NAVER LABS EUROPE

Report about the Habilitation (HDR) thesis "Exploring and Learning from Visual Data" by "Yannis Avrithis"

With Dr Avrithis' words, the manuscript is a journey in computer vision and machine learning research from the early years of Gabor filters and linear classifiers to deep models surpassing human skills in several tasks today. By intermixing rich literature reviews of visual representation and understanding before and after establishment of deep learning as the dominant paradigm with his own contributions to the field, Mr. Avrithis succeeded in giving an interesting manner to discover this dynamic and challenging scientific field.

The manuscript is not a simple concatenation of various contributions, but a real synthesis the related field over several periods, containing deep analysis and reflections on different domains and problems, the proposed solutions are interlinked, the strengths and limitations highlighted and more importantly, potential future directions to be explored are raised.

The manuscript is very dense, it took me some time to go through, but it was a real pleasure to read and I think it can see it as a good textbook for students and researchers on the topic. It consists in three main technical parts, each containing a collection of contributions corresponding to a different period or subject; and a fourth part consolidating the contributions and drawing perspectives on future work. Each part starts with an outline section describing the problems addressed, background knowledge to make the document self-consistent and a concise literature survey. Then more technically detailed chapters discuss the various contributions of Dr Avrithis in the respective field and period.

The first part is dedicated to representations and matching processes for exploring visual data before the deep learning era with several main contributions co-authored by Dr Avrithis. These contributions have in common that it revolves around efficient image search in very large datasets, trying to find solutions that either increase the retrieval performance without decreasing the search speed or speeding up the database mining without losing retrieval accuracy.

Accordingly, in Chapter 3 an Approximate Gaussian Mixture based clustering method to build efficient visual vocabularies for large scale image retrieval is presented. The iterative algorithm dynamically purges components, setting automatically the vocabulary size and uses an approximate nearest neighbor to speed up the clustering process. By exploiting the iterative nature of the algorithm and keeping trace of best neighbors, the method permits to boost both the speed and the precision of the search process itself.
In Chapter 4 describes the Hough pyramid spatial matching algorithm which is linear in the number of correspondences and can be easily integrated as a geometry re-ranking step in any image retrieval engine to increase its flexibility concerning multiple matching surfaces and non-rigid objects and its retrieval performance without losing search speed.

In Chapter 5 the Aggregated Selective Match Kernel is presented which combines ideas from aggregated representations like VLAD and selective match kernels like Hamming Embedding expressed by a common model. The proposed new kernel applies a selectivity function after aggregating descriptors per clusters, producing a more compact visual representation and implicitly handles burstiness by keeping only one representative of all bursty descriptors per cell.

Chapter 6 presents a new and efficient vector quantizer that combines low distortion with fast approximate Nearest Neighbor search in high dimensional spaces. The proposed Locally Optimized Product Quantization, uses a coarse quantizer to index data, but the residuals between data points and centroids are PQ-encoded within each cell by locally optimizing both the space decomposition and the sub-quantizers. It is shown that in the case of multi-indexing the proposed method allows to maintain similar or even better performance than the alternative solutions with significantly lower overhead both in space and in search time.

Chapter 7 is devoted to location and landmark recognition where the solution proposed is to group the images both geographically and visually and for each cluster to build a scene maps which then is used directly for the retrieval. To construct the scene map, all views within a cluster are aligned to a reference view and the visual words clustered by their positions, resulting in spatial codebooks. The scene map of the cluster has the same representation as a single image and hence it can be used in the inverted file indexes yielding effective search. These ideas were integrated into an online search engine (VIRaL) supporting geo-tagging, landmarks recognition and visualization of photo clusters and tourist paths.

With the success of deep learning methods in computer vision and machine learning, Dr Avrithis continued to study visual representation and matching for efficient image indexing and search, but this time designing solutions that relies on deep features and/or deep models. Therefore, in Part II of the manuscript, entitled Exploring Deeper, after an introduction into deep learning and the related literature, several contributions of Dr Avrithis are presented such as advances in manifold search over global or regional CNN representations seen as graph filtering, spatial matching revisited with local features detected on CNN activations or discovering objects from CNN activations in unlabeled image collections.

By exploring the manifold structure of the feature space, in Chapter 10, Dr Avrithis introduces an efficient diffusion process on manifolds of local CNN representations, which can be seen as a recursive form of query expansion. Another contribution presented in this chapter is the Fast Spectral Ranking that exploits a low-rank spectral decomposition of the graph adjacency matrix to express the linear system solution as a sequence of matrix multiplications providing scalable solution and bringing dramatic gains in standard image retrieval benchmarks compared to Euclidean search and Average Query Expansion.

In Chapter 11, Dr Avrithis extends the idea of spatial matching to deep image retrieval where sparse collections of local features are extracted from convolutional activations independently and, to find the geometric transformation between images, they are matched per channels using a RANSAC based fast

spatial matching algorithm. The method is used to rerank top retrieved images according to the number of inliers found. Experiments with different features have shown a consistent performance gain obtained with query-time diffusion using top retrieved images after the spatial reranking.

Chapter 12 presents an unsupervised approach for detecting salient regions in images, where discriminative and frequent patterns are captured within an image database relying on deep features and generalized max pooling. The method first generates 2D feature saliency maps for each image and builds region kNN graphs based on region saliency scores and corresponding deep features extracted from the activation maps. Then the graph centrality score per region is used to form object saliency maps capturing discriminative patterns appearing frequently in the dataset. It is shown that using these maps improves significantly the retrieval especially on datasets where the queried objects are small and severely cluttered in the dataset images.

While Part I and II address instance-level visual search and clustering, with shallow respectively deep visual representations and focuses mainly of efficient indexing solutions, fast spatial matching and re-ranking processes, the third part of the manuscript is devoted to learning visual representations by training deep learning models with limited supervision.

In particular, Chapter 14 describes an unsupervised hard example mining mechanisms that uses the manifold similarity, described in Chapter 10, to guide a deep metric learning process. The main idea is to consider positives pools, with elements that are neighbors according to their manifold similarity but not with Euclidian distance and nearby elements in the Euclidean space lying on different manifolds as hard negatives. The new deep metric learning technique was successfully applied to fine grained classification and object retrieval.

In Chapter 15 an inductive deep version of the classical Label Propagation is presented, where pseudo-labels are inferred by a graph based on the network embedding, and the training alternates between label propagation and updating the embedding. The network uses a label fitting supervised loss and an unsupervised smoothness loss that encouraging consistency between nearby example predictions that is weighted by a predicted class entropy based certainty score.

Chapter 16 considers the problem of few-shot learning where not only the labels but also the amount of available data is limited. There are two main contributions in this chapter, the dense classification method and the neural implanting network. In the former approach, a cosine classifier is adopted where the weight parameters are shared over all spatial positions encouraging the classifier to make correct predictions all over the image and making the activation maps aligned with objects smoother. The second approach consists in implanting convolutional filters in a new processing stream, parallel to a pre-trained network, which are trained it in a few-shot regime yielding new, task-specific features.

Chapter 17 presents a study of adversarial examples which are obtained by imperceptible perturbations of a given input making the model prediction fail and proposes new adversarial examples with higher visual quality obtained by graph filtering where the local smooth perturbations are guided by the input image.

NAVER LABS EUROPE

As only a few articles were exposed in the three technical parts, in the last part further and current contributions are briefly summarized. including methods for video abstraction, spatiotemporal saliency, object proposal and detection, local feature detection and selection, location recognition and active learning. Then M. Avrithis provides a synthesis of the methods described in the manuscript, analyze them in the present context, makes connections between them and highlights their strengths and limitations.

Based on the observations drawn from this analysis, M. Avrithis proposes a four interesting research directions for learning visual representations from data with limited supervision. The first is unsupervised deep metric learning for few-shot and incremental learning, model distillation and self-learning to rank. The second is an end-to-end learning framework using geometrically aligned tensors for category-level tasks where explicit semantic alignment can answer the invariance versus discriminative power dilemma. The third direction proposed is to generalize graph convolutions by using a mixture model for both activations and convolution kernel. A forth direction is to extend manifold similarity, extensively exploited in several contributions, by addressing its scalability issues with quantizers in the spectral embedding space, by extending scalar similarities to geometric transformations found via spatial matching or by computing the graph dynamically per layers according to the geometry. Finally, the fifth direction is considered is the extension of memorizing techniques proposed for classification, such as distillation loss, synaptic plasticity mechanisms and network expansion mechanisms, to metric learning and instance-level tasks.

As the manuscript also shows, M. Avrithis very prolific and has accumulated an impressive amount of work since his dissertation, more than sufficient for a habilitation. He co-authored several dozens of publications, amongst which many of them was presented at main computer vision conferences (CVPR, ICCV, ECCV), and published in top international journals (IJCV, MTAP, CVIU). These woks are well known and well cited (Google scholar h-index being 41), which attest a significant scientific impact within in the community. Many of them have been and are being carried out by supervised or co-supervised students showing the quality of his supervision capabilities. Furthermore, M. Avrithis's contribution to the scientific community beyond the scientific publications is also considerable. He has led or was involved in numerous EU and national projects, he has chaired or participated in the co-organization of important international events and workshops, he regularly reviews for international journals and conferences.

In summary, because of his impressive amount of scientific contribution both in terms of quality and quantity, presented and synthetized in an excellent manner in the manuscript, his recognized impact and presence in the community, his investment in supervision, and his clear vision regarding current and future research directions, I strongly support the attribution of the diploma of habilitation HDR to M. Avrithis.

Gabriela Csurka
Principal Scientiste
NAVER LABS Europe