EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2014)
R. Klein and P. Santos (Editors)

An Approach to Large Scale Interactive Retrieval of Cultural Heritage

Masato Takami[1,2], Peter Bell[1] and Björn Ommer[1]

[1]Heidelberg Collaboratory for Image Processing & IWR, University of Heidelberg, Germany
[2]Robert Bosch GmbH, Corporate Research, Computer Vision Research Lab, Hildesheim, Germany

Abstract

Large scale digitization campaigns are simplifying the accessibility of a rapidly increasing number of images from cultural heritage. However, digitization alone is not sufficient to effectively open up these valuable resources. Retrieval and analysis within these datasets is currently mainly based on manual annotation and laborious preprocessing. This is not only a tedious task, which rapidly becomes infeasible due to the enormous data load. We also risk to be biased to only see what an annotator beforehand has focused on. Thus a lot of potential is being wasted.

One of the most prevalent tasks is that of discovering similar objects in a dataset to find relations therein. The majority of existing systems for this task are detecting similar objects using visual feature keypoints. While having a low processing time, these methods are limited to detect only close duplicates due to their keypoint based representation. In this work we propose a search method which can detect similar objects even if they exhibit considerable variability. Our procedure learns models of the appearance of objects and trains a classifier to find related instances. We address a central problem of such learning-based methods, the need for appropriate negative and positive training samples. To avoid a highly complicated hard negative mining stage we propose a pooling procedure for gathering generic negatives. Moreover, a bootstrap approach is presented to aggregate positive training samples. Comparison of existing search methods in cultural heritage benchmark problems demonstrates that our approach yields significantly improved detection performance. Moreover, we show examples of searching across different types of datasets, e.g., drafts and photographs.

1. Introduction

The increasing number of digitization campaigns of libraries, archives and museums created a massive repository of cultural heritage online. For art historians it is crucial to know the connection between the widespread artworks to understand chronological, stylistics and contextual issues (see the approach of [MBO14]). Especially the history of printmaking can be reconstructed by analyzing the long-term use of the wood blocks and their similar reproductions or variations. Moreover, research on semantic approaches is interested in smaller parts of an illustration as well as on the combination of two or more autonomous illustrations. Therefore, the aim of this work is to provide a search tool which finds same and similar parts in an unlabeled image dataset within a feasible response time.

Focusing on achieving a high recall, we create a classifier, learning a single positive example against a huge set of negatives, based on the Exemplar Support Vector Machine (SVM) approach [MGE11]. The downside of using this learning algorithm is that we do not have a negative training set to train the classifier. This is a very common problem, dealing with large image datasets in art history, which makes it impractical to use support vector approaches straightforwardly. Beyond that, the mining of features from the negative images and the multiple retraining rounds of the classifier model is very time-consuming. A long response time is however an issue to be solved, because it strongly degrades the usability of an online search tool.

To cut the online calculation time down, we create a pool of general negatives offline in a preprocessing step. These negatives are collected from all the images from the dataset and hence can be seen as a representation of the negative feature space. To be more precise, these are the negative support vectors from dummy classifiers, which were created with random query regions. This offline computed pool

DOI: 10.2312/gch.20141307

of general negatives is used as the negative training set for every arbitrary online search request with only small modifications. For that purpose the positive exemplar has to be trained only once against this representation of the negative feature space making the time of mining and retraining negligible. To improve the classifier in a second step, the k best detections for the query feature are used to train an aggregated instance classifier with $k+1$ positives against the pool of general negatives. This aggregated instance classifier can even detect duplicates with high degree of variability. The evaluation shows that our approach obtains a high recall in feasible response time, which makes our method a helpful tool for online searches in unlabeled databases.

2. Related Work

Digitization campaigns created huge amounts of digital data in the field of cultural heritage, which makes the image based search in databases more and more important [BSO13]. The Bayerische Staatsbibliothek for example offers an image based search for a dataset of 4.2 million images from manuscripts, old prints, maps and other works, published between the 9th and 20th century [bay]. Whole images or parts, suggested by the framework, can be used as query objects to search for similar images or parts.

Another web-based tool was created by the Visual Geometry Group from the University of Oxford to search for duplicates in the Bodleian Ballads image set [oxf, BFH*13]. A rectangular region of interest is marked by the user and the framework returns the search result instantly, displaying duplicates found in the dataset. Since this is exactly the task we want to address with our method, we will evaluate our experiments first with this dataset from the Bodleian library. Additionally, we run our method on the Sachsenspiegel dataset to show our method's ability to detect even variations of the query region.

Based on the approach of the bag-of-words method for searching duplicates in images [SZ03], some works deal with the task of a fast image based search in huge databases [CPS*07, PCI*07, CMPM11, BFH*13]. High-dimensional descriptors are computed on the images and assigned to visual words after quantization, which form the description of the image. A search is performed based only on the presence of visual words in an image, which makes the search run fast. Typically, high-dimensional features like Scale-invariant feature transform (SIFT) [Low04] or related features [BTVG06, TL09] are used here, because of their robustness against small shifts, scale changes and affine transformations, which enables them to detect even objects taken from different angles. At the same time, this makes the bag-of-words methods dependent on keypoint matches, which are often not given in drawings, because objects do not look similar on this feature level.

In [YMCO13] a method is introduced to analyze duplicates in medieval images. The intra-class variability is analyzed to achieve a computer-based understanding of medieval images. A classifier is trained to deal with the different kinds of appearances of similar duplicates. However, a labeled negative training set is necessary for this method, which, like for most of the datasets created by digitalization campaigns, is not available.

If we focus more on the detection problem, there are some further interesting approaches based on support vector machines. The Deformable Part Model [FMR08] and the Exemplar SVM [MGE11] methods are some of the leading approaches in detecting objects. But like all approaches using a SVM, the mining of negatives and the multiple retraining rounds are typically very time-consuming, making them improper for the use in quick online searches. One approach, which minimizes the training time is [HCC*13], in which the number of features in the negative training set is reduced by using the overlapping redundant information caused by the dense sliding window sampling. The reduced negative training set allows training to be performed fast in only one single round, showing comparable performance to models created with multiple training rounds. However, the long response time is not the only issue of the SVM approaches. The requirement of a labeled training set is a big challenge not only because many image datasets are unlabeled but also in our search case it is not possible to provide labels. Labeling is only possible, if there is a definition of what is positive and what is negative (e.g. cats vs. dogs). If the user is allowed to mark any arbitrary region in the image dataset as the positive query object, labeling is not possible. But besides these two disadvantages, the SVM based approaches promise high recall on similar duplicates, which is crucial from the art historian's point of view. Therefore, we follow the basic idea of the Exemplar SVM approach in an adapted manner, by also training one single positive exemplar against a huge set of negatives. We introduce a pool of general negatives, which can be used as the negative training set for any arbitrary query object, allowing a fast training in only one single round without the time-consuming iteratively performed negative mining. This pool of general negatives can be created offline in advance leaving only an instant learning and the detecting step as the sole component for the online search. In a further step, we retrain the classifier with additional positives, which we received from our initial classifier. In contrast to learning methods, which use relevance feedback to retrain the classifier [TTLW06, ZWL12], this step is done without any supervision to keep the user interaction as small as possible.

3. Approach

To develop our algorithm we choose two datasets. The first is an image set from the Bodleian library containing early modern ballad prints with often reused illustrations. Being an early mass media, these cheap prints have a high rele-

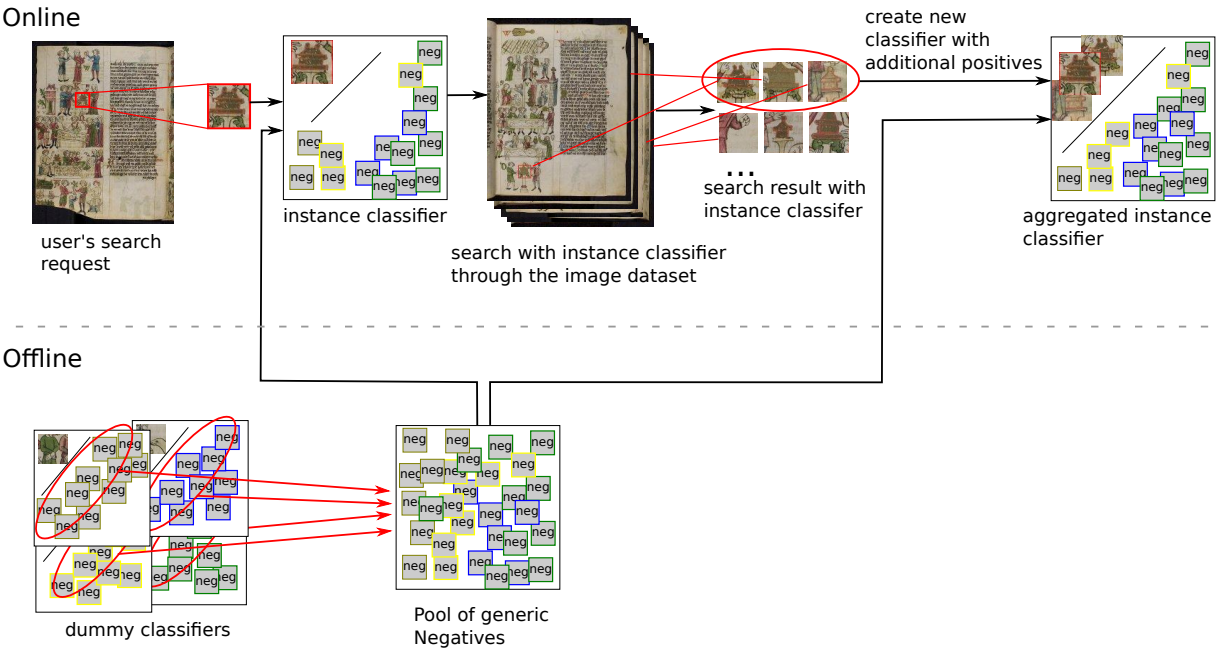


Figure 1: In the offline part, a negative pool with negative features are created. In the online part, first, an instance classifier is created and then by mining positives, an aggregated instance classifier emerges.

vance as cultural heritage and can be seen as a representative example for big databases of early modern prints. The other dataset contains Sachsenspiegel illustrations from Wolfenbüttel, illustrating the prose works of Eike von Repgow composed around 1220-1235. In these images many objects are shown in a somehow standardized appearance. We use this dataset to show how our method is able to detect near duplicates with a wide range of variability to the query object.

To perform a search, typically, a region of interest is chosen in one image, which forms the positive query feature, and the whole dataset is searched for near duplicates of this object. From the view of art historians, a high recall is essential for evaluating the age, when the way of reproducing written works started to change. At the same time, the online search must respond in reasonable time to maintain the usability of this search tool.

As is common for large image datasets in art history, individual objects are not labeled and even more rarely labeled with location information. Therefore, we are confronted with a key problem that is generally existent in the field of visual object detection: When learning a classifier that is to detect a certain class of objects, a set of negatives is required for the currently very popular discriminative approaches such as [FMR08, MGE11]. The problem is not only that many datasets are not labeled. Allowing any arbitrary query region, it is not possible to provide a training set, because a positive feature in one search could be a negative feature in the next search.

The usual way to create a classifier is to train it iteratively in many rounds by mining negatives and updating the model. Because we do not have a negative training set here, we create a pool of generic negatives which can be used as the negative training set for every arbitrary search request. The Exemplar SVM approach trains a classifier for each single exemplar from the positive training set and creates a joint classifier out of them. Without having a positive training set, we instead use the k best detections of our single instance classifier and train one aggregated instance classifier with $k+1$ positives and our pool of general negatives. In Figure 1 an overview is given over the whole algorithm.

3.1. The linear classifier

In our approch, we use one single positive exemplar $\mathbf{x}_E$ against a large set of negatives N_E to learn the classifier's weights $\mathbf{w}$ similar to the Exemplar SVM approach [MGE11].

$$\begin{aligned}\Omega_E(\mathbf{w},b) = ||\mathbf{w}||^2 + C_1 h(\mathbf{w}^T\mathbf{x}_E + b) \\ + C_2 \sum_{\mathbf{x}\in N_E} h(-\mathbf{w}^T\mathbf{x} - b)\end{aligned} \tag{1}$$

The hinge loss function $h(x) = \max(0, 1-x)$ is used here. The exemplar is represented by a rigid grid-like HOG (Histogram of Oriented Gradients) feature [DT05] of fixed size (5x5 cells).

Although having only one positive exemplar, overfitting

is prevented by restricting the learning algorithm to a linear classifier. In contrast to [MGE11], we do not do a negative mining through the negative training set. This iteratively performed mining of features from the negative training images in sliding window manner is very time-consuming. Instead, we use an offline calculated pool of generic negatives to create an instance classifier in a single short retraining round. This instance classifier can be optionally improved by aggregating near duplicates from the image set and retrained with these multiple positives.

3.2. Definition of negatives

For training a linear SVM classifier, a training set is essential. But in the image databases we are facing here, there is no labeled training data available. Furthermore, users are allowed to select any arbitrary region of an image as the positive query object. This means, it is technically not possible to provide a training set, because a positive in one search could be a negative in the next search.

Having one single image dataset the question is how to collect negative features from this dataset to train a classifier for a query region. If we define all features from the image dataset to be part of the negative training set, also near duplicates are included in the negative set. One option to avoid this is to compute the direct Euclidean distance $d = ||\mathbf{x}_E - \mathbf{x}|| \, \forall \mathbf{x} \in N_E$ between the HOG feature $\mathbf{x}_E$ of the positive example and each element from the negative set N_E. Outstanding high scores are interpreted as duplicates by the Grubbs method [Gru69] and eliminated from the negative training set. Using the Grubbs method, it has to be decided how much similarity is allowed in the negative set. If the decision line is too strict, this means features with low HOG similarity are sorted out from the negative training set. In this case, the classifier will perform poorly in distinguishing between near duplicates and wrong detections, because the negative training set only contains negatives which are very much dissimilar to the positive query feature. This results in a classifier, which can provide a high recall, but with a higher probability of detecting false positives. On the other hand, if only features with very high HOG similarity are discarded from the negative training set, there is a risk of having positive near duplicates in the negative training set. This would create a classifier producing many false negatives and therefore low recall.

From the art historian's point of view, a high recall is more important than perfect precision. Therefore, we will continue with the classifier focusing on the achievement of a high recall.

3.3. Pool of generic negatives

We defined the negatives to be features with a certain HOG dissimilarity to the query region, the question is where to get the negatives from. If a negative training set existed, all features from the negative images would be extracted in sliding window manner to train the classifier. Not only that we have no negative training set here, but also a search through every single negative image together with the multiple rounds of retraining is very time-consuming. The idea here is to create a pool of generic negatives, which can be used for every arbitrary classifier and therefore can already be computed offline in advance. This pool should be as general as possible and at the same time not too huge, this means should not contain all features from the whole image database, because this would be not feasible. The support vectors of a classifier are describing a hyperplane in the feature space. To get a general but sparse description of the feature space, the idea is to create classifiers of random query regions and collect the support vectors in a pool. This means, we create N dummy classifiers $c_i \in C, |C| = N$ from random query regions of the image database. Alternatively, classifiers from previous searches on this dataset can be used. We collect all support vectors $\mathbf{s}_{ij} \in S_i$ with j being the number of support vectors from the dummy classifier c_i to get a pool of general negatives $\Psi = \{\cup_{i=1}^{N} S_i\}$. This pool of generic negatives can be used for every arbitrary search request $\mathbf{x}_E$ and therefore can be created offline in advance. Only small adjustments have to be done before training a classifier with this pool of generic negatives, which is not simply a random subset of the whole negative feature space. Compared to just choosing a random subset of features with a certain distance in HOG space, the distance between the collected support vectors are measured in the trained classifiers feature space, which makes them distributed more evenly from the view of a classifier. A comparison is given in the evaluation part. As mentioned in the section before, the negative training set should not contain any feature $\mathbf{s}$, which is describing a near duplicate of the query region $\mathbf{x}_E$. Therefore the HOG similarity $d = |\mathbf{x}_E - \mathbf{s}| \, \forall \mathbf{s} \in \Psi$ between the $\mathbf{x}_E$ and all members of the pool of general negatives Ψ is computed. The Grubbs [Gru69] method is used to find outliers $\mathbf{s}' \in \Psi'$, which have a high HOG similarity and discard them from the negative pool Ψ. The subset $\Psi^* = \Psi - \Psi'$ is then used to train the classifier c_E for the query region $\mathbf{x}_E$. The training finishes within seconds, because only one round of retraining is necessary and no time-consuming negative mining is performed.

3.4. Aggregated instance classifier

As described in chapter 3.1, the instance classifier c_E was created with focus on the achievement of a high recall with a acceptable risk of producing false positives. Searching only for duplicates with a very high similarity, this instance classifier shows a good performance. But facing the task of detecting related instances with a wider range of variability, the number of false positive detections increases, which lowers the usability of this method. At this stage a standard support vector machine classifier is improved by iteratively per-

formed searches for hard-negatives, the negative mining. Because only the negative training set is used here to sharpen the border between the negative and positive feature space, this can be seen as an improvement of the classifier from the negative side. But as already mentioned before, without the availability of a labeled negative training set, hard negatives cannot be identified. Therefore, we decide to improve the classifier from the opposite side, i.e. the positive side. Even though the instance classifier tends to produce some false positives, the best detections are usually true positives. Hence, we choose the k best detections of the instance classifier c_E and aggregate them to the query feature $\mathbf{x}_E$ to train an aggregated instance classifier with this positive training set N_P of $k+1$ positives and the adjusted pool of general negatives Ψ^* in a single training round, optimizing the following convex objective:

$$\Omega_E(\mathbf{w},b) = ||\mathbf{w}||^2 + C_1 \sum_{\mathbf{x}_E \in N_P} h(\mathbf{w}^T \mathbf{x}_E + b) + C_2 \sum_{\mathbf{x} \in \Psi^*} h(-\mathbf{w}^T \mathbf{x} - b). \quad (2)$$

4. Evaluation

To evaluate our algorithm, we choose two image datasets, which contain objects with different levels of variability. The first image dataset we chose from the Bodleian library contains early modern ballad prints. Each of the 916 images contain ballad lyrics and illustrations. The illustrations occur in the image set as strict duplicates or in slightly modified versions. For the experiments on this dataset we focus more on detecting the strict duplicates. Analyzing these different appearances can give an inside view into the early period of mass media. A high recall is therefore crucial from the art historian's point of view.

4.1. Comparison with bag-of-words approach

The Visual Geometry Group of the University of Oxford developed a tool [BFH*13] which is used for searching near duplicates of any arbitrary query region in the Bodleian image dataset. This algorithm is based on the bag-of-words method. High-dimensional keypoints like SIFT features are computed and stored in a codebook for the whole dataset and during each search the keypoint constellation in the query region are compared to the keypoints in the codebook. This has the advantage of a very short responding time. The disadvantage is that it has to be possible to match the keypoints between two near duplicate objects. This similarity might be given for e.g. detecting buildings in photos, taken from many different angles and scales, but not for drawings in the field of cultural heritage, on which we are focusing here. Objects in drawings can be grouped semantically (e.g. couples, portraits), but the high-dimensional keypoint features in those objects are mostly too different to be matched between objects in the same group.

Method	AP_1	AP_2	time
Bag-of-words approach [oxf]	0.766	0.056	-
Exemplar SVM [MGE11]	0.764	0.055	587s
Instance classifier	0.937	0.171	30s
Aggregated instance classifier	0.939	0.348	65s

Table 1: AP_1: Results on the categories, which only considers strict duplicates. AP_2: Results on the high-level categories, which additionally includes semantically similar objects (e.g. different couples holding hands).

To make the comparison between our method and the bag-of-words method as equal as possible, we first only evaluate detections on objects which are very similar to the query object. This is given here in this Bodleian image dataset, because the woodblock printing technique is producing lot of near duplicates. Because of this restriction, which reduces the complexity of the search task, we use our instance classifier for the comparison and not the retrained aggregated instance classifier, which requires a longer response time.

The result is shown in Table 1. Our algorithm, using the pool of general negatives, achieves a higher average precision than the online bag-of-words framework(see Table 1). Our instance classifier takes about half a minute to respond, which is mainly due to the search through all the images, because the training finishes almost instantly.

In Figure 2 the first detections of a search with the instance classifier is shown. It can be seen, that in the lower rows, couples were detected, which vary from the query couple, marked with a red rectangle. This shows that our instance classifier can handle small changes, which is important to analyze the replication of printing plates.

4.2. Performance depending on size

Another advantage of our classifier is, that the performance is mainly independent regarding the size of the exemplar. Figure 3 shows the average precision for search requests of different sizes. While the bag-of-words version [oxf] shows a huge drop in performance, if the diagonal of the region of interest is smaller than 150 pixels, our algorithm shows nearly constant performance. The performance of the bag-of-words approach depends on the existence of a certain amount of keypoints in the search area. If the region of interest contains non or only a very small number of keypoints, it is not possible to describe this area well enough to perform a search. In contrast, our algorithm describes the query region with a HOG feature of a constant size. This explains the high average precision even for very small objects.

Furthermore we compare the performance of the Exemplar SVM [MGE11] on this search task. To provide a negative training set, which is necessary to run the Exemplar SVM algorithm, we consider all images except the positive

Figure 2: The upper image shows the query region marked with a rectangle. The lower image shows the first 40 responses for this search run. It can be seen that the lower rows show detections which has some variability to the query object. In the last row, there are the first wrong detections.

example as part of the negative training set initially and train an Exemplar SVM with one single exemplar. This means our pool of general negatives Ψ forms a subset of this training set N_E, which consists of all features in the image dataset. As described in section 3.2 we throw out all negatives N'_E from the negative training set, to avoid having duplicates of the positive example, resulting in a negative training set $N^*_E = N_E - N'_E$. Although, the Exemplar SVM trains with a bigger negative training set ($\Psi^* \subset N^*_E$), the Exemplar SVM approach shows a drop in performance (see Table 1). This is probably caused by an incorrect negative training set. The Grubbs method [Gru69] is a procedure to find outliers. If we consider a very dense representation of the negative feature space, which we get by sampling densely through the whole image dataset in sliding window manner, a duplicate of the positive query example is hard to identify. For the evaluation of the pool of general negatives, however, this method seems to work properly, because of the sparse representation of the feature space, in which duplicates stick out much stronger from its neighboring features. If not all false negatives are sorted out and therefore $\Psi' \notin N'_E$, the wrong negatives cause a performance drop of the Exemplar SVM approach. This

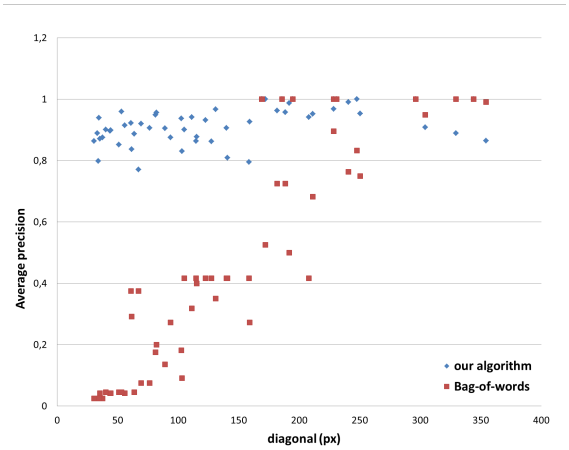


Figure 3: Average precision for different sizes of positive examples. While the bag-of-words approach drops performance drastically for small objects, our algorithm maintains a nearly constant average precision.

comparison has shown that our algorithm based on the idea of the Exemplar SVM shows two essential advantages for arbitrary searches in image databases. First, with only one single training round with the pool of general negatives, our algorithm responds quick compared to the Exemplar SVM. And furthermore, the sparsity of our pool of negatives, helps to sort out false negatives from the training set and therefore result in a superior detection performance.

4.3. Visualization of Models

The superior performance of our method can also be shown directly on the weights. We create multidimensional scaling (MDS) plots, which map the distances between the model's weights in feature space into a two dimensional space. Here we only employ searches for three categories, couples, ships and portraits. First we plot the query images, directly using their HOG Features (see Figure 4). The plot shows no specific clustering of the different types. Next we plot the weights of the classifiers, learned to detect duplicates in the dataset. Figure 5 shows the classifiers trained with the Exemplar SVM method, while Figure 6 shows the instance classifiers of our algorithm. Both methods appear to cluster the three categories better in comparison to the distribution of the raw HOG features in space. Our algorithm shows that its instance classifiers from query images belonging to one category are closer in feature space than the classifiers created with the Exemplar SVM method.

4.4. The aggregated instance classifier

The experiments on the Bodleian dataset already showed that our method is superior in detecting duplicates with bigger variabilities to the query object. Now we evaluate the

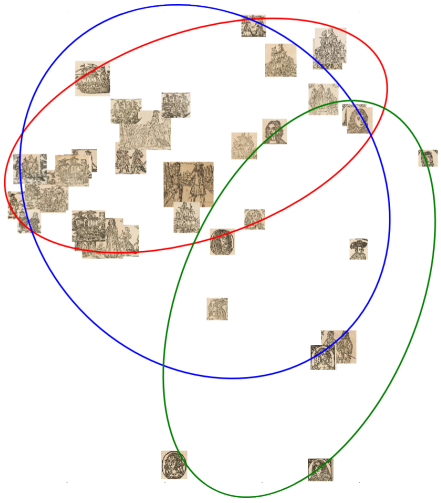


Figure 4: MDS-mapping of HOG-Features.

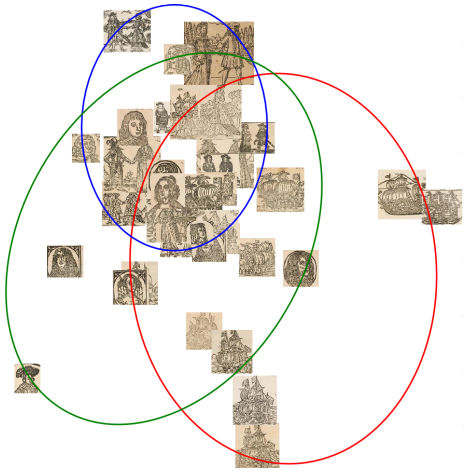


Figure 5: MDS-mapping of classifiers created with the Exemplar SVM. The circles are marking the members of a category: 'couples'-blue, 'ships'-red, 'portraits'-green.

aggregated instance classifier, which should be able to deal with even more variabilities. After running the instance classifier, the best $k = 3$ detections, which are not exact duplicates from each other, are aggregated to create a more robust classifier. First, we evaluate this classifier on the Bodleian dataset again. Until now we only considered strictly real duplicates, which look the same and therefore are created probably with the same printing plate. Now we add a higher-level category, which counts different ships or different 'couples holding hands' to one category. In Figure 7 a subset of the higher-level group for "couple holding hands" is shown. To automate the evaluation of the detections, we labeled the dataset for different objects. An object detection is counted

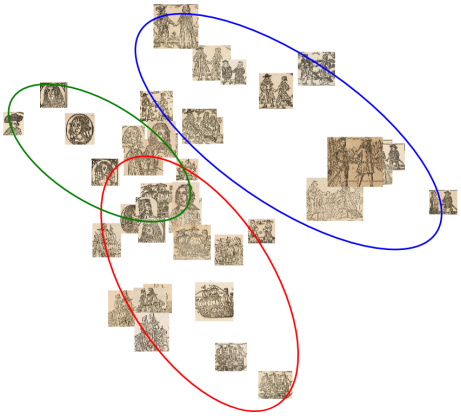


Figure 6: MDS-mapping of instance classifiers created with our algorithm. The circles are marking the members of a category: 'couples'-blue, 'ships'-red, 'portraits'-green.

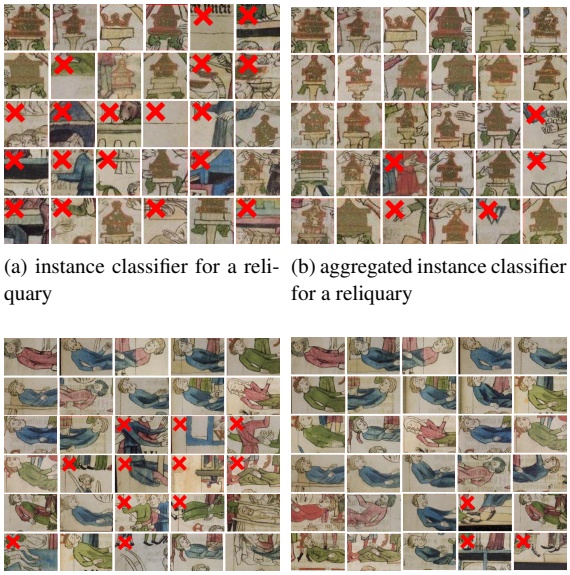
Figure 7: Subset of the higher-level group for 'couple holding hands'.

as correct if $\frac{A_d \cap A_g}{A_d \cup A_g} >= 0.5$ where A_d and A_g is the area of the detection and the ground truth bounding box, respectively.

The evaluation for this higher-level category shows, that our method, using a pool of prototypical negatives is outperforming the other searches regarding the average precision by far (AP_2). Comparing our two classifiers it can be seen that the average precision on the categories, which only consider strict duplicates is almost the same, because the aggregated positives are not including much additional information. On the higher-level category the aggregated instance classifier shows a superior detection behavior taking almost one minute to respond, due to the search run for additional positives. The drop of the bag-of-words approach on this higher-level category is not surprising, because the used keypoint features can not match between members of this group because they are too different.

We make some further experiments on an additional dataset, the Sachsenspiegel image dataset from Wolfenbüttel. In this dataset the variability is greater because everything is drawn by hand. Many situations occur several times in the image dataset, e.g. people with the same gesture like judges at the court, people lying on the floor or objects like reliquaries. To show how much the second training round with aggregated positives reduces false positive detections, we compare the results of the instance classifier to the results of the aggregated instance classifier on the same query re-

gions. In Figure 8 the responses of both classifiers are shown on two examples. It shows that the instance classifier detects true positives first, but then starts to score on more and more false positives, while the aggregated instance classifier is much more robust against false detections. The detections of the reliquaries as well as the detections of people lying on the floor, show how well our method can deal with variabilities. Each drawing is different on the pixel level, because it is drawn by hand. A feature based method would not be able to find keypoint matches between the objects. Our unsupervised learning method can therefore help interpreting images on a higher level, being able to find connections between objects belonging together semantically.



(a) instance classifier for a reliquary (b) aggregated instance classifier for a reliquary

(c) instance classifier for a lying person (d) aggregated instance classifier for a lying person

Figure 8: On the left the first 30 results of the instance classifier are shown. Even though the detections show a wide variability, there are some false positives (marked with a red cross). The results on the right show that the aggregated instance classifier, trained with $k = 3$ aggregated additional positives, exhibits fewer false positives.

5. Discussion and Conclusion

In this work we introduced an online search method, which can be used as a beneficial tool by art historians, to analyze medieval images regarding reproductions or variations. Not only most of those image databases are unlabeled, which makes them difficult to search through with conventional discriminative approaches, but also it is not possible to provide a labeled training set. This is because by allowing any arbitrary region from the whole image dataset to be the query object, a positive feature in one search can be a negative in the next search. We created a pool of general negatives offline, which can be used for any arbitrary object search in unlabeled datasets. The time-consuming multiple rounds of mining negatives and training are replaced by a short single training round, which makes it possible to give a response in feasible time. An additional step, in which the instance classifier is improved by additional positives, reduces the number of false positives and therefore simplifies the interpretation of the results considerably. †

References

[bay] Bayerische Staatsbilbliothek. http://bildsuche.digitale-sammlungen.de/, Online; accessed 01-July-2014. 2

[BFH*13] Bergel G., Franklin A., Heaney M., Arandjelovic R., Zisserman A., Funke D.: Content-based image recognition on printed broadside ballads: The bodleian libraries' imagematch tool. 2, 5

[BSO13] Bell P., Schlecht J., Ommer B.: Nonverbal communication in medieval illustrations revisited by computer vision and art history. *Visual Resources 29*, 1-2 (2013), 26–37. 2

[BTVG06] Bay H., Tuytelaars T., Van Gool L.: Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*. Springer, 2006, pp. 404–417. 2

[CMPM11] Chum O., Mikulik A., Perdoch M., Matas J.: Total recall ii: Query expansion revisited. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (2011), IEEE, pp. 889–896. 2

[CPS*07] Chum O., Philbin J., Sivic J., Isard M., Zisserman A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (2007), IEEE, pp. 1–8. 2

[DT05] Dalal N., Triggs B.: Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 1, IEEE, pp. 886–893. 3

[FMR08] Felzenszwalb P., McAllester D., Ramanan D.: A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (2008), IEEE, pp. 1–8. 2, 3

[Gru69] Grubbs F. E.: Procedures for detecting outlying observations in samples. *Technometrics 11*, 1 (1969), 1–21. 4, 6

[HCC*13] Henriques J. F., Carreira J., Caseiro R., Batista J., Proofs A.: Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In *Computer Vision (ICCV), 2013 IEEE* (2013). 2

[Low04] Lowe D. G.: Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110. 2

[MBO14] Monroy A., Bell P., Ommer B.: Morphological analysis for investigating artistic images. *Image and Vision Computing 32*, 6 (2014), 414–423. 1

† This work has been supported in part by the Ministry of Science, Baden Württemberg and the Heidelberg Academy of Sciences and Humanities.

[MGE11] MALISIEWICZ T., GUPTA A., EFROS A. A.: Ensemble of exemplar-svms for object detection and beyond. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (2011), IEEE, pp. 89–96. 1, 2, 3, 4, 5

[oxf] Bodleian Library Broadside Ballads Search. http://zeus.robots.ox.ac.uk/ballads/page0/, Online; accessed 01-July-2014. 2, 5

[PCI*07] PHILBIN J., CHUM O., ISARD M., SIVIC J., ZISSERMAN A.: Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007), IEEE, pp. 1–8. 2

[SZ03] SIVIC J., ZISSERMAN A.: Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (2003), IEEE, pp. 1470–1477. 2

[TL09] TURCOT P., LOWE D. G.: Better matching with fewer features: The selection of useful features in large database recognition problems. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on* (2009), IEEE, pp. 2109–2116. 2

[TTLW06] TAO D., TANG X., LI X., WU X.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 28*, 7 (2006), 1088–1099. 2

[YMCO13] YARLAGADDA P., MONROY A., CARQUÉ B., OMMER B.: Towards a computer-based understanding of medieval images. In *Scientific Computing and Cultural Heritage*. Springer, 2013, pp. 89–97. 2

[ZWL12] ZHANG L., WANG L., LIN W.: Semisupervised biased maximum margin analysis for interactive image retrieval. *Image Processing, IEEE Transactions on 21*, 4 (2012), 2294–2308. 2