DEEP FEATURE MATCHING FOR DENSE CORRESPONDENCE

Yang Liu *Jinshan Pan* *Zhixun Su*

School of Mathematical Sciences, Dalian University of Technology, Dalian, China

ABSTRACT

Image matching is a challenging problem as different views often undergo significant appearance changes caused by deformation, abrupt motion, and occlusion. In this paper, we explore features extracted from convolutional neural networks to help the estimation of image matching so that dense pixel correspondence can be built. As the deep features are able to describe the image structures, the matching method based on these features is able to match across different scenes and/or object appearances. We analyze the deep features and compare them with other robust features, e.g., SIFT. Extensive experiments on 5 datasets demonstrate the proposed algorithm performs favorably against the state-of-the-art methods in terms of visually matching quality and accuracy.

Index Terms— Deep feature, dense correspondence, scene matching, optical flow, handcrafted feature

1. INTRODUCTION

Establishing meaningful pixel correspondence between images is one of the fundamental problems in computer vision and image processing, which has wide applications, such as semantic motion segmentation and obstacle avoidance [1, 2]. Numerous algorithms have been proposed to promote matching accuracy including optical flow [3], stereo matching [4] and generic scene matching [1].

Existing methods mainly focus on handling images which are captured from different viewpoints but share the same image contents. To establish the correspondence, the pixel information plays a critical role in these methods as pixels are able to describe visual appearance of the object. However, pixel information is not robust to illumination changes.

To overcome this limitation, several methods use image gradient in the optical flow estimation [5]. However, these methods are less effective when there exist large displacement among different views.

Some robust features are developed in optical flow estimation to deal with large displacement, e.g., SIFT features [6]. As the features are designed to characterize the intrinsic geometry structure, the methods based on these features are able to help the estimation of the matching. Moreover, SIFT features are employed to establish the dense correspondence of different scenes [1].

We note that the aforementioned methods mainly use either intensity, gradients, or robust features, e.g., SIFT feature, to solve correspondence problems. However, the features used in the correspondence are handcrafted and limited to one specific kind of structure, e.g., intensity value is sensitive to appearance changes, SIFT feature is able to characterize local gradient structure but less effective for smooth regions.

To address these problems, we in this paper explore the deep features to model image structures to establish the dense correspondence between different scenes. The proposed deep features are more comprehensive and semantic which are able to preserve the image structures while recognizing scene and object more precisely. With the deep features, we propose a robust correspondence method which is able to deal with huge image variations among different scenes. We analyze the deep features and compare them with other robust features, e.g., SIFT. Extensive experimental results demonstrate that the proposed method is more robust and performs favorably against the state-of-the-art correspondence methods in terms of visual quality and accuracy.

2. RELATED WORK

Recent years we have witnessed significant advances in image matching [7]. One of the most popular image matching method is the optical flow method. The seminal framework of optical method is proposed by Horn and Schunk [3]. This method is further improved by [8]. However, these algorithms do not handle illuminance changes. To handle illuminance changes, Brox et al. [5] integrate gradient into the formulation. However, these aforementioned methods usually fail to the cases when there exist large displacement.

To deal with large displacement, Brox et al. [9] incorporate region descriptor or sparse descriptor [10] in the conventional optical flow model [3]. Xu et al. [11] propose an effective sparse feature matching to handle large motion problem. To further help the estimation of optical flow, recent methods [12] develop deep matching methods in the conventional methods. These methods are able to deal with large displacement to some extent. However, these methods do not establish

This work is partially supported by the NSFC (No. 61572099 and 61320106008), the National Science and Technology Major Project (No. ZX20140419 and 2014ZX04001011).

the correspondence between different scenes.

The most typical methods that handle the correspondence between different scenes are based on robust features. Liu et al. [1] develop dense SFIT features and employ optical flow model to handle the correspondence of generic scene images. In [13], deformable spatial pyramids is proposed. Most recently, Shen et al. [14] develop regional foremost to match the internet images. We note that the aforementioned methods are all based on handcrafted features. Developing a robust feature to establish the correspondence between different scenes is very important.

Different from existing methods, we develop a robust image matching method based on deep features.

3. OUR FRAMEWORK

3.1. Proposed deep features

To explore effective features for correspondence, we use the convolutional feature maps from a CNN e.g. VGG-Net [15] as our deep features, which is trained from large datasets (ImageNet [16]) and are able to integrate intensity information, gradient structure, color information, and other information. As demonstrated in [17], the features from first several layers model the low-level information, e.g., gradient and the features from latter layer models the high-level information, e.g., boundaries, and object. In this paper, we use the features from the first fourth layers as our deep features. Since the pooling operator used in the CNNs will gradually decrease the size of feature maps, we resize them using bicubic interpolation [18] so that the feature maps have the same size with input images.

To further increase the robustness and eliminate the influence of illumination intensity changes, we normalize the feature maps to the same scale space. Figure 1 shows the visualization of deep features.

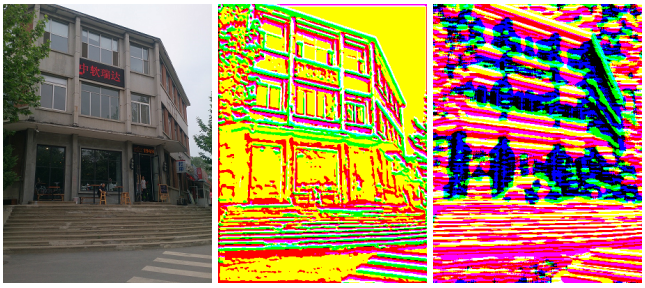
To understand the deep features and their effectiveness, we visualize the feature map for the fourth layer by mapping the features into their corresponding three principal components then display them in RGB space in Figure 1. The results show that the features preserve the image structure well. Compared to the dense SIFT features used in SIFT Flow [1], the SIFT feature has cluttered structure and color, while the CNN feature map is more structured and clear.

3.2. Proposed model

With the deep features, we estimate the dense correspondence by minimizing energy function [19]:

$$E(w) = E_{data}(w) + \lambda E_{reg}(w), \tag{1}$$

where w denotes the optical flow, the data term $E_{data}(w)$ measures the similarity between two images, and $E_{reg}(w)$ is a regularization on w. The concrete forms of $E_{data}(w)$ and $E_{reg}(w)$ are defined as



(a) Input image (b) CNN feature (c) SIFT feature

Fig. 1: Visualization of different features. We project the CNN feature and SIFT feature extracted from (a) into the RGB space and show it in (b), (c), respectively.

$$\begin{aligned} E_{data}(w) &= \sum_x \|I_1(x) - I_2(x + w(x))\|_1, \\ E_{reg}(w) &= \sum_x \gamma(|u(x)| + |v(x)|) \\ &\quad + \sum_{\tilde{x} \in N(x)} (\alpha|u(\tilde{x}) - u(x)|, t) + \\ &\quad (\alpha|v(\tilde{x}) - v(x)|, t). \end{aligned} \tag{2}$$

where x is the spatial coordinate of input feature maps I_1 and I_2, $N(x)$ denotes the neighborhood of x, and $w(x) = (u(x), v(x))$ is the flow at x, and the truncated L_1 norm is used to remove outliers and increase robustness [8, 18].

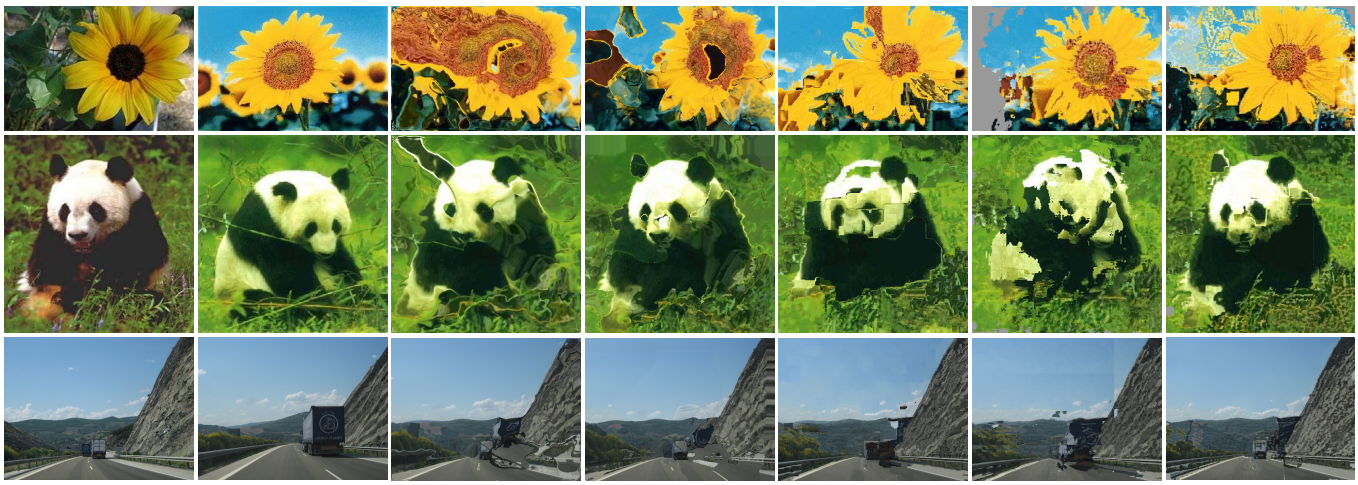
We use BP optimization method [20] to solve (2). To increase the accuracy, we adopt the commonly used coarse-to-fine strategy similar to [1].

3.3. Analysis on proposed method

In this section, we provide analysis on the proposed deep features, discuss its effect on the dense correspondence, and compare with the most related methods [10, 1, 11, 13].

We note that the most commonly used features for correspondence are based on intensity, gradient [10], or robust features [1, 13]. As the intensity information does not model the illuminance changes and gradient does not capture the structures of image scene, the methods based on these two features are less effective for the cases when there exists large displacement (Figure 2(c)). Although robust features are able to model the image structures and methods based on such robust features are able to deal with large displacement. However, these handcraft features are not able to model scale variation well thus leading to unpleasant results in matching (Figure 2(e) and (f)).

In contrast, deep features not only integrate intensity and gradient structure but also other structures and semantic information, e.g., boundaries and contours. These structures are



(a) Reference (b) Input image (c) LDOF [10] (d) MDPOF [11] (e) SIFT Flow [1] (f) DSP [13] (g) Ours

Fig. 2: Effectiveness of proposed CNN-feature in image matching. The SSIM values of the top row of (c)-(g) are 0.4617, 0.4624, 0.3620, 0.3822, and 0.5402 respectively. Our method is able to deal with the scale variation, illuminance changes, and large displacement.

able to model image structures robustly for scale variation, illuminance changes.

Figure 2 shows three examples which include scale variation, illuminance changes, large displacement among different scenes. The proposed method generates better results.

4. EXPERIMENTAL RESULTS

In this section, we present experimental evaluations of the proposed algorithm against several state-of-the-art methods for optical flow methods [10, 11] and different scene correspondence methods [1, 13]. To evaluate our algorithm, we create a dataset from the Caltech-101 [21], the LabelMe Outdoor(LMO) [22], the uni-freiburg image sequences [9], as well as some images created by ourselves and some images from SIFT Flow [1].

Approach/Metric	SSIM SCORE(total:60)	SSIM
SIFT Flow [1]	1	0.4598
DSP [13]	1	0.4183
LDOF [10]	0	0.4873
MDPOF [11]	19	0.5579
Ours	39	0.5848

Table 1: Quantitative evaluation results on 60 pairs of images. The proposed method achieves higher score under the measurement SSIM.

Implementation details: We empirically set the parameters $\alpha = 1, \gamma = 0.01, t = 30$. All of these parameters are fixed in our experiments. We adopt the VGG-Net-19 [15] trained on ImageNet for feature extraction and experimentally employ the 4-th layer as our deep feature after resizing and normalizing.

Evaluation metrics: We use SSIM as the evaluation metrics in our experiments as SSIM is capable of characterizing structure similarity across different appearances.

We randomly choose 60 pairs of images from the proposed dataset for test considering meaningful correspondence. The quantitative evaluation results are shown in Table 1 where the SSIM are employed to measure the matching quality. We count which method defeats others in each image pair under SSIM, then the score increases by one. The results demonstrate that the proposed method outperforms the state-of-the-art scene correspondence methods.

Figure 3 shows some visual examples and matching results from different methods. As the examples contain significant illuminance changes, scale variation, and large displacement, the scene correspondence methods do not generate good matching results as shown in the third and fourth row of Figure 3. In contrast, our method achieves better matching results as the proposed feature is able to model the image structures and is robust to significant illuminance changes, scale variation, etc.

Figure 4 shows more comparisons. The conventional optical flow methods, e.g., LDOF, MDPOF, fails to establish the correspondence between the reference images and input images. The results by these methods contain significant distortion and ringing artifacts. Although SIFT Flow and DSP develop robust features to establish the correspondence, the results still contain significant artifacts. In contrast, our method

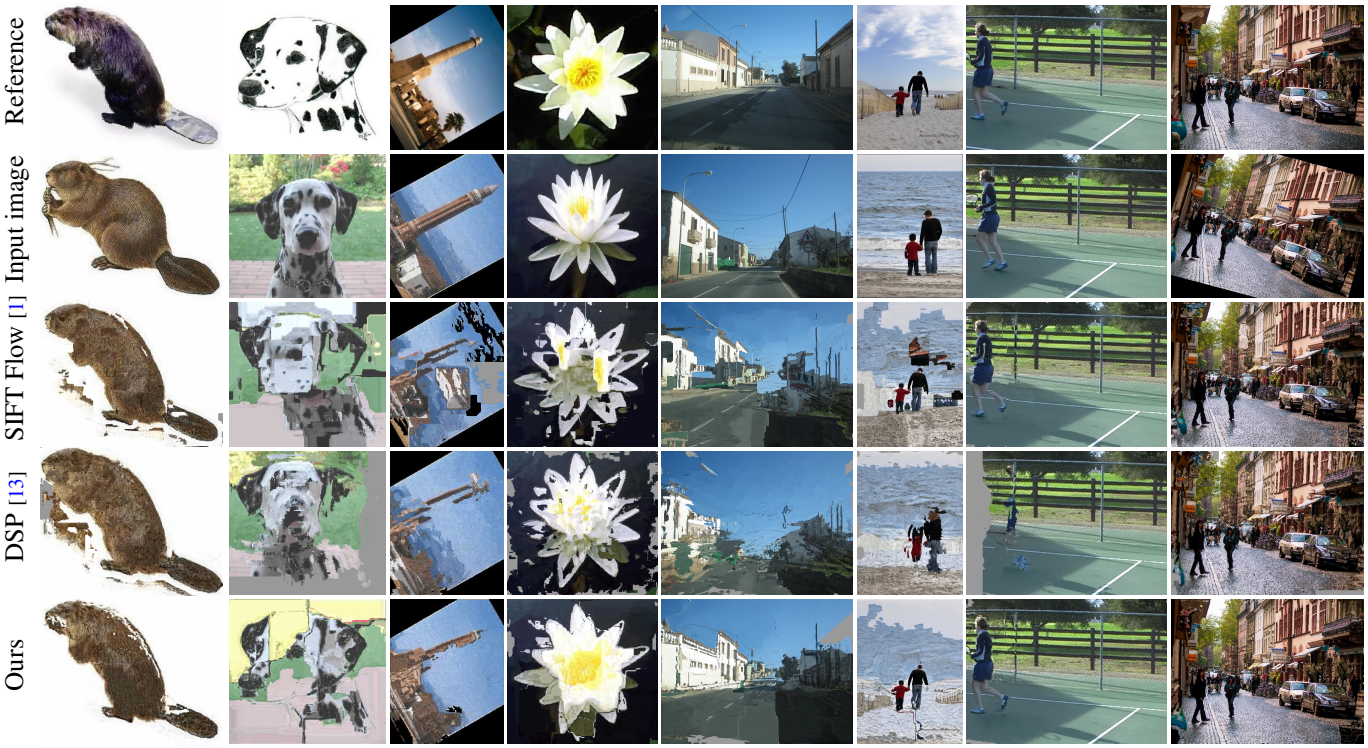


Fig. 3: Comparisons with state-of-the-art methods on the proposed dataset. Compared to SIFT Flow and DSP methods, the proposed method generates better matching results under appearance changes (column 1,2,3), viewpoint changes (column 4), drastic scene variation (column 5), background clutter (column 6), large displacement and distortion (column 7), rotation (column 8).

(a) Reference (b) Input image (c) LDOF [10] (d) MDPOF [11] (e) SIFT Flow [1] (f) DSP [13] (g) Ours

Fig. 4: Comparisons with state-of-the-art methods on the proposed dataset. The proposed method generates better results compared with the state-of-the-art optical flow methods and scene correspondence methods.

generates better results.

5. CONCLUSIONS

We propose a novel image matching method for dense correspondence based on deep features. We show that the deep features are able to model the significant appearance changes caused by deformation, abrupt motion, and occlusion and can establish the correspondence across different scenes and/or object appearances robustly. Extensive experiments testify to the superiority of the proposed method over state-of-the-art algorithms, in both qualitatively and quantitatively.

6. REFERENCES

[1] Ce Liu, J Yuen, and A Torralba, "SIFT Flow: Dense Correspondence across Scenes and Its Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, 2011.

[2] Uwe Franke, Clemens Rabe, and Stefan Gehrig, "6D-Vision : Fusion of Stereo and Motion for Robust Environment Perception," *Lecture Notes in Computer Science*, vol. 3663, pp. 216, 2005.

[3] Berthold KP Horn and Brian G Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[4] Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu, "Constant time weighted median filtering for stereo matching and beyond," *In ICCV*, pp. 49–56, 2013.

[5] Thomas Brox, Nils Papenberg, and Joachim Weickert, "High Accuracy Optical Flow Estimation Based on a Theory for Warping," *In ECCV*, vol. 4, no. May, pp. 25–36, 2004.

[6] David G Lowe, "Distinctive Image Features from," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] Zike Yan and Xuezhi Xiang, "Scene Flow Estimation: A Survey," *arXiv*, pp. 1–51, 2016.

[8] Michael J Black and P Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields," *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[9] Thomas Brox, Jitendra Malik, and C Bregler, "Large displacement optical flow," *In CVPR*, pp. 41–48, 2009.

[10] Thomas Brox and Jitendra Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 500–513, 2011.

[11] Li Xu, Jiaya Jia, and Yasuyuki Matsushita, "Motion detail preserving optical flow estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1744–1757, 2012.

[12] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid, "DeepFlow: Large displacement optical flow with deep matching," *In ICCV*, pp. 1385–1392, 2013.

[13] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman, "Deformable spatial pyramid matching for fast dense correspondences," *In CVPR*, pp. 2307–2314, 2013.

[14] Xiaoyong Shen, Xin Tao, Chao Zhou, Hongyun Gao, and Jiaya Jia, "Regional Foremost Matching for Internet Scene Images," *In SIGGRAPH Asia*, vol. 2, 2016.

[15] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *In ICLR*, pp. 1–14, 2015.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *In CVPR*, pp. 2–9, 2009.

[17] Chao Ma, Jia Bin Huang, Xiaokang Yang, and Ming Hsuan Yang, "Hierarchical convolutional features for visual tracking," *In ICCV*, vol. 11-18-Dece, pp. 3074–3082, 2015.

[18] Deqing Sun, Stefan Roth, and Michael J. Black, "Secrets of optical flow estimation and their principles," *In CVPR*, pp. 2432–2439, 2010.

[19] Philipp Krähenbühl and Vladlen Koltun, "Efficient nonlocal regularization for optical flow," *In ECCV*, vol. 7572 LNCS, no. PART 1, pp. 356–369, 2012.

[20] Alexander Shekhovtsov, Ivan Kovtun, and Václav Hlaváč, "Efficient mrf deformation model for non-rigid image matching," *Computer Vision and Image Understanding*, vol. 112, no. 1, pp. 91–99, 2008.

[21] 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," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[22] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.