Recognition and Analysis of Objects in Medieval Images

Pradeep Yarlagadda, Antonio Monroy, Bernd Carque and Björn Ommer

Interdisciplinary Center for Scientific Computing, University of Heidelberg, Germany
{pyarlaga,amonroy,bcarque,bommer}@iwr.uni-heidelberg.de

Abstract. Rapid and cost effective digitization techniques have lead to the creation of large volumes of visual data in recent times. For providing convenient access to such databases, it is crucial to develop approaches and systems which search the database based on the representational content of images rather than the textual annotations associated with the images. The success of such systems depends on one key component: category level object detection in images.

In this contribution, we study the problem of object detection in the application context of digitized versions of ancient manuscripts. To this end, we present a benchmark image dataset of medieval images with groundtruth information for objects such as 'crowns' in the image dataset. Such a benchmark dataset allows for a quantitative comparison of object detection algorithms in the domain of cultural heritage, as illustrated by our experiments. We describe a detection system that accurately localizes objects in the database. We utilize shape information of the objects to analyze the type-variability of the category and to manually identify various sub-categories. Finally, we report a quantitative evaluation of the automatic classification of object into various sub-categories.

1 Introduction

Large scale digitization efforts in the field of cultural heritage have lead to the accumulation of vast amounts of visual data in recent times. For a systematic access to such collections, it is necessary to develop algorithms that search the database based on the representational content of the images. For this, it is necessary to go beyond a mere analysis of individual image pixels onto a stage where the semantics of images can be modeled and analyzed. In contrast to this semantics based indexing, the current retrieval systems depend almost exclusively on queries which are directed at the textual metadata. Textual annotations provide only limited search options because of the infeasibility of comprehensive manual indexing. To make image databases accessible in a quicker, more reliable and detailed way, semantics based indexing is indeed necessary. The key for such algorithms is category level object detection.

In this contribution, we explore the question of category level object detection in the context of a benchmark dataset for cultural heritage studies.This

dataset is highly significant because of its completeness of late medieval workshop production and also it is the first of its kind to enable benchmarking of object detection and retrieval in pre-modern tinted drawings. We also present a statistical analysis of the variability and relations within object categories i.e medieval crowns. The analysis yields a single 2-d visualization of the diversity found among large numbers of instances of a category. Such a visualization can be augmented to the search results of a semantics based query system and the amount of insight it provides into the database cannot be matched by a text based query system.

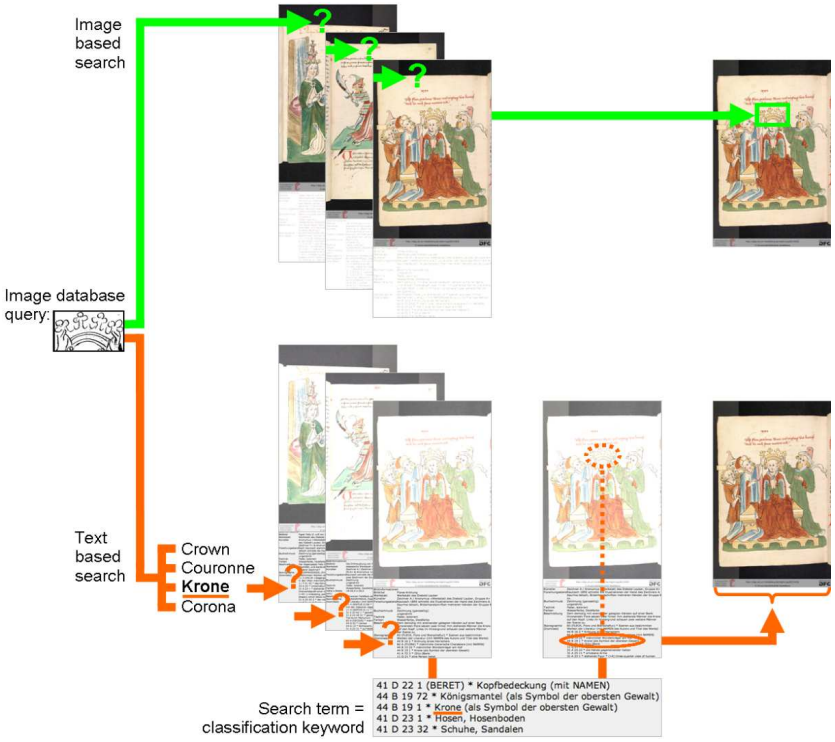


Fig. 1: Text based vs image based retrieval.

2 Related Work

Image databases in the field of cultural heritage are normally made accessible via textual annotations referring to the representational content of the images [1]. Therefore, content-based image retrieval depends on either the controlled vocabularies of the used classification systems or the textual content of free descriptions. In both cases only that can be found what has been considered in

the process of manual indexing; and it can only be found in the specific form in which it has been verbalized. The inevitability of textual descriptions generates numerous problems, for example concerning the scope and detailedness of the taxonomies, their compatibility beyond linguistic [2], professional or cultural boundaries, their focus on specific aspects of the content according to specific scientific interests or not least the qualification and training of the cataloguer. One of the most sophisticated classification systems is ICONCLASS [3]. Yet, despite its high level of differentiation it has severe limits in a global perspective because it was developed only to cover Western art and iconography. Therefore its ability to index for instance transcultural image resources such as the database of the Cluster of Excellence Asia and Europe in a Global Context at the University of Heidelberg [4] is limited. Furthermore, object definition schemes are featuring a very limited differentiation. In our showcase 'crown' the hierarchy of objects ends with this general notion and does not offer varying types of crowns. To focus the object retrieval on subtypes is, in contrast, possible in the case of REALonline, the most important image database in the field of medieval and early modern material culture [5]. Here, the controlled vocabulary contains a few compounds like 'Buegelkrone' or 'Kronhut'. But whereas the main division 'Kleidung–Amtstracht' is searchable in German and in English, these subdivisions are available only in German, thus raising difficulties of translation. Problems such as the lack of detail and connectivity are even greater in the case of heterogeneous databases, which are –like HeidICON [6], Prometheus [7] or ARTstore [8] –generated by the input from different institutional and academic contexts. In such cases, the cataloguing of the image content is almost arbitrary due to the uncontrolled textual descriptions. Finally, a basic problem of all these databases is the fact that –due to the serious efforts of manual indexing in terms of cost and time –the fast-growing number of images that are available in a digital format can hardly be itemized in detail and thus cannot be used efficiently in the long term. To overcome these restrictions, we present a system that directly searches the visual data thereby circumventing the need for detailed textual annotations.

Compared to standard benchmark datasets used in computer vision (e.g. [9, 10]), we present a database with a high degree of background clutter, scale variation, and within-class variability. Being close to the needs in the field of cultural heritage, this image collection is highly challenging for categorization algorithms, e.g. [9], voting methods for detection such as [11, 10, 12], and sliding window based classifiers [13].

3 Benchmarking, Analysis, and Recognition

3.1 Database and Benchmarking

We have assembled a novel, annotated benchmark image dataset for cultural heritage from a corpus of 27 late medieval paper manuscripts, held by Heidelberg University Library [14]. Produced between 1417 and 1477 in three important Upper German workshops, this corpus is rare in its magnitude and, in addition,

offers an exceptional homogeneity concerning its date of origin, its provenance and its technical execution. More than 2,000 half- or full-page tinted drawings illustrate religious and devotional texts, chronicles and courtly epics. Their content has been itemized by means of ICONCLASS, so that we are able to evaluate the capability of the classification system and to detect its desiderata. For this purpose we built a unique dataset of annotations, which covers object categories in a more detailed way than any existing taxonomy, e.g. more than 15 different subtypes of crowns. Thus, the demands on our object retrieval system can be defined precisely. Although our approach is quite generic which can be applied to different object categories, we start from the category which has a high semantic validity since it belongs to the realm of medieval symbols of power [15]. This ensures that our analysis has the highest possible connectivity to research in the humanities, e.g to art history and history with a focus on ritual practices [16] or on material culture.

Fig. 2: Sample images from the late medieval manuscripts.

Breakthroughs entailed by a novel benchmark dataset: Our motivation for introducing a novel benchmark dataset is spurred by the influence the Berkeley Segmentation Dataset (BSDS) [17] has had on the development and evaluation of segmentation algorithms. Before BSDS, measuring segmentation performance was mostly subjective and algorithms were difficult to compare. The new BSDS dataset with its groundtruth annotation has, for the first time, provided an objective performance measure for segmentation. This has stimulated algorithm development which lead to previously unexpected breakthroughs in segmentation performance. The F-measure, which is a suitable metric for comparing the performance of segmentation algorithms, has only seen a slight increase in the years before BSDS. Early segmentation algorithms such as Roberts (1965) [18] and Canny (1986) [19] achieved F-measures of 0.47 and 0.53, respectively. In the

short time since the introduction of BSDS in 2001, contributions such as [20] have increased the performance to 0.7 while human performance stands at 0.79.

Annotating the data: In order to generate groundtruth localizations for objects in the images, we developed an interactive annotation system. Using the expertise of an art historian we have gathered groundtruth annotations. Cubic splines are used to fit a bounding region to the principal curvature of an object. This helps excluding more background from the bounding boxes compared to rectangular bounding boxes.

3.2 Object Analysis

The most basic component for object analysis and object recognition is choosing an appropriate mathematical representation for objects which lays the foundation for recognition and further analysis. We utilize a shape based representation of objects since shape is an important cue in these medieval manuscripts.

Extracting artistic drawings to represent shape: We have discovered from experiments that the images when represented in HSV color space, particularly the saturation component, provide a good starting point for object boundary extraction. Object boundaries are essentially ridges in an image with few pixels thickness. To detect such ridges, we apply a filter which smoothes the image along the direction orthogonal to the ridge and sharpens the image along the direction of the ridge,called the ridge detection filter [21]. It is defined by the following formula.

$$G(x, y, \sigma_x, \sigma_y) = \frac{1}{\pi * \sigma_x^2} * \left(1 - \frac{x^2}{2 * \sigma_x^2}\right) * \exp\left(-\frac{x^2}{{\sigma_x}^2} - \frac{y^2}{{\sigma_y}^2}\right) \tag{1}$$

Coordinates x,y denote image location, σ_x, σ_y determine the support of the ridge filter along the x and y directions. Equation 1 defines the ridge filter assuming that the ridge is oriented along the x-axis. This formula is easily extended for detecting ridges at an orientation θ.

At each point in the image, optimization over the parameters σ_x, σ_y and θ yields the maximal filter response. Images marked 1 and 2 in fig. 1 shows an input image and the result of applying the ridge filter to the input.

Shape representation: Ridges are rerpresented using orientation histograms. We compute these Histograms of Oriented Gradients (HoG) [22] on a dense grid of uniformly spaced cells in the image. We combine histograms from 4 different scales and 9 orientations into a 765 dimensional feature vector.

Automatic discovery of intra-category structure: We capture the relationship between various object instances in the database in a single plot by embedding high dimensional HoG feature vectors into a low dimensional space. Such a plot makes it convenient for researchers from cultural heritage to discover relationships without having to study thousands of images. In a first step pairwise clustering based on HoG descriptors is employed to discover the hierarchical substructure of crowns. Then we compute the pairwise distances for samples

in the vicinity of the cluster prototypes. Thereafter, a distance preserving low-dimensional embedding is computed to project the 765 dimensional feature vectors onto a 2-d subspace that is visualized in fig. 4. This procedure has extracted relationships, variations and substructure of an object category out of hundreds of images and makes these directly apparent.

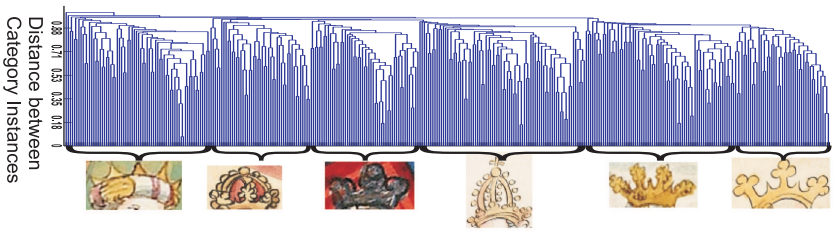


Fig. 3: Hierarchy of substructure in object category 'crown'.

The plot displays two central findings of our recognition system and thus reveal the potential of the approach: i) the high type-variability within a category and ii) the different principles of artistic design. In particular, our clusters for the category 'crown' show that to the simple crown circlet (A) varied elements like arches (B1), lined arches (B2), torus-shaped brims (B3), hats, or helmets are added. Thus, objects provide advanced semantic information concerning e.g. social hierarchies, which is not displayed by the common taxonomies. Since an automated image-based search does not suffer from the desiderata of annotation taxonomies, it becomes a crucial instrument to assist with the detailed differentiation of such subtypes, combining data from large numbers of images and organizing the compositional complexity of objects into a hierarchy of formal variants. Moreover, the clustering and visualization in a MDS-plot (fig. 4) features different principles of artistic design, which are characteristic for different workshops engaged with the illustrations. Group (B) indicates the concise and accurate style, mainly based on definite contours, of the Hagenau workshop of Diebold Lauber [23], group (A) the more delicate and sketchy style of the Swabian workshop of Ludwig Henfflin, and group (C) the particular summary style of the so-called 'Alsatian Workshop of 1418'. This detection of specific drawing styles is a highly relevant starting point to differentiate large-scale datasets by workshops, single teams within a workshop, or even by individual draftsmen.

3.3 Object Recognition

Objects are detected by classifying image regions as object or background using a support vector machine with intersection kernel [24]. This detection algorithm scans the image on multiple scales and orientations. Image regions are represented using the shape representation from subsection 3.2 and a color histogram. The necessary codebook of representative colors is obtained by first quantizing

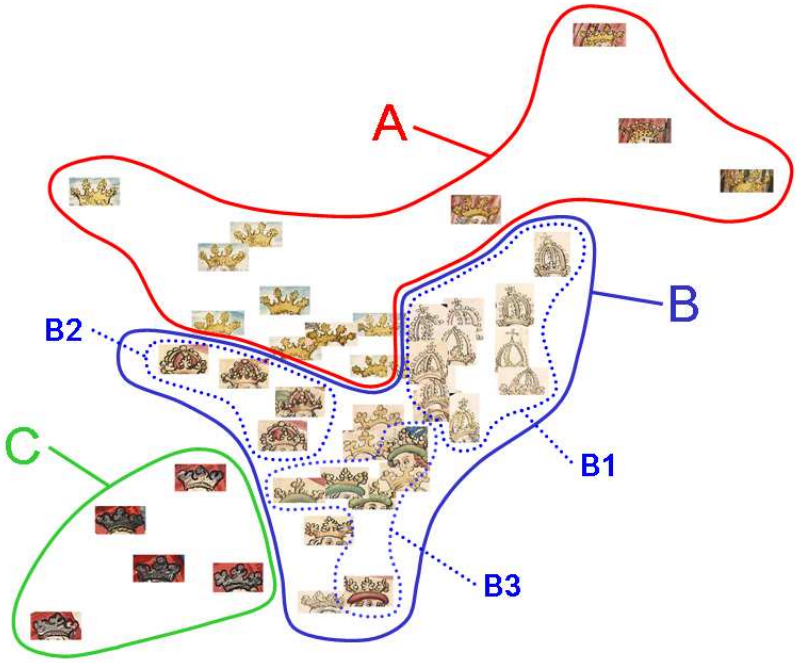


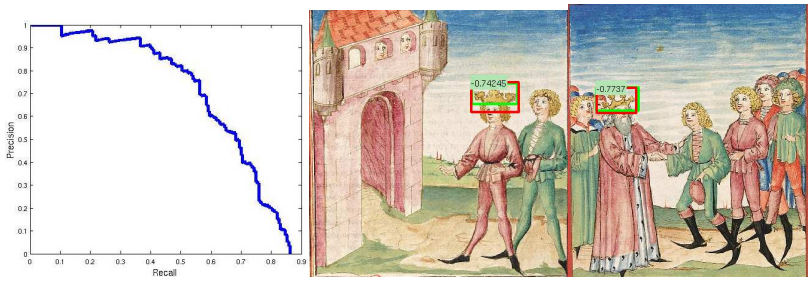
Fig. 4: Visualization of Intra-category variability and substructure of crowns. Group A shows the Swabian workshop of Ludwig Henfflin. Group B shows the Hagenau workshop of Diebold Lauber with the subgroups of crowns with arches (B1), crowns with lined arches (B2) and crowns with torus-shaped brims (B3). Group C shows the Alsatian workshop of 1418.

training image using minimum variance quantization into a set of 100 prototypical clusters per image. The bias towards large, homogenous regions is resolved by clustering all these prototypes into an overall set of 30 prototypical colors. We count an object hypothesis as correct if $\frac{A_h \cap A_g}{A_h \cup A_g} >= 0.4$ where A_h and A_g is the area of the predicted and the groundtruth bounding box, respectively. The precision-recall curve in part a) of fig. 5 shows the detection performance achieved by the presented approach.

The precision recall curves in fig. 5 show scope for improvement as the curves are far from reaching the saturation stage. A closer look at the detection results revealed a lot of false positives in the images which were not sufficiently represented during the training stage of the SVM. To deal with this issue, we have incorporated a bootstrap training procedure to focus on difficult negative samples as is motivated by [25, 26]. Training starts as before by learning an SVM model on all positive training samples and an equally sized, random set of negative samples, i.e. bounding boxes drawn from the background. In the next round, negative samples which are either incorrectly classified by the model or fall inside

(a) (b) (c)

Fig. 5: a) Precision recall curve for crowns obtained from HoG and HoG plus color features. b) Crowns detected in a test image. c) Response of our object detector at each image location.



(a) (b) (c)

Fig. 6: a) Precision recall curve for crowns obtained by using a bootstrapping training procedure. b) and c) Crowns detected in test images along with the SVM scores.

the margin (defined by the SVM classifier) are added to the training set. Also, positive samples which are classified correctly and fall outside the margin are removed from the training set. This process is repeated iteratively until there are no new hard negative samples that can be added to the training set. This iterative training procedure resulted in a significant improvement in the detection performance and the resulting PR curves are presented in fig. 6 along with two examples of detections in test images.

Accurate localization of objects within the images as shown in fig. 5, makes complex representations like battle scenes or coronations with several symbols of power more easily readable. Textual annotations do not provide localization information so that object detection and reasoning about the spatial relationship between objects or about their performative context [27] remains impossible.

In section 3.2, we have presented an unsupervised approach to identify category substructure which has then lead to a visualization (fig. 4) of the different artistic workshops that have contributed to the Upper German manuscripts. Based on this visualization, art historians have provided us with groundtruth

Workshops pred.: correct:	A	B	C
A	0.9836	0.0163	0
B	0.0365	0.9634	0
C	0.0083	0.0083	0.9833

Table 1: Classification results on the crowns from workshops corresponding to groups A, B and C in fig. 4. Columns are the predicted workshop labels and rows are the correct labels. A: Swabian workshop of Ludwig Henfflin, B: Hagenau workshop of Diebold Lauber and C: Alsatian workshop of 1418. The average classification accuracy is 97.67 $\pm$ 1.7 %.

information so that we can conduct a quantitative evaluation: they have labeled all crowns in the dataset with the workshop that they come from based on formal criteria [23]. There are 137 crowns in our dataset that belong to group A (the workshop of Ludwig Henfflin), 106 crowns belong to group B (the workshop of Diebold Lauber) and 23 crowns belong to group C (the Alsatian workshop). We then incorporate a discriminative approach for predicting the workshop that a crown belongs to. This multi-class classification problem is tackled using the features from before and incorporating SVM in a one-versus-all manner. For evaluation, we apply 10-fold cross-validation: In each round, 50 % of the crowns from each group have been used for training and the remaining 50 % of the crowns are used for testing by holding back their labels. The classification results of the crowns according to the workshops are presented in table 1 in the form of a confusion matrix.

4 Discussion and Conclusions

The present case study on the Upper German manuscripts of Heidelberg University Library shows the detection results that can be obtained by state-of-the-art category level object recognition techniques in the context of cultural heritage. It is now possible to automatically discover the substructure of object categories which is, for instance, caused by different subtypes or principles of artistic design. In order to refine our method, we will apply it in a second step to the entire corpus of the Upper German manuscripts and, in a third step, to the remaining c. 5,000 images of the Codices Palatini germanici ([28]), which have, for the most part, not previously been labeled.

References

1. Baca, M., Harpring, P., Lanzi, E., McRae, L., Whiteside, A.: Cataloging Cultural Objects. A Guide to Describing Cultural Works and Their Images. (2006)
2. Kerscher, G.: Thesaurus-Verwendung und internationalisierung in Bilddatenbanken. Kunstchronik **57** (2008) 606–608
3. van Straten, R.: Iconography, Indexing, ICONCLASS. A Handbook. (1994)

4. : (http://www.asia-europe.uni-heidelberg.de/research/heidelberg-research-architecture/hra-databases-1/transcultural-image-database/the-image-database)
5. : (http://www.imareal.oeaw.ac.at/realonline/)
6. : (http://heidicon.ub.uni-heidelberg.de)
7. : (http://www.prometheus-bildarchiv.de)
8. : (http://www.artstor.org)
9. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Intl. Conf. on Comp. Vision and Pat. Rec. (2003)
10. Ferrari, V., Jurie, F., Schmid, C.: From images to shape models for object detection. Intl. Journal of Comp. Vision (2009) 40–82
11. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Europ. Conf. on Comp. Vision. (2004)
12. B.Ommer, Malik, J.: Multi-scale object detection by clustering lines. In: Intl. Conf. on Comp. Vision and Pat. Rec. (2009)
13. Lampert, C., Blaschko, M., Hofmann, T.: Beyond sliding windows: Object localization by efficient subwindow search. In: Intl. Conf. on Comp. Vision and Pat. Rec. (2008)
14. Pietzsch, E., Effinger, M., Spyra, U.: (Digitalisierung und Erschließung spätmittelalterlicher Bilderhandschriften aus der Bibliotheca Palatina. H. Thaller,editor. Digitale Bausteine für die geisteswissenschaftliche Forschung)
15. Schramm, P.E.: Herrschaftszeichen und Staatssymbolik. (1954) 3 vols.
16. Schwedler, G., Meyer, C., Zimmermann, K., eds.: Rituale und die Ordnung der Welt. (2008)
17. Fowlkes, C., Tal, D., Martin, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Intl Conf. on Comp. Vision. (2001)
18. Roberts, L.: Machine perception of three-dimensional solids. Optical and electro-optical information processing (1965) 159–197
19. Canny, J.: A computational approach to edge detection. IEEE Trans.Pat. Analysis and Machine Intelligence (1986) 679–714
20. Arbelaez, P., Fowlkes, C., Maire, M., Malik, J.: Using contours to detect and localize junctions in natural images. In: Intl. Conf. on Comp. Vision and Pat. Rec. (2008)
21. Kovesi, P.D.: MATLAB and Octave functions for computer vision and image processing. (http://www.csse.uwa.edu.au/∼pk/research/matlabfns/)
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Intl. Conf. on Comp. Vision and Pat. Rec. (2005)
23. Saurma-Jeltsch, L.E.: Spätformen mittelalterlicher Buchherstellung. Bilderhandschriften aus der Werkstatt Diebold Laubers in Hagenau. (2001) 2 vols.
24. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: Intl. Conf. on Comp. Vision and Pat. Rec. (2008)
25. Davison, A., Hinkley, D.: Bootstrap Methods and their Application. Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics. (1997)
26. Felzenszwalb, P.F., Girshick, R.B., Mcallester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
27. Petersohn, J.: Über monarchische Insignien und ihre Funktion im mittelalterlichen Reich. Historische Zeitschrift **266** (1998) 47–96
28. : (http://www.ub.uni-heidelberg.de/helios/digi/codpalgerm.html)