Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 1, January 2018, pp. 168-177
DOI: 10.23940/ijpe.18.01.p18.168177

Learning Near Duplicate Image Pairs using Convolutional Neural Networks

Yi Zhang[a,*], Yanning Zhang[a], Jinqiu Sun[b], Haisen Li[a], Yu Zhu[a]

[a]School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, 710072, China
[b]School of Astronautics, Northwestern Polytechnical University, Xi'an, 710072, China

Abstract

In this paper, we illustrate how to learn a general straightforward similarity function from raw image pairs, which is a fundamental task in computer vision. To encode the function, inspired by the recent achievements of deep learning methods, we explore several deep neural networks and adopt one of the suitable networks to our task encoding implementation with several models on benchmark datasets UKBench and Holidays. The adopted network achieves comparable overall results and especially presents the excellent learning ability for global-similar data. Compared to previous approaches, this function eliminates the complex handcrafted features extraction, and utilizes pairwise correlation information by the jointly processing.

Keywords: near duplicate image detection; similarity function; CNN

(Submitted on November 28, 2017; Revised on December 15, 2017; Accepted on December 23, 2017)

1. Introduction

Near duplicate images detection is a widely used fundamental task in computer vision such as stereo matching, image copy detection, and partial tasks on video data. The main task of near duplicate images detection is to learn the similarity between the query image pair. It is quite challenging as the similarity learning results are easily affected by multiple factors, such as changing of viewpoints, illumination, and camera settings. Figure 1 shows several types of image pair samples.

Conventional approaches mainly rely on human designed descriptors to accomplish it. To tackle the easily affected issue above, various descriptors and matching strategies led by SIFT [8] have been developed and achieve good results. However, under these frameworks, the complicated extractions and the expensive storage cannot be ignored. Furthermore, the inner correlation from image-pair is largely under-utilized until the matching stage, which is after the feature extraction. Thus, a demand of how to learn the similarity from raw images while taking advantages of the pairwise correlation information increases.

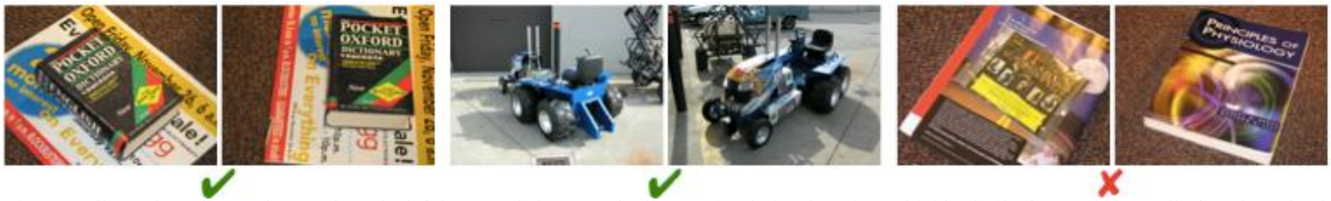


Figure 1. Illustration of several examples. The left image pair is a near duplicate pair, obviously. It is not highly similar but could be easily found out that it is just rotated. The middle image pair is an extreme sample of near duplicate image pair, since it looks totally different as the large-scale view changing. The right image pair, which looks quite similar, but actually it is non-near duplicate image pair.

* Corresponding author.
E-mail address: sophie.zhangy@gmail.com

This paper aims to address the above problems by means of learning a function and provide a straightforward solution without handcrafted features, as illustrated in Figure 2. Inspired by the success of deep learning in recent years, we use convolutional neural networks to encode the function. Thus, we explore and compare various neural network architectures to adopt a suitable model for this task.

To sum up, the contributions of this paper are as follow: (1) We learn an image pair-wise similarity jointly learning function without any handcrafted features. (2) We explore various neural networks and propose a network model to represent the function. (3) We implement the proposed model on benchmark datasets and achieve comparable results on benchmark datasets (*3.29* on UKBench, *61.2%* on Holidays) to conventional handcrafted features, and performance excellent on global-similar samples (on ColumbiaIND).

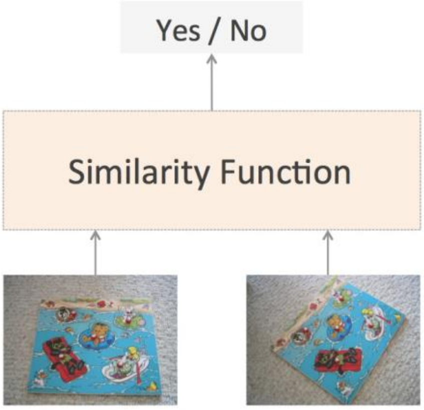


Figure 2. Simply illustration of the task. The sample visualized is a near duplicate image pair with view rotated.

2. Related Work

To measure the pairwise images similarity, previous works mainly focus on the image features extraction and analyzing those features, which can be normally categorized into traditional approaches and deep learning based approaches.

2.1. Traditional Approaches

Most existing approaches can simply conclude into representation-matching pipeline to learn the pairwise similarity, as shown in Figure 3. Generally, local representations are the main trend focus for this pipeline.

To generate local representations, interest points or regions are detected through Dog [15] or Hessian-Affine [16], etc. Subsequently, based on these interest points, SIFT [15] encodes the salient aspects of the image gradient in the neighborhood around, Fisher Vector [19] is a statistics capturing the distribution of a set of vectors, and Vector of Linearly Aggregated Descriptors (VLAD) [7] and their groups represent image in a semantically-richer mid-level. To handle non-rigid distortion, Ke [8] adopted Principal Components Analysis (PCA-SIFT), which is more distinctive and compact than the original SIFT, to represent interest points. In the further work [25], Yu designed an Affine-SIFT to address the affine sensitive problem of original SIFT. In [31], Zhang proposed to extract a binary local descriptor named Edge-SIFT to utilize spatial clues for partial-duplicate detection.

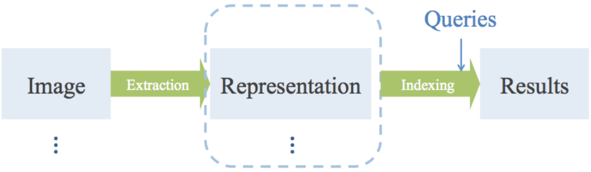


Figure 3. Illustration of conventional solution pipeline

The following stage is matching on extracted features. Based on PCA-SIFT, Ke [8] applied Locality-Sensitive Hashing (LSH) to index those descriptors to complete point set matching method. To avoid the cost in [8], Chum etc. [3, 32, 31] quantize local features into visual words using bag-of-words model and VLAD to encode and aggregate local patch statistics. To address the limitation of BoW, Zhou [33] proposed a system consists of three procedures, which contains BoW, ORGCD and verification.

2.2. Deep Learning Based Approaches

From a certain view, some deep learning based approaches are following the representation-matching strategy. The vectors output by CNN intermediate layers processing on single image can be utilized as global representation for content based image matching [2]. Further works developed several patch-level approaches [17,22]. Different from using CNN as auxiliary cues to BoW, in the further work [24], Yan complementarily integrates SIFT and CNN coequally to present an image in point-level, object level and scene-level. Subsequently, they compress combinations of several methods under different levels with simple PCA on commonly used benchmark datasets and achieve several state-of-art results. Also, CNN is well applied in matching [10] and hashing [9,11,30]. Siamese networks are introduced as an end-to-end framework to process image pairs directly. By doing so, [28, 26, 1] achieve state-of-art results. It suggests that convolutional neural networks are well suited for computing image comparing even for other real-time application.

Most approaches described above mainly rely on single image features to learn the pair-wise similarity. The strong correlation between image pairs, which is underutilized in these methods, could be very useful for near duplicate detection [12]. In the task of image-patches comparing [27], a network named as 2-channel-network allows the patch-pair correlation to be learned at the very start of training. Zbontar [28] proposed a CNN-based patches comparing approach for computing cost in narrow baseline dataset which is the top-performance then. These approaches focus on image-patches comparing, which are much smaller than general images and contain less affine-distortions.

In wide baseline, to handle the non-rigid deformations and repetitive texture in image pairs, Revaud [21] proposed the DeepMatching to tackle non-rigid motion which is hard to process for previous SIFT methods. The strategy is generating sets of 'quadrants' and 'sub-quadrants' until the minimum patch size with the position information to process non-rigid matching. In a novel work [14], the authors use image pair interpolation as the dense correspondence representation which is obtained by training a deep neural network.

3. Method

As mentioned above, we aim to process the raw images directly and jointly without human designed features extraction and output the decision of 'yes or no'. We will describe two suitable basic architectures in the following subsection.

3.1. Architectures

Siamese model From a certain view, actually, Siamese is following the traditional representation-matching strategy. As illustrated in Figure 4, Siamese has two branches that share same weights and same architectures. For example, several sets of convolutional layers, ReLU layers and Pooling layers, the output vectors of each branch are concatenated and come into the top of network to give a final 'yes or no' output. The branches can be regarded as the representation generator and the top decision layers can be viewed as the matching stage.

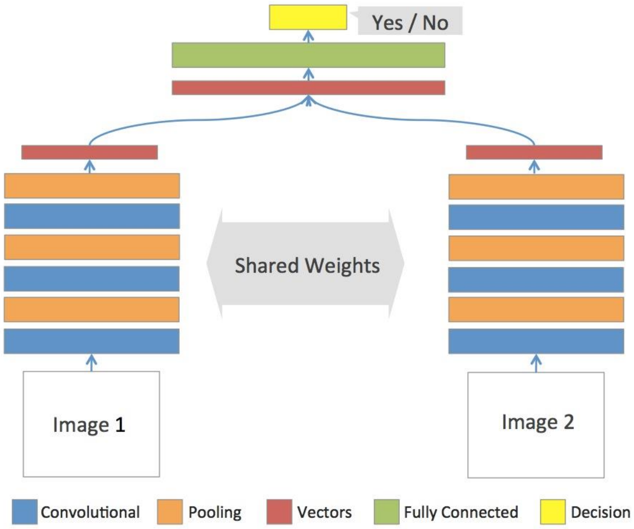


Figure 4. Illustration on architecture of Siamese Network

Double-channel model Different than Siamese, double-channel model has no concept of single image representation generator. It suggests that this model is fundamentally different from the previous methods. This model combines two images into one input, which benefits from the flexibility of deep neural networks, as shown in Figure 5. The output of the bottom intermediate layers, which could be regarded as the extracted feature, is imported into the decision layer. For instance, a fully connected linear decision layer with one output.

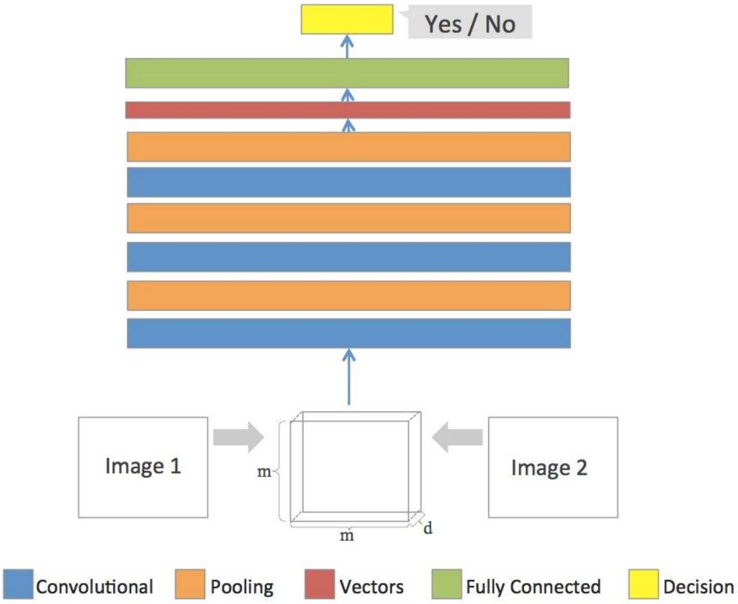


Figure 5. Illustration on architecture of Double-Channel Network

Both these networks process raw images without limitation on image channels number and give decisions directly. Naturally, they offer different trade-offs in terms of efficiency and accuracy. In section 4, we compare the validation accuracy of them. The double-channel shows higher training accuracy. From the prospective of efficiency, the double-channel is faster in training stage, while taking expensive time cost in test stage. Furthermore, the Siamese only takes into account the correlations in a limited way by sharing exactly same architectures and weights while double-channel provides greater flexibility by processing the images jointly from the beginning. Thus, we adopt the more suitable network, double-channel as the baseline for the task.

3.2. Learning

The training process in this work is strongly supervised. Near duplicate images detection using double-channel network can be viewed as a binary-classification issue.

The input $\boldsymbol{I}$ is obtained by

$$I = \begin{bmatrix} r(i_1) \\ r(i_2) \end{bmatrix}$$

where I Î $\mathbb{R}^{m´\ m´\ d}$, i_1, i_2 are the raw-image pair, $r(\cdot)$ is the resize operation, m depends on actual models applied, d refers to the channels of raw data. Specifically, in [27] d $=2$ as for they applied this model on gray scale image patches and the model is named accordingly. In our work, the data is RGB 3-channel natural images, then d=6. $\boldsymbol{I}$ is labeled as ℓ Î {0,1}. For each image pair (i_1, i_2) ∈Î, Î is the set of all pair combinations in T, T refers to the dataset. We tag each raw image i according its group δ_i , then

$$\ell = \begin{cases} d_{i_1} = d_{i_2}, \ell = 1 \\ d_{i_1} \neq d_{i_2}, \ell = 0 \end{cases}$$

With a binary-class softmax regression on the final layer, we measure the likelihood of the query image pair. Assuming the training set (with *training_size = n*) T_{tr} ={(x_1, y_1),…, (x_i, y_i), i = n}, *L* is the layers number of CNN including sets of *convolutional plus pooling* layers, the output by final layer is, λ is the user defined parameter controls the trade-off between minimizing the fit of the network to the training data, W_k is the weights of the *k*-th layer, the loss can be obtained as follow

$$Loss = -\frac{1}{n} \times \sum_{i=1}^{n} y_i' \log f\left(x_i\right) + \lambda \sum_{k=1}^{L} sum\left\|W_k\right\|^2$$

Once the training process is accomplished, the learned parameter sets can offer probability likelihood to a label for the test data. Compared to previous approaches, this model has no explicit descriptors for single image. It significantly reduces the method complexity and hardly requires human participation except labeling the data.

4. Experiments

We apply the models on the benchmark datasets UKBench [18] and Holidays [5]. The details of datasets are shown in Table 1. Images can be rotated, blurred, objects moved etc., from its near duplicate images, broadly ranging from a very large variety of scene types (natural, man-made, water and fire effects, etc.) and images are in high resolution.

Table 1. Details of Datasets

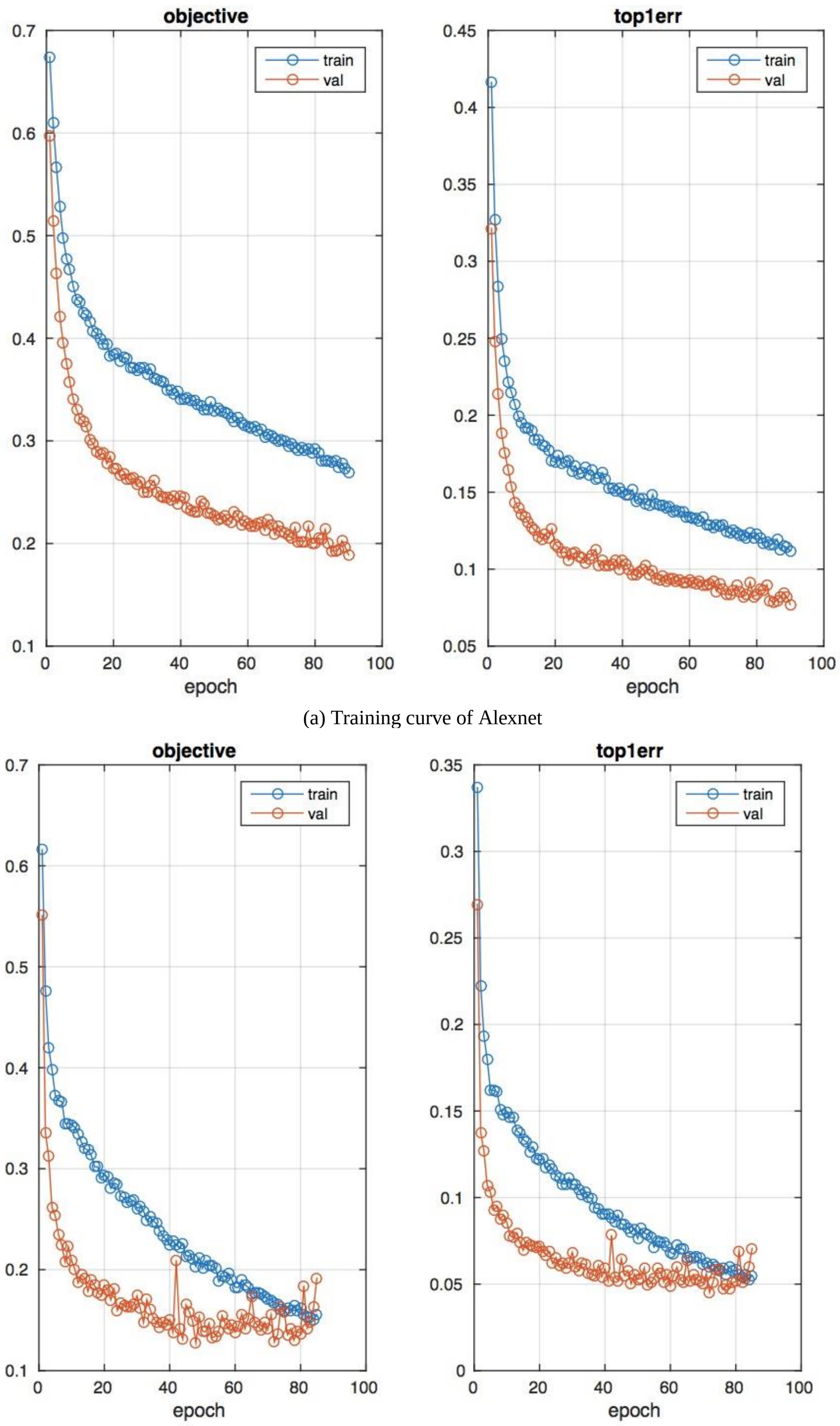
	UKBench	Holidays
Images	10200	1491
Groups	2550	500
Queries	10200	500

The input pair-set is generated from the raw image dataset. Take UKBench as an example; each group contains 4 images. Taking 2 arbitrary images in one group, there are 6 near duplicate image pairs on each group. Then, the positive samples equals to 15300. Also, the negative set consists of image pairs generated by arbitrary images from different groups, with approximately equal size to the positive set.

To value the binary classification precision of two architectures described in section 3, we apply Siamese and double-channels on UKBench (Table 2 shows the accuracy comparison of them). For higher accuracy and jointly processing strategy, we adopt double-channel for further evaluation. We apply Alexnet and VGG16 networks on the double-channel architecture with *learning_rate=*10^{-5}, *batch_size=6*, *epoch=90*. Images are resized to *m=227* and *m=224*, respectively. Both models require about 20hrs. to train, under Matlab on CPU. Testing duration depends on the queries number of datasets. It is quite time costing for evaluation on UKBench due to the 10200 ×10200 combinations.

The training curves of these models on UKBench are illustrated in Figure 6; both models gain the accuracy upper than 99% in validation phases (with train: test ≈ 1:6) with trendy convergence. The training curve of Alexnet is smoother and steadier than VGG16 curve. It suggests that Alexnet, containing 23 layers, is better fitting than VGG16 with 53 layers when training on UKBench. Networks with deeper structures are not necessarily for UKBench. These structures are large enough in the absolute number, but training samples are not sufficient for each class of positive samples and negative samples. It indicates that increasing the number of layers cannot solve the problem, and can even lead to inaccurate fitting.

To compare with conventional approaches, we generate query image pairs set consisting of pairs on each query image with other images. Then, we input all query image pairs into the trained Alexnet-double-channel network. We extract the image-pair probability values in the Softmax layer and rank them for each query image. The numeral comparison is shown in Table 3.



(a) Training curve of Alexnet

(b) Training curve of VGG16

Figure 6. The training curves presenting objective and top-error of Alexnet and VGG16 using double channel model. Both networks achieve about 99.5% on validation sets with trendy convergence. The left curves set (Alexnet) are steady and smooth. The right curves set (VGG16) appears few oscillation phenomenon and instable during late epochs.

Table 2. Validation Accuracy Comparison on Architectures

Networks	Accuracy (%)
Siamese	90.3±
Double-Channel	96.9±

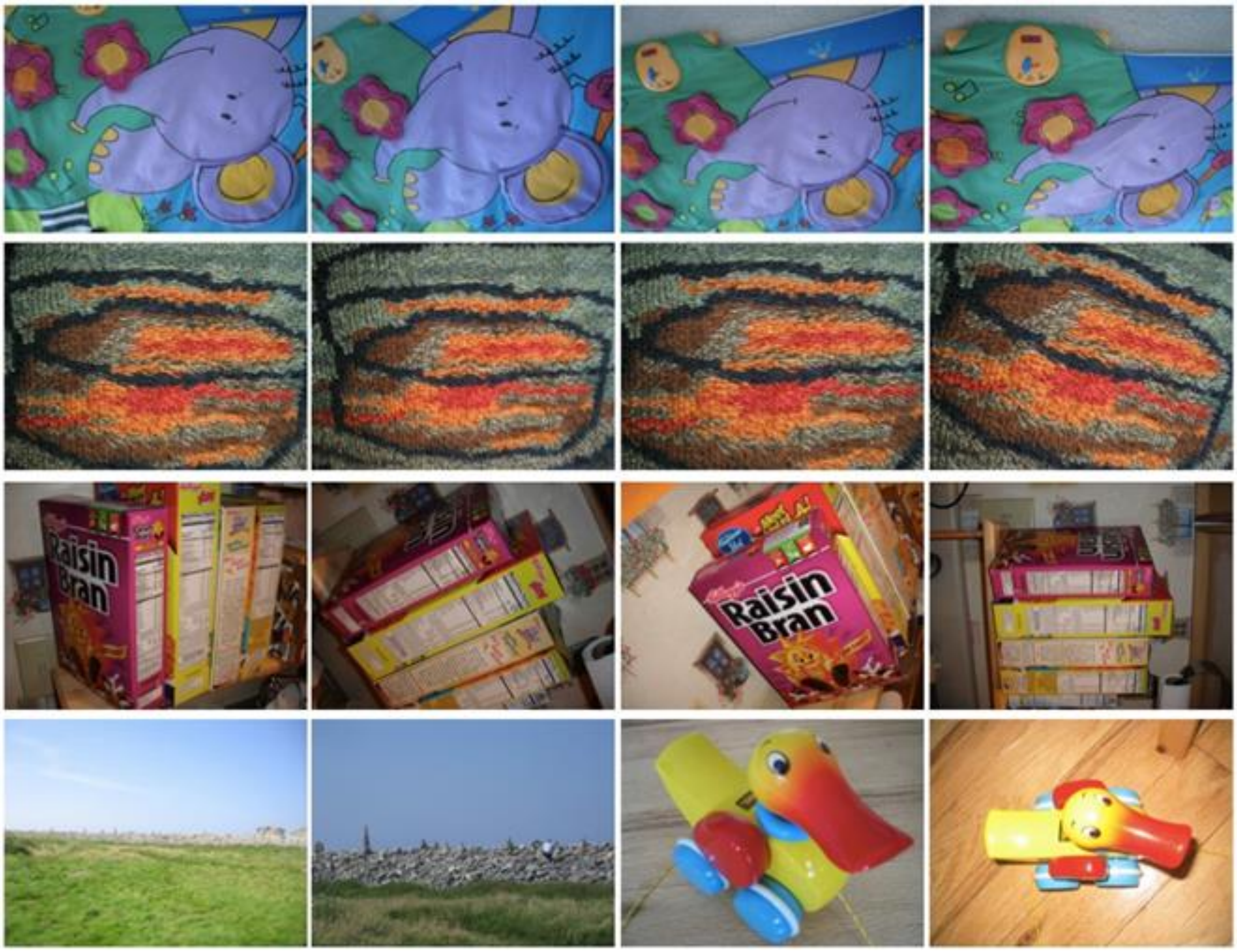


Figure 7. This figure shows variety of image pairs in the datasets. Each row illustrates one group except the last row. The first and second rows are global-similar groups with slightly changing, which human can predicate by one glance. Our model can perform peak score 4.00 and approximately 3.62 in these groups. The third row is a group with large-scale rotation and other changing. Our model achieves comparable results on these groups. The last row shows two extreme pairs with significant changing in views, illumination, and camera settings. Our model gets lower score on them.

The comparison indicates that our model achieves comparable results to conventional approaches. Through analyzing the results distribution, we conclude several typical samples illustrated in Figure 7. We find that though the overall results of the model are just comparable to classic handcrafted descriptors, it still has contribution and makes sense. The model has a strong regularity that scores high on global-similar data and scores lower for the data with obvious change of view, illumination and objects large scale motion.

It is worth mentioning the learning ability on self-pairs. The self-pairs (pair which is positive, consists of image and itself) are 100% returned as positive pairs while not involving in training stage. As mentioned above, the arbitrary images composing input positive sample are 2 images in one group. Though these two images are near duplicate images, they cannot be the exact same. Our model can learn and decide the exact-same pairs from these non-same pairs. Self-pairs can be viewed as extreme global-similar samples. This further illustrates that our model has excellent learning ability on global-similar data.

To prove the global-similar pair learning ability, we apply the Alexnet-double-channel model on ColumbiaIND [29], which is built from TREC-VID 2003 corpus. It consists of 150 near duplicate pairs and 300 non-near duplicate pairs. The near duplicate pairs from ColumbiaIND are quite similar to the duplicate image of the other with slightly differences. The numeral comparison is shown in Table 4. The results indicate the model performance well.

Table 3. Comparison of Results with Hand-Crafted Representation based Methods

Methods	Holidays(mAP)	UKBench
SIFT-BoW[18]	0.597	2.85
SIFT-Soft[4]	-	3.17
VLAD[6]	0.510	3.15
VLAD+SSR[7]	0.557	3.35
Fisher[20]	0.565	3.33
Ours	0.612	3.29

Table 4. Comparison of ERR on Columbia Dataset

	PCA-SIFT[8]	Hashing [3]	SLPM[23]	Liu[13]	Ours
ERR	0.061	0.093	0.075	0.046	0.049

Compare to the top-performance methods [26], our model has limitation to the non-global-similar data. A possible reason is that the training set generated on the datasets is biased. The global-similar positive samples are far more than the positive samples with obvious deformation. It can leads to the learned parameters more fitted to the global-similar pairs. Thus, we believe that the performance could be improved by enriching the training data [27]. In our work, we could increase the data with deformation by manually generating distorted data on original images. Another possible reason is the lack of local information targeting modules. It could be optimized using additional layers focusing on spatial transform and patch or object level processing [1, 21].

5. Conclusions

In this paper, we illustrated how to learn a general straightforward similarity function from raw image pairs. We explored several neural network models and figured out the suitable model. The adopted model achieved comparable overall results and especially showed an excellent learning ability for global-similar data. Compared to conventional approaches, the adopted model eliminates the complex handcrafted features extraction and processed images jointly.

We believe that a more sufficient training set can enhance the performance. Also, any additional layers targeting the spatial transformer would help achieve superior performance.

Acknowledgements

This work was supported by National Natural Science Foundation of China (61231016, 61572405), China 863 (2015AA016 402). Y. Zhang was supported by a Scholarship from University of Sydney.

References

1. H. Altwaijry, E. Trulls, J. Hays, P. Fua and S. Belongie, *"Learning to Match Aerial Images with Deep Attentive Architectures", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3539-3547.
2. A. Babenko, A. Slesarev, A. Chigorin and V. Lempitsky, *"Neural Codes for Image Retrieval"*, in D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, eds., *Computer Vision -- ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, Springer International Publishing, Cham, 2014, pp. 584-599.
3. O. Chum, J. Philbin and A. Zisserman, *"Near Duplicate Image Detection: min-Hash and tf-idf Weighting, Proceedings of the British Machine Vision Conference", Proceedings of the British Machine Vision Conference*, BMVA Press, 2008, pp. 50.1-50.10.
4. J. S. Hare, S. Samangooei and P. H. Lewis, *"Efficient Clustering and Quantisation of SIFT Features: Exploiting Characteristics of the SIFT Descriptor and Interest Region Detectors Under Image Inversion", Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ACM, New York, NY, USA, 2011, pp. 2:1-2:8.
5. H. Jégou, M. Douze and C. Schmid, *"Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search"*, in D. Forsyth, P. Torr and A. Zisserman, eds., *Computer Vision -- ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 304-317.
6. H. Jégou, M. Douze, C. Schmid and P. PÉREZ, *"Aggregating local descriptors into a compact image representation", 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304-3311.
7. H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez and C. Schmid, *"Aggregating Local Image Descriptors into Compact Codes"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34 (2012), pp. 1704-1716.
8. Y. Ke, R. Sukthankar and L. Huston, *"Efficient near-duplicate detection and sub-image retrieval", ACM Multimedia*, 2004, pp.

5.

9. H. Lai, Y. Pan, Y. Liu and S. Yan, *"Simultaneous feature learning and hash coding with deep neural networks", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3270-3278.
10. Y. Li, X. Kong, L. Zheng and Q. Tian, *"Exploiting Hierarchical Activations of Neural Network for Image Retrieval", Proceedings of the 2016 ACM on Multimedia Conference*, 2016, pp. 132-136.
11. K. Lin, H. F. Yang, J. H. Hsiao and C. S. Chen, *"Deep learning of binary hash codes for fast image retrieval", 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 27-35.
12. J. Liu, Z. Huang, H. Cai, H. T. Shen, C. W. Ngo and W. Wang, *"Near-duplicate Video Retrieval: Current Research and Future Trends"*, ACM Comput. Surv., 45 (2013), pp. 44:1-44:23.
13. L. Liu, Y. Lu and C. Y. Suen, *"Variable-length signature for near-duplicate image matching"*, IEEE Trans Image Process, 24 (2015), pp. 1282-96.
14. G. Long, L. Kneip, J. M. Alvarez, H. Li, X. Zhang and Q. Yu, *"Learning Image Matching by Simply Watching Video"*, in B. Leibe, J. Matas, N. Sebe and M. Welling, eds., *Computer Vision -- ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, Springer International Publishing, Cham, 2016, pp. 434-450.
15. D. G. Lowe, *"Distinctive Image Features from Scale-Invariant Keypoints"*, International Journal of Computer Vision, 60 (2004), pp. 91-110.
16. K. Mikolajczyk and C. Schmid, *"Scale & Affine Invariant Interest Point Detectors"*, International Journal of Computer Vision, 60 (2004), pp. 63-86.
17. K. R. Mopuri and R. V. Babu, *"Object level deep feature pooling for compact image representation", 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 62-70.
18. D. Nister and H. Stewenius, *"Scalable Recognition with a Vocabulary Tree", 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, pp. 2161-2168.
19. F. Perronnin and C. Dance, *"Fisher Kernels on Visual Vocabularies for Image Categorization", 2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-8.
20. F. Perronnin, Y. LIU, J. Sánchez and H. Poirier, *"Large-scale image retrieval with compressed Fisher vectors", 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3384-3391.
21. J. Revaud, P. Weinzaepfel, Z. Harchaoui and C. Schmid, *"DeepMatching: Hierarchical Deformable Dense Matching"* International Journal of Computer Vision, 120 (2016), pp. 300-323.
22. L. Xie, R. Hong, B. Zhang and Q. Tian, *"Image Classification and Retrieval Are ONE", Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ACM, New York, NY, USA, 2015, pp. 3-10.
23. D. Xu, T. J. Cham, S. Yan, L. Duan and S. F. Chang, *"Near Duplicate Identification With Spatially Aligned Pyramid Matching"*, IEEE Transactions on Circuits and Systems for Video Technology, 20 (2010), pp. 1068-1079.
24. K. Yan, Y. Wang, D. Liang, T. Huang and Y. Tian, *"CNN vs. SIFT for Image Retrieval: Alternative or Complementary?", Proceedings of the 2016 ACM on Multimedia Conference*, ACM, New York, NY, USA, 2016, pp. 407-411.
25. G. Yu and J.-M. Morel, *"ASIFT: An Algorithm for Fully Affine Invariant Comparison"*, Image Processing On Line, 1 (2011), pp. 11-38.
26. B. Z, Jure and Y. Lecun, *"Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches"*, J. Mach. Learn. Res., 17 (2016), pp. 2287-2318.
27. S. Zagoruyko and N. Komodakis, *"Learning to compare image patches via convolutional neural networks", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4353-4361.
28. J. Žbontar and Y. Lecun, *"Computing the stereo matching cost with a convolutional neural network", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1592-1599.
29. D.-Q. Zhang and S.-F. Chang, *"Detecting Image Near-duplicate by Stochastic Attributed Relational Graph Matching with Learning", Proceedings of the 12th Annual ACM International Conference on Multimedia*, ACM, New York, NY, USA, 2004, pp. 877-884.
30. R. Zhang, L. Lin, R. Zhang, W. Zuo and L. Zhang, *"Bit-Scalable Deep Hashing With Regularized Similarity Learning for Image Retrieval and Person Re-Identification"*, IEEE Transactions on Image Processing, 24 (2015), pp. 4766-4779.
31. S. Zhang, Q. Tian, K. Lu, Q. Huang and W. Gao, *"Edge-SIFT: Discriminative Binary Descriptor for Scalable Partial-Duplicate Mobile Search"*, IEEE Transactions on Image Processing, 22 (2013), pp. 2889-2902.
32. W. L. Zhao and C. W. Ngo, *"Scale-Rotation Invariant Pattern Entropy for Keypoint-Based Near-Duplicate Detection"*, IEEE Transactions on Image Processing, 18 (2009), pp. 412-423.
33. Z. Zhou, Y. Wang, Q. M. J. Wu, C. N. Yang and X. Sun, *"Effective and Efficient Global Context Verification for Image Copy Detection"*, IEEE Transactions on Information Forensics and Security, 12 (2017), pp. 48-63.

Yi Zhang received the Bachelor and Master degree in computer science from Northwestern Polytechnical University, Xi’an, China, where she is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. She visited University of Sydney in Australia as a visiting Ph.D. student from 2015 to 2017. Her current research interests include image recognition and machine learning.

Yanning Zhang received the Bachelor degree from Dalian University of Technology, Dalian, China, and the Ph.D. degree from school of Marine Engineering, Northwestern Polytechnical University, Xi'an, China, in 1996. She is currently a Professor and Dean of School of Computer Science and Engineering, Northwestern Polytechnical University. Her other titles include "Changjiang River Scholar Professor", "Chief of 973 National Defence Project" and "One of First 10,000 Technical Innovation Pioneers of CPC's Organization Department". She has a long-term work on Image Processing, Pattern Recognition, Computer Vision and Intelligent Information Processing.

Jinqiu Sun received her Bachelor, Master and Doctoral degrees from School of Astronautics, Northwestern Polytechnical University, Xi'an, China. She is currently an associate professor in School of Astronautics, Northwestern Polytechnical University. She leads some projects and funds focusing on motion background estimation and object detection etc. Her current research interests include image processing, pattern recognition and related problems.

Haisen Li received the Bachelor, Master and Doctoral degree from Northwestern Polytechnical University, Xi'an, China. His current research interests include image processing, sparse representation, and related problems.

Yu Zhu received the Bachelor, Master and Doctoral degree from Northwestern Polytechnical University, Xi'an, China. He is currently an associate research fellow in School of Computer Science and Engineering. His current research interests include image processing, image reconstruction, and related problems.