Research and Teaching Statement

Yannis Avrithis

April 25, 2018

1 Research statement

2015-today

Deep learning has recently revolutionized most *computer vision* problems, with *convolutional neural networks* (CNN) most commonly used to learn representations from data given a particular task. In *image retrieval*, CNNs are typically trained to extract a vector representation of images, such that different views of the same object are mapped to similar representations. In this context, our *diffusion on region manifolds* [C103,R10][1] introduces a recursive form of *query expansion* that explores the image manifold efficiently online through solving a linear system.

Fast spectral ranking [C106,R11] reproduces or improves the results of the previous work, while manifold exploration is now mostly done offline via an explicit embedding. Online search is reduced to dot product similarity. We show that this is equivalent to linear graph filtering of a sparse signal in the frequency domain, and we introduce a scalable offline computation of an approximate Fourier basis of the graph. As standard image retrieval datasets appear to saturate, we revisit them and introduce a new benchmark "*Revisited Oxford and Paris*" [C107,R16], which facilitates further research.

A particular application of image retrieval is *visual location recognition*. In [C101,R13], we introduce a *panorama-to-panorama* matching process for location recognition from CNN representations of street-view images. It achieves near-perfect performance on a standard benchmark. Another application is *visual clustering* of unstructured image collections. In [C99], we introduce an extremely efficient method for *web-scale image clustering* using quantized CNN representations. It achieves clustering of a collection of 100M images in less than an hour on a single machine.

Image classification is the most common problem involving both computer vision and machine learning methods, and it was the first where the power of deep learning has been demonstrated. It has been very common to learn mid-level discriminative parts for this problem, even before deep learning. Our *automatic discovery of discriminative parts* [C104,R9] casts this learning as a quadratic assignment problem, allowing the use of a number of known relaxations and optimization algorithms. It is based on pre-trained networks for feature extraction and classifies images using a part-based encoding. An unsupervised version of this work is applied both to classification and instance retrieval [C102,R12]. A related work [C105,R14] introduces *unsupervised object discovery* from regional CNN activations of an entire dataset, applied to instance retrieval. It eliminates the impact of background clutter and captures patterns that are both discriminative and common in the dataset. The dataset is indexed only focusing on discovered objects, in the form of a saliency map and a set of detected regions.

Building on the findings of previous work, particularly measuring similarity on manifolds, we introduce *mining on manifolds* [C108,R15], an unsupervised method for learning representations. We follow standard *metric learning* frameworks, where we define positive examples to be distant points on a single manifold, and negative examples to be nearby points on different manifolds. Both types of examples are revealed by disagreements between Euclidean and manifold similarities. This appears to be a unique way of improving visual representations by just looking at image collections.

[1] http://people.rennes.inria.fr/Ahmet.Iscen/diffusion.html

2009-2015

Image matching based on *local features* and *descriptors* has been one of the subjects of our earlier research, with particular emphasis on applications to *sub-linear indexing/retrieval* and *mining* in large image collections. According to the popular *bag-of-words* (BoW) model, a *visual codebook* encodes feature appearance in the descriptor space. Sub-linear indexing is due to the sparsity induced by a fine codebook. Research directions have been *e.g.* the construction of large scale codebooks, encoding and search for descriptors, the geometry of local features and the required memory footprint.

Visual codebook construction at a scale of *e.g.* 10^6 visual words is so far constrained to variants of k-means. Our *approximate Gaussian mixtures*[2] (AGM) [C91] is the first clustering method to apply a Gaussian mixture model (GMM) at this scale, employing *approximate nearest neighbor* (ANN) search in the EM algorithm to make complexity linear in the number of data points. The size of the codebook is dynamically estimated, resulting in less tuning and higher retrieval performance at the same cost with *approximate k-means* (AKM). There are numerous applications and extensions, in particular to supervised classification.

Beyond BoW model, performance may be increased by aggregating descriptors into a global image representation instead of quantizing. Such models are *Fisher vectors* and *vector of locally aggregated descriptors* (VLAD), and have been successful with small codebooks for tasks like recognition or retrieval. Descriptors may also be encoded into binary vectors as in *Hamming embedding* (HE) model. We are the first to explore such models with large codebooks for retrieval in [C93,J25], developing a common model that incorporates VLAD and HE as special cases, and achieving significant performance gain at a small memory cost[3]. The cost can be further reduced by *early burst detection* [C97].

With the use of encoded descriptors, the problem of image retrieval boils down to ANN search in high dimensions, where points are compressed. One of the most successful search methods in the compressed domain is *product quantization* (PQ), which decomposes the space into a Cartesian product of subspaces and independently applies vector quantization to each. An improvement is *optimized product quantization* (OPQ), which additionally optimizes subspace decomposition. Our method *locally optimized product quantization*[4] (LOPQ) [C95] applies these ideas locally and achieves lower distortion at nearly the same memory and search cost. It is far superior to all known methods in datasets of up to one billion vectors. In 2017, using a CNN image representation, Yahoo! Research has chosen LOPQ to index and provide a "similar image search" functionality on its entire Flickr collection (>10B images).[5]

Further research into using ideas related to PQ and the more recent *inverted multi-index* for clustering have resulted in *dimensionality-recursive vector quantization*[6] (DRVQ) [C94]. Traditionally, in the assignment step of the k-means algorithm, one needs to search for the nearest centroid for each data point. In approximate k-means (AKM), this search is accelerated by being approximate. In DRVQ, we rather start from centroids and construct a distance map over the entire space. Thus, search reduces to a lookup operation. The result is a clustering algorithm that is orders of magnitude faster than even AKM, and practically constant in the number of data points.

Feature *geometry* is traditionally only considered in a sequential process of *spatial verification* that is typically slow and only applied to a short list of top-ranking images. There is an increasing interest in *geometry indexing*. Existing methods are either not invariant or limited to local or weak constraints. Exploiting local feature shape (affine parameters), our *feature map hashing*[7] (FMH) [C85] has been the first to encode global feature geometry in the index, while enjoying invariance to affine transforms. It has been tested on up to 50K images, with query times of milliseconds.

Acknowledging that query times in practice depend on spatial verification alone, we have developed *Hough pyramid matching*[8] (HPM) [C89,J22], a very simple, generic *spatial matching* model. It accommodates for multiple surfaces or *flexible* objects, improving precision over any rigid motion model, including homography. While existing models are at least quadratic in the number of correspondences in the worst case, HPM is

[2] http://image.ntua.gr/iva/research/agm/
[3] http://image.ntua.gr/iva/research/asmk/
[4] http://image.ntua.gr/iva/research/lopq/
[5] https://yahooresearch.tumblr.com/post/158115871236/introducing-similarity-search-at-flick
[6] http://image.ntua.gr/iva/research/drvq/
[7] http://image.ntua.gr/iva/research/feature_map_hashing/
[8] http://image.ntua.gr/iva/research/relaxed_spatial_matching/

linear, reaching thousands of image matches per second.

A popular way to reduce the memory footprint of the index is *feature selection* through matching of multiple views of the same object or scene. We have developed a novel feature selection method based on our *alignment score* of [C85], that has scaled FMH up to 1M images [J23]. We have also applied HPM for this purpose: in particular, we have introduced SymCity[9] [C90], the first method to select features from *single views* by detecting symmetries and repeating patterns, with running times of just a few milliseconds.

A large part of our retrieval experiments make use of datasets originating from *community photo collections*. Such datasets are frequently accompanied by additional, non-visual information, like tags or geo-tags. At the same time they typically exhibit high redundancy, for instance tourist photos in city scenes. *Image clustering* is a popular way to deal with redundancy, though pairwise matching may be prohibitive in datasets of millions of images. Most existing methods are limited to clusters of popular images like landmarks.

In [C86], we have exploited geo-tags and sub-linear indexing to cluster an 1M dataset in only a few hours on a single processor. We employ *kernel vector quantization*, guaranteeing that isolated images are preserved, and that all images in a cluster share at least a rigid object or surface with one particular reference image. By projecting them on that image plane, we construct a *scene map*[10] [C86] for each cluster. Now, indexing scene maps instead of individual images not only saves space, but increases recall as well. This has a similar effect to *query expansion* methods, only now the process is off-line and query times are not affected.

We have developed a number of applications of the above results, including automated *location estimation* and geo-tagging, *recognition* of thousands of landmarks and points of interest, and *visualization* of photo clusters, landmarks and tourist paths on an on-line map. These results have been published in [J19] and are available in our application VIRaL[11]. On the other hand, our implementations have given rise to `ivl`[12], a general purpose, full-header template C++98 library that extends standard C++ syntax towards mathematical notation. Often resembling a new language, `ivl` targets concise, readable, yet efficient code.

All methods mentioned so far rely on the early vision task of detecting a sparse set of *local features* in images. The ability of a matching process to withstand changes in viewpoint or lighting crucially depends on the quality of detected features. Though most existing detectors treat features as regions of elliptic shape, we have recently developed a novel detector[13] of regions of arbitrary shape and scale [C84] starting with unstable, *single-scale edges*. Given the same input, our WαSH detector[14] [C92,C96,J24,J26] uses *weighted α-shapes* and outperforms most existing detectors in matching and retrieval experiments.

Generalizing the previous results and starting with single-scale image gradient, we compute the exact *weighted distance transform* and *weighted medial axis* and partition the image similarly to *topological watershed*. The resulting *medial feature detector*[15] (MFD) [C88] is similar in performance with WαSH, but provides pixel-accurate features.

We have also worked on *spatio-temporal* feature detection in video sequences. Along with appropriate spatio-temporal descriptors, a baseline BoW model and a nearest neighbor classifier, we have outperformed all relevant methods in *human action recognition*[16] [C83]. The detector itself is an extension of earlier *spatio-temporal saliency* models[17], which employ a competition approach across different feature channels and scales. Such models have been applied to a wide range of problems, including *visual attention modelling* [B8], *salient event detection* [B6,C82,J20], *movie summarization* [C65,C77,C82,J21], *human action recognition* [C62], *sports video classification* [C39,C44,J18], *image denoising* [C63], and *video coding* [C55]. Prior models have been published in [C31,C35,J10].

[9] http://image.ntua.gr/iva/research/symcity/
[10] http://image.ntua.gr/iva/research/scene_maps/
[11] http://viral.image.ntua.gr/
[12] http://image.ntua.gr/ivl/
[13] http://image.ntua.gr/iva/research/edge_based_feature_detection/
[14] http://image.ntua.gr/iva/research/wash/
[15] http://image.ntua.gr/iva/research/medial_features/
[16] http://image.ntua.gr/iva/research/spatiotemporal_feature_detection/
[17] http://image.ntua.gr/iva/research/visual_saliency/

1993-2009

A large part of earlier work concerns *object detection* and *image understanding*. In particular, we have modelled local feature geometry by means of multi-scale triangulation to develop a *logo detector* [C87] able to recognize thousands of classes. Local feature *grouping* has been the basis of a *region-of-interest* detector employed for *generic object detection* [C74]. Another result is *human face detection* based solely on color and shape on image partitions, applied to *face indexing* in video [C16,C17,J5], as well as *broadcast news parsing* [C15]. Modelling *facial expressions* has been the subject of [C12].

A number of generic approaches based on image partitions have followed. One important example is integrated *segmentation* and *region labelling* [C36,C81,J13] along with a spatio-temporal extension [C75]. Image *classification* using a *region codebook* [C71,C80,J16] is another example. We have also worked on the interaction of the two approaches by means of *visual context* [C49,C61,C76,J17]. In all cases, recognition is based on *region descriptors*. Descriptor *extraction* for moving objects has been the subject of [C26], while descriptor *fusion* has been studied using neural networks [C41] and support vector machines [C48].

Even earlier research has focused on deriving a global *image representation* from local descriptors. Here we combine image partitions obtained by different criteria, including color and motion statistics, giving rise to a sparse, multidimensional histogram. This work has been first published in [C2,C4] where we analyze the temporal evolution of video sequences in the descriptor space to detect extremal points for automated *video abstraction*. Introducing a correlation measure on sets of video frames, we have then attacked the same problem by means of *combinatorial optimization* [C7,C14,J1,J3]. Adding a *depth map* partition in the case of stereoscopic sequences, we have been able to compute a highly accurate and temporally consistent object support [B1,C13,J2]. Finally, we have applied the same representation to *video retrieval* [C8,C10].

On another research track, we have investigated *shape representation* for shape-based object matching, recognition, and retrieval. In particular, we have focused on *invariance* to rigid geometric transformations. While a large body of research at that time had been on extracting low-dimensional invariants with very limited discriminative power, we have followed a *normalization* approach whereby the entire shape information is retained, apart from six degrees of freedom relevant to affine transformations. This allows using any shape matching or recognition method. Interestingly, high performance has been achieved with a baseline matching approach in [C9,C18,J4]. Matching has been subsequently extended to *curvature scale space* in [C73]. Application to *object tracking* has been the subject of [C26,C51].

There are several other individual publications on diverse problems that do not exactly fit the topics discussed above, for instance *personalized image and video retrieval* [J12,J14], using *domain knowledge* and *taxonomies* in video classification [C25,J8,J15] and annotation [C38], *video archiving* [C23,J7], *ultrasonic imaging* [C6], *remote sensing* [C3], and *optical character recognition* [C1].

2 Teaching statement

As a Visiting Professor, I am teaching since 2017 one postgraduate course in *Research in Computer Science* (SIF) interdepartmental master, organized by a consortium of computer science universities and graduate schools in Brittany, including University of Rennes 1, University of Southern Brittany (UBS), ENS Rennes, National Institute of Applied Sciences, Rennes (INSA) and CentraleSupélec:

- *Deep learning for vision*[18], where I am the course responsible. The course studies learning visual representations for common computer vision tasks including matching, retrieval, classification, and object detection. Related problems are discussed including indexing, nearest neighbor search, clustering and dimensionality reduction. The course discusses well-known methods from low-level description to intermediate representation and their dependence on the end task. It then studies a data-driven approach where the entire pipeline is optimized jointly in a supervised fashion, according to a task-dependent objective. Deep learning models are studied in detail and interpreted in connection to conventional models. The focus of the course is on recent, state of the art methods and large scale applications.

[18] https://sif-dlv.github.io/

As an Adjunct Professor, I have been teaching two undergraduate courses between 2005 and 2014 at the *Electrical and Computer Engineering School* of the *National Technical University of Athens*:

- *Signals and systems*[19], where I have assisted in lectures and prepare exercise material. The course syllabus includes signal and system properties, convolution, correlation, sampling, quantization, Fourier series, discrete and continuous time Fourier transforms, Laplace and Z transforms, time and frequency analysis of linear, time-invariant systems, stability, and discrete Fourier transform.
- *Image and video analysis*[20], where I have assisted in lectures and conduct a weekly laboratory, using Matlab. The syllabus includes image sampling and quantization, two-dimensional transforms, image filtering, edge detection, enhancement and restoration, image and video coding and compression, and JPEG and MPEG standards. The laboratory includes additional topics, in particular Hough transform, corner and local feature detection, descriptor extraction, image retrieval, template matching and motion analysis.

Between 2009 and 2012, I have also given a number of informal lectures within the *Image and Video Analysis*[21] research team at NTUA:

- *Machine learning*, including linear classification and regression, kernel methods, graphical models, mixture models, EM, linear and non-linear PCA, decision trees, randomized forests, and boosting.
- *Linear optimization*, including convex polyhedra, the simplex method, duality, network flow problems, interior point methods, and topics in integer programming.

Besides these lectures, I have put together a *reading group*[22], studying recent computer vision and machine learning bibliography along with background material.

[19] http://cvsp.cs.ntua.gr/courses/systems/
[20] http://image.ntua.gr/courses_static/dip/
[21] http://image.ntua.gr/iva/
[22] http://image.ntua.gr/iva/reading_group/