Computer Vision and Image Understanding xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Hierarchical semantic image matching using CNN feature pyramid

Wei Yu[a], Xiaoshuai Sun[a], Kuiyuan Yang[b], Yong Rui[b], Hongxun Yao[*,a]

[a] School of Computer and Technology, Harbin Institute of Technology, Harbin 150090, China
[b] Microsoft Research, Beijing 100080, China

ARTICLE INFO

Keywords:
CNN feature
Image matching
Hierarchical framework
Dense correspondence
Visualization

ABSTRACT

Image matching remains an important and challenging problem in computer vision, especially for the dense correspondence estimation between images with high category-level similarity. The effectiveness of image matching largely depends on the advance of image descriptors. Inspired by the success of Convolutional Neural Network(CNN), we propose a hierarchal image matching method using the CNN feature pyramid, named as CNN Flow. The feature maps output by different layers of CNN tend to encode different information of the input image, such as the semantic information extracted from higher layers and the structural information extracted from lower layers. This nature of CNN feature pyramid is suitable to build the hierarchical image matching framework, which detects the patterns of different levels in an implicit coarse-to-fine manner. In particular, we take advantage of the complementarity of different layers using guidance from higher layer to lower layer. The high-layer features present semantic patterns to cope with the intra-class variations, and the guidance from high layers can resist the semantic ambiguity of low-layer features due to small receptive fields. The bottom-level matching utilize the low-layer features with more structural information to achieve finer matching. On one hand, extensive experiments and analysis demonstrate the superiority of CNN Flow in image dense matching under challenging variations. On the other hand, CNN Flow is demonstrated through various applications, such as fine alignment for intra-class object, scene label transfer and facial expression transfer.

1. Introduction

Image matching is a central topic in computer vision, which aims to estimate the dense pixel-level correspondences between two images. The pixel-level correspondence can solve many vision problems, such as motion estimation (Brox and Malik, 2011; Liu et al., 2011b), label propagation (Gould and Zhang, 2012; Rubinstein et al., 2012), fine-grained categorization (Gavves et al., 2015; Krause et al., 2014; 2015) and object modeling (Barnes et al., 2010b; Jiang and Yu, 2009; Li et al., 2011).

The image matching problem can be divided into two sub-problems considering the variations between two images, i.e. instance-level matching and category-level matching. The instance-level matching is considered as the simplest image matching problem, where the two images are assumed to be of the same object varied by motion, or the same scene with affine transformations, such as dynamic scenes in video sequences. The instance-level correspondence can be established even under the brightness constancy assumption, such as the classic optical flow (Bruhn et al., 2005; Horn and Schunck, 2004; Revaud et al., 2016; Weinzaepfel et al., 2013). The category-level matching is a more complicated problem, since the two images are about objects/scenes of same category with larger and more challenging variations (Hosni et al., 2013; Matas et al., 2004; Okutomi and Kanade, 1993; Qiu et al., 2014; Yang et al., 2014). These variations arise from not only the changes in illumination and viewpoint, but also the appearance variations due to the appearance of different object instances, such as cars with various shapes and colors, and cats with different poses and furs. Therefore, category-level matching focuses more on overcoming the intra-class variability in shape and other visual properties. Meanwhile, the object instances may be captured from different viewpoints, placed at different spatial locations or imaged at different scales, which make category-level matching problem extremely challenging. In this paper, we are interested in category-level image matching: aligning different instances of same object category under cluttering background.

In the matching task, the similarity of potential locations is used to estimate the correspondence between pixels in two images, which is typically measured by appearance and geometric constraint. Generally, object appearance is described by image features extracted at each location, while geometric constraint is preselected such as smoothness and small displacements (Kim et al., 2013; Liu et al., 2008; Yu et al., 2013). Obviously, the image features are required with different invariance abilities for different image matching problems. In instance-

* Corresponding author.
E-mail addresses: w.yu@hit.edu.cn (W. Yu), xiaoshuaisun@hit.edu.cn (X. Sun), kuyang@microsoft.com (K. Yang), yongrui@microsoft.com (Y. Rui), h.yao@hit.edu.cn (H. Yao).

https://doi.org/10.1016/j.cviu.2018.01.001
Received 13 April 2017; Received in revised form 17 October 2017; Accepted 3 January 2018

level matching, the intensity can be used directly to match two images, such as adjacent frames in video sequences, since there only exists small variations caused by motion. Further, the more powerful feature is helpful to achieve image matching with affine transformations (Liu et al., 2011b), such as SIFT (Lowe, 2004).

Since the training datasets with millions of labeled images are widely available (Everingham et al., 2012); (Russakovsky et al., 2014), a growing number of bigger and deeper Convolutional Neural Networks (CNN) can be effectively trained by using efficient GPU implementations (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Szegedy et al., 2015). With the rebirth of Deep CNN, a series of feature extractors stacked from low-level to high-level can be automatically learned from large-scale training data in an end-to-end manner (Simonyan et al., 2013; Zeiler and Fergus, 2014). The learned feature extractors not only achieve superior performance on the classification task, but also show great ability in handling the variations under the same category. With the gradual abstraction through multiple layers, CNN overcomes the gap between raw pixels and semantic labels. The feature extractors from different layers tend to encode different kinds of information in an image. In lower layers, the feature extractors are more close to image and contains more local structural details about the image. Thus, the low-layer features describe the simple patterns, such as edges and blobs. Meanwhile, the feature extractors from higher layers are more close to semantic categories, which care more about semantic information but less structural details. Thus, the high-layer features present the complex patterns with similar semantic, such as the object parts or even object.

Inspired by the powerful CNN architecture, we propose a hierarchal image matching method using CNN feature pyramid, named as CNN Flow. CNN Flow is established followed the complementarity of CNN features of different layers, since the similarity measuring is key to correspondence estimation. On one hand, similarity defined by high-layer feature can overcome the intra-class variability, which is helpful to achieve the coarse matching in semantic level. On the other hand, similarity defined by low-layer feature aims to achieve fine matching, which is close to hand-designed low-level feature. Although the low-layer features suffer from the problem of semantic ambiguity due to small receptive fields, the guidance from high-layer similarity is introduced to resist this issue.

In CNN Flow, we aims to achieve the matching results at different levels with more suitable feature. For ease of explanation, we aim to estimate the flow field from *source image* to *targeted image*, as illustrated in Fig. 1. The top-level matching attempts to estimate the category-level correspondence, since high-layer feature presenting semantic patterns can cope with the intra-class variations. The bottom-level matching attempts to achieve fine-grained local structure matching, while middle-level matching establishes part-level correspondences. From top to bottom, the matching of higher layer will guide correspondence establishing of lower layer on spatial neighbor and matching reliability. Even if two images of same category are obvious bottom-level similar, high-level matching still produces helpful coarse flow field and guides low-level matching along with the reasonable direction.

2. Related work

2.1. Image matching

The image matching methods aim to discover the dense correspondence relationships among images, which removes intra-class variability and canonicalizing pose. The traditional image matching approaches focused on instance-level matching, where the given images are of same scene but with slight view point changes. Generally, the objective consists of data term and smoothness term. The matching cost on pixels is defined as data term for accurate pixel correspondence estimation. In addition, dense correspondence between two images is a nontrivial problem with spatial regularity, thus the smoothness term measures the similarity of displacement for neighboring pixels. To estimate dense correspondence, image matching is casted as a graph optimization problem, where pixels are nodes, and edges between neighboring nodes reflect spatial constraints between them. In particular, the Markov Random Field (MRF) model leads a way to solve this problem through combining with powerful optimization techniques, such as graph cut (Boykov et al., 2001) and belief propagation (Sun et al., 2003). Other graph-based matching algorithms (Cho and Lee, 2012; Duchenne et al., 2011) attempt to find category-level feature matches by leveraging a flexible graph representation of images, but they commonly handle sparsely sampled or detected features due to their computational complexity.

More approaches attempt to estimate pixel-level correspondence between two images with challenging variations. The variations introduce the problem of efficiency where the spatial searching range spreads even from neighbors to the whole image. In order to avoid massive computational cost, recent approaches have to consider how to balance efficiency and effectiveness, and the coarse-to-fine strategy is a feasible way to produce acceptable correspondences matching. Patch Match algorithm (Barnes et al., 2010a) is an implicit coarse-to-fine strategy, which attempt to estimate dense correspondences based on randomized search technology. Some reliable matchings are estimated firstly, then the reliable matchings will guide nearby locations'matching. SIFT Flow (Liu et al., 2011b) pioneered the idea of dense correspondences across different scenes, which improves the flow field between two images by using SIFT features. The matching process

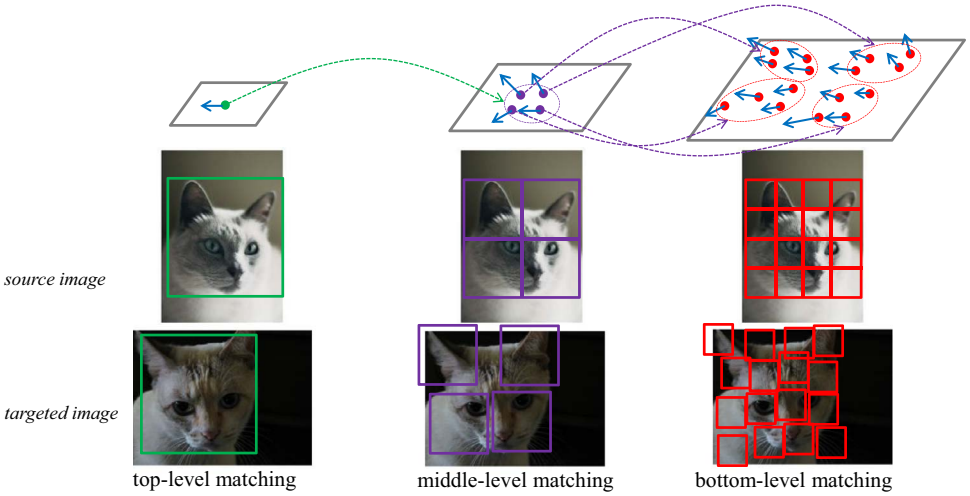

Fig. 1. A brief illustration of matching two images (*source image* and *targeted image*). Each column illustrates the image matching of different levels. In first row, parallelogram denotes the CNN feature map of *source image*, where dot represents the example feature. Line with arrow denotes the estimated flow vector of the corresponding feature, while curve with arrow denotes the guidance from high-level correspondence to low-level correspondence. In second row, rectangle shows the receptive field of CNN feature in first row. Third row shows the receptive fields of matching feature in *targeted image*.