THE SHORTEST MATCHING PATH BASED ON NOVEL CYCLE CONSISTENCY

Wei Yu, Yi Tao, Hongxun Yao

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
yuwei.hit@outlook.com, taoyisummer@gmail.com, h.yao@hit.edu.cn

ABSTRACT

Category-level image matching is extremely challenging due to various intra-class variations. To tackle the large variations, we propose an algorithm to jointly estimate the dense correspondence for image set, which reformulates image set alignment into the problem of shortest path searching. We propose a novel tri-image cycle-consistency to measure the matching "distance" between two image, which is further used to improve the pair-wise dense correspondence. Meanwhile, we utilize CNN feature pyramid to achieve pair-wise image matching hierarchically. Extensive experiments and analysis demonstrate the superiority of our method in matching images with challenging variations.

Index Terms— Joint matching, Hierarchical matching, Shortest path searching, CNN feature pyramid

1. INTRODUCTION

Image matching aims to estimate dense correspondence (flow field) between two images, which is a fundamental component for many vision problems, such as label propagation [1], fine-grained classification [2] and object discovery [3]. Currently, most methods focus on the image pair from same semantic category, where larger and more challenging variations are required to be handled [4, 5, 6, 7, 8]. These variations arise from not only the changes in illumination and viewpoint, but also the intra-class appearance variations of different instances. To cope with large variations, we attempt to improve category-level image matching along two main directions.

On one hand, we improve image matching in the joint way. An image set of the same category is jointly matched and large variations between two images are mitigated by inserting intermediate images. To discover the intermediate images, we propose a novel cycle-consistency and reformulate the problem of joint image alignment as the problem of shortest path searching.

On the other hand, we improve image matching in the hierarchical way. CNN feature pyramid is used to match images from top level to bottom level. The learned feature extractors from CNN demonstrate their ability in resisting intra-category variations not only on the classification task [9, 10] but also on other vision tasks [11, 12]. With multi-level feature extractors of a pre-trained CNN, a CNN feature pyramid is built for each image. The matching process is carried out in a hierarchical way. The top-level features are used to match large patterns without distractions from details, while bottom-level feature are used to match structural details. The higher level matching results are used to guide the lower level matching.

The architecture of our method is illustrated in Fig.1. Both hierarchical way and joint way are unified into a single framework, where the matching process is achieved hierarchically, and each level is carried out in a joint way with the help of intermediate images.

2. RELATED WORK

Inspired by the classical optical flow algorithm, SIFT flow [4] is proposed to match images from different scenes by replacing pixel intensities with SIFT descriptors. DSP [5] extends SIFT flow by building a hierarchically connected pyramid of grid graphs, and regularizes matching consistency on different scales. PatchMatch [13] leverages the idea of random search to speed up the time costing matching process. ProposalFlow [8] takes advantage of bottom-up object proposals which can matches image patches that are more likely to be meaningful objects.

Research on joint image alignment is pioneered by works on estimating affine transformations for an image set, where variations only come from affine transformations with some sensor noises. Congealing [14] aims to make a set of images more similar via a continuous series of allowable transformations, and performs well on the digit and other simple datasets. Inspired by low rank and sparse recovery, RASL [15] utilizes parametric transformations to align images to the common subspace. Collection Flow [16] aims to model the common appearance model of an image set with low-rank subspace. FlowWeb [7] models the connections between images based on cyclic constraint across multiple images. The output of existing matching algorithm is used as initial correspondences, and the compositions of correspondences are refined as the result of post-processing.

The feature extractors learned from CNN has been proved to be robust to large intra-category variations through many vision tasks [11, 12]. Long *et al.* [6] investigate whether the features of one convolutional layer can learn correspondence

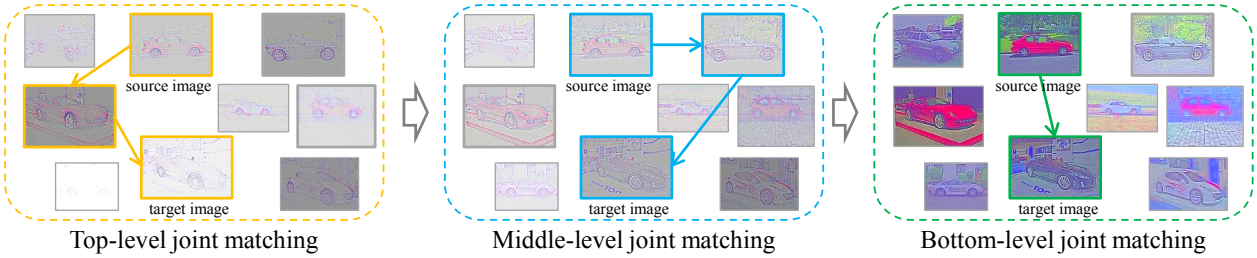

Fig. 1. Illustration of the hierarchical and joint image matching framework. The visualizations reconstructed from feature maps represent corresponding images on different matching levels. Matching from top level to bottom level is illustrated from left to right, and matching results from higher levels guide the matching of their next levels. In each matching level, joint image matching is carried out by accessing an image set of the same category. A fully-connected graph is constructed for the set of images where weights of edges are calculated by cycle-consistency. The weight measures the reliability degree of flow field between two images. From source image to target image, a shortest path is found in the graph to decompose large variations along the path. (Best viewed electronically)

between images and show positive answer by several well-designed experiments.

3. METHOD

3.1. Cycle-consistency

Cycle-consistency is a key concept to introduce the joint matching, which has been successfully used in shape matching [17, 18] and co-segmentation [19, 20]. The cycle-consistency aims to measure the difficulty degree of two images to be matched. Basically, two images (I_i and I_j) can be matched in two directions, from i to j and from j to i. When one pixel can match back in the two directions, the matching of this pixel is more reliable since it is with less ambiguities. Formally, let $w_{i,j}$ denote the estimated flow field from I_i to I_j, and $w_{i,j}(p)$ be the flow vector at pixel p. Then the bi-image cycle-consistency for pixel p is defined as:

$$c_{i,j}(p) = ||w_{i,j}(p) + w_{j,i}(q)||_1, \tag{1}$$

where $q = p + w_{i,j}(p)$ is matched pixel in I_j from direction of i to j. $c_{i,j}(p) = 0$ means the pixel can be matched back in the directions and the pixel satisfies bi-image cycle-consistency. The tri-image cycle-consistency is defined as:

$$c_{i,k,j}(p) = ||w_{i,k}(p) + w_{k,j}(r) + w_{j,i}(q)||_1, \tag{2}$$

where pixel p in I_i is matched to pixel $r = p + w_{i,k}(p)$ in I_k, pixel r in I_k is matched to pixel $q = r + w_{k,j}(r)$ in I_j, then pixel q in I_j is matched back to I_i. If $q + w_{j,i}(q)$ equals to p, then $c_{i,k,j}(p)$ equals to 0 which means the cycle-consistency.

Suppose there are large variations between I_i and I_j, and I_k is the intermediate image between I_i and I_j which can produce reliable flow fields $w_{i,k}$ and $w_{k,j}$. However, $w_{j,i}$ is hard to be reliable due to large variations between I_i and I_j, which further makes the cycle-consistency being calculated as unreliable. To address the issue, we propose a novel tri-image cycle-consistency based on bi-image cycle-consistency, which bypasses the matching between I_i and I_j and can be defined as:

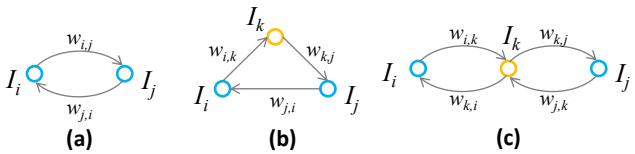

Fig. 2. The illustration of different cycle-consistency. (a) Bi-image cycle-consistency, (b) Tri-image cycle-consistency, and (c) Proposed tri-image cycle-consistency.

$$c_{i,k,j}(p) = c_{i,k}(p) + c_{k,j}(r) \tag{3}$$

where $c_{i,j,k}(p) = 0$ means $c_{i,k}(p) = 0$ and $c_{k,j}(r) = 0$. That means pixel p in I_i is reliably matched to pixel r in I_k, and pixel r in I_k is reliably matched to pixel q in I_j. Then it can be deduced that pixel p can be reliably matched to q through the intermediate pixel r.

Fig. 2 illustrates the three cycle-consistency definitions. In our definition, the intermediate I_k acts as a bridge and bypasses the matching between I_i and I_j.

Based on cycle-constitency, we further define distance between images which will be used for intermediate image finding in joint matching. When I_i is directly matched to I_j, their distance is defined based on bi-image cycle-consistency:

$$d(i,j) = \sum_{p \in I_i} B_2(p), \tag{4}$$

where $B_2(p) = 0$ if $c_{i,j}(p) \leq t$ and equals 1 otherwise, t is a threshold to tolerate small unreliable matching and set as $0.05 \cdot \max(h, w)$ following [7], where h and w are the height and width of feature map respectively.

When I_i is matched to I_j through the intermediate image I_k, the distance is defined based on the proposed tri-image cycle-consistency:

$$d(i,k)\oplus d(k,j) = \sum_{p\in I_i} B_3(p) \tag{5}$$

where $B_3(p) = 0$ if $c_{i,k,j}(p) \leq t$ and equals 1 otherwise.

3.2. Joint matching in a level

Given an image set $\mathcal{I} = \{I_i\}_1^N$, each image I_i is represented as a CNN feature pyramid $F_i = F_i^1, ..., F_i^L$, where F_i^l is the feature map of level l. To do joint matching at level l, we design an objective function considering both pair-wise information and image-set information, i.e.,

$$\mathcal{E}^l = \mathcal{E}_P^l + \mathcal{E}_J^l \tag{6}$$

where $\mathcal{E}_P^l$ and $\mathcal{E}_J^l$ denote the pair-wise matching energy and the joint-image matching energy on the l^{th} level.

The pair-wise matching energy $\mathcal{E}_P$ measures the matching energy of all image pairs at the l^{th} level:

$$\mathcal{E}_P^l = \sum_{i\neq j}[E_D(w_{i,j}^l) + \alpha E_S(w_{i,j}^l) + \beta E_G(w_{i,j}^l|w_{i,j}^{l+1})] \tag{7}$$

where E_D, E_S, E_G are the data term, smoothness term and guidance term respectively. The data term E_D measures the distance between features of the l^{th} level along with the flow field $w_{i,j}^l$. The smooth term E_S measures the distance between flow vectors that are spatially close. The guidance term E_G leverages the flow field of level $l+1$ as guidance for matching level l. These three terms are defined following the existing methods [21].

The joint-image matching energy measures the length of matching path for all image pairs, and is formulated as:

$$\mathcal{E}_J^l = \sum_{i\neq j} D^l(i,j) \tag{8}$$

where $D^l(i,j)$ is the length of alignment path from I_i to I_j, which may pass through multiple intermediate images.

3.3. Optimization

Starting from the top level, we optimize the matching objective function level by level. And a two-step optimization is introduced to minimize the matching energy of each level, namely, pair-wise step and joint-image step.

The pair-wise step aims to estimate the flow fields for all image pairs, both higher-level flow field and the features of current level are used. Dual-layer loopy belief propagation [4] to minimize the pair-wise matching energy Eq.7, and the guidance from higher level is inputted as the form of message in optimization. The horizontal flow and vertical flow are separated in the message passing, and both smoothness term and guidance term are decoupled, which significantly improve computational efficiency.

Algorithm 1 Search for the shortest alignment path

Input: A set of flow fields w^l estimated from pair-wise step
Output: A set of refined flow fields w^l

1: Initialize the distance matrix D^l caculated by Eq.4
2: **for** k = 1 to N **do**
3: **for** i = 1 to N **do**
4: **for** j = 1 to N **do**
5: **if** $D^l(i,j) > d(i,k)\oplus d(k,j)$ **then**
6: $D^l(i,j) = d(i,k)\oplus d(k,j)$
7: **for** $p \in I_i$ **do**
8: **if** $B_2(p) = 1 \&\& B_3(p) = 0$ **then**
9: $r = p + w_{i,k}^l(p)$
10: $w_{i,j}^l(p) = w_{i,k}^l(p) + w_{k,j}^l(r)$
11: **return** the refined w^l

Joint-image step aims to refine the flow fields with intermediate images. The Floyd algorithm [22] is used to search the shortest alignment path between each image pair through inserting the intermediate images. The detailed implementation of joint-image step is presented in Alg.1. From line 7 to 10, $w_{i,j}^l$ is updated by replacing unreliable flow vectors with reliable ones using the composition of $w_{i,k}^l$ and $w_{k,j}^l$ generated by intermediate image k.

4. EXPERIMENT

4.1. Implementation details

In this section, we verify the proposed method on two widely used benchmarks: PASCAL part [7] and Caltech101 [23]. Images of PASCAL part dataset are object regions cropped from original images of PASCAL dataset, where the variations are mainly confined to a small portion of the objects. Caltech101 are with cluttering background, where the image numbers of different categories are extremely unbalance. For fair comparison, we sample 40 images in sequence for the categories with more than 40 images, since the size of 40 images has led to remarkable performance as shown in [7].

We utilize the VGG-16 model to build the CNN feature pyramid Since some adjacent layers are of same feature map size, we only keep one layer from the set of layers with same feature map size. The layers (c5_1, c4_2, c3_3, c2_1, c1_2) are selected to build the CNN feature pyramid. The tradeoff weights for defining pair-wise matching energy are empirically set as: $\alpha = 2^{l-1}$, $\beta = 0.005{\cdot}\alpha$ for the l^{th} level.

4.2. Matching results on PASCAL part

Following the same setting as [7], we evaluate our method on two tasks on this dataset. The weighted IoU(Intersection

Table 1. Matching accuracy on PASCAL part.

	Methods	IOU	PCK
Pair-wise Matching	DSP [5]	0.39	0.17
	ProposalFlow [8]	0.41	0.17
	Ours(Hierarchical way)	0.42	0.20
Joint Matching	Congealing [14]	0.38	0.11
	RASL [15]	0.39	0.16
	CollectionFlow [16]	0.38	0.12
	FlowWeb [7]	0.43	0.26
	Ours	0.46	0.27

over Union) is used to measure the accuracy between transferred part segments and ground-truth ones on the target image, where the weights determined by the size of each part. The PCK($\alpha = 0.05$) metric is used to measure the matching accuracy between predicted keypoints and ground-truth ones. Tab.1 reports the quantitative results, where our method consistently improves existing methods on both tasks. We also evaluate the matching results estimated by our method only in hierarchical way.

4.3. Matching results on Caltech101

Following the experimental protocol of [5], we compare our method with other baselines on the task of rough matching. The metrics of LT-ACC(Label Transfer Accuracy) and IoU are used to quantitatively evaluate the predicted masks and ground-truth ones. Tab. 2 reports the performance averaged over all 101 categories, where our method consistently outperforms other methods by a significant margin. Fig. 3 shows the matching results warped by different methods. Our method shows the superior performance on benchmark with cluttering elements, and it is also powerful to cope with the intra-class variations between objects, such as scale and pose(the first example), rotation(the second example) and the appearance(the third example).

Table 2. Matching accuracy on Caltech101.

	Methods	LT-ACC	IoU
Pair-wise Matching	PatchMatch [13]	0.62	0.35
	SIFTFlow [4]	0.70	0.41
	DSP [5]	0.78	0.49
	ProposalFlow [8]	0.77	0.49
	Ours(Hierarchical way)	0.80	0.55
Joint Matching	FlowWeb [7]	0.81	0.54
	Ours	0.83	0.59

4.4. The proposed tri-image cycle consistency

In particular, we show the effectiveness of our posed tri-image cycle-consistency. Fig. 4 shows the comparison of joint matching using different tri-image cycle-consistency in 8 categories of Caltech101.

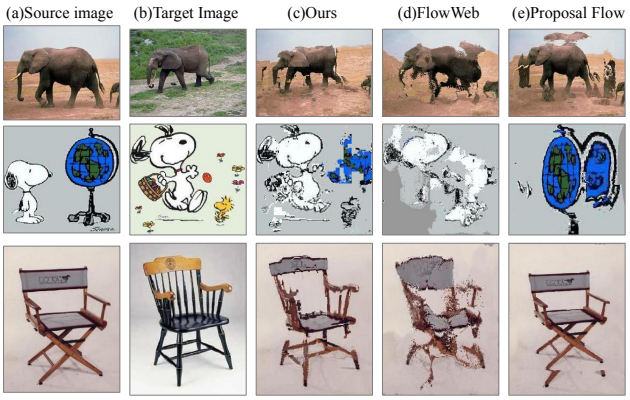

Fig. 3. Examples of rough matching results on Caltech101. Each row shows that (a) source image is warped to (b) target image using different methods (c)-(e)

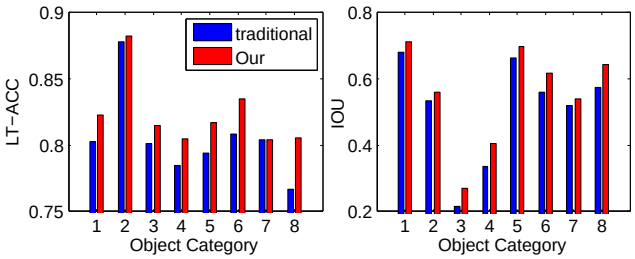

Fig. 4. The comparison of joint matching between using the traditional tri-image cycle-consistency and using our tri-image cycle-consistency.

5. CONCLUSION

In this paper, large variations in category-level image matching are tackled through a hierarchical and joint way. To further deal with large variations, joint matching is used during matching in each level. Extensive experiments demonstrate the superiority of our method under challenging intra-class variations. Currently, graph construction for bottom levels is computationally costing due to the larger size of feature maps. Since bottom levels receive rich guidance from top levels and need fewer intermediate images by empirical study, we will study the possibility to remove joint matching in bottom level for computational efficiency.

6. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China under Project No. 61472103.

7. REFERENCES

[1] Stephen Gould and Yuhang Zhang, "Patchmatchgraph : Building a graph of dense patch correspondences for label transfer," in *European Conference on Computer Vision*, 2012, pp. 439–452.

[2] Jonathan Krause, Hailin Jin, Jianchao Yang, and Fei Fei Li, "Fine-grained recognition without part annotations," in *Computer Vision and Pattern Recognition*, 2015, pp. 5546–5555.

[3] M. Rubinstein, A. Joulin, J. Kopf, and Ce Liu, "Unsupervised joint object discovery and segmentation in internet images," in *Computer Vision and Pattern Recognition*, 2013, pp. 1939–1946.

[4] Ce Liu, Jenny Yuen, and Antonio Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–94, 2011.

[5] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman, "Deformable spatial pyramid matching for fast dense correspondences," in *Computer Vision and Pattern Recognition*, 2013, pp. 2307–2314.

[6] Jonathan Long, Ning Zhang, and Trevor Darrell, "Do convnets learn correspondence?," *Neural Information Processing Systems*, vol. 2, pp. 1601–1609, 2014.

[7] Tinghui Zhou, Jae Lee Yong, S. X. Yu, and A. A. Efros, "Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences," in *Computer Vision and Pattern Recognition*, 2015, pp. 1191–1200.

[8] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce, "Proposal flow," in *Computer Vision and Pattern Recognition*, 2016.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, no. 2, pp. 2012, 2012.

[10] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[11] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Computer Vision and Pattern Recognition*, 2012, pp. 3626 – 3633.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[13] Connelly Barnes, Eli Shechtman, B. Goldman Dan, and Adam Finkelstein, "The generalized patchmatch correspondence algorithm," in *European Conference on Computer Vision*, 2010, pp. 29–43.

[14] Erik G. Learned-Miller, "Data driven image models through continuous joint alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 236–250, 2006.

[15] Yigang Peng, A. Ganesh, J. Wright, Wenli Xu, and Yi Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2233–2246, 2012.

[16] Steven M. Seitz, "Collection flow," in *Computer Vision and Pattern Recognition*, 2012, pp. 1792–1799.

[17] Andy Nguyen, Mirela Ben-Chen, Katarzyna Welnicka, Yinyu Ye, and Leonidas Guibas, "An optimization approach to improving collections of shape maps," *Computer Graphics Forum*, vol. 30, no. 5, pp. 1481C1491, 2011.

[18] Qi Xing Huang and Leonidas Guibas, "Consistent shape maps via semidefinite programming," *Computer Graphics Forum*, vol. 32, no. 5, pp. 177C186, 2013.

[19] Fan Wang, Qixing Huang, and Leonidas J. Guibas, "Image co-segmentation via consistent functional maps," in *International Conference on Computer Vision*, 2013, pp. 849–856.

[20] Fan Wang, Qixing Huang, Maks Ovsjanikov, and Leonidas J. Guibas, "Unsupervised multi-class joint image segmentation," in *Computer Vision and Pattern Recognition*, 2014, pp. 3142–3149.

[21] Wei Yu, Kuiyuan Yang, Yalong Bai, Hongxun Yao, and Yong Rui, "Dnn flow: Dnn feature pyramid based image matching," in *British Machine Vision Conference*, 2014, pp. 109.1–109.10.

[22] Robert W. Floyd, "Algorithm 97: Shortest path," *Communications of the Acm*, vol. 5, no. 6, pp. 345–345, 1962.

[23] Li Fei-Fei, Rob Fergus, and Pietro Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *CVIU*, vol. 106, no. 1, pp. 59–70, 2007.