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

Graz University of Technology

Faculty of Computer Science and Biomedical Engineering
Institute for Computer Graphics and Vision

Prof. Dr. Horst Bischof

Inffeldgasse 16, 2nd floor
8010 Graz
Austria

Mail: bischof@icg.tugraz.at
Phone: ++43 (0) 316 873 5014
Fax: ++43 (0) 316 873 5050

DVR: 008 1833 UID: ATU 574 77 929

Fabian Allain
Head of Schooling
University Rennes 1
France

Graz, February, 29,2020

Subject: Report Ioannis Avrithis

Dear Dr. Fabian Allain,

It is a great pleasure for me to write an evaluation report for Ioannis Avrithis.

Let me first introduce myself. I am professor of computer vision and vice rector of research at Graz University of Technology. My area of research is computer vision and visual computing. I have published more than 750 peer reviewed papers and have received numerous awards (>20), among them the Jan Koenderink Award, Pattern Recognition Award, two times the German Pattern Recognition award and also the award of the British machine vision society. I personally know the work of Dr. Avrithis since quite some time. We see us regularly at conferences but we have never worked together and I do not have any other conflicts of interest.

Dr. Avrithis is an excellent and visionary researcher. Let me present the review of his HDR-Thesis entitled “Exploring and learning from Visual Data” which summarizes the research Dr. Avrithis has been conducting with his collaborators during the last 10 years. Let me state that this is an impressive piece of work.

The work is structured in four parts, with a total of 19 chapters. Most of the chapters are papers that have been published in A* conferences/journals (17 high level papers in eight years is very impressive track record, besides that many other publications have been done, this shows the scientific productivity of Dr. Avrithis). In this report, I do not evaluate the papers which have been reviewed by the conferences and journals and presented in premier venues. The review is based on the introductory chapters of the various Parts and the Part IV which is a summary and outlook.

First of all, the author should be congratulated that he a but together a coherent piece of work. Reading the thesis was a pleasure. In this ten years’ computer vision has changed significantly and made tremendous progress (due to the raise of deep learning). Nevertheless, the author is able to present a coherent piece of work. This manuscript studies solutions to related problems before and after the establishment of deep learning as the dominant paradigm in representing and understanding visual data.

Part I contains methods before the deep learning area. It addresses instance-level visual search and clustering, building on shallow (hand crafted features) visual representations and matching processes. The representation is obtained by (at time) state of the pipeline of local features, hand-

crafted descriptors and visual vocabularies. Various improvements in the pipeline are introduced, including the construction of large scale vocabularies, spatial matching for geometry verification, representations beyond vocabularies, and nearest neighbor search. Applications are centered around large photo collections.

In Part II similar topics than in Part I are addressed but building on deep visual representations and matching processes. A particular focus is put on the manifold structure of the feature space. The manifold of images (and their representations) is a powerful concept. The representation is obtained by deep parametric models learned from visual data. Contributions are made in spatial and spectral graph filtering. Spatial matching is addressed CNN activations. Region proposals (object detection) from CNN activations over an unlabeled image is covered.

Part III is devoted to learning the deep visual representations. The main focus is on limited or no supervision. Methods for category-level and instance-level recognition with limited supervision are covered. Various contributions are made in the area of metric learning, semi-supervised learning and a few-shot learning. A method to improve the visual quality of adversarial examples is introduced.

Part IV, as mentioned above, is a summary of the main contributions. It puts the contributions of the thesis in context of the present developments. It also discusses other papers by the author and how they related to the thesis. The final chapter outlines a roadmap of future research (with topics for at least ten more years of research).

So in summary Dr. Avrithis has proven to have an extraordinary ability to perform cutting edge research and to achieve scientific results with a considerable impact on the field. I fully recommend without any hesitation the acceptance of the HDR thesis.

Yours sincerely,

Univ. Prof. Dr. Horst Bischof

Report on the HDR dissertation of Yannis Avrithis, titled

"Exploring and learning from visual data"

Patrick Pérez
Scientific Director Valeo.ai

After twenty years of research in the field of computer vision, the mentoring of twelve PhD students and the publication of numerous articles in tier-one conferences and journals (including eighteen articles in CVPR/ICCV/ECCV since 2009), Yannis Avrithis applies today to the "Habilitation à Diriger les Recherches" (HDR). To this end, he submits a comprehensive manuscript where he presents a selection of his scientific contributions on visual data analysis, puts them in context, draws future research directions and shares his views of the field.

By essence, HDR dissertations come in a wide range of forms, but I have to say that the one of Yannis Avrithis stands out in several ways. Indeed, the research that is presented covers no fewer than twenty years, and two decades over which computer vision has undergone tremendous metamorphoses along with its sister discipline, machine learning. Already very active in the 2000's, the field has witnessed a complete transfiguration in the 2010's with the revival of trained artificial neural networks, now termed deep learning (DL) and the associated explosion of real-life applications. Having been through such a revolution as a researcher is an incredible experience, but it makes rather challenging the writing of a synthetic dissertation. To attack this challenge, Yannis Avrithis chooses the most ambitious path: the one of writing a comprehensive, superb document, which takes the reader through a long and fascinating journey. The less-versed reader will feel privileged to have such a guide, getting a lucid and pedagogical overview of this scientific adventure; The young specialist will discover that the pre-DL era is very rich and has still a lot to offer for the one curious to revisit it; And the senior researcher in the field will be grateful for all the new insights into a seemingly familiar material. For all of them, the journey will be all the more pleasant since no detail has been neglected: the organization of the manuscript, its typesetting, the scientific and literary style, everything has been treated with the greatest care.

As a result, Yannis Avrithis' dissertation amounts to more than two hundred dense pages, organized in nineteen chapters and four parts: "Exploring", "Exploring deeper", "Learning" and "Beyond". Each of the first three parts is based on a selection of works that share some high-level theme, each chapter describing mostly one work into some details (related art, motivation, approach, experiment summary, discussion). In each of these three parts, a first "Outline" chapter summarizes the rationale of the part and its content. This effectively allows a hierarchical reading, where chapters 1, 2, 8, 13, 18 and 19 form a stand-alone synthetic version of the whole manuscript. Conditioned on this back-bone, the reader can then zoom in other chapters, depending on interest and time, since they are mostly (conditionally) independent, apart from the background on graph filtering (Chap. 9) which is required in five subsequent chapters. This neat organization of the dissertation is explained in introductory Chapter 1, which also presents the approach taken by Yannis Avrithis for his HDR dissertation and offers a clear executive summary of it. In the rest of this report, each of the four parts is briefly evoked before providing an overall assessment of the presented work.

Part 1 (Exploring) presents a set of works developed in the period 2008-2014 and focused on shallow visual descriptions for the comparison of images, be it for example-based retrieval, matching or clustering, hence for instance-level exploration of visual data collections at large. After a concise yet thorough overview of the classic image representation pipeline based on hand-crafted features, aggregated or not, and its use for indexing, search and geometric matching, Yannis Avrithis introduces several methodological contributions to the field. The first one, approximate Gaussian mixtures (AGMs) is motivated by the need for learning very large visual vocabularies (dictionaries of local descriptors) in some applications. Combining EM with approximate nearest neighbors (ANNs) search, the technique is not only scalable but also estimates automatically the number of components, a very desirable property. A second contribution, Hough pyramid matching (HPM), concerns the problem of precise image comparison, crucial for instance in the geometric verification step that large scale search systems conduct on the short list of answers that fast global matching returns through global indexing. This original extension of the generalized Hough transform achieves massive speed-ups while being robust. In the 2010's, histogramming methods on visual words were extended by a number of teams. Yannis Avrithis explains how he contributed to this effort with aggregated selective match kernel (ASMK), which subsumes both Hamming embedding and VLAD, two very popular methods at the time, and delivered best performance before the advent of deep representations (with which it is now combined). The last methodological contribution of this part deals with the key problem of ANN for large scale search of representations (whether local or global). Vector quantization techniques are among the most powerful to this end (with hashing being another class of approaches). Building on product quantization (which partitions vectors and quantizes sub-vectors independently with specific dictionaries), Yannis Avrithis introduces locally optimized product quantization (LOPQ) in 2014, achieving remarkable large-scale results, including in real-world systems. The last chapter of the part is dedicated to a complete system for exploring landmarks in a geo-tagged dataset, Visual Image Retrieval and Localization (VIRaL). Combining some of the previous contributions with so-called "scene maps" derived from several grouped images, it can be tried online. While the contributions in this part do not rely on modern DL techniques, they are still mostly relevant (AGM and LOPQ in particular) since they are agonistic to the type of representations they manipulate.

Part 2 (Exploring deeper) presents fairly recent works where instance-level image exploration addressed in the previous part is now based on pre-trained deep features. Owing to the compactness and power of such representations, one important theme in this part is the construction and the exploitation of a graph that captures the manifold of the visual data, whether at the region level or at the image level. Also, beside search and retrieval, the distinct problem of object discovery is addressed here. The Outline chapter summarizes the advent of learned deep representations for category-level object recognition and detection, as well as for visual comparison and retrieval. This chapter is remarkably clear given the amount of material covered in little space. Before introducing the selected contributions of this part, Yannis Avrithis provides in a dedicated chapter the necessary tools and notations to deal with filtering on graphs in the context of data exploration. This is classic material exposed with a great deal of clarity, and its exploitation in subsequent chapters is briefly summarized at this stage, which makes reading very easy. The next chapter on "Searching on manifolds" makes a heavy use of these tools. It presents two contributions, regional diffusion and fast spectral ranking, that both rely on a k-NN graph of image regions, built in deep feature space. These methods can accommodate new queries (not part of the original graph) and exploit various forms of graph diffusion, which can be seen as query expansion. As a result, scalable image search with strong performance is obtained. With deep spatial matching (DSM), Yannis Avrithis revisits the challenging problem of geometric matching for visual search. The key insight is that, due to

their spatial sparsity and their complementarity, the various activation maps that classic CNNs associate to an input image can be turned virtually for free into multiple local feature detectors and exploited for local matching. While this matching is less accurate than classic one due to the extreme compactness of the representation, it is extremely efficient, benefits from existing deep pipelines without modification and provides good enough pre-ranking for subsequent diffusion that yields excellent final results. Off-the-shelf feature maps also lie at the heart of the last contribution of this Part, on object discovery. In the context of a given image collection, they allow the unsupervised extraction of salient regions that lead to improved retrieval of small objects in cluttered scenes, a particularly challenging task.

In Part 3 (Learning), Yannis Avrithis discusses some of his most recent works, where deep representations are explicitly learned, compensating the low level of supervision by an exploitation of the structure of the data at hand. Also, instance-level problems that are the focus of previous parts are complemented to some extent by category-level tasks. As in previous parts, Yannis Avrithis starts by presenting the relevant concepts (various low-supervision regimes, metric and manifold learning) with a quick overview of the associated literature, before moving to a selection of his contributions. The first one, Mining on Manifold (MoM), is an *unsupervised* deep metric learning approach. Given an initial deep representation and the associated graph for the dataset, the positive and negative examples that metric learning requires are provided by distances on this graph, rather than by human or algorithmic annotation. Despite being less supervised, the approach is competitive for image retrieval and fine-grain recognition. The second contribution attacks semi-supervised learning for object recognition, whereby only part of training examples are labelled. Two classic semi-supervised learning ideas are neatly combined: exploiting the manifold structure of the dataset to propagate labels from annotated examples to others, and retraining from pseudo-labels (labels inferred by current model on unlabeled examples). Excellent results are reported on small-scale benchmarks. A third type of low supervision is then addressed: few shot learning (learning new categories with only few examples, past the standard learning of other base classes). On this problem, which is regaining momentum with the outburst of deep representations, Yannis Avrithis makes two original contributions: “dense classification” that exploits local activations instead of spatial pooling, and “implanting” that allows grafting new, dedicated filters to base CNN for the benefit of recognizing the new classes under few-shot supervision. Both contributions bring performance boost on classic miniImageNet splits. Using again a graph, though intra-image this time (Laplacian graph), Yannis Avrithis shows in the last chapter how to compute more imperceptible adversarial perturbations, by focusing on textures and structures. Contrary to other classic adversarial perturbations, these image-dependent smooth perturbations cannot be revealed by magnification.

In Part IV (Beyond), Yannis Avrithis wraps up the dissertation with two rich chapters. The first one completes the picture drawn so far by a rapid evocation of other past and current works of his. Among the latter, the work on active deep learning for instance revisits in a very innovative manner active learning, a type of learning that DL has neglected so far despite its important real-world applications (learning under annotation budget). This chapter closes with a complete synthetic recap of the work, which sheds further light on the impressive coherence of Yannis Avrithis’ work over 20 years and the depth of his views of the domain. These qualities shine in the last chapter where Yannis Avrithis shares with the reader his thoughts regarding a number of promising research directions for visual data analysis and exposes his grand vision: “Just as there are mechanisms to automatically translate more computing power to better performance, the same should happen with storage capacity.” These questions that revolve around how to memorize the data, to capture at best its structure and to blur the boundaries

between train and test times, are fundamental. Yannis Avrithis' HDR dissertation already contributes to them and should stimulate the interest of the community in the future.

To summarize, Yannis Avrithis presents in a dissertation of exceptional quality an overview of his research over twenty years. Broadly speaking, all his work addresses the key question of how to represent visual data in order to solve various analysis tasks (retrieval, discovery, recognition, detection of objects or scenes, at the instance or the category level). Such representations can be partly or entirely learned, which requires some form of supervision and/or structure knowledge. In this broad context, several important themes traverse the work of Yannis Avrithis; they include:

- How to capture and exploit the underlying structure (manifold) of visual data, in particular when in presence of an image collection;
- How to make the extraction of the representation and its use scalable, i.e., suited to large scale applications such as image retrieval or automatic object discovery in very big image collections;
- When learning, how to deal with limited supervision, a problem of tremendous practical importance.
- How to reconcile shallow and deep approaches for visual data representation.
- How to combine the insights from image retrieval (storage, indexing, explicit memory, visual comparison, train and test coming together) with those from recognition (deep representation, implicit memory, transferability and generalization).

The breadth, the depth, the creativity and the quality of Yannis Avrithis' research qualify him for the habilitation. It is therefore my pleasure to recommend wholeheartedly the defense of his HDR thesis.

Paris, 1 March 2020

Patrick Pérez