Nuances in visual recognition

Gavves, E.

Publication date
2014

7

CONCLUSIONS

"Imagination is more important than knowledge. For knowledge is limited to all we now know and understand, while imagination embraces the entire world, and all there ever will be to know and understand."

– Albert Einstein (1879-1955)

7.1 SUMMARY OF CHAPTERS

In this thesis we aimed for exploring the importance of discovering fine nuances for visual understanding. We started from rigid objects, such as buildings and landmarks, continued with more generic object categories and finished with fine-grained bird and dog species. Before revisiting the questions posed in the introduction, we will give a brief overview of the chapters.

In Chapter 2, we address the incoherence problem of the visual words in bag-of-words vocabularies. Different from existing work, which assigns words based on closeness in descriptor space, we focus on identifying pairs of independent, distant words-the visual synonyms- that are likely to host image patches of similar visual reality. We focus on landmark images, where the image geometry guides the detection of synonym pairs. Image geometry is used to find those image features that lie in the nearly identical physical location, yet are assigned to different words of the visual vocabulary. Defined in this way, we evaluate the validity of visual synonyms. We also examine the closeness of synonyms in the ℓ_2 normalized feature space. We show that visual synonyms may successfully be used for vocabulary reduction. Furthermore, we show that combining the reduced visual vocabularies with synonym augmentation, we perform on par with the state-of-the-art bag-of-words approach, while having a 98% smaller vocabulary.

In Chapter 3 we consider the practical applicability of codebooks for retrieving images containing known objects. Where many have demonstrated the benefits of very large codebooks, containing up to 1 million codewords, we aim for small, yet robust codebooks. We propose supervised codebook construction via selection of the informative codewords, without the use of spatial cues. As the complete combinatorics of informative codeword selection is computationally intractable, we formalize the selection as a near-optimal cross-entropy approximation. More specifically, we consider the combinations of codewords that store the essential visual information to be rare. Importance sampling recovers informative codeword combinations. In contrast to recent selections based on geometric correspondences between codewords, our method is applicable to geometric and non-geometric image queries as we demonstrate on challenging known object datasets containing famous buildings and product logos. By its effectiveness in finding the most informative codes, the resulting tiny codebooks are very small with compression ratios to as much as 99.5% of a full codebook. Despite their small size, tiny codebooks improve retrieval rates over full codebooks, even when challenging sets of distractor images are added. The informative codewords appear to correspond to geometrically and visually meaningful details in the images, without being instructed to do so, which is bound to help greatly in the understanding of image search selections.

In Chapter 4 we consider the limiting factors of fast and effective classifiers for large sets of images with respect to their dependence on the number of images analyzed *and* the dimensionality of the image representation. Considering the growing number of images as a given, we aim to reduce the image feature dimensionality in this chapter. We propose reduced linear kernels that use only a portion of the dimensions to reconstruct a linear kernel. We formulate the search for these dimensions as a convex optimization problem, which can be solved efficiently. Different from existing kernel reduction methods, our reduced kernels are faster and maintain the accuracy benefits from non-linear embedding methods that mimic non-linear SVMs. We show these properties on both the Scenes and PASCAL VOC 2007 datasets. In addition, we demonstrate how the reduced kernels can be combined with Fisher vector and non-linear embeddings, leading to high accuracy. What is more, without using any labeled examples the selected and weighed kernel dimensions appear to correspond to visually meaningful patches in the images.

The aim of the Chapter 5 is fine-grained categorization without human interaction. Different from prior work, which relies on detectors for specific object parts, we propose to localize distinctive details by roughly aligning the objects using just the overall shape. Then, one may

proceed to the differential classification by examining the corresponding regions of the alignments. More specifically, the alignments are used to transfer part annotations from training images to unseen images (supervised alignment), or to blindly yet consistently segment the object in a number of regions (unsupervised alignment). We further argue that for the distinction of subclasses, distribution-based features like color Fisher vectors are better suited for describing localized appearance of fine-grained categories than popular matching oriented intensity features, like HOG. They allow capturing the subtle local differences between subclasses, while at the same time being robust to misalignments between distinctive details. We evaluate the local alignments on the CUB-2011 and on the Stanford Dogs datasets, composed of 200 and 120, visually very hard to distinguish bird and dog species. In our experiments we study and show the benefit of the color Fisher vector parameterization, the influence of the alignment partitioning, and the significance of object segmentation on fine-grained categorization. We show how rough alignments naturally blend with off-the-shelf object hypothesis algorithms resulting in fully automatic fine-grained categorization. The proposed local alignments set a new state-of-the-art on both the fine-grained birds and dogs datasets, even without any human intervention. What is more, the local alignments reveal what appearance details are most decisive per fine-grained object category.

In Chapter 6 we aim for segmentation and classification of objects. We propose codemaps that are a joint formulation of the classification score and the local neighborhood it belongs to in the image. We obtain the codemap by reordering the encoding, pooling and classification steps over lattice elements. Other than existing linear decompositions who emphasize only the efficiency benefits for localized search, we make three novel contributions. As a preliminary, we provide a theoretical generalization of the sufficient mathematical conditions under which image encodings and classification becomes locally decomposable. As first novelty we introduce ℓ_2-normalization for arbitrarily shaped image regions, which is fast enough for semantic segmentation using our Fisher codemaps. Second, by exploiting the independence of codemaps to the number of classifiers employed, we introduce locality into attribute learning. Third, we focus on fixed lattices, as is the case in object classification. We propose kernel pooling with codemaps, which embeds nonlinearities for object classification by explicit or approximate feature mappings in a recursive manner, arriving at a feedforward architecture. Results demonstrate that ℓ_2-normalized Fisher codemaps improve the state-of-the-art in semantic segmentation for PASCAL VOC2011 and VOC2012. Furthermore, by injecting locality, we notably improve the state-of-the-art in fine-grained classification using attributes on the challenging Caltech-UCSD 200-2011 birds. Finally, we show that the feedforward architecture of codemaps improves classification accuracy for free in an object classification setting tested on PASCAL VOC 2007. We conclude that codemaps are accurate and efficient, enabling us to exploit locality with strong state-of-the-art features to benefit a variety of tasks.

7.2 GENERAL CONCLUSION

The aim of this thesis was to study the fine nuances that could lead to a better visual understanding. Several aspects of this problem have been explored, each of them contributing one more piece to the puzzle.

We began with the question: *is the machine representation of physically identical elements constant?* Our findings support the conclusion that the state-of-the-art image representations using large codebooks can be rather unstable, especially when one attempts to focus on the fine nuances; when going too deep, nuances that correspond to the same physical elements will at some point become too different and in the end will be treated differently. By employing geometry the detected nuances are being described in a stable way, but they are not necessarily optimal as image representations. Hence, we posed our second question: *what set of nuances matters*

most for object search? We found that by treating object search from an importance sampling perspective, we are able to discover the few visual nuances that are particularly accurate for certain objects. Encouraged by the result, we investigated the question of *what set of nuances matters most for object classification*. We expressed the problem in terms of regularized least squares. We found that the nuances are not only descriptive among categories, but they are also capable of explaining the contribution of less descriptive nuances. For our next question we addressed *which visual nuances discriminate fine-grained object categories*, where discovering the important nuances is vital and not just helpful for successful classification. Our findings support the idea that fine-grained object categories usually share similar shapes and poses, generally found in their common-super category. Exploiting their common spatial characteristics, fine-grained objects can be examined by their local details from corresponding object regions. We ended up with a highly accurate recognition machine, in a problem considered difficult even for (non-expert) humans. Last, we questioned *whether we can decompose the interpretation of image nuances into local, reusable components, allowing for exact region-level nuance reconstruction*. The results showed that this is possible, enabling us to use strong, state-of-the-art features for local search. This in turn lead to excellent semantic segmentation and object classification results.

As a future direction, we note that human recognize easily about 30,000 basic categories [12]. Experts will recognize many, many more fine-grained sub-categories. For so many categories a number of them will appear similar to the point where distinguishing nuances are subtle. Harvesting enough training data in order to uncover such subtle nuances for so many categories would be rather difficult. This is especially true when we would refine them further with respect to their location in the image.

To ease the learning on so many categories, one possibility would be to use knowledge transfer [98], where one could exploit prior experience to guide the machine in future learning. Another possibility would be to use zero-shot learning [70], where one could use an intermediate layer of representations for recognition. In [90] it was discovered that the exploitation of non-visual patterns, such as object-object relationships, is an interesting avenue for zero-shot classification. Moreover, as fine-grained classification approaches instance search, in [128] evidence is presented that in instances for which only a single image is available, locality makes all the more sense. Last, as recent advances of deep learning architectures have revealed, learning features from data is the ultimate way of representing the appearance of fine nuances. All in all, computer vision is bound to discover more flexible ways of learning the visual nuances that facilitate a reliable and practical recognition of such an extensive palette of categories.

We close this thesis by revisiting our fundamental question, *i.e. what are the fine details, the nuances, we need to discover to make sense of a visually complex image*. We note that there is no universal answer, it simply depends on what we are looking for. In fact, to find the right nuances we need to see them for what they are and not for what we want them to be. To understand the Penrose triangle of the front page picture, for example, as the back page reveals we need to rotate the triangle a bit and see it from a different perspective.