Group Invariant Deep Representations for Image Instance Retrieval

Olivier Morère[*,1,2], Antoine Veillard[*,1], Lin Jie[2]
Julie Petta[3], Vijay Chandrasekhar[2], Tomaso Poggio[4]

Abstract—**Most image instance retrieval pipelines are based on comparison of vectors known as global image descriptors between a query image and the database images. Due to their success in large scale image classification, representations extracted from Convolutional Neural Networks (CNN) are quickly gaining ground on Fisher Vectors (FVs) as state-of-the-art global descriptors for image instance retrieval. While CNN-based descriptors are generally remarked for good retrieval performance at lower bitrates, they nevertheless present a number of drawbacks including the lack of robustness to common object transformations such as rotations compared with their interest point based FV counterparts.**

In this paper, we propose a method for computing invariant global descriptors from CNNs. Our method implements a recently proposed mathematical theory for invariance in a sensory cortex modeled as a feedforward neural network. The resulting global descriptors can be made invariant to multiple arbitrary transformation groups while retaining good discriminativeness.

Based on a thorough empirical evaluation using several publicly available datasets, we show that our method is able to significantly and consistently improve retrieval results every time a new type of invariance is incorporated. We also show that our method which has few parameters is not prone to overfitting: improvements generalize well across datasets with different properties with regard to invariances. Finally, we show that our descriptors are able to compare favourably to other state-of-the-art compact descriptors in similar bitranges, exceeding the highest retrieval results reported in the literature on some datasets. A dedicated dimensionality reduction step –quantization or hashing– may be able to further improve the competitiveness of the descriptors.

Index Terms—**Invariance theory, group invariance, deep learning, convolutional neural networks, image instance retrieval, global image descriptors, compact descriptors.**

I. Introduction

Image instance retrieval is the discovery of images from a database representing the same object or scene as the one depicted in a query image. The first step of a typical retrieval pipeline starts with the comparison of vectors representing the image contents known as *global image descriptors*. Good quality descriptors is key to achieving good retrieval performances.

Deep learning neural networks are fast becoming the dominant approach for image classification due to their remarkable performance for large scale image classification [1], [2]. In their recent work, Babenko et al. [3] propose using representations extracted from Convolutional Neural Nets (CNN) as a global descriptor for image retrieval, and show promising initial results for the approach. In our recent work [4], [5], we also show how stacked Restricted Boltzmann Machines (RBM) and supervised fine-tuning can be used for generating extremely compact hashes from global descriptors obtained from CNNs for large scale image-retrieval.

While CNN based descriptors are progressively replacing Fisher Vectors (FV) [6] as state-of-the-art descriptors for image instance retrieval, we have shown in our recent work thoroughly comparing both types of descriptors [7] that the use of CNNs still presents a number of significant drawbacks compared with FVs. One of them is the lack of invariance to transformations of the input image such as rotations: the performance of CNN descriptors quickly degrade when the objects in the query and the database image are rotated differently.

In this paper, we propose a method to produce global image descriptors from CNNs which are both compact and robust to such transformations. Our method is inspired from a recent invariance theory (subsequently referred to as *i-theory*) for information processing in sensory cortex [8], [9], [10]. The theory is an information processing model explaining how feedforward information processing in a sensory cortex can be made robust to various types of signal distorsions. In particular, it provides a practical and mathematically proven way for computing invariant object representations with feedforward neural networks.

After showing that CNNs are compatible with the *i-theory*, we propose a simple and practical way to apply the theory to the construction of global image descriptors which are robust to various types of transformations of the input image at the same time. Through a thorough empirical evaluation based on multiple publicly available datasets, we show that our method is able to significantly consistently improve retrieval results while keeping dimensionality low. Rotations, translations and scale changes are studied in the scope of this paper but our approach is extensible to other types of transformations.

II. Related Work

Since the winning submission of Krizhevsky et al. in the ImageNet 2012 challenge [1], deep CNNs are now considered as the mainstream go-to approach for large-scale image classification. They are also known to achieve state-of-the-art results with many other visual recognition tasks such as face recognition [11], [12], pedestrian detection [13] and pose estimation [14].

* O. Morère and A. Veillard contributed equally to this work.
[1] Université Pierre et Marie Curie
[2] A*STAR Institute for Infocomm Research
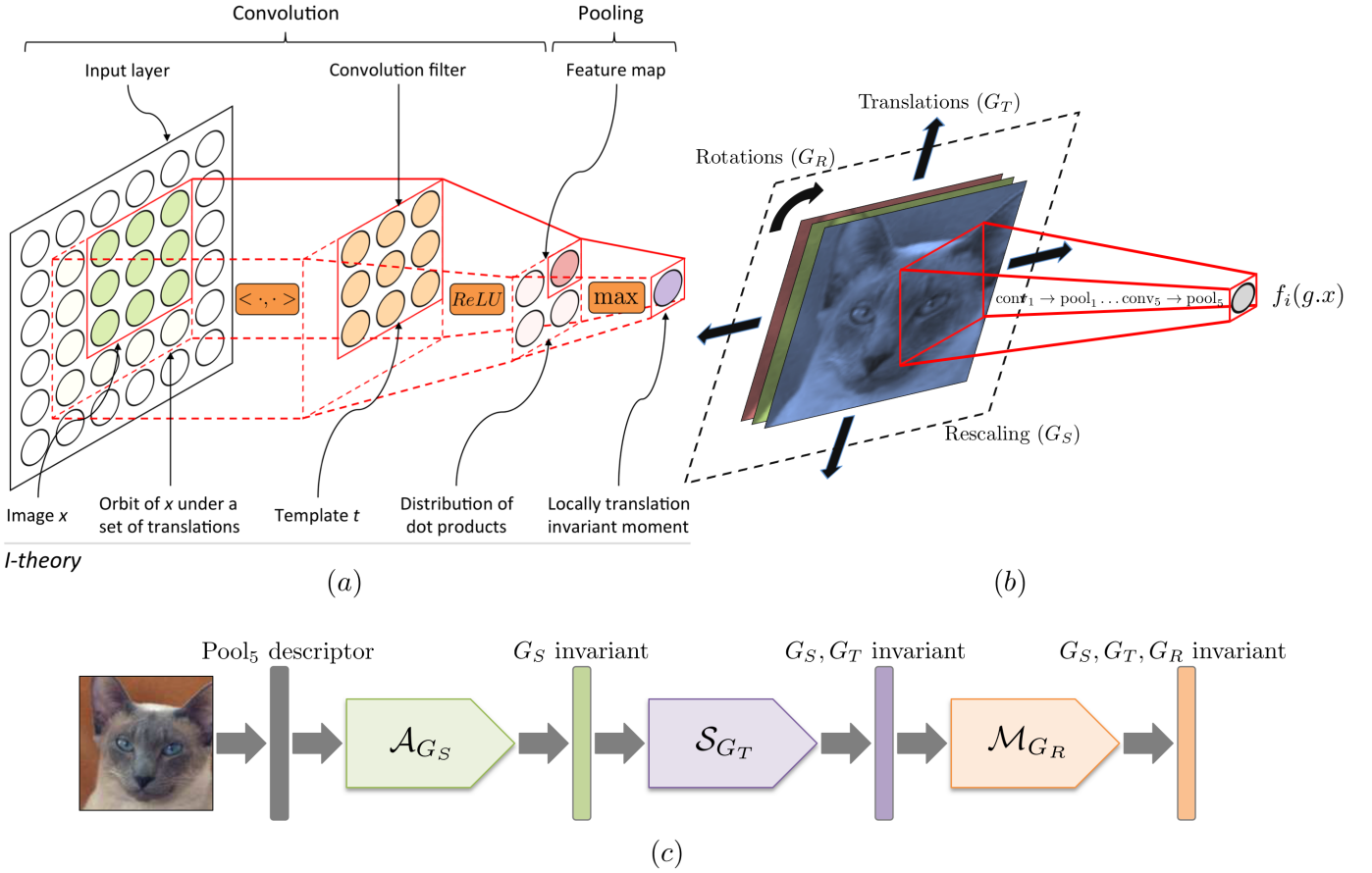[3] CentraleSupélec
[4] CBBM, LCSL, IIT and MIT

Fig. 1. (a) A single convolution-pooling operation from a CNN schematized for a single input layer and single output neuron. The parallel with *i-theory* shows that the universal building block of CNNs is compatible with the incorporation of invariance to local translations of the input according to the theory. The network architecture is responsible for the invariance properties while back-propagation provides a practical way to learn the templates from data. (b) A specific succession of convolution and pooling operations learnt by the CNN (depicted in red) computes the *pool5* feature f_i for each feature map i from the RGB image data. A number of transformations g can be applied to the input x in order to vary the response $f_i(g.x)$. (c) Our method takes inspiration from the *i-theory* to create compact and robust global image descriptors from CNNs. Starting with raw *pool5* descriptors, it can be used to stack-up an arbitrary number of transformation group invariances while keeping the dimensionality under control. The particular sequence of transformation groups and statistical moments represented on the diagram was shown to produce the best performing hashes on average in our study but other arbitrary combinations are also able to improve retrieval results.

CNN began to be applied to the instance retrieval problem as well, although there is comparatively less work on CNN-based descriptors for instance retrieval compared to large-scale image classification. Razavian et al. [15] evaluate the performance of CNN model of [1] on a wide range of tasks including instance retrieval, and show initial promising results. Babenko et al. [3] show that a pre-trained CNN can be finetuned with domain specific data (objects, scenes, etc.) to improve retrieval performance on relevant data sets. The authors also show that the CNN representations can be compressed more effectively than their FV counterparts for large-scale instance retrieval. In [4], [5], we show how sparse high-dimensional CNN representations can be hashed to very compact representations (32-1024 bits) for large scale image retrieval with little loss in matching performance.

Meanwhile, we showed in our recent evaluation work [7] that CNN-based descriptors still suffer from a number of drawbacks including the lack of robustness to certain transformations of the input data. The *i-theory* proposed as the guideline method in this paper was recently used to successfully compute robust representations for face recognition [16] and music classification [17].

The contributions of our work can be summarized as follows.

- A method based on *i-theory* for creating robust and compact global image descriptors from CNNs.
- The ability to iteratively incorporate different group invariances, each new addition leading to consistent and significant improvements in retrieval results.
- A set of highly competitive global descriptors for image instance retrieval compared with other state-of-the-art compact descriptors at similar bitrates.
- A low risk of overfitting: few parameters and many reasonable settings which can generalize well across all datasets.
- A thorough empirical study based on several publicly available datasets.

III. INVARIANT GLOBAL IMAGE DESCRIPTORS

A. I-theory in an Nutshell

Many common classes of image transformations such as rotations, translations and scale changes can be modeled by the action of a group G over the set E of images. Let $x \in E$ and a group G of transformations acting over E with group action $G \times E \to E$ denoted with a dot (.). The orbit of x by G is the subset of E defined as $O_x = \{g.x \in E | g \in G\}$. It can be easily shown that O_x is globally invariant to the action of any element of G and thus any descriptor computed directly from O_x would be globally invariant to G.

The *i-theory* predicts that an invariant descriptor for a given object $x \in E$ is computed in relation with a predefined template $t \in E$ from the distribution of the dot products $D_{x,t} = \{< g.x, t >\in \mathbb{R} | g \in G\} = \{< x, g.t >\in \mathbb{R} | g \in G\}$ over the orbit. One may note that the transformation can be applied either on the image or the template indifferently. The proposed invariant descriptor extracted from the pipeline should be a histogram representation of the distribution with a specific bin configuration. Such a representation is mathematically proven to have proper invariance and selectivity properties provided that the group is compact or at least locally compact [8].

In practice, while a compact group (e.g. rotations) or locally-compact group (e.g. translations, scale changes) is required for the theory to be mathematically provable, the authos of [8] suggest that the theory extends well (with approximate invariance) to non-locally compact groups and even to continuous non-group transformations (e.g. out-of-plane rotations, elastic deformations) provided that proper class-specific templates can be provided. Recent work on face verification [16] and music classification [17] apply the theory to non-compact groups with good results. Additionally, the histograms can also be effectively replaced by statistical moments (e.g. mean, min, max, standard deviation, etc.).

B. CNNs are i-theory Compliant Networks

All popular CNN architectures designed for image classification such as *AlexNet* [1] and *OxfordNet* [2] share a common building block: a succession of convolution-pooling operations designed to model increasingly high-level visual representations of the data. The highest level visual features may then be fed into fully connected layers acting as a classifiers.

As shown in detail on Figure 1 (a), the succession of convolution and pooling operations in a typical CNN is in fact a way to incorporate local translation invariance strictly compliant with the framework proposed by the *i-theory*. The network architecture provides the robustness such as predicted by the invariance theory while training via back propagation ensures a proper choice of templates. In general, multiple convolution-pooling steps are applied (5 times in both *AlexNet* and *OxfordNet*) resulting in increased robustness and higher level templates. Note that the iterative composition of local translation invariance approximately translates into robustness to local elastic distortions for the features at the *pool5* layer.

In this study, instead of the popular first fully-connected layer (*fc6*) which is on average the best single CNN layer to use as a global out-of-the-box descriptor for image retrieval [7], we decide to use the locally invariant *pool5* as a starting representation for our own global descriptors and further enhance their robustness to selected transformation groups in a way inspired from *i-theory*.

C. Transformation Invariant CNN Descriptors

For every feature map i of the *pool5* layer ($0 \leq i < 512$ in the case of the presently used *OxfordNet*), we denote $f_i(x)$ the corresponding feature obtained from the RGB image data through a succession of convolution-pooling operations. Note that the transformation f_i is non-linear and thus not strictly a mathematical dot product with a template but can still be viewed as an inner product.

In this study, we propose to further improve the invariance of *pool5* CNN descriptors by incorporating global invariance to several transformation groups. The specific transformation groups considered in this study are translations G_T, rotations G_R and scale changes G_S. As shown on Figure 1 (b), transformations g are applied on the input image x varying the output of the *pool5* feature $f_i(g.x)$ accordingly.

The invariant statistical moments computed from the distributions $\{f_i(g.x) | g \in G\}$ with $G \in \{G_T, G_R, G_S\}$ are averages, maxima and standard deviations, respectively:

$$\mathcal{A}_{G,i}(x) = \frac{1}{\int_G dg} \int_G f_i(g.x) dg \tag{1}$$

$$\mathcal{M}_{G,i}(x) = \max_G (f_i(g.x)) \tag{2}$$

$$\mathcal{S}_{G,i}(x) = \frac{1}{\int_G dg} \sqrt{\int_G f_i(g.x)^2 dg - (\int_G f_i(g.x) dg)^2} \tag{3}$$

with corresponding global image descriptors obtained by simply concatenating the moments for the individual features:

$$\mathcal{A}_G(x) = (\mathcal{A}_{G,i}(x))_{0 \leq i < 512} \tag{4}$$

$$\mathcal{M}_G(x) = (\mathcal{M}_{G,i}(x))_{0 \leq i < 512} \tag{5}$$

$$\mathcal{S}_G(x) = (\mathcal{S}_{G,i}(x))_{0 \leq i < 512} \tag{6}$$

In principle, G is always measurable and of finite measure as required since G_T and G_S must be restricted to compact subsets due to image border effects.

An interesting aspect of the *i-theory* is the ability to chain multiple types of group invariances one after the other as already demonstrated in [17]. In this study, we construct descriptors invariant to several transformation groups by successively applying the method to different transformation groups as shown on Figure 1 (c). For instance, following scale invariance with average by translation invariance with standard deviation for feature i would correspond to:

$$\max_{g_t \in G_T} \left(\frac{1}{\int_{g_s \in G_S} dg_s} \int_{g_s \in G_S} f_i(g_t g_s.x) dg_s \right) \tag{7}$$

One may note that the operations are sometimes commutable (e.g. $\mathcal{A}_G$ and $\mathcal{A}_{G'}$) and sometimes not (e.g. $\mathcal{A}_G$ and $\mathcal{M}_{G'}$) depending on the specific combination of moments.

IV. Image Instance Retrieval

A. Evaluation Framework

We evaluate our invariant image descriptors in the context of image instance retrieval. As our starting representation, we use the *pool5* layer from the 16 layers *OxfordNet* [2] with a total dimensionality of 25088 organized in 512 feature maps of size 7×7.

Similarly to our evaluation work in [7], the rotated input images are padded with the mean pixel value from the ImageNet data set. The step size for rotations is 10 degrees yielding 36 rotated images per orbit. For scale changes, 10 different center crops geometrically spanning from 100% to 50% of the total image have been taken. For translations, the entire feature map is used for every feature, resulting in an orbit size of $7 \times 7 = 49$.

We evaluate the performances of the descriptors against four popular data sets: *Holidays*, *Oxford buildings (Oxbuild)*, *UKBench (UKB)* and *Graphics*. The four datasets are chosen for the diversity of data they provide: *UKBench* and *Graphics* are object-centric featuring close-up shots of objects in indoor environments. *Holidays* and *Oxbuild* are scene-centric datasets consisting primarily of outdoor buildings and scenes.

INRIA Holidays. The INRIA Holidays dataset [18] consist of personal holiday pictures. The dataset includes a large variety of outdoor scene types: natural, man-made, water and fire effects. There are 500 queries and 991 database images. Variations in lighting conditions are rare in this data set as the pictures from the same location are taken at the same time.

Oxbuild. The Oxford Buildings Dataset [19] consists of 5062 images collected from Flickr representing landmark buildings in Oxford. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. Note that the set contains 55 queries only.

UKBench. The University of Kentucky (UKY) data set [20] consists of 2550 groups of common objects. There are 4 images representing each. Only the object of interest present in each image. Thus, there is no foreground or background clutter within this data set. All 10200 images are used as queries.

Graphics. The Graphics data set is part of the Stanford Mobile Visual Search data set [21], which notably was used in the MPEG standard: Compact Descriptors for Visual Search (CDVS) [22]. The data set contains different categories of objects like CDs, DVDs, books, software products, business cards, etc. The query images include foreground and background clutter that would be considered typical in real-world scenarii, e.g., a picture of a CD might contain other CDs in the background. This data set distinguishes from the other ones as it contains images of rigid objects captured under widely varying lighting conditions, perspective distortion, foreground and background clutter. Query images are taken with heterogeneous phone cameras. Each query has two relevant images. There are 500 unique objects, 1500 queries, and 1000 database images.

B. Pairwise Matching Distance

Image instance retrieval starts with the construction of a list of database images ordered according to their pairwise matching distance with the query image. With CNN descriptors, the matching distance is strongly affected by commonly encountered image transformations. As shown in our previous evaluation work [7], a rotation of the query image by 10 degrees or more causing a sharp drop in results. This particular issue is much less pronounced with the popular Fisher vectors, largely due to the use of interest point detectors.

Figure 2 provides an insight on how adding different types of invariance with our proposed method will affect the matching distance on different image pairs of matching objects. With the incorporation of each new transformation group, we notice that the relative reduction in matching distance is the most significant with the image pair which is the most affected by the transformation group.
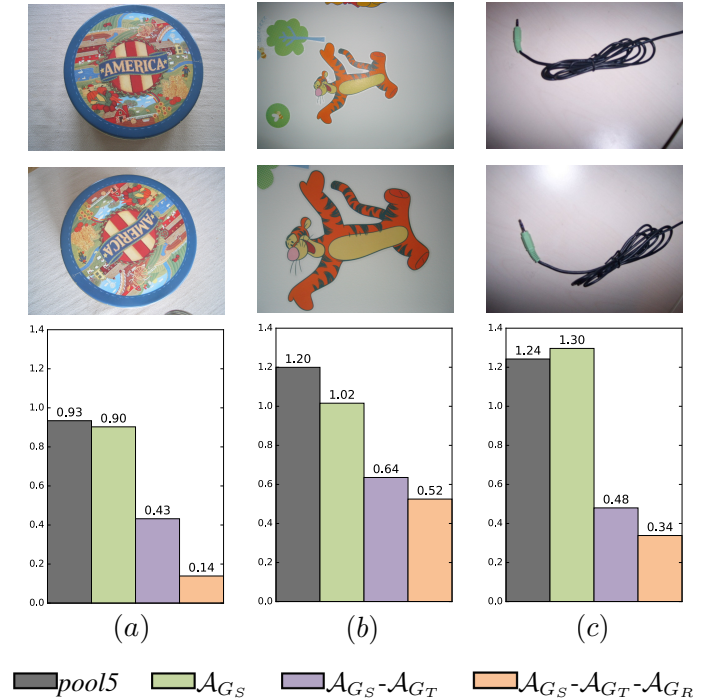

Fig. 2. Distances for 3 matching pairs from *UKBench*. For each pair, 4 pairwise distances (L_2-normalized) are computed corresponding to the following descriptors: *pool5*, $\mathcal{A}_{G_S}$, $\mathcal{A}_{G_S}$-$\mathcal{A}_{G_T}$ and $\mathcal{A}_{G_S}$-$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$. Adding scale invariance makes the most difference on (b), translation invariance on (c), and rotation on (a) which is consistent with the scenarii suggested by the images.

C. Transformations, Order and Moments

Our first set of results summarized in Table I study the effects of incorporating various transformation groups and using different moments. *Pool5* which is the starting point of our descriptors and *fc6* which is considered the best off-the-shelf descriptor [7], [15] are provided as baselines. Table I also provides results for all possible combinations of transformation groups for average pooling (order does not matter as averages commute) and for the single best performer which is $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ (order matters).

TABLE I
RETRIEVAL RESULTS (MAP) FOR DIFFERENT SEQUENCES OF TRANSFORMATION GROUPS AND MOMENTS.

SEQUENCE	DIMS	DATASET			
		Oxbuild	Holidays	UKB	Graphics
pool5	25088	0.427	0.707	0.823(3.105)	0.315
fc6	4096	0.461	0.782	0.910(3.494)	0.312
$\mathcal{A}_{G_S}$	25088	0.430	0.716	0.828(3.122)	0.394
$\mathcal{A}_{G_T}$	512	0.477	0.800	0.924(3.564)	0.322
$\mathcal{A}_{G_R}$	25088	0.462	0.779	0.954(3.718)	0.500
$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$	512	0.418	0.796	0.955(3.725)	0.417
$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_S}$	512	0.537	0.811	0.931(3.605)	0.430
$\mathcal{A}_{G_R}$-$\mathcal{A}_{G_S}$	25088	0.494	0.815	0.959(3.752)	0.552
$\mathcal{A}_{G_S}$-$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$	512	0.484	0.833	0.971(3.819)	0.509
$\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$	512	**0.592**	**0.838**	**0.975(3.842)**	**0.589**
$\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$	512 **bits**	0.523	0.787	0.958(3.741)	0.552

Results are computed with the mean average precision (mAP) metric. For reference, 4×Recall@4 results are also provided for UKBench (between parentheses). G_T, G_R, G_S denote the groups of translations, rotations and scale changes respectively. Note that averages commute with other averages so the sequence order of the composition does not matter when only averages are involved. Best results are achieved by choosing specific moments. $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ corresponds to the best average performer for which a binarized version is also given. *fc6* and *pool5* are provided as a baseline.
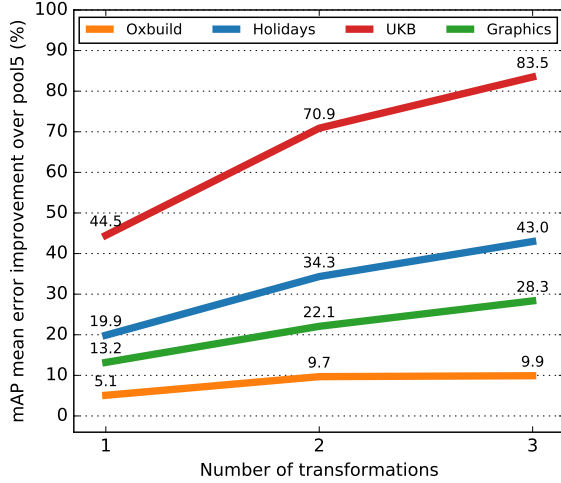
Fig. 3. Results from Table I for the 7 strategies using averages only (rows 3 to 9) expressed in terms of improvement in mAP over *pool5*, and aggregated by number of invariance groups. Improvements range from +5% on Oxbuild using 1 transformation to +83.5% on UKBench using 3 transformations. On all 4 datasets, results clearly improve with the amount of groups considered.

First, we can immediately point out the high potential of *pool5*. Although it performs notably worse than *fc6* as-is, a simple average pooling over the space of translations $\mathcal{A}_{G_T}$ makes it both better and 8 times more compact than *fc6*. As shown in Figure 3, accuracy increases with the number of transformation groups involved. On average, single transformation schemes perform 21% better compared to *pool5*, 2-transformations schemes perform 34% better, and the 3-transformations scheme performs 41% better.

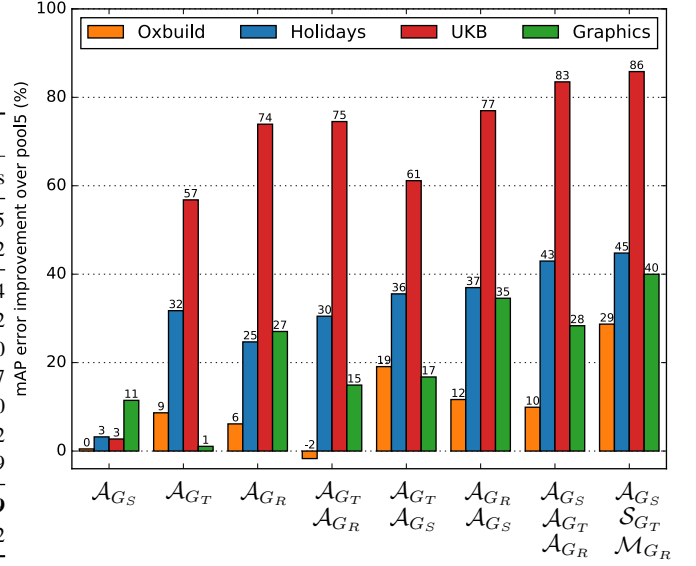Selecting statistical moments different than averages can

Fig. 4. Results from Table I expressed in terms of improvement in mAP over *pool5*. Most strategies yield significant improvements over *pool5* on most datasets. The average improvement is 68% for the best strategy.

further improve the retrieval results. In Figure 4, we observe that $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ performs roughly 17% better (average results over all datasets) than $\mathcal{A}_{G_S}$-$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$. Notably, the best combination corresponds to an increase in the orders of the moments: $\mathcal{A}$ being a first-order moment, $\mathcal{S}$ second order and $\mathcal{M}$ of infinite order. A different way of stating this fact is that a more invariant representation requires a higher order of pooling.

Overall, $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ improves results over *pool5* by 29% (Oxbuild) to 86% (UKBench) with large discrepancies according to the dataset. Better improvements with UKBench can be explained with the presence of many rotations in the dataset (smaller objects taken under different angles) while Oxbuild consisting mainly of upright buildings is not significantly helped by incorporating rotation invariance.

D. Compact Binary Hashes

TABLE II
RETRIEVAL PERFORMANCE (MAP) COMPARING OUR METHOD TO OTHER STATE-OF-THE-ART METHODS.

METHOD	REP. SIZE	DATASET		
	#dim (size in bits)	Oxbuild	Holidays	UKB
Our results	512 (512)	0.523	**0.787**	**0.958**
OxfordNet [23]	256 (1024)	**0.533**	0.716	0.842
OxfordNet [23]	256 (256)	0.436	0.578	0.693
T-embedding [24]	256 (2048)	0.472	0.657	0.863
T-embedding [24]	128 (1024)	0.433	0.617	0.850
VLAD+CSurf [25]	128 (1024)	0.293	0.738	0.830
mVLAD+Surf [25]	128 (1024)	0.387	0.718	0.875

Only methods within a comparable range of bitrates are selected.

As shown in Table I, a simple binarization strategy (thresholding at dataset mean) applied to our best performing descriptor $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ degrades retrieval performance only very marginally and is in fact sufficient to produce a hash that compares favourably with other state-of-the-art methods. On average over the four datasets, the hash performs 39% better than *pool5*, while having a 1568 times smaller representation size. Compared to *fc6* feature, the hash performs 26% better while being 256 times smaller.

As mentioned, the invariant hash also performs well compared to other state-of-the-art. In Table II, we compare our invariant hash against other approaches designed to produce compact representations with comparable bit sizes, 512 being considered fairly compact by current standards [23]. In addition, note that the hashing scheme used in this experiment is very naive and better hashes (better retrieval results and/or higher compression rates) could most certainly be obtained by applying more sophisticated hashing methods such as in our previous work [26], [5] based on stacked RBMs and metric finetuning.

V. CONCLUSION

We proposed a novel method based on *i-theory* for creating robust and compact global image descriptors from CNNs for image instance retrieval. Through a thorough empirical study, we show that the incorporation of every new group invariance property following the method leads to consistent and significant improvements in retrieval results.

Our method has a number of parameters (sequence of the invariances and choice of statistical moments) but experiments show that many default and reasonable settings produce results which can generalise well across all datasets meaning that the risk of overfitting is low.

This study also confirms the high potential of the feature pyramid (*pool5*) as a starting representation for high-performance compact hashes instead of the more commonly used first fully connected layer (*fc6*). Our method produces a set of descriptors able to compare favourably with other state-of-the-art compact descriptors at similar bitrates even without dedicated compression/hashing step.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 1–9.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *arXiv preprint arXiv:1409.1556*, 2014, pp. 1–10.

[3] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural Codes for Image Retrieval," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[4] V. Chandrasekhar, L. Jie, O. Morère, A. Veillard, and H. Goh, "Compact Global Descriptors for Visual Search," in *Proceedings of IEEE Data Compression Conference (DCC)*, Snowbird, Utah, March 2015.

[5] J. Lin, O. Morère, J. Petta, V. Chandrasekhar, and A. Veillard, "Tiny descriptors for image retrieval with unsupervised triplet hashing," *CoRR*, vol. abs/1511.03055, 2015. [Online]. Available: http://arxiv.org/abs/1511.03055

[6] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale Image Retrieval with Compressed Fisher Vectors," in *Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, June 2010, pp. 3384–3391.

[7] V. Chandrasekhar, J. Lin, O. Morère, H. Goh, and A. Veillard, "A practical guide to cnns and fisher vectors for image instance retrieval," *CoRR*, vol. abs/1508.02496, 2015. [Online]. Available: http://arxiv.org/abs/1508.02496

[8] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations in hierarchical architectures," *arXiv:1311.4158*, 2013.

[9] F. Anselmi and T. Poggio, "Representation learning in sensory cortex: a theory." *CBMM memo n 26*, 2010.

[10] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Magic materials: a theory of deep hierarchical architectures for learning sensory representations." *CBCL paper*, 2013.

[11] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1701–1708.

[12] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Computer Vision and Pattern Recognition (CVPR), 2014 Conference on*. IEEE, 2014, pp. 1891–1898.

[13] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *International Conference on Computer Vision (ICCV)*, Dec 2013, pp. 2056–2063.

[14] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1653–1660.

[15] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2014, pp. 512–519.

[16] Q. L. Liao, J. Z. Leibo, and T. Poggio, "Learning invariant representations and applications to face verification," in *NIPS*, 2013.

[17] C. Zhang, G. Evangelopoulos, S. Voinea, L. Rosasco, and T. Poggio, "A deep representation for invariance and music classification," in *ICASSP*, 2014.

[18] H. Jégou, M. Douze, and C. Schmid, "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search," in *Proceedings of European Conference on Computer Vision (ECCV)*, Berlin, Heidelberg, October 2008, pp. 304–317.

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," in *Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, June 2007, pp. 1–8.

[20] D. Nistér and H. Stewénius, "Scalable Recognition with a Vocabulary Tree," in *Computer Vision and Pattern Recognition (CVPR)*, New York, USA, June 2006, pp. 2161–2168.

[21] V. Chandrasekhar, D.M.Chen, S.S.Tsai, N.M.Cheung, H.Chen, G.Takacs, Y.Reznik, R.Vedantham, R.Grzeszczuk, J.Back, and B.Girod, "Stanford Mobile Visual Search Data Set," in *Proceedings of ACM Multimedia Systems Conference (MMSys), San Jose, California, February 2011*, 2011.

[22] *MPEG CDVS (Compact Descriptors for Visual Search) Benchmark. Stanford Digital Repository*.

[23] A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, "A baseline for visual instance retrieval with deep convolutional networks," in *International Conference on Learning Representations, May 7-9, 2015, San Diego, CA*. ICLR, 2015.

[24] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 3310–3317.

[25] E. Spyromitros-Xioufis, S. Papadopoulos, I. Y. Kompatsiaris, G. Tsoumakas, and I. Vlahavas, "A comprehensive study over vlad and product quantization in large-scale image retrieval," *Multimedia, IEEE Transactions on*, vol. 16, no. 6, pp. 1713–1728, 2014.

[26] J. Lin, O. Morère, V. Chandrasekhar, A. Veillard, and H. Goh, "Deephash: Getting regularization, depth and fine-tuning right," *CoRR*, vol. abs/1501.04711, 2015. [Online]. Available: http://arxiv.org/abs/1501.04711