Image Hub Explorer: Evaluating Representations and Metrics for Content-based Image Retrieval and Object Recognition

Nenad Tomašev and Dunja Mladenić

Institute Jožef Stefan
Artificial Intelligence Laboratory
Jamova 39, 1000 Ljubljana, Slovenia
nenad.tomasev@ijs.si, dunja.mladenic@ijs.si

Abstract. Large quantities of image data are generated daily and visualizing large image datasets is an important task. We present a novel tool for image data visualization and analysis, Image Hub Explorer. The integrated analytic functionality is centered around dealing with the recently described phenomenon of *hubness* and evaluating its impact on the image retrieval, recognition and recommendation process. Hubness is reflected in that some images (*hubs*) end up being very frequently retrieved in 'top k' result sets, regardless of their labels and target semantics. Image Hub Explorer offers many methods that help in visualizing the influence of major image hubs, as well as state-of-the-art metric learning and hubness-aware classification methods that help in reducing the overall impact of extremely frequent neighbor points. The system also helps in visualizing both beneficial and detrimental visual words in individual images. Search functionality is supported, along with the recently developed hubness-aware result set re-ranking procedure.

Keywords: image retrieval, object recognition, visualization, k-nearest neighbors, metric learning, re-ranking, hubs, high-dimensional data

1 Introduction

Image Hub Explorer is the first image collection visualization tool aimed at understanding the underlying *hubness* [1] of the k-nearest neighbor data structure. Hubness is a recently described aspect of the well known *curse of dimensionality* that arises in various sorts of intrinsically high-dimensional data types, such as text [1], images [2] and audio [3]. Its implications were most thoroughly examined in the context of music retrieval and recommendation [4]. Comparatively little attention was given to emerging hubs and the skewed distribution of influence in image data. One of the main goals of the Image Hub Explorer was to enable other researchers and practitioners to easily detect hubs in their datasets, as well as test and apply the built-in state-of-the-art hubness-aware data mining methods.

2 Resources

Image Hub Explorer is an analytics and visualization tool that is built on top of the recently developed *Hub Miner* Java data mining library that is focused on evaluating various types of kNN methods. Additional resources on the use of Image Hub Explorer (`http://ailab.ijs.si/tools/image-hub-explorer/`) and the Hub Miner library (`http://ailab.ijs.si/nenad\_tomasev/hub-miner-library/`) are available online. This includes the demo video: `http://youtu.be/LB9ZWuvm0qw`.

3 Related Work

3.1 Hubs in High-dimensional Data

The concentration of distances [5] in intrinsically high-dimensional data affects the distribution of neighbor occurrences and causes *hubs* to emerge as centers of influence in form of very frequent neighbor points. The kNN hubs often act as semantic singularities and are detrimental for the analysis [6]. Different representations and metrics exhibit different degrees of neighbor occurrence distribution skewness [2]. Selecting an appropriate feature representation paired with an appropriate distance measure is a non-trivial task and very important for improving system performance. This is what Image Hub Explorer was designed to help with.

Hubness-aware methods have recently been proposed for instance selection [7], clustering [8], metric learning [9][4], information retrieval [10], classification [11] and re-ranking [12]. Most of these methods are implemented and available in Image Hub Explorer.

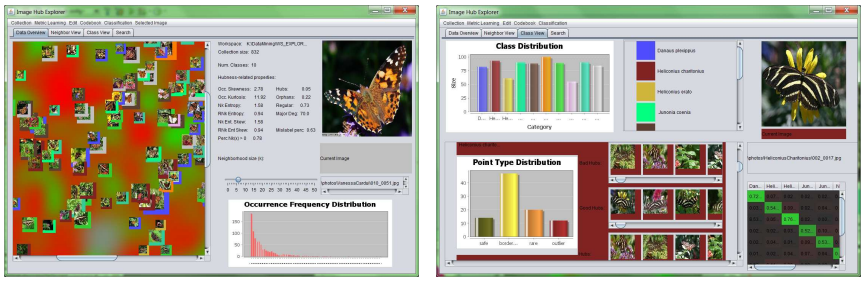
3.2 Visualizing Image Collections

Visualization plays an essential role in examining large image databases. Several similarity-based visualization approaches have been proposed [13][14] and ImagePlot (http://flowingdata.com/2011/09/18/explore-large-image-collections-with-imageplot/) is a typical example. What these systems have in common is that they mostly focus on different ways of similarity-preserving projections of the data onto the plane, as well as selection strategies that determine which images are to be shown. Some hierarchical systems are also available [15]. These systems allow for quick browsing through large collections, but they offer no support for examining the distribution of influence and detection of emerging hub images.

4 System Components and Functions

Image Hub Explorer implements several views of the data, to facilitate easier analysis and interpretation. All images in all views can be selected and the information is shared among the views and updated automatically. The desired neighborhood size k is controlled by a slider and its value can be changed at any time.

Metric Learning plays an important role in the analysis. For any loaded data representation, many different metrics can be employed. Image Hub Explorer supports 7



(a) MDS Data Overview (b) Class Comparisons

Fig. 1. Screenshots of some selected Image Hub Explorer functions

primary metrics and 5 secondary metrics that are learned from the primary distance matrices. This includes two recently proposed hubness-aware distance measures: $simhub_s$ [9] and mutual proximity [4].

Data Overview gives a quick insight into the hub-structure of the data (Fig. 1(a)). The most influential image hubs are projected onto a visualization panel via multi-dimensional scaling (MDS) and the background coloring is determined based on the nature of their occurrences - green denotes the beneficial influences, red the detrimental ones. The occurrence frequency distribution is shown, followed by a set of statistics describing various aspects of neighbor occurrences and kNN set purity.

Neighbor View offers a deeper insight into the local neighbor structure. For each selected image, its k-neighbors and reverse k-neighbors are listed and any selected image can be inserted into the local subgraph visualization panel. This allows the user to visualize all kNN relations among a selected set of images as a graph, with distance values displayed on the edges.

Class View allows the user to compare the point type distributions among different classes, as well as a way to quickly select and examine the top hubs, good hubs and bad hubs for each class separately(Fig. 1(b)). Additionally, the global class-to-class occurrence matrix can be used to determine which classes cause most label mismatches in k-neighbor sets and which classes these mismatches are directed at.

Search, Re-ranking and Classification : Apart from simple browsing, image search is an important function in examining large image databases. Image Hub Explorer supports image queries, for which a set of k most similar images from the database is retrieved. Image Hub Explorer implements 8 different kNN classification methods to help with image labeling, as well as a hubness-aware result set re-ranking procedure [12].

Feature Assessment for quantized feature representations can easily be performed in Image Hub Explorer, as it calculates the occurrence profile for each visual word (codebook vector) and determines which features help in increasing the intra-class similarity and which increase the inter-class similarity. Beneficial and detrimental features and texture regions can be visualized on each image separately.

5 Applicability

The Image Hub Explorer system can also be used to visualize other data types, when rectangular nodes are shown instead of the loaded image thumbnails. Only the feature visualization and image search functions are restricted to working with image data specifically.

Acknowledgements. This work was supported by the Slovenian Research Agency, the ICT Programme of the EC under XLike (ICT-STREP-288342), and RENDER (ICT-257790-STREP).

References

1. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research **11** (2010) 2487–2531
2. Tomašev, N., Brehar, R., Mladenić, D., Nedevschi, S.: The influence of hubness on nearest-neighbor methods in object recognition. In: Proceedings of the 7th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP). (2011) 367–374
3. Aucouturier, J., Pachet, F.: Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences **1** (2004)
4. Schnitzer, D., Flexer, A., Schedl, M., Widmer, G.: Local and global scaling reduce hubs in space. Journal of Machine Learning Research (2012) 2871–2902
5. François, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE Transactions on Knowledge and Data Engineering **19**(7) (2007) 873–886
6. Radovanović, M., Nanopoulos, A., Ivanović, M.: Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In: Proc. 26th Int. Conf. on Machine Learning (ICML). (2009) 865–872
7. Buza, K., Nanopoulos, A., Schmidt-Thieme, L.: Insight: efficient and effective instance selection for time-series classification. In: Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining - Volume Part II. PAKDD'11, Berlin, Heidelberg, Springer-Verlag (2011) 149–160
8. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: The role of hubness in clustering high-dimensional data. IEEE Transactions on Knowledge and Data Engineering **99**(PrePrints) (2013) 1
9. Tomašev, N., Mladenić, D.: Hubness-aware shared neighbor distances for high-dimensional k-nearest neighbor classification. Knowledge and Information Systems (2013)
10. Tomašev, N., , Rupnik, J., Mladenić, D.: The role of hubs in cross-lingual supervised document retrieval. In: Proceedings of the PAKDD Conference. PAKDD 2013 (2013)
11. Tomašev, N., Mladenić, D.: Nearest neighbor voting in high dimensional data: Learning from past occurrences. Computer Science and Information Systems **9** (2012) 691–712
12. Tomašev, N., , Leban, G., Mladenić, D.: Exploiting hubs for self-adaptive secondary re-ranking in bug report duplicate detection. In: Proceedings of the ITI conference. ITI 2013 (2013)
13. Nguyen, G.P., Worring, M.: Similarity based visualization of image collections. In: In Intl Worksh. Audio-Visual Content and Information Visualization in Digital Libraries. (2005)
14. Nguyen, G.P., Worring, M.: Interactive access to large image collections using similarity-based visualization. J. Vis. Lang. Comput. **19**(2) (April 2008) 203–224
15. Tomašev, N., Fortuna, B., Mladenić, D., Nedevschi, S.: Ontogen extension for exploring image collections. In: Proceedings of the 7th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP). (2011)