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[3,5],
Lingyu 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 the emerging requirements for video analysis, the motion picture experts group (M-PEG) initiated the compact descriptor for video analysis (CDVA) standard in 2015. The current development of CDVA technologies rely on SIFT descriptors, while deep-learning based representation has not been taken into account. The semantic information contained in deep learning features would be highly desirable for video analysis. In this paper, we incorporate deep-learning features into the CDVA evaluation framework and study its effectiveness in combination with handcrafted features. We propose a nested invariance pooling (NIP) method to obtain compact and robust CNN descriptors. In our method, the CNN descriptors are generated by applying three different pooling operations on intermediate feature maps of CNNs in a nested way. With such nested pooling, the final compact CNN descriptors are robust to rotation and scale variation. The experimental results show that the proposed CNN descriptors can outperform the state-of-the-art CNN descriptors and the handcrafted descriptors of CDVA with 11.3%, 4.7% in terms of mAP. Meanwhile the bitrate costs can be reduced to 1/300 of original CNN features (*e.g., 'pool5' in VGG-16*). Besides, we also explore the combination of CNN and handcrafted descriptors, which can achieve 10.5% mAP improvements compared to handcrafted descriptors.

1. Introduction

Image/video retrieval refers to searching the image/videos from a database representing the same objects or scenes as the one depicted in a query image/video. Nowadays, camera equipped mobile devices are facilitating mobile visual search applications. Typically, a mobile visual search system transmits query images/videos from the mobile end to the server to perform search. With the aim of saving the transmission and storage costs, there is great concern on how to efficiently represent the image/video content and facilitate the search procedure by transmitting the descriptors. In view of this, in 2009 MPEG started the standardization of Compact Descriptors for Visual Search (CDVS)[1] that provides a standardized bitstream of descriptors and the descriptor extraction process, ensuring the interoperability between mobile and server toward mobile image-based retrieval applications. In 2015, MPEG released the official version of CDVS. The core building blocks of CDVS consist of local descriptor compression and global descriptor aggregation, which generate compressed SIFT descriptors and scalable compressed Fisher Vector (SCFV).

Recently, MPEG moved forward to look into the standardization of Compact Descriptors for Video Analysis (CDVA)[2]. Intuitively, CDVA is a more challenging problem than CDVS as the introduced temporal dimension. One of the possible solutions to handle the temporal domain information is first detecting keyframes from both query and database videos, and then translating video retrieval into a keyframe based image retrieval task. Finally, the keyframe level matching results are combined together for each database video, and then ranked at video level to identify the truly matched one.

Besides the handcrafted descriptors adopted by CDVS/CDVA, recent work in [3, 4, 5, 6, 7, 8] started to apply deep learning descriptors for image retrieval, motivated by the remarkable success of Convolutional Neural Networks (CNNs) [9, 10]. First initial study [3, 4] proposed using the output of fully connected layer of CNNs as a global descriptor, and promising results have been reported over traditional handcrafted descriptors like Fisher Vector (FV) based on local SIFT. Recent work [5, 6] show that max pooling of intermediate feature maps of CNNs (*e.g.*, the last pooling layer named *pool5* from here on) is an effective representation and higher performance can be achieved compared to using the fully connected layers. Babenko et al. in their recent work [7] show that Average Pooling (*i.e.* Sum Pooling) of intermediate feature maps performs better than Max Pooling when the representation is PCA whitened.

In the context of compact descriptors for video retrieval, there are remaining issues for importing CNNs descriptors. Firstly, the compactness of CNNs descriptors is a critical issue. The more compact the CNNs descriptors are, the faster the video descriptors can be transmitted and compared. Secondly, the original representation of CNNs lacks invariance to other geometric transformations like rotation and scale variation. Very recent work [8] proposed Regional Maximum Activation of Convolutions (R-MAC), which is the state-of-the-art CNN based method for image retrieval, to compute average of max pooled features over a set of multi-scale regions of feature maps. R-MAC is invariant to scale in some extent, while its performance quickly degrades when the objects in the query or database are rotated. Finally, it is an open question on whether or not CNNs can totally replace handcrafted CDVS descriptors (e.g., SCFV [11]) for image/video retrieval.

To tackle the above issues, there are three contributions in this work:

(1) We propose a Nested Invariance Pooling (NIP) method to obtain the compact and discriminative global descriptions based on CNNs. In particular, NIP progressively encodes translation, scale and rotation invariances into CNNs, resulting in better quality descriptors that achieve more than 10% mAP improvements over state-of-the-art CNN descriptors in CDVA datasets. Moreover, we show that the proposed NIP descriptors and the conventional global descriptors based on handcrafted features are complementary to each other, and the combination of them is validated to achieve the state-of-the-art performance in video matching and retrieval tasks.

(2) The pooling strategy in the NIP implicitly achieves feature compression by significantly reducing the dimension of the CNN features. In particular, compared to the raw CNN features, the compression ratio is 1/300. Such compact representation can bring further benefits to the feature transmission and video analysis.

(3) As the first effort in proposing the deep learning based feature descriptors in the on-going CDVA standard, this work not only leads to an effective way to further improve the video analysis performance, but also provides useful evidence on the

complementary properties of the deep learned and handcrafted features, which are very helpful to the future development of CDVA.

The rest of paper is organized as follows. In Sect.2, we introduce the MPEG-CDVA Evaluation Framework. In Sect.3, we propose compact deep invariant descriptors, including Nested Invariance Pooling method and complementary descriptors with CNNs and handcrafted features. We discuss the experimental results of proposed approaches in Sec.4, and conclusions are drawn in Sect.5

2. MPEG-CDVA Evaluation Framework

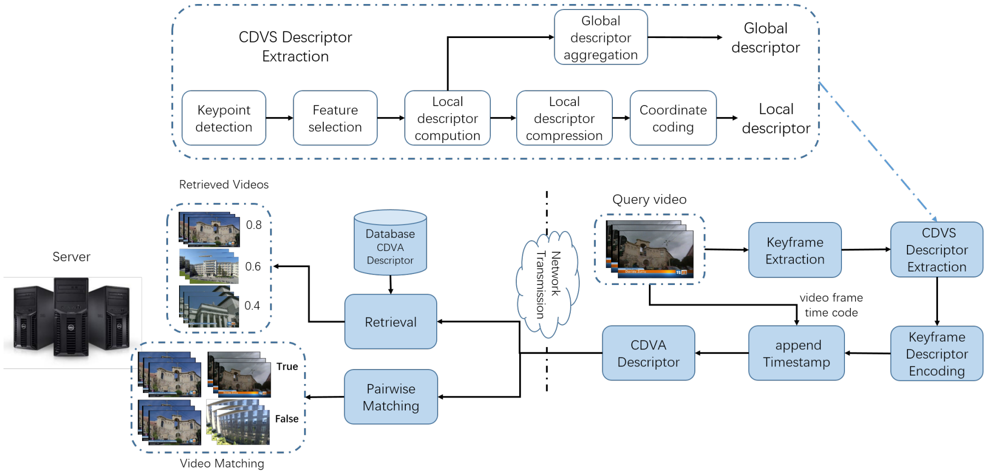

Figure 1: Illustration of the current development of CDVA application. The pipeline shows that the CDVA descriptors are extracted from a input video at client side and transmitted to the server side for retrieval or matching tasks.

The on-going MPEG-CDVA standard [2] aims to standardize the bitstream of compact video descriptors for efficient and interoperable video analysis. The requirements of CDVA descriptors contain two aspects: Firstly, the descriptors should be compact enough due to the bandwidth constraints in mobile application. Secondly, the descriptors should be equipped with invariance capabilities to handle geometric transformations such as rotation and scale variations in real-world scenarios. At the 115th MPEG meeting, the CDVA Experimental Model (CXM0.2) was released, which was implemented based on CDVS reference software TM14.2.

Compact Descriptor for Visual Search. The MPEG-7 CDVS [1] standardized descriptors are extracted for the key frames in CDVA. The normative blocks of the CDVS standard is illustrated in Fig. 1, which include the extraction of local and global descriptors. The local descriptor conveys the detailed visual characteristics and the global descriptor summarizes the feature distribution of the whole image. For local descriptor, a low-complexity transform coding scheme is adopted to compress the SIFT descriptor. For global descriptor, the selected raw local descriptors are aggregated into a scalable compressed fisher vector (SCFV), such that high matching accuracy can be achieved with extremely low memory requirements for mobile visual search. In particular, CDVS supports six operating points from 512 B to 16KB to meet different bandwidth constraints. In CXM0.2, the 4KB operating point is used due to its good balance between the performance and bitrate [1].

Compact Descriptor for Video Analysis In Fig. 1, the evaluation framework details the pipeline of CDVA descriptors extraction, transmission, and video analysis. At the client side, color histogram comparisons are used for keyframe identification. Subsequently, CDVS descriptors (including global and local descriptors) are generated from the keyframes and packed into the CDVA descriptors. At the server side, identical operations are performed for video descriptors extraction and video analysis is carried out given the received CDVA descriptor of the query videos.

For video pairwise matching, the matching score is the maximum score as the product of the local and global descriptor matching: $score = \max s_{i\_local} * s_{i\_global}$. For video retrieval, the global matching score is used to get the initial set, then the local descriptor matching score is used to obtain the final retrieved videos.

3. Compact Deep Invariant Global Descriptors

3.1 Nested Invariance Pooling (NIP)

In this work, we consider two successful networks, named AlexNet [9] and VGG-16 [10] that separately corresponds to small and large scale networks, for the construction of compact deep invariant descriptors. Both networks are pre-trained on ImageNet ILSVRC classification. The feature maps in CNN encodes the maximum "local" response of each of the convolutional filters [8], thus we consider the last convolutional layer (*i.e.*, *pool5* in VGG-16). The activation feature maps output by *pool5* is $w*h*c$, where w, h and c denote width and height of each feature map, and the number of feature maps. In our experiments, we fix the input image size as $640*480$, and the corresponding feature map size is $w = 20$ and $h = 15$. These feature maps are fed as the input to NIP.

We perform approximate pooling method to reduse the dimensionality of features. Let us denote i as the i_{th} feature map of *pool5* layer, $f_i(x)$ as the corresponding unit's output and m as the number of pixels in i_{th} feature map. Accordingly, the pooling scheme can be formulated as:

$$X_{t,i,n}(x) = \frac{1}{m}(\sum_{j=0}^{m-1} f_i(t_i.x^n))^{\frac{1}{n}}, \tag{1}$$

where t_i is the pooling parameter. The pooling method will vary by the different value of n. In particular, $n = 1$ means average pooling, and $n = 2$ is analogous to standard deviation and we call it squareroot pooling. When n tends to $+\infty$, it means max-pooling.

To equip NIP descriptors with several transformation invariances, we need perform pooling operate one after the other. For instance, following scale invariance with average ($n = 1$) by translation invariance with max-pooling ($n \rightarrow +\infty$) is done by:

$$\max_{j\in[0,m_t-1]}(\frac{1}{m_s}(\sum_{i=0}^{m_s-1} f_i(t_{i+1}, t_i, x)) \tag{2}$$

The pipeline of nested invariance pooling is visualized in Fig. 2, where r and s separately represents number of rotations and scales, and $w' * h'$ represents the size of sampled regions. The nested invariance pooling includes 5 steps. Firstly, we extract c feature maps. Then, we sample regions $(w' * h')$ with s different positions

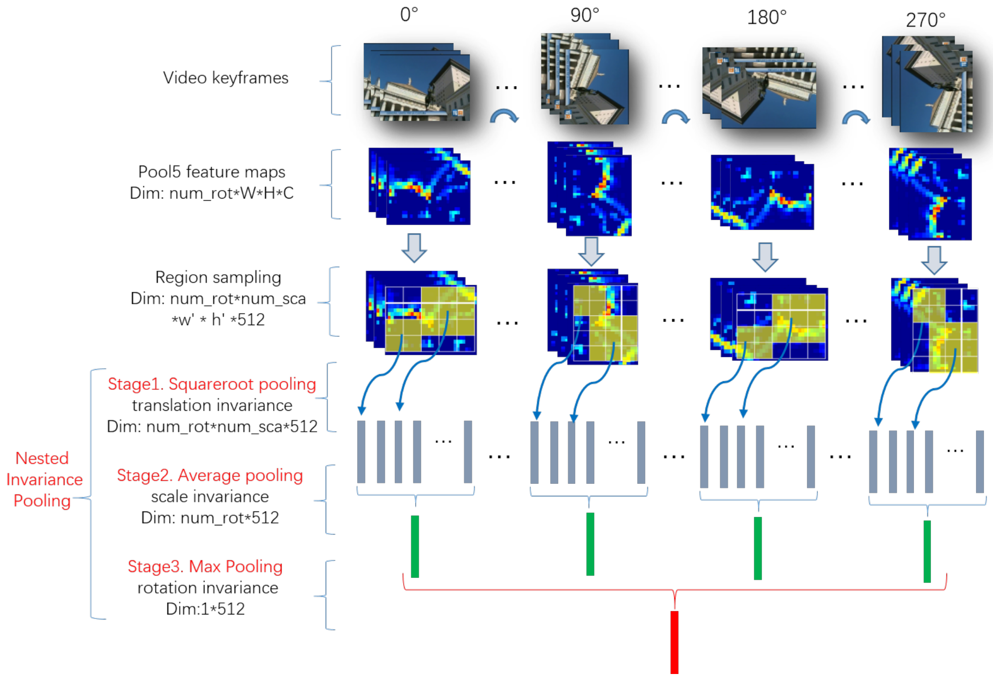

Figure 2: Nested Invariance pooling on *pool5* layer of VGG-16.

and scales in each feature maps and perform squareroot pooling. After this, the features' robustness to translation invariance are improved. The corresponding global image descriptors are obtained after each pooling step by concatenating the individual features: $X_{t,n}(x) = (X_{t,i,n}(x))_{0 \leq i < c}$. Then the pooled features is firstly L2 normalized, subsequently PCA transformed and whitened with a pre-trained PCA matrix. After this, we perform average pooling on all the region vectors. With aggregating different scales regions, the features are encoded with certain scale invariance. Nest, to make the features robust to rotation changes, we rotate the input images with certain degrees and repeat the above operations. Finally, max pooling is performed over features $(r * c)$ of different directions to get the final compact descriptor (c).

Feature vectors computed from different positions, scales and rotated regions in feature maps are separately encoded with translation, scale and rotation invariances. To equip those three invariances to the final compact descriptor, we adopt a nested pooling way. The basic idea of our nested invariance pooling exactly coincides with invariance theory [12, 13], which states how to compute group invariant transformations with feedforward neural networks. According to the theory, convolution or pooling operation is basically an inner product, and if the filters t or feature maps f_i in convolution or pooling stage is invariant to a specific transformation, the inner product of them $t \times f_i \rightarrow f_{i+1}$ after convolution or pooling operation can also obtain corresponding invariance capabilities. Namely, the output f_{i+1} has the same invariance capabilities as f_i.

Empirically, we found that the nested pooling method can represent the image in a highly compact way. Compared to *pool5* features, the final descriptors' compression rate is up to 1/300 (from $w' * h' * c$ to c).

3.2 Complementary Descriptions with CNNs and handcrafted features

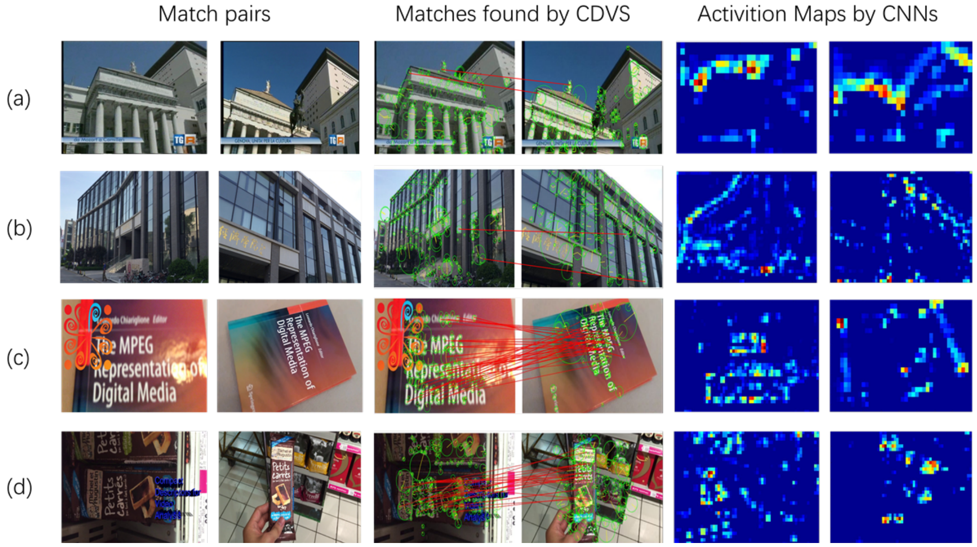

Figure 3: Examples illustrate the strength and weakness of CNNs and CDVS descriptors for keyframe matching. Each row respectively shows a match pair, matches found by CDVS based image matching after geometric consistency check, and activation maps (*pool5*) produced by CNNs that used as the input for NIP aggregation. (a) and (b) denote the success cases for CNNs but failure cases for CDVS, while (c) and (d) are in the opposite way.

We step forward to analyze the strength and weakness of CNNs descriptors in the context of image matching/retrieval, compared to CDVS descriptors built upon handcrafted local invariant features like SIFT.

On the one hand, thanks to the built-in design of scale and rotation invariances in SIFT detector, CDVS descriptors are robust to a wide range of rigid scale distortions as well as rotation distortions within 360 degrees, both in 2D plane. Compared to SIFT, NIP aggregated from activation maps by CNNs relatively preserves limited capabilities of transformation invariance, which probably fails to handle severe scale/rotation changes. As shown in Fig. 3, there exists strong rotation distortion in (c) and severe scale distortion in (d). Dense matches by CDVS descriptors show that they are not sensitive to these distortions. Due to rotation and scale changes, the highest neuron responses (brightest pixels) in CNNs activation maps comes from different parts of the objects, implying that the pooled features over these activation maps are mismatched. NIP can alleviate this issue to some extent, but still underperform compared with SIFT.

On the other hand, the invariance property of SIFT cannot be perfectly guaranteed in case of out-of-plane scale/rotation distortions in 3D plane, such as viewpoint changes when capturing building pictures, which probably leads to drastic appearance/shape changes. Furthermore, the burstiness of similar SIFTs in images with repeated structures (e.g. buildings) would cause redundancy, and confuse geometric consistency check. A typical example is shown in Fig. 3 (a) and (b), we observe the number of matches found by CDVS descriptors is very few. On the contrary, C-NNs are tolerant of out-of-plane transformations, and good at capturing image-level salient object features and ignoring bursty visual elements. As seen in Fig. 3 (a) and (b), the highest neuron responses in CNNs activation maps are always located on the same parts of the buildings even there is extreme viewpoint change in (b). Their similarities can be further enlarged by NIP.

In view of the phenomenon that CDVS and CNNs are complementary to each other, we propose a simple combination strategy to better capture the benefits of both CNNs and CDVS descriptors. In particular, we adopt a weighted sum fusion method in similarity scoring stage instead of simply concatenating the NIP descriptors and CDVS descriptors. It can be formulated as follows:

$$S_{total}(r,q) = S_c(r,q) * \alpha + (1-\alpha)S_t(r,q), \tag{3}$$

where α is the weight. S_c and S_t represent the matching score of NIP and matching score of CDVS descriptors, respectively. In the experiments, the parameter α is empirically set to be 0.75.

4. Experimental Results

Experimental setup: We compare NIP with the CDVA Experiment Model (CXM0.2) baseline as well as the recent works [7][8] that also propose compact CNN descriptors with the pooling method. All experiments are conducted on Tianhe HPC platform, where each node is equipped with 2 processors (2x12 cores, Xeon E5-2692V2) @2.2GHZ, and 64GB RAM.

Dataset: The MPEG-CDVA dataset [14] includes 9974 query and 5127 reference videos. As shown in Fig. 4, those videos 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 keyframes) in terms of user-generated content, broadcast are used as distracters. For matching task, there are 4693 matching and 46930 non-matching pairs.

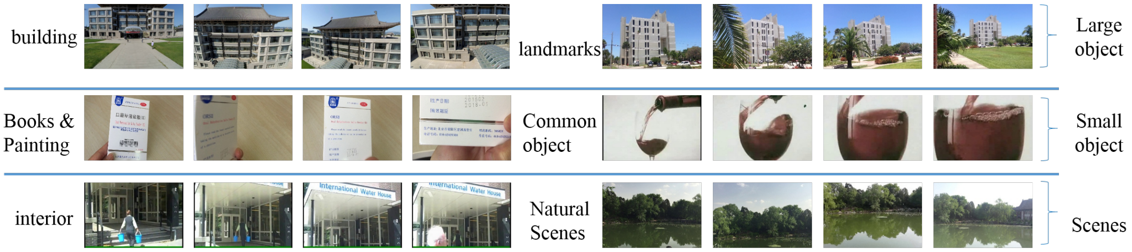

Figure 4: Examples illustrate the CDVA dataset

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,

$$JI = \frac{[T_{start}, T_{end}] \bigcap [T'_{start}, T'_{end}]}{[T_{start}, T_{end}] \bigcup [T'_{start}, T'_{end}]}, \tag{4}$$

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

4.1 Performance Comparison of Deep Descriptors

The results are obtained by incorporating the proposed scheme into the CDVA evaluation framework. The similarity scores of NIP descriptors are evaluated by L2 Norm distance. Table 1 shows that NIP improves performance over pool5 by 18.1% on mAP and 17.5% on Precisian@R. The matching performance improves 9.7% on TPR and 7.0% on Localization accuracy. Compared with R-MAC, more than 5% improvements on mAP can be achieved, the reason of which can be attributed to the rotations or scale changes in the dataset (objects taken under different angles). CXM 0.2 provides the baseline performance of the combination of CDVS global and local descriptors. Our proposed NIP descriptors present significant performance improvements on retrieval (4.7% on mAP) and matching (5.3% on TPR) tasks.

Table 1: Performance comparison between baseline and NIP

	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	0.597	2 KB
NIP	0.768	0.736	0.879	0.597	2 KB

4.2 Combination of NIP and SCFV

In order to validate the efficiency of the combination of NIP and CDVS descriptors, we set up three experiments for comparison:

(1) Only NIP descriptors. For matching, if NIP matching score is larger than a given threshold, then we record the matched interval. For retrieval, the result is obtained by sorting NIP matching score without reranking operation.

(2) NIP descriptors for retrieval and CDVS local descriptors for rerank. For matching, if the NIP matching score exceeds the given threshold, then we use CDVS local descriptors for further matching. 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 and CDVS global descriptors. For both matching and retrieval, the score is defined as the weighted sum of matching score of NIP and CDVS global descriptors. If the score exceeds the threshold, then we record the matched interval. Specifically, there is no reranking operation in retrieval.

As for NIP descriptors generation, we adopt AlexNet and VGG-16. Typically, the feature of larger scale network tends to be more discriminative since the layers can go deeper (VGG-16 is larger than Alexnet). Note that those two models are only pretrained on ImageNet dataset and no finetuning is performed on CDVA dataset. Dimensions of NIP descriptors of VGG-16 and AlexNet are 512 and 256 respectively. Assuming that a float occupies 32bit, such that the size of NIP is 1KB for AlexNet and 2KB for VGG-16. Both global and local descriptors' size are 2KB in CXM0.2.

It is observed that the improvements of NIP+CDVS global descriptors are quite significant. Specifically, for VGG-16 network, mAP improvements exceeds 10%. Under FPR=1% constraint, the TPR can also get more than 5% improvements.

Table 2: Performance of the combination of NIP 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

Compared to only NIP descriptors, the incorporation of rerank with CDVS local descriptors (NIP + CDVS local) degrades the performance in terms of mAP, TPR, and localization accuracy. Since we only use local descriptors to rerank the top 500 candidates returned by NIP, in some cases large perspective or lighting condition variations can greatly affect the representation of handcrafted descriptors. Therefore some semantic similar references rank higher by NIP will be moved behind by local descriptors reranking.

The experimental results show that the reasonable combination of CDVS and NIP can greatly improve the performance and integrates the advantages of both of CNN and handcrafted descriptors.

4.3 Discussion

Our large scale experiments also demonstrate that the CNNs descriptors and traditional handcrafted descriptors are complementary. The performance gains on combination of them are quite significant compared to either of them. It also explains the handcrafted descriptors are still necessary in feature representation. Besides, in Table 3, the complexity cost of combination of NIP and global descriptors is close to original CXM0.2.

Table 3: Complexity comparison between baseline and NIP

	CXM0.2	NIP VGG-16	NIP VGG-16 +CDVS local	NIP VGG-16 +CDVS global
Retrieval Time Cost(s) (Per query)	38.63	9.15	19.97	39.45

In the future, we aims to compress the high quality NIP descriptors with proper binarization method, such as Weighted Component Hash [15], Affinity Preserving Quantization [16] or more sophisticated method DeepHash [17], as the binarized representation has obvious advantages in Hanming distance based fast retrieval and wireless transmission.

5. Conclusion

In this paper, we have proposed a compact CNN descriptor for image/video retrieval, which is robust to multiple geometric transformations. The novelty of this paper lies in that, three different pooling operations which aim to efficiently reduce the feature size and incorporate global geometric invariance, are applied on the intermediate feature maps of CNN in a nested way. The experimental results demonstrate that the nested invariance pooling can significantly reduce the bitrate of features and obtain better performance. Additionally, the combination of NIP and handcrafted descriptors can fuse the high-level and low-level vision characteristics for more efficient representation. The results and insights of this paper provide valuable information for the future development of MPEG-CDVA standardization.

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

[3] 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.

[4] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[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. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "A baseline for visual instance retrieval with deep convolutional networks," *arXiv preprint arXiv:1412.6574*, 2014.

[7] 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.

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

[9] 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.

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

[11] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision.* Springer, 2010, pp. 143–156.

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

[13] 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.

[14] "Evaluation framework for compact descriptors for video analysis - search and retrieval," *ISO/IEC JTC1/SC29/WG11/N15338*, 2015.

[15] 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.

[16] Z. Wang, L.-Y. Duan, T. Huang, and G. Wen, “Affinity preserving quantization for hashing: A vector quantization approach to compact learn binary codes,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[17] 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.