

# Robust Spatial Matching as Ensemble of Weak Geometric Relations

Xiaomeng Wu  
 wu.xiaomeng@lab.ntt.co.jp  
 Kunio Kashino  
 kashino.kunio@lab.ntt.co.jp

NTT Corporation  
 3-1, Morinosato Wakamiya Atsugi-shi  
 Kanagawa, Japan 243-0198

## Abstract

This appendix discusses the relation between the relative position and the translation coherences in Section A, presents additional details about the relation between current spatial matching methods and our method in Section B, and shows many more examples of retrieval results to allow a comprehensive comparison.

## A Relative Position vs. Translation

In this study, we did not directly impose any constraint on the translation coherence because it was well incorporated in Eq. 10. Here, we show evidence of this inference. Recall Eq. 10:

$$h_{\mathbf{v}}(c_a, c_b) = \left[ \max \left( \|\mathbf{v}(p_b|p_a) - \mathbf{v}(q_b|q_a)\|_2, \|\mathbf{v}(p_a|p_b) - \mathbf{v}(q_a|q_b)\|_2 \right) < \varepsilon_{\mathbf{v}} \right] \quad (\text{A})$$

which can be decomposed and reformulated as:

$$\mathbf{v}(p_b|p_a) \approx \mathbf{v}(q_b|q_a) \quad (\text{B})$$

$$\mathbf{v}(p_a|p_b) \approx \mathbf{v}(q_a|q_b). \quad (\text{C})$$

Given Eq. 9, we can rewrite Eq. B as:

$$\mathbf{M}(p_a)^{-1}(\mathbf{t}(p_b) - \mathbf{t}(p_a)) \approx \mathbf{M}(q_a)^{-1}(\mathbf{t}(q_b) - \mathbf{t}(q_a)). \quad (\text{D})$$

If we multiply both sides by  $\mathbf{M}(p_a)$ , we obtain:

$$\mathbf{t}(p_b) - \mathbf{t}(p_a) \approx \mathbf{M}(c_a)(\mathbf{t}(q_b) - \mathbf{t}(q_a)) \quad (\text{E})$$

where  $\mathbf{M}(c) = \mathbf{M}(p)\mathbf{M}(q)^{-1}$  is the between-image linear transformation. Likewise, we obtain Eq. F from Eq. C.

$$\mathbf{t}(p_a) - \mathbf{t}(p_b) \approx \mathbf{M}(c_b)(\mathbf{t}(q_a) - \mathbf{t}(q_b)). \quad (\text{F})$$

Equations E and F give us:

$$\mathbf{M}(c_a) \approx \mathbf{M}(c_b). \quad (\text{G})$$

Note that Eq. **G** serves as an alternative form of the scaling and rotation coherences discussed in Sections 3.3.2 and 3.3.3. Now, we rewrite Eq. **E** as:

$$\mathbf{t}(p_a) - \mathbf{M}(c_a)\mathbf{t}(q_a) \approx \mathbf{t}(p_b) - \mathbf{M}(c_a)\mathbf{t}(q_b). \quad (\text{H})$$

Exploiting Eq. **G**, we can replace the  $\mathbf{M}(c_a)$  on the right side with  $\mathbf{M}(c_b)$  such that:

$$\mathbf{t}(p_a) - \mathbf{M}(c_a)\mathbf{t}(q_a) \approx \mathbf{t}(p_b) - \mathbf{M}(c_b)\mathbf{t}(q_b) \quad (\text{I})$$

$$\mathbf{t}(c_a) \approx \mathbf{t}(c_b). \quad (\text{J})$$

where  $\mathbf{t}(c) = \mathbf{t}(p) - \mathbf{M}(c)\mathbf{t}(q)$  as presented in Section 3.3.2. We can see that Eq. **J** is literally the coherence of between-image translations. To minimize the sensitivity to parameters, we decided not to impose this arguably redundant constraint on correspondence pairs but to use the relative position constraint represented by Eq. **10** only.

## B Related Research vs. Our Method

In this section, we discuss the close technical relation between current spatial matching methods and the four fundamental classes of geometric coherences described in Section 3.

### B.1 RANSAC

The RANSAC algorithm proposed by [Philbin et al.](#) computes a geometric transformation  $\mathbf{F}(c)$ , called a hypothesis, from each correspondence  $c$ . All hypotheses  $\{\mathbf{F}(c)\}$  are verified by counting the inliers that inversely fit the transformation. More strictly, given a hypothesis  $\mathbf{F}(c_a)$  computed from a correspondence  $c_a$ , another correspondence  $c_b$  is determined as an inlier if:

$$\left\| \begin{bmatrix} \mathbf{t}(p_b) \\ 1 \end{bmatrix} - \mathbf{F}(c_a) \begin{bmatrix} \mathbf{t}(q_b) \\ 1 \end{bmatrix} \right\|_3 \approx 0, \quad (\text{K})$$

which can be rewritten as:

$$\begin{aligned} \mathbf{t}(p_b) &\approx \mathbf{M}(c_a)\mathbf{t}(q_b) + \mathbf{t}(c_a) \\ &\approx \mathbf{M}(c_a)\mathbf{t}(q_b) + \mathbf{t}(p_a) - \mathbf{M}(c_a)\mathbf{t}(q_a) \end{aligned} \quad (\text{L})$$

and in consequence:

$$\begin{aligned} \mathbf{t}(p_b) - \mathbf{t}(p_a) &\approx \mathbf{M}(c_a)(\mathbf{t}(q_b) - \mathbf{t}(q_a)) \\ &\approx \mathbf{M}(p_a)\mathbf{M}(q_a)^{-1}(\mathbf{t}(q_b) - \mathbf{t}(q_a)). \end{aligned} \quad (\text{M})$$

If we multiply both sides by  $\mathbf{M}(p_a)^{-1}$ , we obtain:

$$\mathbf{M}(p_a)^{-1}(\mathbf{t}(p_b) - \mathbf{t}(p_a)) \approx \mathbf{M}(q_a)^{-1}(\mathbf{t}(q_b) - \mathbf{t}(q_a)) \quad (\text{N})$$

$$\mathbf{v}(p_b|p_a) \approx \mathbf{v}(q_b|q_a). \quad (\text{O})$$

We can see that Eq. **O** is exactly the same as Eq. **B**, i.e. an asymmetric version of our relative position coherence defined in Eq. **10**.

## B.2 Hough Transform-Based Methods

Jegou et al.'s method constitutes a disjunction of scaling and rotation constraints. Some studies assume that the dataset contains no zoomed or rotated images, i.e.  $\forall c \in C, \mathbf{M}(c) = \mathbf{I}_2$  with  $\mathbf{I}_2$  being an identity matrix. For example, Zhang et al. set up a 2D Hough space spanned by (unnormalized) translations of correspondences, but that does not support scaling or rotation invariance. Avrithis and Tolias's method equals a conjunction of scaling, rotation and translation constraints. Section A explains the close technical relation between the relative position and the translation coherences.

## B.3 Spatial Context Methods

Yang and Newsam's method achieves spatial matching by using the neighborhood constraint described as Eq. 4 only. Liu et al.'s method and Wu and Kashino's method equal a conjunction of Equations 4 and 12. Tolias et al.'s method performs in much the same way as Liu et al.'s method except that the neighborhood coherence is not taken into account in the former case. Wu et al. exploited the spatial order of nearby features sorted on the axes of the original (rather than a normalized) image space. The geometric constraint can be understood as a weak and unnormalized approximation of the relative position coherence discussed in Section 3.3.4.

## C Retrieval Result Visualization and Comparison

Figures A-D compare the bag-of-visual-words (BOVW) method, Wu and Kashino's method, Avrithis and Tolias's HPM and our method. The top row shows the query, and the others show the top five results returned by various methods. Correspondences are highlighted in colors according to their contribution to the image similarity. Specifically, the colors indicate: TF-IDF weights of visual words for BOVW; degree centralities of correspondences (if we treat  $\hat{G}$  in Eq. 3 as a graph with vertices being correspondences and edges indicating the geometric constraint) for Wu and Kashino's method and our method; cumulative level affinities (see the original paper for more detail) of correspondences for HPM. The contribution is normalized for each result. The correspondences with the largest contribution are shown in red and those with the smallest contribution in blue.

We can see that the direct matching of local features led to massive mismatches when the images contained repeated patterns, e.g. building facades and windows (Figures A and C), finely-textured patterns, e.g. foliage and sand (Fig. B), and minute letters (Fig. D). The BOVW, Wu and Kashino's method and HPM were all influenced by these mismatches. In contrast, our method showed much greater discriminative power in terms of these clutters.

Basically, our method imposes a stronger constraint than Wu and Kashino's method and so successfully rejected more mismatches than the latter. This can be observed if we look at the correspondences in the same images returned by the two methods. Compared with HPM, our method provides not only a higher discriminative power but also a greater flexibility as regards feature detection errors. For Fig. D as an example, our method successfully identified more true correspondences than HPM for the same images returned by both methods.

## References

- Yannis S. Avrithis and Giorgos Tolias. Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision*, 107(1):1–19, 2014.
- Herve Jegou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.
- Zhen Liu, Houqiang Li, Wengang Zhou, and Qi Tian. Embedding spatial context information into inverted file for large-scale image retrieval. In *ACM Multimedia*, pages 199–208, 2012.
- James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.
- Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.
- Giorgos Tolias, Yannis Kalantidis, Yannis S. Avrithis, and Stefanos D. Kollias. Towards large-scale geometry indexing by feature selection. *Computer Vision and Image Understanding*, 120:31–45, 2014.
- Xiaomeng Wu and Kunio Kashino. Image retrieval based on anisotropic scaling and shearing invariant geometric coherence. In *ICPR*, pages 3951–3956, 2014.
- Zhong Wu, Qifa Ke, Michael Isard, and Jian Sun. Bundling features for large scale partial-duplicate web image search. In *CVPR*, pages 25–32, 2009.
- Yi Yang and Shawn Newsam. Spatial pyramid co-occurrence for image classification. In *ICCV*, pages 1465–1472, 2011.
- Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, pages 809–816, 2011.

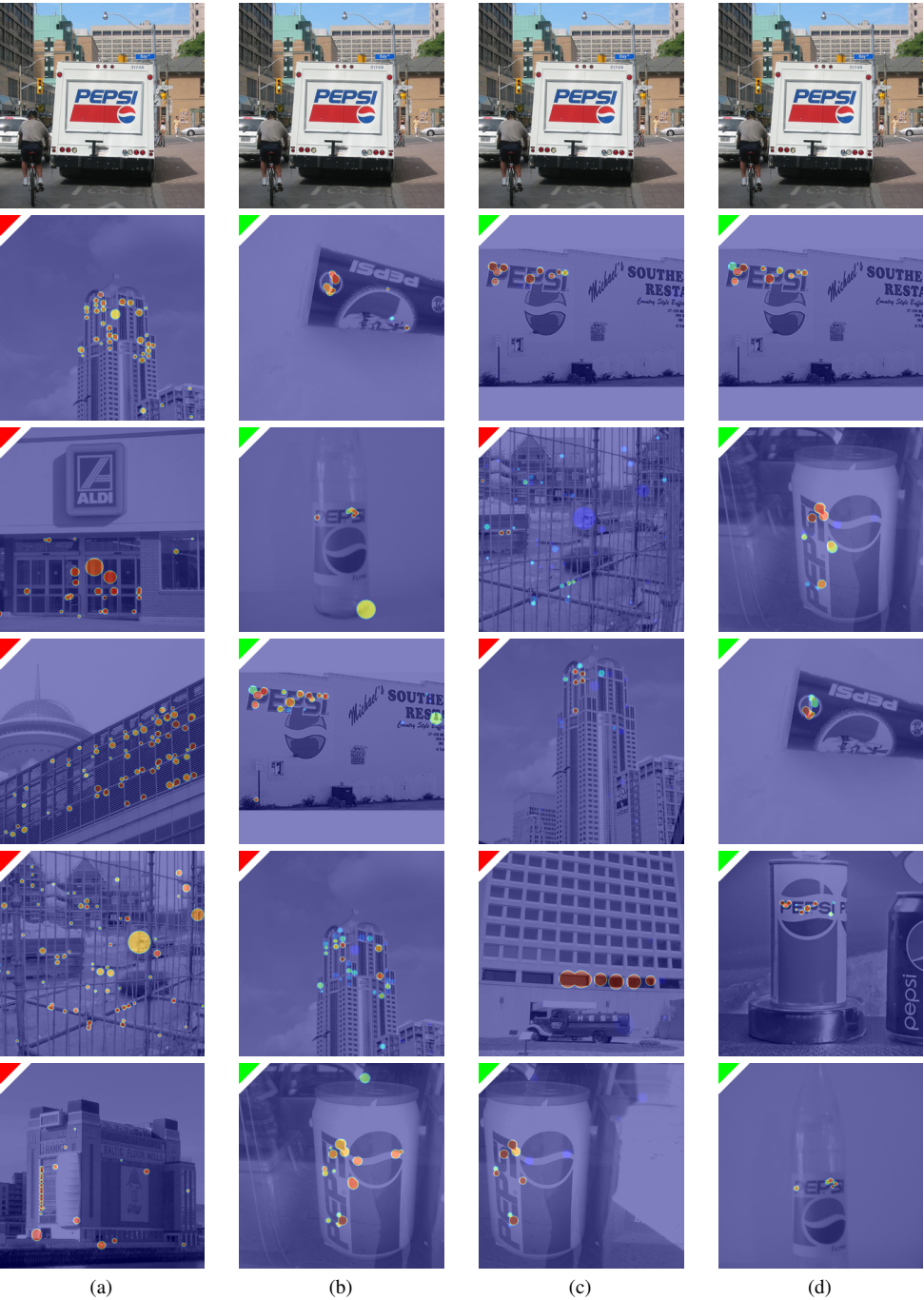


Figure A: Comparison of (a) BOVW proposed by [Sivic and Zisserman](#), (b) [Wu and Kashino's](#) method, (c) HPM proposed by [Avrithis and Tolias](#) and (d) our method. The green and red colors of the upper-left corners of the images indicate positive and negative results, respectively. Correspondences identified by the methods are highlighted in color.

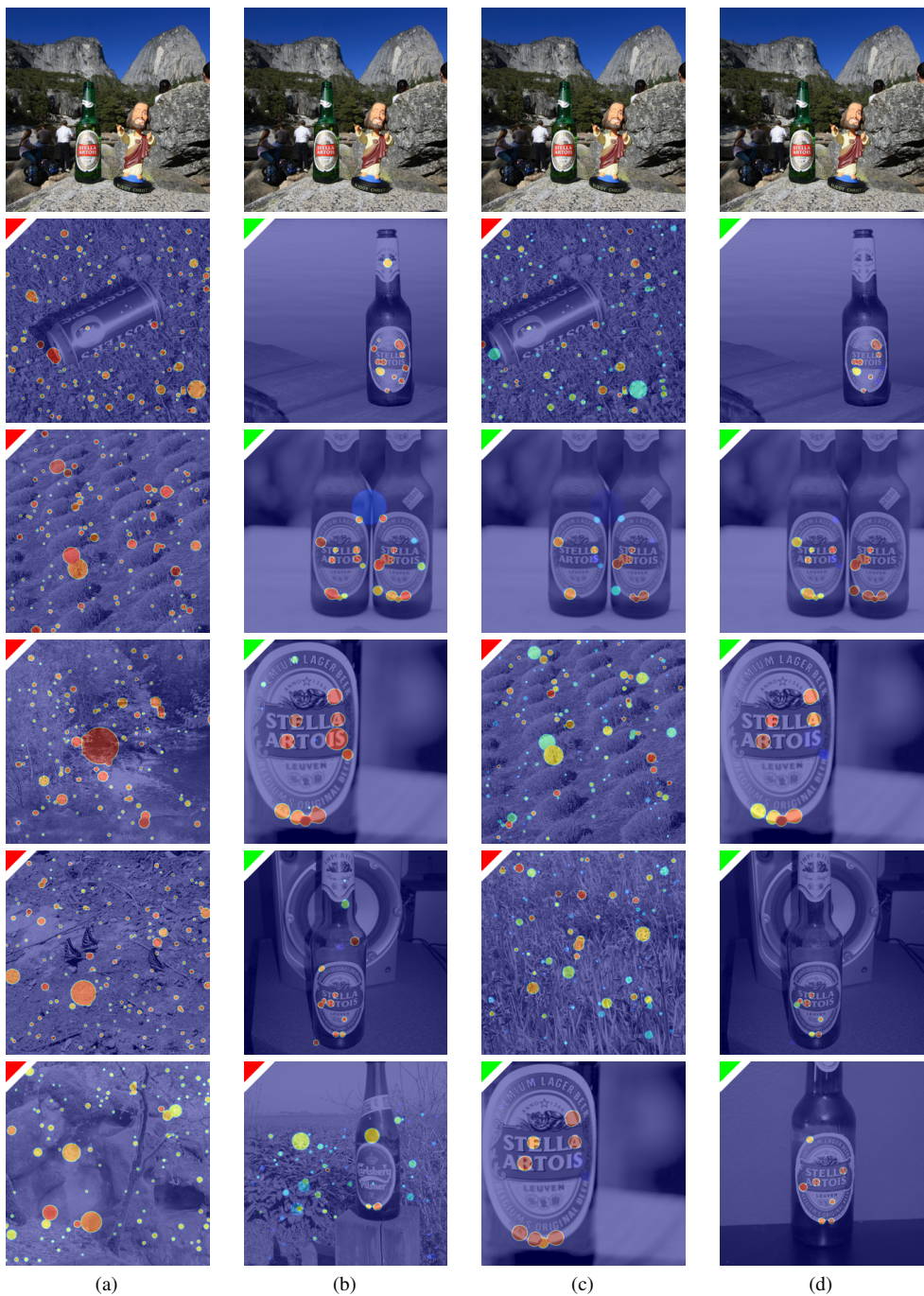


Figure B: Comparison of (a) BOVW proposed by [Sivic and Zisserman](#), (b) [Wu and Kashino's](#) method, (c) HPM proposed by [Avrithis and Tolias](#) and (d) our method. The green and red colors of the upper-left corners of the images indicate positive and negative results, respectively. Correspondences identified by the methods are highlighted in color.



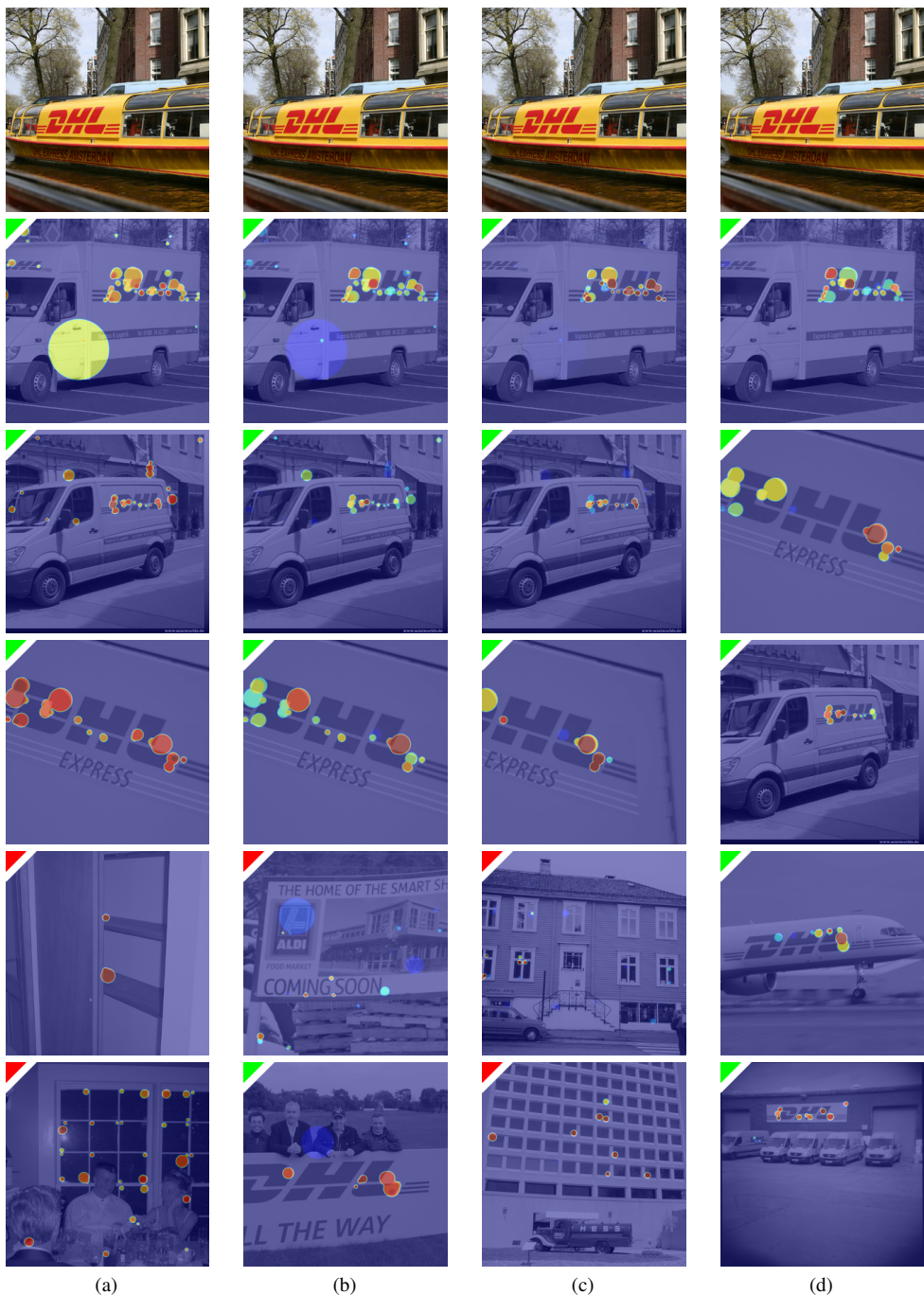


Figure C: Comparison of (a) BOVW proposed by [Sivic and Zisserman](#), (b) [Wu and Kashino](#)'s method, (c) HPM proposed by [Avrithis and Tolias](#) and (d) our method. The green and red colors of the upper-left corners of the images indicate positive and negative results, respectively. Correspondences identified by the methods are highlighted in color.



Figure D: Comparison of (a) BOVW proposed by [Sivic and Zisserman](#), (b) [Wu and Kashino's](#) method, (c) HPM proposed by [Avrithis and Tolias](#) and (d) our method. The green and red colors of the upper-left corners of the images indicate positive and negative results, respectively. Correspondences identified by the methods are highlighted in color.