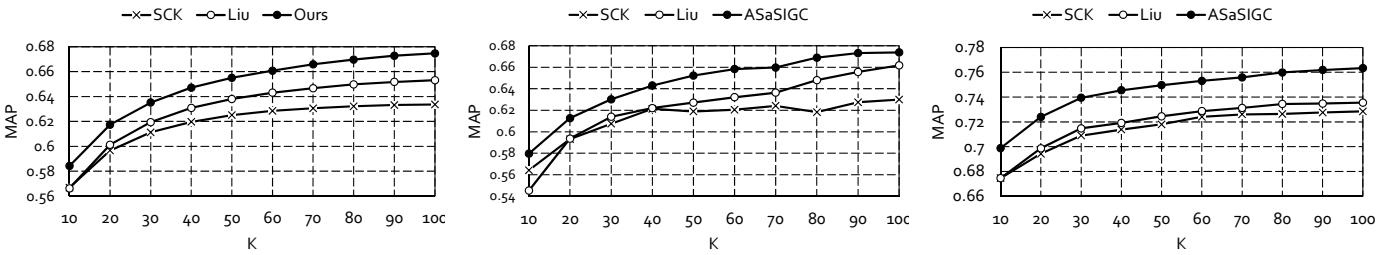
To the best of our knowledge, our approach's MAP of 67.5% for FL32 is the highest yet reported for the retrieval protocol of this dataset, and is more than 8% higher than the second highest reported value [17]. Romberg et al. [17] have reported the MAPs of LO-RANSAC obtained using FL32 and OB under the same setting as ours. The best reported MAPs are 56.8% for FL32 and 72.9% for OB, both of which are much lower than those obtained with our approach. RANSAC has been known to perform poorly when the percentage of true inliers falls much below 50%. This situation commonly occurs with real datasets, and so may explain the surprisingly low performance of LO-RANSAC. RANSAC has also been known to perform poorly when the percentage of true inliers falls much below 50%.

One may notice that the MAPs on the OB dataset are inferior to some of those reported in previous studies [15], [13], [16]. This is because we are not introducing the pre- or post-processing methods, including query expansion, Hamming embedding, soft assignment, and database augmentation, into the retrieval system, as has been done by previous studies. Because the proposed approach is fully compatible with the above methods, we believe that a combination of our approach with these advanced techniques can further improve the performance in the image retrieval tasks.

D. Discussion

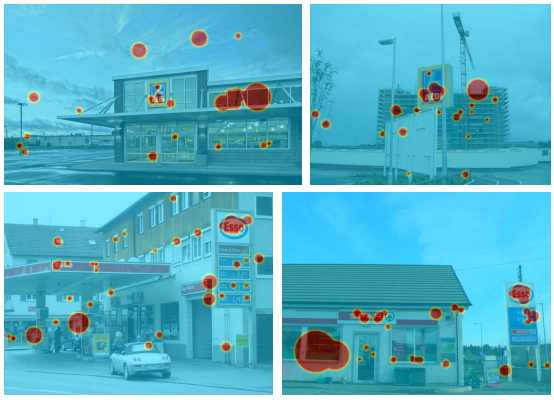
Figure 3 shows examples of the features matched by BOW and the tuples matched by our approach. Both approaches extracted true responses corresponding to the logos. BOW also provided a lot of false responses. In contrast, our approach precisely matched the objects and rejected every false response. Please note the small size of the logos and the coherent shapes of the matches in Fig. 3b.

Figure 4 shows the false responses detected by SCK. It is known that local features such as SIFT are indiscriminating as regards pictures of dense characters, e.g. the newspaper in the second example. Although SCK took the local feature co-occurrence into consideration, it still detected some false responses. In contrast, our approach rejected all false responses because we enforced an additional geometric constraint over the confusing tuples of interest points.



(a) Flickr Logos 32. (b) Holiday. (c) Oxford Buildings.

Figure 2. Relationship between MAP and k used in k-NN.



(a) BOW

(b) Our approach.

Figure 3. BOW versus our approach.

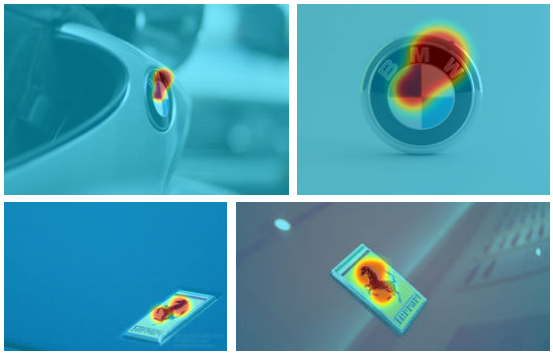
Figure 4. False responses obtained using SCK. All these responses were rejected successfully by our approach.

Figure 5 compares our approach with Liu's approach. Our approach successfully extracted true responses in Fig. 5a even when the viewpoints varied significantly. Liu's approach obtained 0 responses from these image pairs. Our approach is not only more robust as regards viewpoint changes but also achieves higher discriminative power than Liu's approach. Liu's approach provided a lot of false responses in Fig. 5b, and all the false responses were rejected by our approach.

Fig. 6a shows two examples of the false responses mismatched by our approach. In the top row, both the sheet in the left image and the tablecloth in the right image have checkered patterns. In the bottom row, the dotted curtain in the left image is very similar to the sphere-shaped object in the right image. The two pairs are both unrelated but the matched responses make sense. From our experiments, we also found some pairs of images that contain the same object but that are labled by human annotators as unrelated. Figure 6b shows two examples. In the top row, our approach precisely matched the target logo in the left image and rejected false responses patentially as a result of the similar wrappings of the different snacks. In the bottom row, notice the very coherent shapes of the matched tuples and the clear difference between the viewpoints of the two images.

V. Conclusion

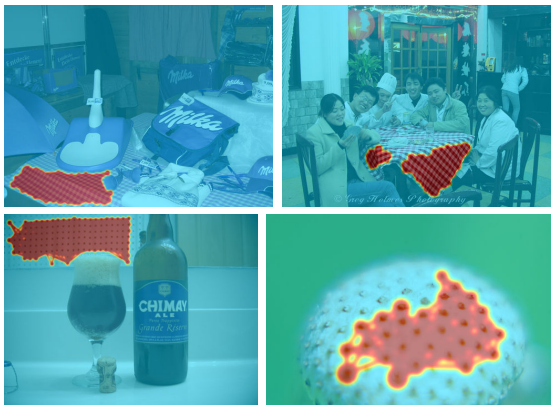
We have proposed an approach for modeling the second-order geometric coherence between local features by extending the application of the Hessian-based affine adapatation. We presented comparisons of the proposed approach and state-of-the-art related studies. The results show that our approach is highly discriminative and more robust to 6-DOF affine transformations. Our experimental results showed that the Hessian-based affine adaptation became unstable if the viewpoint change was too large. We regard this issue as a future subject for investigation. We also intend to test the extension of our approach on a large scale and examine the selectivity of local feature tuples.



(a) True responses obtained using our approach.

(b) False responses obtained using Liu's approach.

Figure 5. Liu's approach versus ours.



(a) False responses obtained using our approach.

(b) True responses missed by human being.

Figure 6. Our approach versus human.

REFERENCES

[1] K. Lebeda, J. Matas, and O. Chum, "Fixing the locally optimized ransac," in *BMVC*, 2012, pp. 1–11.

[2] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. S. Avrithis, "Scalable triangulation-based logo recognition," in *ICMR*, 2011, p. 20.

[3] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *ICCV*, 2011, pp. 1465–1472.

[4] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *CVPR*, 2011, pp. 809–816.

[5] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *CVPR*, 2009, pp. 25–32.

[6] Z. Liu, H. Li, W. Zhou, and Q. Tian, "Embedding spatial context information into inverted file for large-scale image retrieval," in *ACM Multimedia*, 2012, pp. 199–208.

[7] G. Carneiro and A. D. Jepson, "Flexible spatial configuration of local image features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2089–2104, 2007.

[8] J. Gao, Y. Hu, J. Liu, and R. Yang, "Unsupervised learning of high-order structural semantics from images," in *ICCV*, 2009, pp. 2122–2129.

[9] T. Lindeberg, "Scale-space," in *Wiley Encyclopedia of Computer Science and Engineering*. John Wiley & Sons, Inc., 2007.

[10] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[11] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *CVPR*, 2008.

[12] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol, "Scalable logo recognition in real-world images," in *ICMR*, 2011, p. 25.

[13] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV (1)*, 2008, pp. 304–317.

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

[15] O. Chum, A. Mikulík, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *CVPR*, 2011, pp. 889–896.

[16] 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.

[17] S. Romberg and R. Lienhart, "Bundle min-hashing," *International Journal of Multimedia Information Retrieval*, pp. 1–17, 2013.

[18] C. J. van Rijsbergen, *Information Retrieval*. Butterworth, 1979.