Robust Feature Matching via Multiple Descriptor Fusion

Yuan-Ting Hu Yen-Yu Lin
Academia Sinica
128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan
r01922042@ntu.edu.tw yylin@citi.sinica.edu.tw

Abstract

We present a novel approach to boost image matching performance by fusing multiple local descriptors in the homography space. Traditional matching methods find correspondences based on a single descriptor and the performance becomes unstable due to the goodness of the chosen descriptor. To address this problem, our method uses multiple descriptors and select a good descriptor for matching each feature point. Specifically, we project every correspondence into the homography space, where correct correspondences tend to gather together due to the similarity of their homographies. Then kernel density estimation is applied to measure the density in the homography space and verify the correctness of correspondences. The proposed approach is comprehensively compared with the state-of-the-art methods and the promising results manifest its effectiveness.

1. Introduction

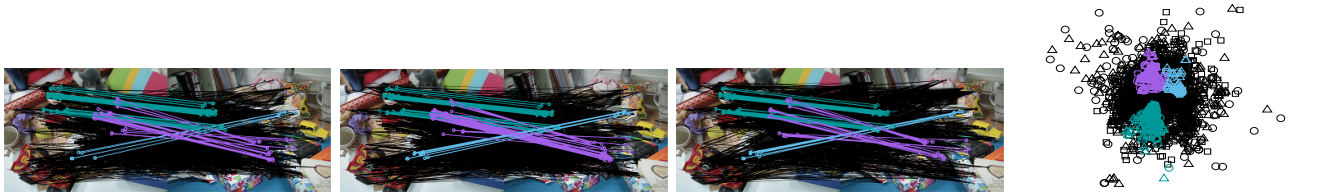
Image matching is a key component of image content analysis. It is one of the critical stages in widespread image processing and computer vision applications, such as panoramic stitching [19], object recognition [14], and image retrieval [9]. The development of powerful descriptors, such as [14, 2, 20, 11, 17], has gained significant progress on matching challenging images. However, the goodness of a descriptor is usually *image-dependent*. There is in general no such a single descriptor that is sufficient for dealing with all kinds of variations in feature matching. Without any prior knowledge about images, using only one descriptor becomes insufficient and unreliable to conquer the wild image matching problems.

To address the aforementioned problem, this paper proposes an unsupervised approach to improve the quality of image matching with the use of multiple, complementary descriptors. Two challenges arise in this scenario. First, features extracted by distinct descriptors are of different dimensions and with diverse scales of statistics. How to effectively fuse heterogeneous descriptors becomes a challenge. Second, image matching in general is an unsupervised task. The goodness of descriptors is hard to determine without ground truth. When feature matchings by different descriptors present, how to identify correct ones from them is another problem. For the first challenge, we use homography space as the unified domain for descriptor selection. For the second challenge, motivated by the observation that correct matchings are highly consistent with each other and gather together in the homography space as shown in Figure 1, we carry out density estimation for identifying correct correspondences. We introduce an unsupervised approach to overcome the two problems so that it can generate more accurate correspondences by leveraging complementary descriptors.

2. Related work

Image matching with geometric verification Image matching through geometric layout checking is one of the most effective ways for identifying correct correspondences. RANSAC [10] is a classic method for removing outliers through geometric checking, but RANSAC becomes time-consuming when dealing with large number of outliers. Graph matching [6, 13] finds a mapping that maximizes the coherent relationship between two sets of feature points. However, as pointed out in [22], graph matching is less robust in multiple object matching. Clustering based techniques [5, 25] and voting schemes [21, 4] have also been explored. Our work is inspired by approach in [4], which casts the voting process as a kernel density estimation problem. Nonetheless, our work can be distinguished from [4] by leveraging multiple descriptors to increase the quality and the number of matched points.

Image matching with multiple descriptor fusion Using multiple descriptors is a feasible way for improving performance since different descriptors can catch diverse visual cues. Mortensen et al. [18] proposed to *concatenate* SIFT [14] and shape contexts [1]. However, simple feature concatenation may lead to suboptimal performance. Works



(a) SIFT (b) DAISY (c) RI (d) The homography space

Figure 1. (a)-(c)are the matching results of SIFT, DAISY and RI, respectively. (d) gives the 2D visualization of correspondences in the homography space via classical multi-dimensional reduction [8]. Each correct matching is colored according to the common object it resides in, and wrong matchings are in black.

in [3, 24] try to address this problem by using *kernel matrices* or *energy functions*. Although both of them can serve as the unified domains for descriptor fusion, these approaches tune or learn weights for descriptor combination. Thus, such methods need training or validation data for determining the weights. In contrast, the plausible correspondences by various descriptors can be identified by our method in a fully unsupervised manner.

3. The proposed method

We aim to match two given images I^P and I^Q, which come with two sets of detected feature points, $U^P = \{u_i^P\}_{i=1}^{N^P}$ and $U^Q = \{u_i^Q\}_{i=1}^{N^Q}$, respectively. The support region of each feature $u_i \in U^P \cup U^Q$ is assumed to be an ellipse in this work. The elliptical support region is decided during feature detection. We use Hessian-Affine detector [16] for its efficiency and high repeatability. Multiple descriptors are employed to characterize each feature point. The center and the described appearances of feature u_i are respectively denoted by $\mathbf{x}_i$ and $\{\mathbf{f}_{i,m}\}_{m=1}^M$, where M is the number of the employed descriptors. For each feature u_i^P in I^P, we find the set of the most similar r correspondences (or matchings), $\mathcal{C}_{i,m} = \{(u_i^P, u_{i_k}^Q \in I^Q)\}_{k=1}^r$, with descriptor m, *i.e.* $\|\mathbf{f}_{i,m}^P - \mathbf{f}_{i_k,m}^Q\|$. Since total M descriptors are adopted, at most $r \times M$ matchings of u_i^P are kept in $\mathcal{C}$ after removing duplicated matchings. Namely,

$$\mathcal{C} = \bigcup_{i=1}^{N^P} \mathcal{C}_i, \text{ where } \mathcal{C}_i = \bigcup_{m=1}^{M} \mathcal{C}_{i,m}. \quad (1)$$

Our goal is to detect the correct correspondence for each $u_i^P \in U^P$ in $\mathcal{C}$ if it exists.

3.1. Homographies of feature correspondences

A homography in this work refers to the geometric transformation of a feature correspondence. After the elliptical region of feature u_i is detected by the detector, it can be specified by mapping a circular region centered on the origin via the affine transformation defined as:

$$T(u_i) = \begin{bmatrix} A(u_i) & \mathbf{x}_i \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \quad (2)$$

where $\mathbf{x}_i \in \mathbb{R}^{2\times 1}$ is the feature center, and $A(u_i) \in \mathbb{R}^{2\times 2}$ is a non-singular matrix which accounts for the scale, the shape, and the orientation of u_i. After normalization with the transformation $T(u_i)^{-1}$, all the adopted descriptors can be applied to u_i, and generate $\{\mathbf{f}_{i,m}\}_{m=1}^M$. Refer to [16] for the details about the calculation of $T(u_i)$. Note that $T(u_i)$ for mapping the elliptical region of feature u_i to a circular region centered at the origin is not unique and we follow [16] to find $T(u_i)$.

For a correspondence between $u_i^P \in U^P$ and $u_j^Q \in U^Q$, its homography can be derived as

$$H_{ij} = T(u_j^Q) * T(u_i^P)^{-1} \in \mathbb{R}^{3\times 3}. \quad (3)$$

H_{ij} is a 6-dof (degrees of freedom) affine homography. Thus, it can be viewed as a point in the 6-dimensional homography space $\mathcal{H}$.

Consider two correspondences $c = (u_i^P, u_j^Q) \in \mathcal{C}$ and $c' = (u_{i'}^P, u_{j'}^Q) \in \mathcal{C}$ and their homography H_{ij} and $H_{i'j'}$. We use the *reprojection error* [5] to measure their geometric dissimilarity. Specifically, the *projection error* of (u_i^P, u_j^Q) with respect to $H_{i'j'}$ is then calculated by

$$d_{err}(u_i^P, u_j^Q, H_{i'j'}) = \|\mathbf{x}_j^Q - \rho(H_{i'j'} \begin{bmatrix} \mathbf{x}_i^P \\ 1 \end{bmatrix})\|, \quad (4)$$

where function $\rho : \mathbb{R}^3 \to \mathbb{R}^2$ is defined as $\rho(\begin{bmatrix} a & b & c \end{bmatrix}^T) = \begin{bmatrix} \frac{a}{c} & \frac{b}{c} \end{bmatrix}^T$. The projection error $d_{err}(u_i^P, u_j^Q, H_{i'j'})$ is the induced error when changing the homography from H_{ij} to $H_{i'j'}$ on correspondence (u_i^P, u_j^Q). The *reprojection error* between correspondences c and c' is then defined as

$$d(c, c') = \frac{1}{4}\big(d_{err}(u_i^P, u_j^Q, H_{i'j'}) + d_{err}(u_j^Q, u_i^P, H_{i'j'}^{-1}) + d_{err}(u_{i'}^P, u_{j'}^Q, H_{ij}) + d_{err}(u_{j'}^Q, u_{i'}^P, H_{ij}^{-1})\big). \quad (5)$$

We will use the reprojection error to measure the geometric dissimilarity between correspondences in $\mathcal{C}$.

3.2. Homography as a unified representation

It can be observed that the homography of a feature correspondence in Eq. (3) is *descriptor-independent*, and can

hence serve as the domain for descriptor fusion. That is, each descriptor determines its own candidate correspondences as shown in Eq. (1), while all the candidate correspondences are represented by the corresponding homographies, each of which can be treated as a point in the homography space $\mathcal{H}$ as shown in Figure 1(d). In this way, the dissimilarity between correspondences that are generated by different descriptors can be measured through Eq. (5). We use this property to fuse various features and we select good features by measuring the geometric distribution in this domain.

3.3. Correct matching identification

The goal at this stage is to decide the correct correspondence for each u_i^P in $\mathcal{C}_i$, if it exists. We tackle this issue based on the observation that correct matchings are similar and hence get together in the homography space (see Figure 1). It implies that the density in the homography space can verify the correctness of correspondences. Correct matchings will form high density regions in the homography space. Therefore, we utilize *kernel density estimation* to identify correct correspondences and our kernel density estimator is defined as

$$\hat{f}(c_i) = \sum_{c_j \in \mathcal{C}, c_j \neq c_i} k(c_i, c_j) = \sum_{c_j \in \mathcal{C}, c_j \neq c_i} \exp\left(-\frac{d(c_i, c_j)}{\sigma}\right), \tag{6}$$

where σ is empirically set to the average reprojection error from each correspondence to its nearest neighbor. It follows that each correspondence $c_i \in \mathcal{C}$ is predicted via its density $\hat{f}(c_i)$. For each feature point u_i^P in image I^P, we pick its correspondence as the one that has the highest density in $\mathcal{C}_i$ (cf. Eq. (1)).

The average running time of our approach on Co-reg dataset [7] is about 3.27 seconds including measuring the dissimilarities of correspondences and estimating the density in the homography space. It is tested on a single PC with an Intel i7-4770 CPU and 16GB memory, and there are around 1,100 detected feature points in each image of Co-reg dataset.

4. Experimental results

In this section, we first describe the experimental setup and then we present two sets of experiments.

4.1. Experimental setup

Adopted baselines We adopt the five state-of-the-art matching algorithms, including SM [12], RRWM [6], ACC [5], HV [4] and VFC [15]. As their original setting, the five methods use only a single descriptor. Besides, we implement two fusion baselines that employ multiple descriptors, including *Ranking* and *Ratio*. In baseline Ranking, we find the first nearest neighbors of all feature points in image I^p with a specific descriptor, and rank the yielded correspondences according to the descriptor distances. For each feature point in I^p, we determine its correspondence by using the descriptor that has the highest rank at this point. In baseline Ratio, we match each point in I^p to its first two nearest neighbors by a specific descriptor, and compute the distance ratio between the two matches. The smaller the ratio is, the more confident the descriptor is at this point. We find the correspondence of this point by the descriptor with the smallest distance ratio.

Adopted descriptors We adopt five descriptors to construct the initial candidate set, $\mathcal{C}$, namely SIFT [14], LIOP [23], DAISY [20], GB [2] and raw intensities (RI) for their complementary properties. The RI descriptor describes the support region by storing the gray-level intensities in a raster scan order. We set r for establishing initial correspondences as 1 in our approach, Ranking and Ratio, while r in the five single-descriptor baselines is set to 5 as used in [4]. Therefore, though we use multiple descriptors, the size of the initial candidate sets in our approach and the adopted baselines are basically the same for fair comparison, except there are some duplicates being removed.

Evaluation metrics The performance of a matching algorithm on an image pair is measured by *precision* and *recall* which are defined as

$$\text{PRECISION} = \frac{n\text{TP}}{n\text{TP} + n\text{FP}} \quad \text{and} \quad \text{RECALL} = \frac{n\text{TP}}{n\text{TP} + n\text{FN}}, \tag{7}$$

where nTP and nFP are the numbers of correctly and wrongly detected correspondences by a matching method, respectively. nFN is the number of correct correspondences that are not detected. Then the 1-precision vs. recall curve (PR curve) can be drawn.

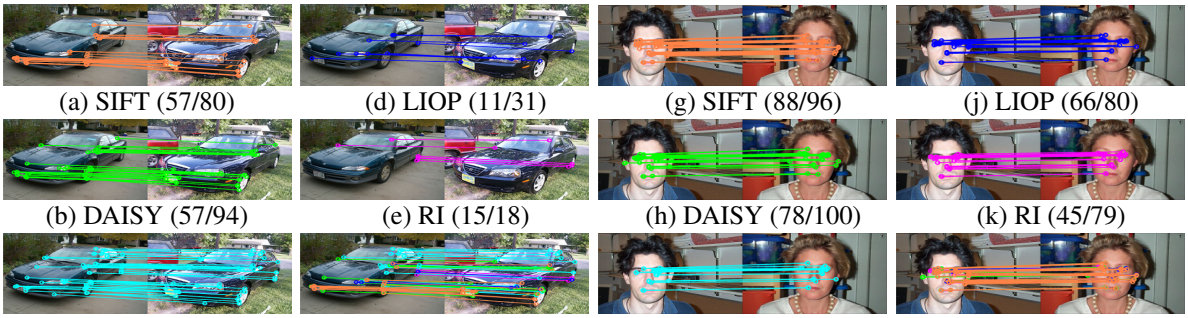
Besides, *mean average precision* (mAP) is used to summarize the performance of each algorithm on a dataset. The average precision on an image pair is calculated by averaging the precisions with different numbers of returned correspondences. *Mean accuracy* (mAcc) on a dataset is also used, where accuracy [6] is defined as

$$\text{ACCURACY} = \frac{n\text{TP}}{n\text{TC}} \tag{8}$$

where nTC is the number of points that have correct correspondences in the candidate set, $\mathcal{C}$. The candidate sets for single-descriptor and multi-descriptor baselines are different. Thus, mAcc is used only for the comparisons between our approach and multi-descriptor baselines.

4.2. Single descriptor vs. multiple descriptors

In this subsection, we show the advantages of using multiple descriptors over a single descriptor. In Figure 2, there are two pairs of results on `car` and `face` of object dataset. In `car`, GB finds more correct correspondences. However, in `face`, SIFT outperforms other descriptors. The



(a) SIFT (57/80) (d) LIOP (11/31) (g) SIFT (88/96) (j) LIOP (66/80)
(b) DAISY (57/94) (e) RI (15/18) (h) DAISY (78/100) (k) RI (45/79)
(c) GB (79/121) (f) Ours (122/154) (i) GB (38/64) (l) Ours (109/120)

Figure 2. The matching results by using a single descriptor and multiple descriptors on object dataset, `car` ((a) $\sim$ (f)) and `face`((g) $\sim$ (l)). The accuracy (nTP / nTC) are shown in the brackets. Only correct matchings are drawn. The color indicates which descriptor is used for matching.

results point out that the goodness of descriptors is image-dependent, and no a single descriptor can outperform others in all the cases. Our method take multiple, complementary descriptors into account, and can select a good descriptor for matching each point. We further use colors to specify which descriptors our approach selects for matching each feature point. Specifically, SIFT is drawn in orange, LIOP in blue, DAISY in green, RI in magenta, and GB in cyan. We can unsupervisedly pick good descriptors, such as SIFT for most of the points in `face` and GB for the upper part of the car in `car`.

4.3. Evaluation on two benchmarks

The performance evaluation is conducted on object dataset [6] and Co-reg dataset [7]. Object dataset [6] gathers 30 pairs of images and each pair contains different object instances of the same category. Co-reg dataset consists of 6 image pairs. There are multiple common objects in every pair. They jointly serve as a good test bed for performance evaluation.

We compare our method with the state-of-the-art approaches, and the results of each image pair are reported in the form of PR curves in Figure 3 for Co-reg dataset. The performances on the two datasets are summarized in the forms of mAP and mAcc in Table 1 and Table 2. Note that we don't compare mAcc with the single descriptor baselines because the denominators of the accuracy in Eq. (8), namely nTC, are different. Because HV, SM, ACC and VFC use only a single descriptor, we manually pick the best descriptor for them for the clarity of PR curves. Thus, their performances may be overestimated in this sense. Their recalls are limited because their candidate sets are constructed by only a single descriptor. Note that Co-reg dataset contains multiple common objects with different transformations, thus VFC is less able to rank correct matchings properly due to its assumption on one smooth vector field. Fusion baselines Ranking and Ratio, which use the five descriptors as our approach does, can increase the recall with the aid of multiple descriptors in most cases, but their precision and mAcc is unsatisfactory. On the contrary, our method can not only effectively match multiple objects, but further improve the performance of matching in both recall and precision by leveraging multiple descriptors.

5. Conclusion

We have presented a simple but effective matching approach that can leverage multiple, complementary descriptor. Specifically, the correspondences yielded by all descriptors are firstly projected into the homography space, in which we select good descriptors in an unsupervised way. Kernel density estimation is then employed to identify correct matchings. The proposed approach is featured with its high flexibility in the sense that it can work with any elliptical region detectors as well as heterogeneous descriptors. Our approach has been evaluated and compared with the state-of-the-art approaches on two benchmark datasets. The experimental results demonstrate that our approach can enhance the matching quality in both precision and recall.

Acknowledgement

This work was supported in part by grants III 104-EC-17-A-24-1170, MOST 103-2221-E-001-026-MY2, and MOST104-2628-E-001-001-MY2.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *TPAMI*, 24(4):509–522, 2002. 1

[2] A. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, 2001. 1, 3

[3] A. Bosch, A. Zisserman, and X. Muñoz. Representing shape with a spatial pyramid kernel. In *Proc. ACM Conf. Image and Video Retrieval*, 2007. 2

[4] H.-Y. Chen, Y.-Y. Lin, and B.-Y. Chen. Robust feature matching with alternate Hough and inverted Hough transforms. In *CVPR*, 2013. 1, 3, 5

Table 1. Performance in mAP on Co-reg dataset [7] and object dataset [6].

	Co-reg dataset					Object dataset				
mAP (%)	SIFT	LIOP	DAISY	RI	GB	SIFT	LIOP	DAISY	RI	GB
SM [12]	55.30	38.74	46.71	34.57	12.13	27.86	16.40	22.80	11.13	27.74
RRWM [6]	66.55	49.40	53.38	40.23	10.24	30.98	16.83	26.73	21.04	29.31
ACC [5]	60.28	29.83	36.49	15.10	8.88	21.81	7.41	17.88	7.51	16.16
HV [4]	60.12	43.97	50.06	37.14	12.08	27.01	16.72	23.49	15.37	31.54
VFC [15]	31.11	11.79	16.29	4.51	1.77	21.44	9.26	18.74	4.46	21.60
Ranking			48.48					19.38		
Ratio			53.61					18.62		
Ours			**71.14**					**37.08**		

Table 2. Performance in mAcc on Co-reg dataset [7] and object dataset [6].

mAcc (%)	Co-reg dataset	Object dataset
Ranking	72.89	42.63
Ratio	67.29	43.07
Ours	**89.27**	**60.74**

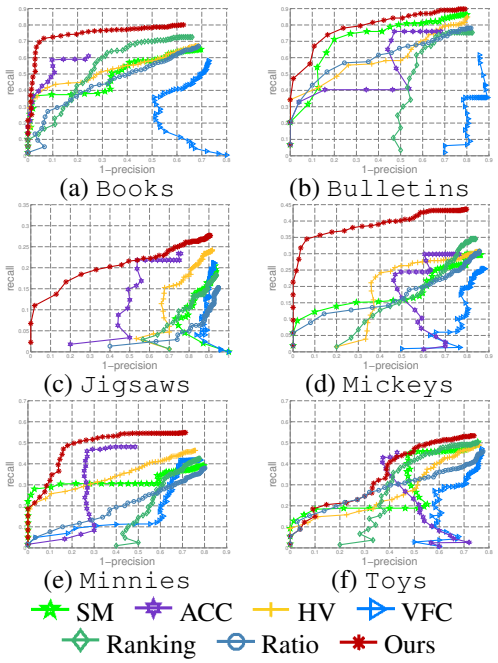


Figure 3. PR curves of various approaches on the six image pairs of Co-reg dataset.

[5] M. Cho, J. Lee, and K. Lee. Feature correspondence and deformable object matching via agglomerative correspondence clustering. In *ICCV*, 2009. 1, 2, 3, 5

[6] M. Cho, J. Lee, and K. Lee. Reweighted random walks for graph matching. In *ECCV*, 2010. 1, 3, 4, 5

[7] M. Cho, Y. Shin, and K. Lee. Co-recognition of image pairs by data-driven Monte Carlo image exploration. In *ECCV*, 2008. 3, 4, 5

[8] T. F. Cox and M. A. A. Cox. *Multidimentional Scaling*. Chapman & Hall, London, 1994. 2

[9] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideals, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008. 1

[10] M. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 1

[11] D. Hauagge and N. Snavely. Image matching using local symmetry features. In *CVPR*, 2012. 1

[12] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *CVPR*, 2005. 3, 5

[13] M. Leordeanu, R. Sukthankar, and M. Hebert. Unsupervised learning for graph matching. *IJCV*, 96(1):28–45, 2012. 1

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 3

[15] J. Ma, J. Zhao, J. Tian, A. Yuille, and Z. Tu. Robust point matching via vector field consensus. *TIP*, 23(4):1706–1721, 2014. 3, 5

[16] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004. 2

[17] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *TPAMI*, 27(10):1615–1630, 2005. 1

[18] E. Mortensen, H. Deng, and L. Shapiro. A SIFT descriptor with global context. In *CVPR*, 2005. 1

[19] R. Szeliski and H.-Y. Shum. Creating full view panoramic image mosaics and environment maps. In *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 1997. 1

[20] E. Tola, V. Lepetit, and P. Fua. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *TPAMI*, 32(5):815–830, 2010. 1, 3

[21] G. Tolias and Y. Avrithis. Speeded-up, relaxed spatial matching. In *ICCV*, 2011. 1

[22] C. Wang, L. Wang, and L. Liu. Improving graph matching via density maximization. In *ICCV*, 2013. 1

[23] Z. Wang, B. Fan, and F. Wu. Local intensity order pattern for feature description. In *ICCV*, 2011. 3

[24] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, 2013. 2

[25] W. Zhang, X. Wang, D. Zhao, and X. Tang. Graph degree linkage: Agglomerative clustering on a directed graph. In *ECCV*, 2012. 1