DNN Flow: DNN Feature Pyramid based Image Matching

Wei Yu[1]
w.yu@hit.edu.cn
Kuiyuan Yang[2]
kuyang@microsoft.com
Yalong Bai[1]
ylbai@mtlab.hit.edu.cn
Hongxun Yao[1]
h.yao@hit.edu.cn
Yong Rui[2]
yongrui@microsoft.com

[1] Harbin Institute of Technology
Harbin, China
[2] Microsoft Research
Beijing, China

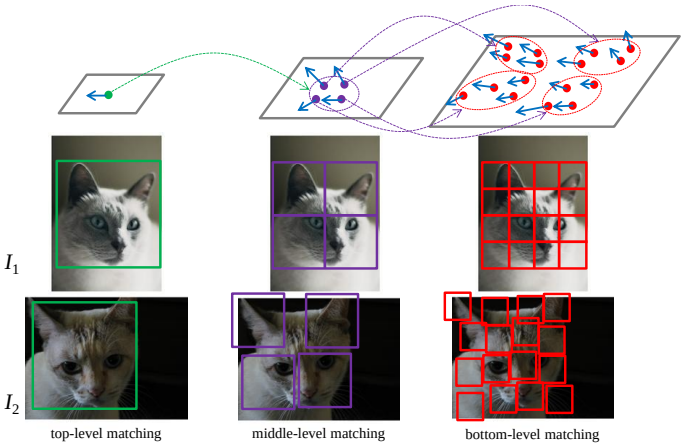

Figure 1: Matching I_1 and I_2 using DNN flow. Each column shows the matching of different levels. In first row, parallelogram denotes the DNN feature image of I_1, where dot represents the feature at that location. Line with arrow denotes the flow vector of the corresponding feature, while curve with arrow denotes guidance from high level to low level. In second row, the color rectangles show I_1's patches covered by the DNN features. Third row shows I_2's matching patches corresponding to the patches of second row.

As a fundamental problem in computer vision, image matching is the cornerstone for many vision problems, such as motion estimation [2], label propagation [3] and object modeling [1]. The goal of image matching is to find the corresponding pixels between two images. Based on the variations between the two images, we roughly divide image matching into two categories, i.e., instance-level matching and category-level matching. Compared to instance-level matching, category-level matching tries to match two images with more challenge variations, which belong to same category. Category-level matching aims to overcome the intra-class variability in shape and other visual properties, such as cars with various shapes and colors and cats with different poses and furs.

In this paper, we propose a DNN feature based image matching approach, which focuses on category-level matching. Recently, Deep Neural Network (DNN) has shown great ability in handling the variations under the same category. The ability comes from the gradual abstraction through several layers, where low layer detects simple patterns, such as edges and blobs, middle layer detects object parts and high layer detects objects. Considering the ability of DNN feature in handling semantic variations, we propose a novel image matching method based on DNN feature pyramid, named as DNN Flow. DNN Flow utilizes DNN features of different layers to achieve coarse to fine matching. As shown in Figure 1, top level matching attempts to achieve object level matching since top level features detect patterns at object level, middle level matching establishes correspondences at part level, finally bottom level matching achieves fine level matching through small patterns.

The main advantage of DNN flow is to utilize more targeted feature to achieve the matching goal of each level. The top level feature with semantic invariance helps to discriminate inter-class variance and stand intra-class variance. Therefore, top level feature is robust to fight against various visual variance. Even if two images of the same category are obvious similar at bottom level, high-level matching still produces helpful coarse flow field and guides low-level matching along with the reasonable direction.

For given two images I_1, I_2, and corresponding DNN feature pyramids F_1 and F_2, let $p = (x, y)$ be the grid coordinate in the feature pyramid, and $F_1(p,i)$ denotes the feature at p on the i^{th} level. Let w_i be the flow field on the i^{th} level, and $w_i(p) = (u_i(p), v_i(p))$ be the flow vector at p, where $u_i(p)$ and $v_i(p)$ are horizontal flow vector and vertical flow vector respectively.

Then, the DNN Flow's matching objective function can be formulated as:

$$E(w_i|w_{i-1},i) = \sum_p (E_D(p,w_i) + \alpha \sum_{q\in\varepsilon(p,i)} E_S(p,q,w_i) + \beta E_{SD}(p,w_i,w_{i-1})) \quad (1)$$

$$E_D(p,w_i) = |F_1(p,i) - F_2(p+w_i(p),i)| \quad (2)$$

$$E_S(p,q,w_i) = |u_i(p) - u_i(q)| + |v_i(p) - v_i(q)| \quad (3)$$

$$E_{SD}(p,w_i,w_{i-1}) = |u_i(p) - \widetilde{u}_{i-1}(p)| + |v_i(p) - \widetilde{v}_{i-1}(p)| \quad (4)$$

where E_D, E_S, E_{SD} are the data term, smoothness term and small displacement term respectively, $\varepsilon(p,i)$ is the neighborhoods of p on the i^{th} level, $(\widetilde{u}_{i-1}, \widetilde{v}_{i-1})$ is the w_{i-1} mapped to i^{th} level based on mapping of DNN. E_D measures the similarity between the correspondence features on the same level. E_S leverages the geometric prior that neighbors' flow vectors should be similar. E_{SD} uses the flow field of upper level to guide the optimization of low-level flow field.

We build a four-level pyramid to estimate dense correspondences. The DNN used for extracting DNN feature is learned by supervised back propagation on ILSVRC2012 training set, which contains eight layers with weights: five convolutional layers followed by three fully-connected layers. Three max-pooling layers are used following the first, seconde and fifth convolutional levels. The output of fifth convolutional layer is adopted as top-level feature, while the outputs of second and first convolutional layer are adopted as two mid-level features. In order to extract bottom-level feature for each pixel, the dense output of first convolutional layer is adopted as bottom-level feature through adjusting stride.

The performance of DNN Flow is demonstrated based on three experiments: rough image dense matching, fine object alignment and label transfer. The experiments are designed repectively on different datasets. Three image matching approaches, PatchMatch, SIFT FLow and DSP, are compared with DNN Flow in all experiments. The selected approaches are based on local feature or hierarchical local feature. In order to quantitatively evaluate image matching, two evaluation metrics are introduced into experiment: label transfer accuracy (**LT-ACC**) metric and intersection over union.

[1] Yan Li, Leon Gu, and Takeo Kanade. Robustly aligning a shape model and its application to car alignment of unknown pose. *PAMI*, 33 (9):1860–1876, 2011.

[2] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*. 2008.

[3] Michael Rubinstein, Ce Liu, and William T Freeman. Annotation propagation in large image databases via dense image correspondence. In *ECCV*. 2012.