Compact Deep Invariant Descriptors for Video Retrieval

Yihang Lou[1,2], Yan Bai[1,2], Jie Lin[4], Shiqi Wang[3,5], Jie Chen[2,5], Vijay Chandrasekhar[3,4], Ling-Yu Duan[2,5], Tiejun Huang[2,5], Alex Chichung Kot[3,5], Wen Gao[1,2,5]
[1]SECE of Shenzhen Graduate School, Peking University, Shenzhen, China
[2]Institute of Digital Media, Peking University, Beijing, China
[3]Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore
[4]Institute for Infocomm Research, A*STAR, Singapore
[5]NTU-PKU Joint Research Institute

Abstract

With emerging demand for large-scale video analysis, the Motion Picture Experts Group (MPEG) initiated the Compact Descriptor for Video Analysis (CDVA) standardization in 2014. In this work, we develop novel deep-learning features and incorporate them into the well-established CDVA evaluation framework to study its effectiveness in video analysis. In particular, we propose a Nested Invariance Pooling (NIP) method to obtain compact and robust Convolutional Neural Network (CNNs) descriptors. The CNNs descriptors are generated by applying three different pooling operations to the feature maps of CNNs in a nested way towards rotation and scale invariant feature representation. In particular, the rational, advantages and performance on the combination of CNNs and handcrafted descriptors are provided to better investigate the complementary effects of deep learnt and handcrafted features. Extensive experimental results show that the proposed CNNs descriptors outperform both state-of-the-art CNNs descriptors and canonical handcrafted descriptors adopted in CDVA Experimental Model (CXM) with significant mAP gains of 11.3% and 4.7%, respectively. Moreover, the combination of NIP derived deep invariant descriptors and handcrafted descriptors not only fulfills the lowest bitrate budget of CDVA, but also significantly advances the performance of CDVA core techniques.

1. Introduction

Image/video retrieval refers to searching for the images/videos representing the same objects or scenes as the one depicted in a query image/video. Nowadays, camera equipped mobile devices have significantly facilitated mobile visual search applications. Typically, a mobile visual search system transmits query images/videos from the mobile end to the server side where searching is performed. To effectively reduce query transmission and save storage, in 2009, MPEG started the standardization of Compact Descriptors for Visual Search (CDVS) [1] which came up with a normative bitstream of descriptors as well as standardized descriptor extraction process for mobile image retrieval and augmented reality applications. In Sep. 2015, the CDVS standard was published [2].Recently, towards large-scale video analysis, MPEG has moved forward to standardize Compact Descriptors for Video Analysis (CDVA) [3]. To deal with content redundancy of temporal dimension, CDVA Experimental Model (CXM) currently adopted key frames based feature representation, which translates the problem of video retrieval into a multi-keyframes based images retrieval task, in which the frame level matching results are combined for video retrieval and matching.

2375-0359/17 $31.00 © 2017 IEEE
DOI 10.1109/DCC.2017.31

In CDVS, handcrafted local and global descriptors have been successfully standardized in a compact manner (*e.g.*, 512B, 1KB, 2KB, 4KB, 8KB, and 16KB), where local descriptors capture the invariant characteristics of local image patches and the global descriptors reflect the aggregated statistics of local descriptors. Although handcrafted descriptors have achieved great successes in CDVS standard [1] and CDVA experimental model, how to leverage promising deep learning features remains an open issue in MPEG CDVA Ad-hoc group. Many recent works [4, 5, 6, 7] have shown the advantages of deep learning features for image retrieval, which may be attributed to the remarkable success of Convolutional Neural Networks (CNNs) [8, 9]. As initial study, [4] proposed to apply the output of fully connected layers of CNNs to form a global descriptor, which outperformed canonical handcrafted descriptors like SIFT based Fisher Vector (FV) [10]. [5] has further shown that the max pooling of feature maps of CNNs (*e.g.*, the last pooling layer, named *pool5*) can generate more effective representations than using fully connected layers. Babenko *et al.* [6] demonstrated that the sum pooling of feature maps performs better than max pooling, where the PCA whitening is applied. More recently, Regional Maximum Activation of Convolutions (R-MAC) [7] was proposed, which averages max pooled features over a set of multi-scale regions of interest (ROI) in feature maps. Though R-MAC is scale invariant to some extent, it would underperform when a query object undergoes rotation.

In the context of compact descriptors for video retrieval, there exist important practical issues with CNNs descriptors. First, the raw representation of CNNs features is less invariant to geometric transformations like rotation and scale changes. Second, the compactness of CNNs descriptors has to be improved much. The more compact CNNs descriptors are, the faster the descriptors can be transmitted and compared. Third, more insights should be given on whether there is great complementarity between CNNs and conventional handcrafted descriptors for better performance.

To tackle the above issues, we make the following three contributions:

(1) We propose a novel Nested Invariance Pooling (NIP) method to obtain compact and discriminative deep invariant descriptors, which encodes translation, scale and rotation invariances into CNNs features. By incorporating NIP derived CNNs descriptors into the CDVA evaluation framework, we have achieved more than 10% mAP improvement over existing CNNs descriptors. Besides, over extensive CDVA benchmark, we have shown that the proposed deep invariant descriptors and canonical handcrafted descriptors are complementary to each other, and their combination can significantly improve the performance in video matching and retrieval.

(2) The proposed nested invariance pooling method has greatly reduced the dimensionality of CNNs features, which contributes to the compactness of video descriptors. In particular, the compact descriptors by combining NIP derived CNNs descriptors and CDVS descriptors have fulfilled the lowest bit rate of CDVA, i.e., 16K Bps.

(3) This work has been adopted by CDVA Ad-hoc group as technical reference to setup new core experiments [11], which opens up future exploration of deep-learning techniques in the video descriptor standard development. The latest core experiments involve compact deep feature representation, deep learning model compression, etc.

The rest of this paper is organized as follows. In Section 2, we brief the MPEG CDVA. In Section 3, we present the nested invariance pooling for extracting compact

deep invariant descriptors. In Section 4, we investigate the combination of CNNs descriptors (deep features) and CDVS descriptors (canonical handcrafted features) for performance improvements. Finally, we give experimental results in Section 5, and conclusions are drawn in Section 6.

2. MPEG CDVA

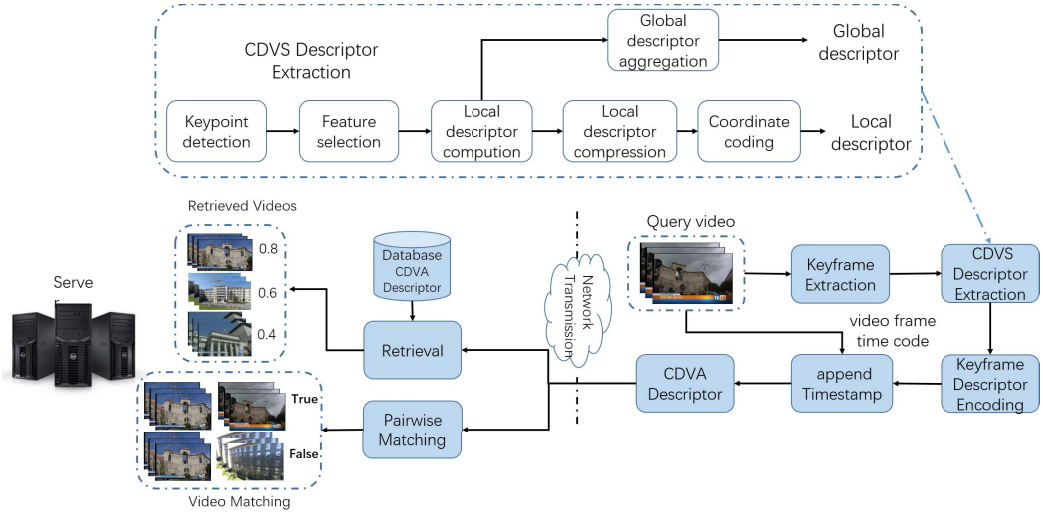

Figure 1: Illustration of MPEG CDVA evaluation framework.

MPEG CDVA [3] aims to standardize the bitstream of compact video descriptors for efficient video analysis, in which the interoperability of descriptors is expected to facilitate large-scale video processing and analysis. The requirements of CDVA standard incur two main technical aspects of descriptor compactness and robustness against geometric transformations. The first is to save transmission, storage, and computing cost. The latter is to improve the performance in the presence of rotation and scale variations. In the 115th MPEG meeting, the CDVA Experimental Model (CXM0.2) [12] was released, which relies on CDVS reference software TM14.2 [2] to implement the extractor of key frame level compact handcrafted features.

CDVS based Handcrafted Descriptors. The MPEG-7 CDVS [1] standardized descriptors are extracted from video key frames. The normative blocks of the CDVS standard are illustrated in Fig. 1, involving the extraction of local and global descriptors. For local descriptors, a low-complexity transform coding is applied to compress the SIFT descriptors. For global descriptors, selected raw local descriptors are aggregated into a Scalable Compressed Fisher Vector (SCFV), with competitive matching accuracy as well as extremely low memory footprint. In particular, CDVS supports six operating points from 512 B to 16KB to target different bandwidth constraints. Readers are referred to [1] for more technical details. In CDVA CXM0.2, the 4KB descriptor of CDVS is adopted to implement the frame-level handcrafted features, in which the operating point of 4KB empirically produces nice trade off between performance and bit rates.

CDVA Descriptors Evaluation Framework. In Fig. 1, the evaluation framework details the pipeline of CDVA descriptors extraction, transmission, and video analysis (i.e., retrieval and matching). At client side, color histogram comparison is

applied to identify key frames for each video clip. The standardized CDVS descriptors (including global and local descriptors) are extracted from key frames, which can be packed to form CDVA descriptors [13]. The temporal redundancy can be further removed by inter-frame coding. At server side, the same extraction process is applied to database videos to generate CDVS features, which follow the index structure of CDVS reference software. Accordingly, video matching and retrieval is performed given the CDVA descriptors of query videos. Readers are referred to Section 5 for more details of the experimental setup in CDVA evaluation framework. It is worthy to note that the subsequent deep invariant descriptors will be incorporated into the CDVA framework to improve performance. The complementary effects of handcrafted and deep learnt features will be carried out in this framework.

3. Compact Deep Invariant Global Descriptors

Translation Invariance of CNNs. Existing well-known CNNs architectures like AlexNet [8] and VGG-16 [9] share a common building block: a succession of convolution and pooling operations, which in fact provides a way to incorporate local translation invariance. For instance, a convolutional filter learned cat face always responses to an image with cat face no matter where it is located in the image, and pooling operation (*e.g.*, max pooling in AlexNet and VGG-16) over the activation feature maps further captures the most salient feature of the cat face, which is naturally invariant to object translation.

Invariance Theory in a Nutshell. Recently, Anselmi and Poggio [14] proposed the Invariance theory trying to understand invariance property of more complicated transformations underneath neural networks. Basically, the Invariance theory predicts that an invariant descriptor for a given image $x \in E$ is computed in relation with a predefined template $t \in E$ (*e.g.*, convolutional filter in CNNs) from the distribution of the dot products $D_{x,t} = \{< g.x, t > \in \mathbb{R} | g \in G\} = \{< x, g.t > \in \mathbb{R} | g \in G\}$, where g denotes a transformation performed either on images or templates (denotes with a dot (.)) within a group G, such as rotations (G_R), translations (G_T) and scales (G_S).

In practice, invariant descriptors are derived by any form of statistical moments (*e.g.*, mean, max, standard deviation, etc.) computed from the distribution $D_{x,t}$. More details on mathematical proofs are referred to [14].

One may note that the convolution-pooling operations in CNNs is strictly compliant with the invariance theory, where convolution operation over translated images is equivalent to $< g.x, t >$, and pooling operation is in line with statistical moments computation. Recent work on face verification [15] and music classification [16] successfully applied this theory to practical applications.

Nested Invariance Pooling. Towards descriptors invariant to translation, scale and rotation for video retrieval, we proposed a Nested Invariance Pooling (NIP) method which is inspired by the invariance theory.

We consider both AlexNet and VGG-16, which are pre-trained on ImageNet classification dataset. We build our NIP descriptors starting from the already locally translation invariant *pool5* feature maps. Given an input video key frame, the activation feature maps output by *pool5* layer is $W \times H \times C$, where W and H respectively

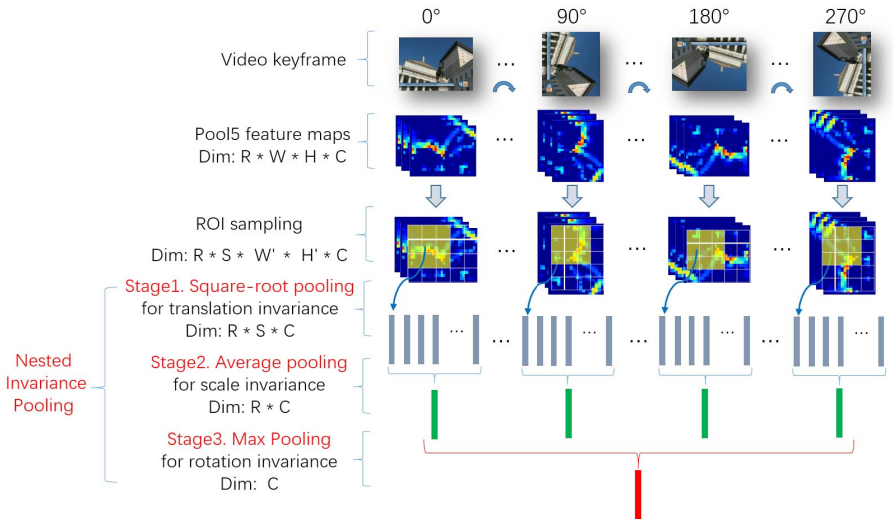

Figure 2: Nested Invariance pooling on feature maps from *pool5* layer of VGG-16 network.

denote width and height of feature map, and C the number of feature maps (e.g., $C = 256$ for AlexNet and 512 for VGG-16).

Transformations $g \in G$ are applied to the input image x by varying the output of the *pool5* feature maps, denoted as $f_i(g.x)$ for i_{th} feature map with size $W \times H$. The transformation groups in this work are translations G_T, rotations G_R and scale changes G_S. Accordingly, the NIP pooling scheme with $G \in \{G_T, G_R, G_S\}$ is:

$$\mathcal{X}_{G,i,n}(x) = \left(\frac{1}{m} \sum_{j=0}^{m-1} f_i(g_j.x)^n \right)^{\frac{1}{n}}, \tag{1}$$

where m represents the number of transformations in group G. The pooling operation has an order parameter n defining the statistical moments, where $n = 1$ is average pooling while $n \to +\infty$ on the other extreme is max-pooling. $n = 2$ is analogous to square-root pooling. The corresponding global descriptors are obtained after each pooling step by concatenating the moments for all feature maps:

$$\mathcal{X}_{G,n}(x) = (\mathcal{X}_{G,i,n}(x))_{0 \le i < C}. \tag{2}$$

Equation 1 shows the pooling strategy for single type of transformation invariance, and the work in [16] has shown that it is possible to chain multiple types of transformation invariances. We apply this principle to our NIP descriptors by making them invariant to several transformations. For instance, following scale invariance with average ($n = 1$) by rotation invariance with max-pooling ($n \to +\infty$) is done by:

$$\max_{j \in [0, m_s - 1]} \left(\frac{1}{m_t} \sum_{i=0}^{m_t - 1} f_i(g_{r,j}.g_{s,i}.x) \right). \tag{3}$$

Fig. 2 shows the extraction pipeline for compact deep invariant global descriptors using NIP. Given an input image, we rotate it by R times and extract *pool5* feature maps ($W \times H \times C$) for each rotated image. Then we perform multi-scale uniform ROI sampling on each feature map, resulting in a 5-D structure ($R \times S \times W^{'} \times H^{'} \times C$), where S denotes the number of sampled ROIs with varied scales. Subsequently, NIP

performs global n-norm pooling over translations ($W^{'} \times H^{'}$), then scales (S) and finally rotations (R) in a nested way, resulting in a C-dimensional global descriptor. One may note that the choice of n impacts retrieval performance, we empirically found pooling with increasing moments works well, *e.g.*, start with square-root pooling ($n = 2$) for translations and average pooling ($n = 1$) for scales, end with max pooling ($n = +\infty$) for rotations. We leave theoretical analysis of pooling moments in future work.

Finally, NIP descriptors can be further improved by post-processing techniques such as PCA whitening [6, 7]. In particular, NIP is firstly L2 normalized, followed by PCA projection and whitening with a pre-trained PCA matrix. The whitened vectors are L2 normalized and compared with cosine similarity.

4. Complementary nature of CNNs and CDVS descriptors

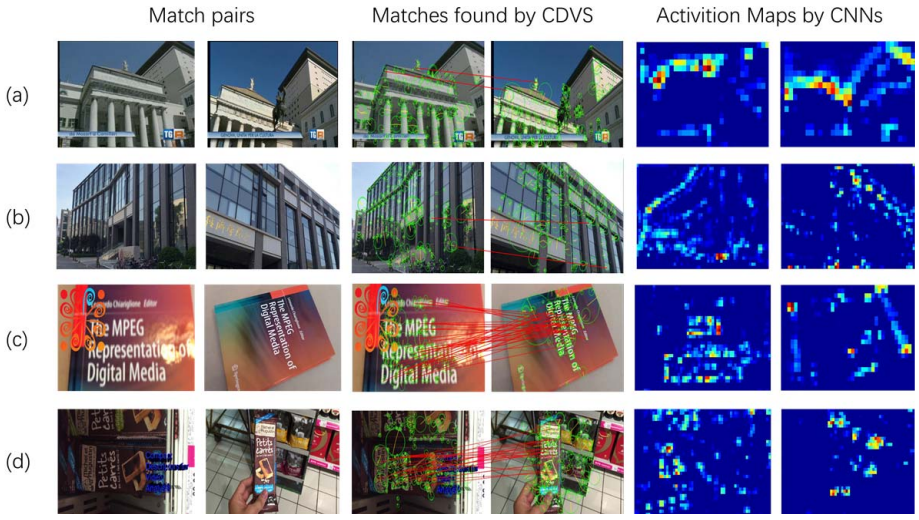

Figure 3: Key frame matching examples to illustrate the strength and weakness of deep CNNs and CDVS descriptors. In (a) and (b), CNNs work well but CDVS fails, while in (c) and (d) CDVS work but CNNs fail.

We step forward to analyze the strength and weakness of deep CNNs descriptors in the context of image matching/retrieval, compared to CDVS descriptors built upon handcrafted local invariant features. Let's discuss the room of fulfilling complementary effects between CNN-based descriptors and handcrafted CDVS descriptors:

- The CNNs descriptor is extracted in a dense manner, while CDVS descriptor works on local patches at sparse interest points detected by LoG or DoG.
- CDVS primarily work in textured regions as the interest point detector fires around blobs. CNN-based feature maps excel in low-texture regions.
- As shown in [17], CNNs feature maps in shallow layers represent Gabor like filters, and complex low level features are represented by individual neuron spikes at deeper layers. CDVS works well in compressing and aggregating shallow features, while CNNs representation tends to aggregate deeper and richer features.

Fig. 3 shows matching examples of CDVS and CNNs descriptors. Building images contain large untextured blocks like walls, and contain bursty repetitive small-scale features like windows, where CNNs descriptors are more robust.

In this work, we propose to leverage the benefits of both CNNs and CDVS descriptors. Instead of simply concatenating the NIP derived deep descriptors and CDVS handcrafted descriptors, we apply the weighted similarity scores as follows:

$$S_{total}(r,q) = S_c(r,q) * \alpha + (1-\alpha)S_t(r,q), \tag{4}$$

where α is the weighting factor. S_c and S_t represent the matching score of NIP and CDVS descriptors, respectively. In this work, α is empirically set to 0.75.

5. Experimental Results

Experimental setup: We compare NIP derived descriptors with the CDVA Experiment Model (CXM0.2) baseline as well as the recent works using CNNs descriptors [6][7]. All experiments are conducted on Tianhe HPC platform, where each node is equipped with 2 processors (24 cores, Xeon E5-2692) @2.2GHZ, and 64GB RAM. **Dataset:** The MPEG CDVA [1] dataset includes 9974 query and 5127 reference videos, which contain large objects (*e.g.*, buildings, landmarks), small objects (*e.g.*, paintings, books, products) and scenes (*e.g.*, interior or natural scenes). For retrieval experiments, 8476 videos with more than 1000 hours (about 1.2 million key frames) in terms of user-generated content, broadcast are used as distracters. For matching task, there are 4693 matching and 46930 non-matching pairs.

Figure 4: Examples from the CDVA dataset.

Evaluation Metrics: For video retrieval, the performance is evaluated by the mean Average Precision (mAP) as well as the precision at a given cut-off rank R for a single query (Precisian@R). For matching, the performance is evaluated by the True Positive Rate (TPR) and Jaccard Index. In particular, the TPR at 1% False Positive Rate (FPR) is setup in experiments, and the temporal localization accuracy of a video pair is calculated by Jaccard Index,

$$JI = \frac{[T_{start}, T_{end}] \bigcap [T'_{start}, T'_{end}]}{[T_{start}, T_{end}] \bigcup [T'_{start}, T'_{end}]}, \tag{5}$$

where $[T_{start}, T_{end}]$ denotes the ground truth and $[T'_{start}, T'_{end}]$ is the matched interval.

5.1 Performance Comparison of CNNs and CDVS Descriptors

The comparison experiments work in the CDVA evaluation framework. Table 1 shows NIP has significantly improved the retrieval performance over pool5 (CNNs features)

[1]The MPEG CDVS/CDVA evaluation framework including test dataset is available upon request at http://www.cldatlas.com/cdva/dataset.html

by 18.1% in mAP and 17.5% in Precisian@R. The matching performance improvements are 9.7% on TPR and 7.0% on Localization accuracy. Moreover, the compression ratio of NIP derived deep descriptors reaches up to 300 compared with the raw pool5 descriptors. Compared with the state-of-the-art R-MAC, +5% mAP improvements is achieved, which is mainly attributed to the improved robustness against the rotations or scale changes in the dataset (the key frames capture an object or scene from different angles). Note that CXM 0.2 provides the baseline performance of handcrafted CDVS descriptor including local and global descriptors. The proposed NIP derived deep invariant descriptors produce significant performance improvements on retrieval (4.7% on mAP) and matching (5.3% on TPR) over CXM 0.2.

Table 1: Comparison of NIP with state-of-the-art CNNs and CDVS descriptors

	mAP	Precisian@R	TPR@FPR=0.01	Localization Accuracy	Descriptor Size
CXM0.2	0.721	0.712	0.836	0.544	4 KB
Pool5	0.587	0.561	0.782	0.527	600 KB
R-MAC	0.713	0.681	0.870	0.583	2 KB
NIP	0.768	0.736	0.879	0.597	2 KB

5.2 Combination of NIP CNNs and CDVS Descriptors

To validate the effectiveness of combining NIP CNNs and CDVS descriptors, we set up three groups of retrieval and matching experiments for comparison:

(1) NIP CNNs descriptors. For matching, by simply thresholding NIP matching score we determine the matched interval between two video clips. For retrieval, the result is obtained by sorting NIP matching scores (re-ranking is not applied).

(2) NIP CNNs descriptors plus CDVS local descriptors. For matching, if the NIP matching score exceeds a threshold, then we use CDVS local descriptors for verification. For retrieval, NIP matching score is used to select the top 500 candidates list, and then we use CDVS local descriptors for reranking.

(3) Combination of NIP CNNs and CDVS global descriptors. For both matching and retrieval, the similarity score is defined as the weighted sum of matching scores of NIP and CDVS global descriptors. If the score exceeds a threshold, then we record the matched interval. Specifically, there is no reranking in retrieval.

As for NIP CNNs descriptors generation, we adopt AlexNet and VGG-16. Typically, larger network tends to generate more discriminative feature maps since the layers can go deeper (VGG-16 is larger than Alexnet). Dimensions of NIP CNNs descriptors of VGG-16 and AlexNet are 512 and 256 respectively. If each dimension occupies 4 bytes, the sizes of NIP CNNs descriptors are 1KB for AlexNet and 2KB for VGG-16. Note that the size of both global and local descriptors is 2KB in CXM0.2.

As shown in Table 2, the improvements of NIP+CDVS global descriptors are significant compared to CXM0.2. Specifically, for VGG-16 network, mAP improvements exceeds 10%. With FPR=1%, TPR get more than 5% improvements.

It is worthy to mention that, the incorporation of reranking with CDVS local descriptors (NIP + CDVS local) degrades the performance. Since we only use local descriptors to rerank the top 500 candidates, large perspective or lighting condition

Table 2: Performance of combining NIP CNNs and CDVS descriptors

	mAP	Precisian@R	TPR@FPR=0.01	Localization Accuracy	Descriptor Size
CXM0.2	0.721	0.712	0.836	0.544	4 KB
NIP VGG-16	0.768	0.736	0.879	**0.597**	2 KB
NIP VGG-16 +CDVS local	0.754	0.741	0.841	0.552	4 KB
NIP VGG-16 +CDVS global	**0.826**	**0.803**	**0.886**	0.583	4 KB
NIP Alex	0.670	0.641	0.804	0.571	1 KB
NIP Alex +CDVS local	0.728	0.718	0.834	0.549	3 KB
NIP Alex +CDVS global	0.772	0.751	0.823	0.567	3 KB

variations may greatly affect the representation of handcrafted descriptors, which could damage the ranking quality from the initial retrieval.

Overall, it is shown that a combination of CDVS and NIP CNNs descriptors can greatly improve performance and well demonstrates the positive complementary effects of handcrafted descriptors and deep invariant descriptors.

5.3 Discussion

On the recent 116th MPEG meeting in Oct. 2016, MPEG CDVA Ad-hoc group has adopted the proposed scheme into core experiments [11] for investigating more practical issues when dealing with deep learning features in the well-established CDVA evaluation framework. First, it is valuable to study how to further improve the performance by optimizing the combination of deep features and handcrafted features. Second, further compressing NIP derived CNNs descriptors (*e.g.*, binary codes) are worth investigating, like Weighted Component Hash [18], DeepHash [19]. Third, as CNNs incur huge number of network model parameters (say, over 10 millions), how to effectively and efficiently compress the CNNs models is a promising direction.

6. Conclusion

We have proposed a compact and discriminative CNNs descriptor for video retrieval, which is robust to multiple geometric transformations. The proposed novel pooling method can greatly reduce the CNNs feature size and improve the geometric invariance of deep descriptors, which works on the intermediate feature maps of CNNs in a nested way. Experimental results demonstrate that the NIP derived CNNs descriptors significantly outperforms CDVS and state-of-the-art CNNs descriptors, with comparable or even smaller descriptor size. More importantly, the promising video matching and retrieval performance by combining NIP deep invariant descriptors and the state-of-the-art standardized handcrafted descriptors CDVS has provided valuable insights for the ongoing CDVA standardization efforts.

Acknowledgments. This work was supported by grants from National Natural Science Foundation of China (U1611461, 61661146005, 61390515) and National High-tech R&D Program of China (2015AA016302). This research is partially supported

by the NTU-PKU Joint Research Institute, that is sponsored by a donation from the Ng Teng Fong Charitable Foundation. Ling-Yu Duan is the corresponding author.

References

[1] L.-Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, "Overview of the MPEG-CDVS standard," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 179–194, 2016.

[2] G. C. Stavros Paschalakis, Gianluca Francini, "Test Model 14: Compact Descriptors for Visual Search," *ISO/IEC JTC1/SC29/WG11/W15372*, 2011.

[3] "Call for Proposals for Compact Descriptors for Video Analysis (CDVA)-Search and Retrieval," *ISO/IEC JTC1/SC29/WG11/N15339*, 2015.

[4] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European Conference on Computer Vision*. Springer, 2014, pp. 584–599.

[5] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36–45.

[6] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1269–1277.

[7] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] V. Chandrasekhar, J. Lin, and O. Morère, "A practical guide to cnns and fisher vectors for image instance retrieval," *Signal Processing*, vol. 128, pp. 426–439, 2016.

[11] M. B. Werner Bailer, Massimo Balestri, "Description of core experiments in cdva," *ISO/IEC JTC1/SC29/WG11/W16510*, 2016.

[12] W. B. Massimo Balestri, Miroslaw Bober, "Cdva experimentation model (cxm) 0.2," *ISO/IEC JTC1/SC29/WG11/W16274*, 2015.

[13] D. M. Chen, M. Makar, A. F. Araujo, and B. Girod, "Interframe coding of global image signatures for mobile augmented reality," pp. 33–42, 2014.

[14] F. Anselmi and T. Poggio, "Representation learning in sensory cortex: a theory," Center for Brains, Minds and Machines (CBMM), Tech. Rep., 2014.

[15] Q. Liao, J. Z. Leibo, and T. Poggio, "Learning invariant representations and applications to face verification," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, 2013.

[16] C. Zhang, G. Evangelopoulos, S. Voinea, L. Rosasco, and T. Poggio, "A deep representation for invariance and music classification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6984–6988.

[17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[18] L.-Y. Duan, J. Lin, Z. Wang, T. Huang, and W. Gao, "Weighted component hashing of binary aggregated descriptors for fast visual search," *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 828–842, 2015.

[19] J. Lin, O. Morere, V. Chandrasekhar, A. Veillard, and H. Goh, "Deephash: Getting regularization, depth and fine-tuning right," *arXiv preprint arXiv:1501.04711*, 2015.