Graph-based Methods for the Automatic Annotation and Retrieval of Art Prints

Gustavo Carneiro[*]
Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisbon, Portugal
gcarneiro@isr.ist.utl.pt

ABSTRACT

The analysis of images taken from cultural heritage artifacts is an emerging area of research in the field of information retrieval. Current methodologies are focused on the analysis of digital images of paintings for the tasks of forgery detection and style recognition. In this paper, we introduce a graph-based method for the automatic annotation and retrieval of digital images of art prints. Such method can help art historians analyze printed art works using an annotated database of digital images of art prints. The main challenge lies in the fact that art prints generally have limited visual information. The results show that our approach produces better results in a weakly annotated database of art prints in terms of annotation and retrieval performance compared to state-of-the-art approaches based on bag of visual words.

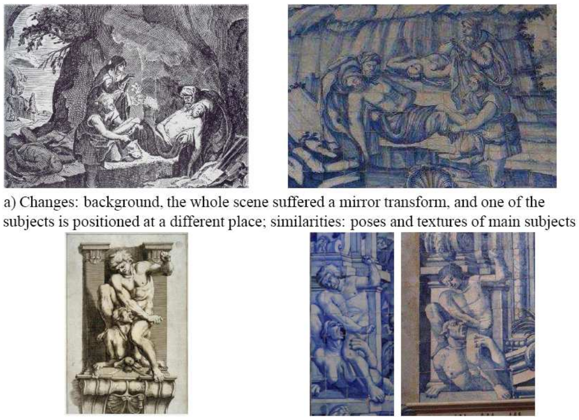
H.3.3Information storage and retrievalInformation search and retrieval|Retrieval models I.4.8Image Processing and Computer VisionScene Analysis|Object Recognition G.3Probability and StatisticsProbabilistic Algorithms

General Terms

Algorithms, Measurement, Experimentation

Keywords

Content-based image retrieval and annotation, Art image annotation and retrieval, Graph-based Learning methods



a) Changes: background, the whole scene suffered a mirror transform, and one of the subjects is positioned at a different place; similarities: poses and textures of main subjects

b) Changes: background; similarities: poses and textures of main subjects

Figure 1: Examples of how art print images (left) are altered in the process of becoming a tile panel (right).

1. INTRODUCTION

The analysis of digital images taken from artistic productions is a emerging field of research in the areas of information retrieval, computer vision, digital image analysis and machine learning. There are several applications being developed in this area, such as: the system designed by Google to identify paintings [1]; the artistic identification methodologies designed to classify Van Gogh's brush strokes [27]; the model to classify brushstrokes [48]; and the approach to discover, recover, represent and understand cultural heritage artifacts [21]. Therefore, the techniques developed in this area will be an essential tool for systems designed to help art historians in the task of analyzing art production.

The analysis of digital images of prints is particularly important for art historians because printmaking methods have been intensively used over the last five centuries with the goal of replicating paintings produced by artists from all over the world. This intense use of printmaking techniques is associated with the the fast and cheap production of paper and advancements in graphical arts, which started in the XVth century. As a result, prints of artistic paintings have reached a vastly superior number of people compared to the original paintings. Consequently it is worth understanding the importance of art prints in art history since they have influenced and served as a source of inspiration for generations of artists. Therefore, the proper classification, retrieval and annotation of art prints constitute an important activity for art historians to understand the art produced in the last five centuries.

The influence of prints on visual arts can be evidenced in several artistic productions, such as the influence of Japanese art prints on impressionist artists in the XIXth century [3]. More relevant to our work is the influence of prints on artistic tile painters in Portugal [9, 12], as evidenced by the large number of panels found in Churches, public buildings and palaces. In the context of artistic tile panels,

[*] This work was supported by the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds and Project PRINTART (PTDC/EEA-CRO/098822/2008). This project is also funded by the European Commission; Marie Curie IIF, contract number PIIF-GA-2009-236173 (IMASEG3D).

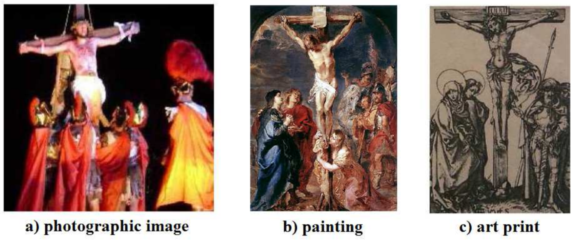
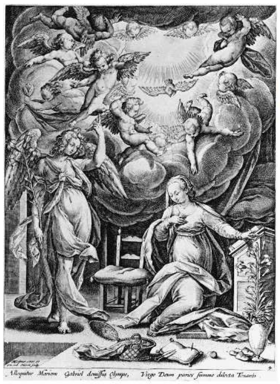


Figure 2: Loss of visual information from the photographic image in (a), to the painting in (b), to the art print in (c).



Annotation

Theme: Annunciation
Subjects: Mary
Gabriel
dove
lilly
vase
wing
angels floating in the air

Figure 3: Example of the manual annotation of an art print image produced by an art historian.

the analysis of digital images of prints is extremely important for understanding how the Portuguese tile panel artists were influenced by such prints, which has the potential to furnish relevant information to art historians in Portugal. It is important to mention that art prints were generally used as references to produce the panels, but the artists often produced an impression of the print (i.e., not an exact copy of it - see Fig. 1). For this reason, the task of discovering the influence of one or several prints in the composition of a tile panel requires an expert with a specialized visual knowledge of the current databases of prints together with peculiar abilities in relating the tile composition with the prints. Therefore, an automatic system that can retrieve a set of print images related to a given artistic work can can be useful for art historians.

Compared to photographic digital images and paintings, art print images loses important visual information [11]. For example, Fig. 2 shows three images of the theme *the Crucifixion of Jesus Christ*, which displays a large sensorial gap [40] between the photographic, painting and art print images. Notice how the low level information (texture, color, depth, etc.) is lost moving from the photographic image to the art print image. This loss of visual information reduces the effectiveness of current image analysis techniques, which usually work with photographic digital images [14]. Art image analysis methodologies can also be used for art prints [42, 27], but the great majority of these techniques have been developed for the analysis of digitized images of paintings, which contain richer visual information than prints. However, note that Li and Wang [29] have developed a system that analyzes ancient Chinese paintings, which are similar to art prints.

In this work, we present a new method for the analysis of art prints. The first goal of the work is to automatically produce global annotations to previously unseen test images using a statistical model whose parameters are estimated using a database of manually annotated training images. The second goal is to retrieve un-annotated images given specific visual classes using the statistical model described above. The art print images in this database are constrained in the following ways: 1) they were created between the centuries XV and XVII, and 2) they are of religious themes. These constraints are relevant for discovering the influences suffered by artistic tile panel painters, which is the long-term goal of this project. The manual annotation (Fig. 3) has been produced by art historians, who identified the image theme and the relevant subjects present in the print (weak annotation without relation between subjects and image regions). The method proposed in this paper follows graph-based learning algorithms, which have the assumption that the visually similar images are likely to share the same annotations [50]. Note that this assumption is important to uncover the aforementioned influence of art prints over other artistic productions (and also over other art prints). Specifically, we explore the following graph-based algorithms [6, 35, 50]: label propagation [18, 28, 34, 46], random walk [11], stationary solution using a stochastic matrix [30], and combinatorial harmonic [19]. We adapt each of those techniques to a bag of visual words (BOV) approach, and we compare their performance with BOV approaches that use the following classifiers: support vector machines (SVM) [45], and random forests (RF) [7]. Note that BOV approaches with these two classifiers can be regarded as the state-of-the-art methods for image retrieval and annotation. The experimental setup uses a database of art prints (which will soon be available for the information retrieval and computer vision communities) containing 307 images and 22 labels, where one label represents a multi-class problem with 7 classes, and the other labels represent binary problems. The results show that graph-based methods produce better results than BOV models (using SVM and RF classifiers) in terms of retrieval and annotation performance.

This paper is organized as follows. Section 2 introduces relevant works in photographic and art image retrieval, and graph-based learning. Section 3 describes the proposed methodology, while Sec. 4 outlines the implementation of our approach. The experiments are presented in Sec. 5, and the paper is concluded in Sec. 6.

2. LITERATURE REVIEW

In this section, we provide a brief review of papers in the areas of photographic and art image retrieval and annotation. We also review a few relevant graph-based learning methods.

Currently in photographic image retrieval and annotation, the most successful methods [15] are based on the the bag of visual words representation [13] combined with SVM [45] or multiple kernel learning (MKL) classifiers [41]. This methodology is effective when the number of visual classes is relatively small (faster training and inference procedures) and the number of training images per class is relatively large (better generalization). Another constraint of this methodology is that the introduction of new images and new labels require a full (re-)training of relevant classifiers. Unfortunately, both constraints are too restrictive for our case where the number of visual classes can be large (with a limited number of training images per class), and the introduction of new images and labels to the database can happen frequently. In photographic image analysis, there is a trend to get around the problem of the high number of visual classes with the use of machine learning methods that finds a sub-space of smaller dimensionality for classification [23]. However, the dynamic nature of our problem, where new classes are regularly introduced into the training database, is still an issue in this area of research.

The area of art image retrieval has attracted the attention of researchers in the fields of computer vision and machine learning [27, 33, 42]. The main focus of the papers is on the artistic identification problem, where the goal is to classify original and fake paintings of a given artist [4, 32, 39] or to produce stylistic analysis of paintings [20, 24, 25]. Most of the methods above can be regarded as adaptations from the content-based image retrieval systems [14], where the emphasis is placed on the characterization of brush strokes using texture or color. The ancient Chinese painting classification studied by Li and Wang [29] is more similar to ours in the sense that they deal with the multi-class classification of painting styles. Finally, the work on the automatic brushwork annotation by Yelizaveta et al. [48] is also similar to ours given that the authors are dealing with multi-class classification of brush strokes. Different from the papers above, our method addresses not only a multi-class, but also a multi-label problem [16].

Graph-based learning has been thoroughly studied by the information retrieval community to rank web pages on the World Wide Web [6, 8, 35]. Essentially, a graph is built where vertexes represent web pages and the edge weights are denoted by the existence of hyper-links. Analysis algorithms based on random walks in this graph have been designed to rank the vertexes (i.e., web pages) in terms of their importance in this network. These graph-based techniques have received considerable attention by the machine learning community for the problem of semi-supervised learning [50]. The random walk algorithm has also been studied in the domains of unsupervised image segmentation [19] and multi-class classification [43]. Finally, random walk algorithms have also been explored in the area of image retrieval [11, 18, 22, 26, 28, 30, 34, 46], where the main advantages of such approaches are: 1) the ability to use visual and non-visual cues in the random walk procedure, 2) the potential to extend the method to large-scale databases, and 3) the relatively facility to adapt the method to dynamic problems, where new images and labels are continuously introduced into the database.

3. METHODOLOGY

Assume that a training set of annotated images is available, and is represented as follows: $\mathcal{D} = \{(I_i, \mathbf{x}_i, \mathbf{y}_i)\}_{i=1..N}$ with $\mathbf{x}_i$ representing the feature vector of image I_i and $\mathbf{y}_i$ denoting the annotation of that image. An annotated test set is also available and is represented by $\mathcal{T} = \{(\widetilde{I}_i, \widetilde{\mathbf{x}}_i, \widetilde{\mathbf{y}}_i)\}_{i=1..P}$, but note that the annotation in the test set is used only for the purpose of methodology evaluation. Each annotation $\mathbf{y}$ represents L multi-class and binary problems, so $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_L] \in \{0,1\}^M$, where each problem is denoted by $\mathbf{y}_l \in \{0,1\}^{|\mathbf{y}_l|}$ (with $|\mathbf{y}_l|$ denoting the dimensionality of $\mathbf{y}_l$, where binary problems have $|\mathbf{y}_l| = 1$, and multi-class problems have $|\mathbf{y}_l| > 1$ and $\|\mathbf{y}_l\|_1 = 1$), and M represents the dimensionality of the annotation vector (i.e., $\sum_{l=1}^{L} |\mathbf{y}_l| = M$). In summary, binary problems involve an annotation that indicates the presence or absence of a class, while multi-class annotation regards problems that one (and only one) of the possible classes is present.

Following the notation introduced by Estrada et al. [17], who applied random walk algorithms for the problem of image de-noising, let us define a random walk sequence of k steps as
$T_{r,k} = [(\mathbf{x}_{(r,1)}, \mathbf{y}_{(r,1)}), ..., (\mathbf{x}_{(r,k)}, \mathbf{y}_{(r,k)})]$, where each $\mathbf{x}_{(r,l)}$ belongs to the training set $\mathcal{D}$, and r indexes a specific random walk. Our goal is to estimate the probability of annotation $\mathbf{y}$ for a test image $\widetilde{\mathbf{x}}$, as follows [43]:

$$p(\mathbf{y}|\widetilde{\mathbf{x}}) = \frac{1}{\mathcal{Z}_T} \sum_{r=1}^{R} \sum_{k=1}^{K} p(T_{r,k}|\widetilde{\mathbf{x}})^{\frac{1}{k}} p(\mathbf{y}|\mathbf{x}_{(r,k)}). \qquad (1)$$

In (1), $\mathcal{Z}_T$ is a normalization factor, $p(\mathbf{y}|\mathbf{x}_{(r,k)}) = \delta(\mathbf{y} - \mathbf{y}_{(r,k)})$ (with $\delta(.)$ being the Dirac delta function, which means that this term is one when $\mathbf{y} = \mathbf{y}_{(r,k)}$), the exponent $\frac{1}{k}$ means that steps taken at later stages of the random walk have higher weight,

$$\begin{aligned} p(T_{r,k}|\widetilde{\mathbf{x}}) =& p([(\mathbf{x}_{(r,1)}, \mathbf{y}_{(r,1)}), ..., (\mathbf{x}_{(r,k)}, \mathbf{y}_{(r,k)})]|\widetilde{\mathbf{x}}) \\ =& \prod_{j=2}^{k} p(\mathbf{x}_{(r,j)}|\mathbf{x}_{(r,j-1)}, \widetilde{\mathbf{x}}) p(\mathbf{y}_{(r,j)}|\mathbf{y}_{(r,j-1)}) \qquad (2) \\ & p(\mathbf{x}_{(r,1)}|\widetilde{\mathbf{x}}) p(\mathbf{y}_{(r,1)}) \end{aligned}$$

with the derivation made assuming a Markov process and that the training labels and features are independent given the test image, $p(\mathbf{y}_{(r,j)}|\mathbf{y}_{(r,j-1)}) = s_y(\mathbf{y}_{(r,j)}, \mathbf{y}_{(r,j-1)})$ with $s_y(\mathbf{y}_{(r,j)}, \mathbf{y}_{(r,j-1)}) = \frac{1}{\mathcal{Z}_y} \sum_{m=1}^{M} \lambda_m \times \mathbf{y}_{(r,j)}(m) \times \mathbf{y}_{(r,j-1)}(m)$ (λ_m is the weight associated with the label $\mathbf{y}(m) \in \{0,1\}$ and $\mathcal{Z}_y$ is a normalization factor), $p(\mathbf{y}_{(r,1)}) = \frac{1}{N}$, and $p(\mathbf{x}_{(r,j)}|\mathbf{x}_{(r,j-1)}, \widetilde{\mathbf{x}})$ and $p(\mathbf{x}_{(r,1)}|\widetilde{\mathbf{x}})$ are defined in Sec. 4.2.

We propose the use of class mass normalization [51] to determine the annotation of image $\widetilde{\mathbf{x}}$. The class mass normalization takes into consideration the probability of a class annotation and the proportion of samples annotated with that class in the training set. Specifically, we have

$$\widehat{\mathbf{y}} = \left[\sum_{i=1}^{N} \mathbf{y}_i(m) \times \max(p(\mathbf{y}_i(m)|\widetilde{\mathbf{x}}) - p(\mathbf{y}(m)), 0) \right]_{m=1..M}, \qquad (3)$$

where $p(\mathbf{y}(m)) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i(m)$ (m indicates the m^{th} dimension of label vector $\mathbf{y}$). The use of class mass normalization makes the annotation process more robust to imbalances in the training set with respect to the number of training images per visual class. Notice that $\widehat{\mathbf{y}}$ in (3) represents the confidence that the image represented by $\widetilde{\mathbf{x}}$ is annotated with the labels $\widehat{\mathbf{y}}(m)$ for $m = \{1, .., M\}$. Finally, we further process $\widehat{\mathbf{y}}$ for multi-class problems as follows:

$$\begin{aligned} & \forall l \in \{1, ..., L\}, \text{ with } |\widehat{\mathbf{y}}_l| > 1 \\ & \mathbf{y}_l^* = \begin{cases} \min(\lfloor \widehat{\mathbf{y}}_l / \max(\widehat{\mathbf{y}}_l) \rfloor, 1), & \text{if } \max(\widehat{\mathbf{y}}_l) > 0.5 \\ \{0\}^{|\mathbf{y}_l|}, & \text{otherwise} \end{cases}, \qquad (4) \end{aligned}$$

and for binary problems we define:

$$\begin{aligned} & \forall l \in \{1, ..., L\}, \text{ with } |\widehat{\mathbf{y}}_l| = 1 \\ & \mathbf{y}_l^* = \begin{cases} 1, \widehat{\mathbf{y}}_l > 0.5 \\ 0, \text{otherwise} \end{cases}. \qquad (5) \end{aligned}$$

As a result, the final annotation for image $\widetilde{\mathbf{x}}$ is represented by $\mathbf{y}^* = [\mathbf{y}_1^*, ..., \mathbf{y}_L^*]$.

The retrieval problem is defined as the most relevant image returned from the database of test images $\mathcal{T}$ given the m^{th} visual class $\mathbf{y}(m)$ for $m \in \{1, ..., M\}$, as in:

$$\mathbf{x}^*_{\mathbf{y}(m)} = \max_{\widetilde{\mathbf{x}} \in \mathcal{T}} p(\widetilde{\mathbf{x}}|\mathbf{y}(m)), \qquad (6)$$

where $p(\widetilde{\mathbf{x}}|\mathbf{y}(m)) = p(\mathbf{y}(m)|\widetilde{\mathbf{x}}) p(\widetilde{\mathbf{x}}) / p(\mathbf{y}(m))$ with $p(\widetilde{\mathbf{x}}) =$ constant and $p(\mathbf{y}(m))$ defined in (3). Furthermore, $p(\mathbf{y}(m)|\widetilde{\mathbf{x}})$ is defined as in (1) replacing the last term by $p(\mathbf{y}(m)|\mathbf{x}_{(r,k)}) = \delta(\mathbf{y}(m) - \mathbf{y}(m)_{(r,k)})$.

4. IMPLEMENTATION

In this section we provide the details of the data set used, the image representation based on bags of visual words, and the implementation of the following algorithms: label propagation, random walk, stationary solution, and combinatorial harmonic.

4.1 Data Set

The data set consists of 307 annotated images with one multi-class problem (theme with seven classes) and 21 binary problems (see Fig. 5). All images have been collected from the Artstor digital image library [2], and annotated by art historians (Fig. 3 shows an example of a manual annotation). For the experiments in Sec. 5, we run a 10-fold cross validation in order to show the results, and for each run, divide the data set into a training set $\mathcal{D}$ with 276 images (90% of the data set) and a test set $\mathcal{T}$ with 31 images (10% of the data set).

4.2 Image Representation

The images are represented with the bag of visual words model [13], where each visual word is formed using a collection of scale invariant feature transform (SIFT) local descriptors [31]. The visual vocabulary is built using the vocabulary tree proposed by Nistér and Stewénius [36]. The SIFT descriptor consists of a feature transform applied to an image patch, which extracts a histogram of gradient orientations weighted by the gradient magnitudes. In this work, the image patch location (in the image) and scale (patch size) are randomly determined [37], and we generate 1000 descriptors per image. Then, the vocabulary is built by gathering the descriptors from all images and running a hierarchical clustering algorithm with three levels, where each node in the hierarchy has 10 descendants (this hierarchy is a directed tree, where each node has at most 10 edges) [36]. This results in a directed tree with $1 + 10 + 100 + 1000 = 1111$ vertexes, and the image feature is formed by using each descriptor of the image to traverse the tree and record the path (note that each descriptor generates a path with 4 vertexes). The histogram of visited vertexes is weighted by the node entropy (i.e., vertexes that are visited more often receive smaller weights). As a result, an image I is represented by the histogram $\mathbf{x} \in \Re^{1111}$.

The probability of the transition of feature vector $\mathbf{x}_{(r,j)}$ given $\mathbf{x}_{(r,j-1)}$ and $\widetilde{\mathbf{x}}$ of (2) is then defined as:

$$p(\mathbf{x}_{(r,j)}|\mathbf{x}_{(r,j-1)}, \widetilde{\mathbf{x}}) = s_x(\mathbf{x}_{(r,j)}, \mathbf{x}_{(r,j-1)}) s_x(\mathbf{x}_{(r,j)}, \widetilde{\mathbf{x}}), \quad (7)$$

and the transition probability between two feature vectors is

$$p(\mathbf{x}_{(r,1)}|\widetilde{\mathbf{x}}) = s_x(\mathbf{x}_{(r,1)}, \widetilde{\mathbf{x}}), \quad (8)$$

with $s_x(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\mathcal{Z}_x} \sum_{d=1}^{1111} \min(x_i(d), x_j(d))$ where $\mathcal{Z}_x$ is a normalization factor. Consequently, we can define the following adjacency matrix to be used in the graph-based algorithms below:

$$\mathbf{W}(j,i) = s_y(\mathbf{y}_i, \mathbf{y}_j) \times s_x(\mathbf{x}_i, \mathbf{x}_j) \times s_x(\mathbf{x}_j, \widetilde{\mathbf{x}}), \quad (9)$$

with $\mathbf{W}(i,i) = 0$ for all $i \in \{1, ..., N\}$.

In Fig. 4 it is shown a small part of the graph which takes into account the image features and annotation of the training set described by the adjacency matrix $\mathbf{W}$ in (9). In this part of the graph, we take a training image represented by $\widetilde{\mathbf{x}}$ shown at the center (note the enlarged image in the figure), and display the graph structure around it. Notice that the neighboring images in the graph tend to be similar in terms of their visual information or their annotation keywords (for instance, notice in the graph that most of the neighboring images are of the theme *The Annunciation*, which is the theme of the central image). In order to show this graph, we use a variant of the multidimensional scaling algorithm for visualization (MDS) [5].

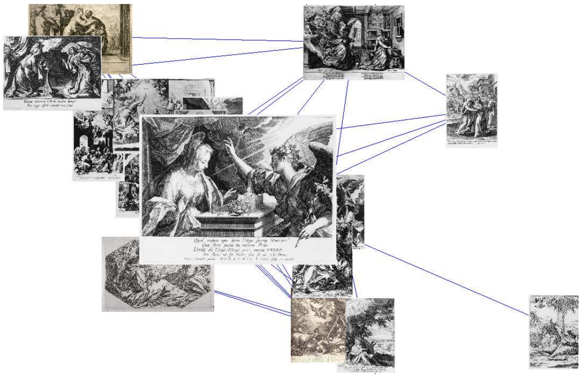


Figure 4: Network structure in the training set shown using a variant of the MDS algorithm [5]. The large image in the center represents a training image with its most similar images (in visual an annotation terms) closer in the graph.

4.3 Label Propagation

The annotation and retrieval using the label propagation model consists of a one step random walk procedure that uses only the visual similarity, as follows:

$$p(\mathbf{y}|\widetilde{\mathbf{x}}) = \frac{1}{\mathcal{Z}_{LP}} \sum_{r \in \mathcal{N}} p(\mathbf{x}_{(r)}|\widetilde{\mathbf{x}}) p(\mathbf{y}|\mathbf{x}_{(r)}), \quad (10)$$

where $\mathcal{N} \in \mathcal{D}$ denotes the set of $|\mathcal{N}|$ nearest neighbors relative to $\widetilde{\mathbf{x}}$ using the proximity measure (8), and $\mathcal{Z}_{LP}$ represents a normalization factor. The definition of $p(\mathbf{y}|\widetilde{\mathbf{x}})$ in (10) is used in the annotation (Equations 4 and 5) and retrieval (6) of test images.

4.4 Random Walk

The random walk uses the adjacency matrix $\mathbf{W}$ in (9) in order to build the probability transition matrix as follows:

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}, \quad (11)$$

where the diagonal matrix $\mathbf{D}(i,i) = \sum_j \mathbf{W}(i,j)$, which makes the row sum of $\mathbf{P}$ one. The initial distribution vector takes into account the similarity between the test image $\widetilde{\mathbf{x}}$ and all images in the database, as in $\mathbf{u} = [s_x(\mathbf{x}_1, \widetilde{\mathbf{x}}), ..., s_x(\mathbf{x}_N, \widetilde{\mathbf{x}})]^T$, where $\mathbf{u}$ is normalized in order to have $\|\mathbf{u}\|_1 = 1$. The random walk starts with the selection of a training image (say i^{th} training image) by sampling the distribution $\mathbf{u}$. Then, the distribution denoted by $\pi_i^T \mathbf{A}$ (with π_i a vector of zeros with a one at the i^{th} position) is used to select the next training image. After $R = 10$ steps of the random walk, a list of visited training images $T_{r,k}$ is formed, and the test image annotation is produced by (4) and (5), where the number of random walks is $K = 100$. The retrieval is produced as described in (6).

4.5 Stationary Solution

The stationary solution estimates the result of a random walk with a large number of steps [26]. The adjacency matrix $\mathbf{W}$ in (9) is used to build the following normalized transition matrix:

$$\mathbf{S} = \mathbf{D}^{-0.5}\mathbf{W}\mathbf{D}^{-0.5}, \quad (12)$$

with $\mathbf{D}$ defined in (11). This solution exploits the eigenvector centrality (i.e., the eigenvector of $\mathbf{S}$ associated with the eigenvalue 1), in order to determine the rank of a node (recall that a node represents a database image). This rank denotes the likelihood of visiting the node after a large number of random steps using a random starting point, where the decision to visit graph nodes is based on the edge weights [26].

Assuming a random initial distribution of the vertexes denoted by the vector $\mathbf{v}^{(0)}$, and that at each iteration of the random walk, the distribution used for the decision process is based on the weighted edges and on the probability vector $\mathbf{u}$ in (11) weighted by $(1-\alpha)$, with $0 \leq \alpha < 1$, we compute the stationary vector as follows [49]:

$$\begin{aligned} \mathbf{v}^{(t)} &= (\alpha \mathbf{S})^t \mathbf{v}^{(0)} + (\mathbf{I} + \alpha \mathbf{S})(1-\alpha)\mathbf{u} \Rightarrow \\ \mathbf{v}^{(\infty)} &= (1-\alpha)(\mathbf{I} - \alpha \mathbf{S})^{-1}\mathbf{u}, \end{aligned} \tag{13}$$

with $\mathbf{I}$ denoting the identity matrix.

Finally, in order to produce the annotation defined in (4) and (5), and the retrieval in (6), we define the probability of label given a test image, as follows:

$$p(\mathbf{y}|\widetilde{\mathbf{x}}) = \frac{1}{\mathcal{Z}_{SS}} \sum_{i=1}^{N} \mathbf{v}_i^{(\infty)} \delta(\mathbf{y} - \mathbf{y}_i), \tag{14}$$

where $\mathbf{v}_i^{(\infty)}$ is the i^{th} component of $\mathbf{v}^{(\infty)}$ defined in (13), and $\mathcal{Z}_{SS}$ represents a normalization factor.

4.6 Combinatorial Harmonic

The solution based on combinatorial harmonic follows the notation of semi-supervised labeling [51] and image segmentation [19] problems. Consider the following extension of the adjacency matrix (9), which includes the similarity between the test image and database images:

$$\widetilde{\mathbf{W}} = \begin{bmatrix} \mathbf{W} & \widetilde{\mathbf{u}} \\ \widetilde{\mathbf{u}}^T & 0 \end{bmatrix}, \tag{15}$$

where $\mathbf{W}$ is the adjacency matrix in (9) and $\widetilde{\mathbf{u}}$ is the un-normalized initial distribution vector $\mathbf{u}$ defined in (11). The goal is then to find the distribution $\mathbf{f}_U \in \Re^N$ ($\|\mathbf{f}_U\|_1 = 1$), which represents the probability of first reaching each of the training images in a random walk procedure, where the labeling matrix representing the training images is denoted by $\mathbf{F}_L = \mathbf{I}_N$, where $\mathbf{I}_N$ is an $N \times N$ identity matrix. In order to find $\mathbf{f}_U$, we minimize the following energy function:

$$E(\mathbf{F}) = \frac{1}{2} \left\| [\mathbf{F}_L, \mathbf{f}_U] \widetilde{\mathbf{L}} \begin{bmatrix} \mathbf{F}_L^T \\ \mathbf{f}_U^T \end{bmatrix} \right\|_2^2, \tag{16}$$

where $\mathbf{F} = [\mathbf{F}_L, \mathbf{f}_U]$, and $\widetilde{\mathbf{L}} = \widetilde{\mathbf{D}} - \widetilde{\mathbf{W}}$ is the Laplacian matrix computed from the the adjacency matrix $\widetilde{\mathbf{W}}$ (15), with diagonal matrix $\widetilde{\mathbf{D}}(i,i) = \sum_j \widetilde{\mathbf{W}}(i,j)$. This Laplacian matrix can be divided into the same blocks as in $\widetilde{\mathbf{W}}$, that is

$$\widetilde{\mathbf{L}} = \begin{bmatrix} \mathbf{L}_W & \mathbf{B} \\ \mathbf{B} & \mathbf{L}_U \end{bmatrix}. \tag{17}$$

Hence, in order to find $\mathbf{f}_U$, we solve the following optimization problem [19]:

$$\begin{array}{ll} \text{minimize} & E(\mathbf{F}) \\ \text{subject to} & \mathbf{F}_L = \mathbf{I}_N \end{array} \tag{18}$$

which is convex given that $\widetilde{\mathbf{L}}$ is positive semi-definite. Eq. 18 is solved by setting $\frac{\partial E(\mathbf{F})}{\partial \mathbf{f}_U} = 0$, which leads to the following analytical solution: $\mathbf{f}_U^T = -\mathbf{L}_U^{-1}\mathbf{B}^T\mathbf{I}_N$

The annotation defined in (4) and (5), and the retrieval in (6) are computed using the following probability of label given test image:

$$p(\mathbf{y}|\widetilde{\mathbf{x}}) = \frac{1}{\mathcal{Z}_{CH}} \sum_{i=1}^{N} \mathbf{f}_U(i) \delta(\mathbf{y} - \mathbf{y}_i), \tag{19}$$

where $\mathbf{f}_U(i)$ is the i^{th} component of $\mathbf{f}_U$, and $\mathcal{Z}_{CH}$ is a normalization factor.

5. EXPERIMENTS

In this experiment we compare the model presented in (3) to models based on SVM [45] and on RF [7] classifiers. For the RF model, we build $L = 22$ independent classifiers (one for the multi-class theme classification and the others for the binary problems - see Sec. 4.1), where each classifier is defined as $p(\mathbf{y}_l|\widetilde{\mathbf{x}}, \theta_{RF}(l))$, with $\theta_{RF}(l)$ representing the parameters of the random forests classifier for the l^{th} classification problem (recall that $l = 1, ..., L$). We obtain $\widehat{\mathbf{y}}$ using the notation defined in (3), but replacing $p(\mathbf{y}_i(m)|\widetilde{\mathbf{x}})$ by $p(\mathbf{y}_l(m)|\widetilde{\mathbf{x}}, \theta_{RF}(l))$. The main parameters of the random forests, which are the number and height of trees, are determined with cross validation, where the training set $\mathcal{D}$ is further divided into a training and validation sets of 90% and 10% of $\mathcal{D}$, respectively. The annotation is performed with (4) and (5), and the retrieval is done with (6). For the SVM, we train $M = 28$ classifiers using the one-versus-all training method. Specifically, we train the following classifiers $p(\mathbf{y}(m)|\widetilde{\mathbf{x}}, \theta_{SVM}(m))$, for $m \in \{1, ..., M\}$, and the annotation confidence $\widehat{\mathbf{y}}$ is produced by (3), replacing $p(\mathbf{y}_i(m)|\widetilde{\mathbf{x}})$ by $p(\mathbf{y}(m)|\widetilde{\mathbf{x}}, \theta_{SVM}(m))$. The main parameter of the support vector machine, which is the penalty factor for the slack variables, is also determined with cross validation, where the training set $\mathcal{D}$ is divided into a training and validation sets of 90% and 10% of $\mathcal{D}$, respectively. Also, we perform the test image annotation with (4) and (5), and the retrieval with (6). Note that these two models (RF and SVM) roughly represent the state-of-the-art approaches for image annotation and retrieval problems explained in Sec. 2.

In the results below, we use the following acronyms: LP$_i$ for label propagation with i nearest neighbors in (10), SS for stationary solution, CH for combinatorial harmonic, RF for random forests and SVM for support vector machines.

5.1 Retrieval

We measure the performance of the system in terms of retrieval using the precision and recall measures [10]. For each annotation class $\mathbf{y}(m)$ belonging to the set of classes in the test set $\mathcal{T}$ find the Q test images that produce the maximum values for (6). Out of those Q images, let the set $\mathcal{A}$ be the images for which $\mathbf{y}(m) = 1$ (note that $|\mathcal{A}| \leq Q$). Also, let $\mathcal{B} \subset \mathcal{T}$ be the set of all test images that have $\mathbf{y}(m) = 1$. Then, the precision and recall are computed as follows:

$$precision_R = \frac{|\mathcal{A}|}{Q}, \text{ and } recall_R = \frac{|\mathcal{A}|}{|\mathcal{B}|}. \tag{20}$$

The performance is computed with the mean average precision [10] (MAP), which is defined as the average precision over all queries, at the ranks where the recall changes. The results in Table 1 (first column) show the MAP average and standard deviation for the 10-fold cross validation experiment. Figure 6 (left) displays the retrieval performance of the CH solution as a function of the number

Table 1: Average ± standard deviation of the retrieval and annotation performance over the 10-fold cross validation experiment (the best performance for each measure is highlighted).

Models	MAP	Mean per-class Recall	Mean per-class Precision
LP_1	$0.33 \pm .03$	$0.43 \pm .04$	$0.43 \pm .04$
LP_{10}	$0.33 \pm .03$	$0.57 \pm .05$	$0.31 \pm .03$
LP_{20}	$0.33 \pm .02$	$0.58 \pm .06$	$0.31 \pm .02$
RW	$0.31 \pm .03$	$0.51 \pm .06$	$0.30 \pm .04$
CH	$\mathbf{0.35 \pm .03}$	$\mathbf{0.66 \pm .05}$	$0.32 \pm .03$
SS	$0.33 \pm .03$	$0.57 \pm .03$	$0.30 \pm .02$
RF	$0.30 \pm .03$	$0.23 \pm .03$	$0.40 \pm .06$
SVM	$0.22 \pm .01$	$0.12 \pm .01$	$\mathbf{0.52 \pm .03}$

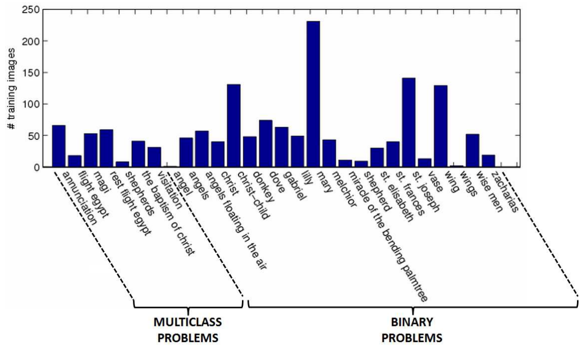


Figure 5: Number of training images per class.

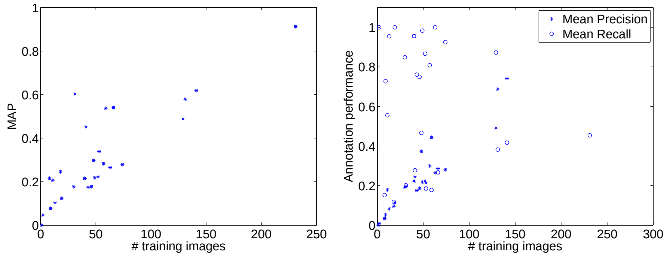


Figure 6: Performance of the retrieval (left) and annotation (right) as a function of the number of training images using the CH algorithm.

of training images. Notice that the retrieval performance is positively correlated with the number of training images. Figure 7 shows the the top retrieval results for four annotation classes using the CH algorithm.

5.2 Annotation

The performance of the annotation procedure is evaluated by comparing the results of the system in (4) and (5) with the manual annotation of the ground truth (recall that the set $\mathcal{T}$ also contains the manual annotation) [10]. For each annotation problem (binary or multi-class) indexed by $l \in 1, ..., L$, assume that there are w_H manually annotated images in $\mathcal{T}$, and the system annotates w_{auto}, of which w_C are correct. The precision and recall are computed as:

$$precision_A = \frac{w_C}{w_{auto}}, \text{ and } recall_A = \frac{w_C}{w_H}. \qquad (21)$$

Then, the values of $precision_A$ and $recall_A$ are averaged over the set of binary and multi-class problems. The results in Table 1 (last two columns) show the average and the standard deviation of the per-class precision and recall for the 10-fold cross validation. Fig. 6 (right) displays the annotation performance (mean per class precision and recall) of the CH algorithm as a function of the number of training images. Notice the positive correlation between precision and recall in terms of the number of training images. Figure 8 shows the annotation produced by the CH algorithm in four test images.

6. DISCUSSION AND CONCLUSIONS

In this work we presented a graph-based model for the annotation of art images. The retrieval experiment (Tab. 1) shows that the CH model produces better results than current state-of-the-art approaches based on SVM and RF. Also notice that the LP methods also show competitive results, indicating that simple models can also lead to powerful methods. The annotation results in Tab. 1 show that the CH approach produces the best performance in terms of recall, but SVM is better in terms of precision (but notice the poor result of SVM in terms of recall). This happens because SVM rarely classifies positively the test images with respect to each label, but whenever it does so, the annotation is often correct. We believe that this happens due to the limited number of training images per class to estimate the parameters of the SVM model. We plan to improve our method by modeling the dependencies between labels using, for example, structural learning methods [44, 47]. This would prevent the following two issues observed in Fig. 8: 1) use of too many labels in the annotation, and 2) presence of pairs of annotations that should never appear together (e.g., the presence of *wise men* in prints of theme *Annunciation* should not be allowed). The incorporation of structural learning in the methodology should be assessed with more appropriate measures, such as the one described by Nowak et al. [38]. We also intend to investigate other image features for image representation, such as wavelets and curvelets.

7. ACKNOWLEDGMENTS

The author would like to thank Duarte Lázaro and Rosário Carvalho for their help with the art history issues. I would like to acknowledge the use of code written by T. Joachims and by A. Vedaldi. I also would like to thank David Lowe, Benito Navarrete, and the members of the PrintArt project for valuable suggestions during the development of this work.

8. REFERENCES

[1] http://www.google.com/mobile/goggles/#artwork.

[2] http://www.artstor.org.

[3] F. Baumann, M. Karabelnik, and et al. *Degas Portraits*. London: Merrell Holberton, 1994.

[4] I. Berezhnoy, E. Postma, and H. van den Herik. Computerized visual analysis of paintings. In *Int. Conf. Association for History and Computing*, pages 28–32, 2005.

[5] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag, 2005.

[6] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Internet Techn.*, 5(1):231–297, 2005.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

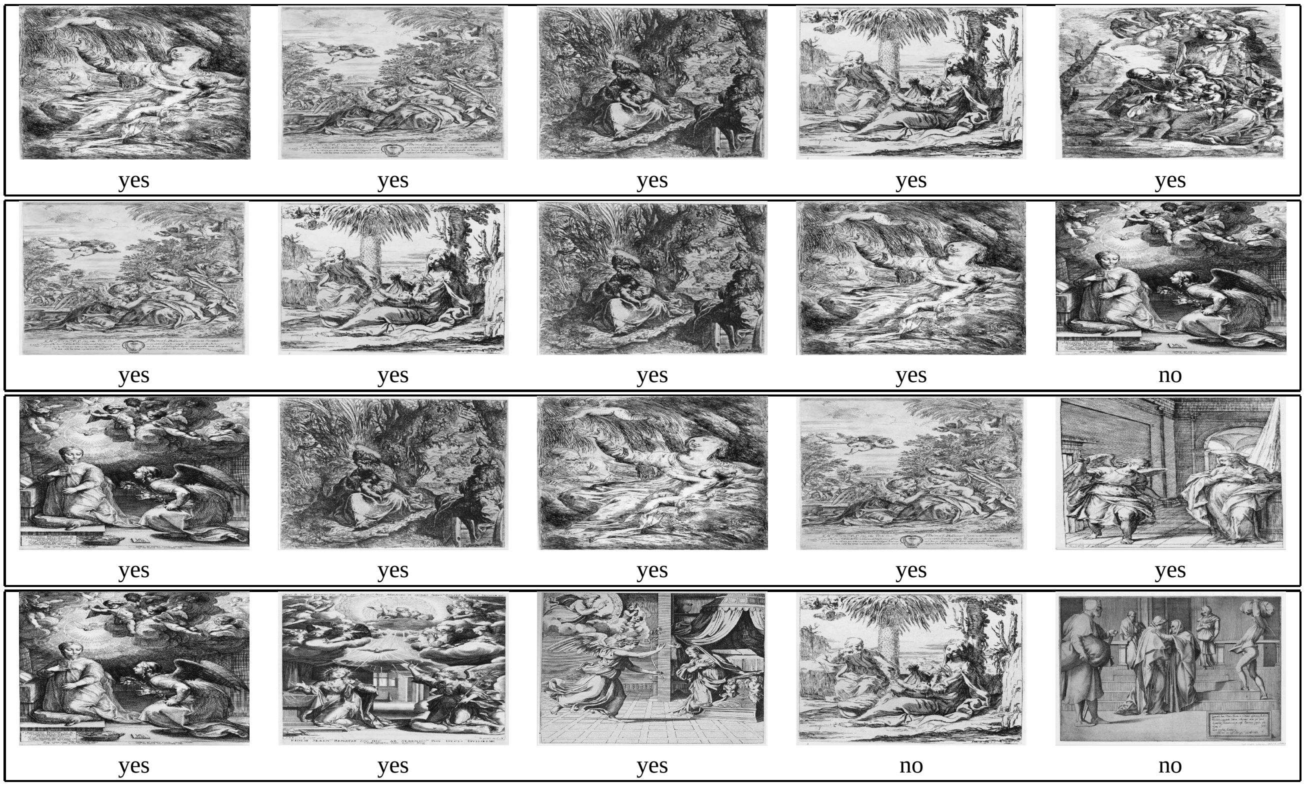


Figure 7: Retrieval results of the CH algorithm on the test set. Each row shows the top five matches to the following queries (from top to bottom): *‘rest on the flight into Egypt’***,** *‘Christ child’***,** *‘Mary’***, and** *‘dove’***. Below each image, it is indicated whether the image is manually annotated with the class.**

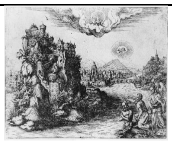
				
Human Annotation	Theme: Annunciation angels floating, dove, Gabriel Lilly, Mary, wing	Theme: Magi Christ-child, Mary, st. Joseph wise men	Theme: Baptism of Christ angels, Christ, dove, st. Frances, wing	Theme: Rest flight Egypt angels, angels floating, Christ-child donkey, Mary, miracle...palm tree st. Joseph, wing
Comb. Harm. Annotation	Theme: Annunciation angels floating, Christ-child, dove, Gabriel, Lilly, Mary, shepherd, vase, wing, wise men	Theme: Magi angels, angels floating, Christ, dove, Gabriel, Lilly, Melchior, st. Elizabeth st. Frances, vase, wing, wise men Zacharias	Theme: Baptism of Christ angel, angels, angels floating Christ, dove, st. Frances, wing, wings	Theme: Rest flight Egypt angels floating, Christ-child, donkey, Mary, miracle...palm tree, st. Joseph vase, wing, wings

Figure 8: Comparison of CH annotations with those of a human subject on test images.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextualweb search engine. In *Computer Networks and ISDN Systems*, pages 107–117, 1998.

[9] T. Campos. Application des regles iconographiques aux azulejos portugais du xviieme siecle. In *Europalia*, pages 37–40, 1991.

[10] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410, 2007.

[11] G. Carneiro and J. Costeira. The automatic annotation and retrieval of digital images of prints and tile panels using network link analysis algorithms. In *Proceedings of the SIE - Computer Vision and Image Analysis of Art II*, 2011.

[12] R. Carvalho. O programa artistico da ermida do rei salvador do mundo em castelo de vide, no contexto da arte barroca. *Artis -Revista do Instituto de Historia da Arte da Faculdade de Letras de Lisboa*, (2):145–180, 2003.

[13] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[14] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008.

[15] K. V. de Sande, T. Gevers, and A. Smeulders. The university of amsterdamâĂŹs concept detection system at imageclef 2009. In *CLEF working notes 2009*, 2009.

[16] O. Dekel and O. Shamir. Multiclass-multilabel classification with more classes than examples. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[17] F. Estrada, D. Fleet, , and A. Jepson. Stochastic image denoising. In *BMVC*, 2009.

[18] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR (2)*, pages 1002–1009, 2004.

[19] L. Grady. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(11):1768–1783, 2006.

[20] D. Graham, J. Friedenberg, D. Rockmore, and D. Field. Mapping the similarity space of paintings: image statistics and visual perception. *Visual Cognition*, 18(4):559–573, 2010.

[21] S. Griffin. Recovering the past through computation: new techniques for cultural heritage. In *Multimedia Information Retrieval*, pages 13–14, 2010.

[22] X. He, W.-Y. Ma, and H.-J. Zhang. Imagerank: Spectral techniques for structural analysis of image database. In *IEEE International Conference on Multimedia and Expo*, pages 25–28, 2003.

[23] D. Hsu, S. Kakade, J. Langford, and T. Zhang. Multi-label prediction via compressed sensing. In *Neural Information Processing Systems*, 2009.

[24] J. Hughes, D. Graham, and D. Rockmore. Stylometrics of artwork: uses and limitations. In *Proceedings of SPIE: Computer Vision and Image Analysis of Art*, 2010.

[25] S. Jafarpour, G. Polatkan, I. Daubechies, S. Hughes, and A. Brasoveanu. Stylistic analysis of paintings using wavelets and machine learning. In *European Signal Processing Conference*, 2009.

[26] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1877–1890, 2008.

[27] C. Johnson, E. Hendriks, I. Berezhnoy, E. Brevdo, S. Hughes, I. Daubechies, J. Li, E. Postma, and J. Wang. Image processing for artistic identification: Computerized analysis of vincent van goghâĂŹs brushstrokes. *IEEE Signal Processing Magazine*, pages 37–48, 2008.

[28] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *CVPR*, pages 1719–1726, 2006.

[29] J. Li and J. Wang. Studying digital imagery of ancient paintings by mixtures of stochastic models. *IEEE Trans. Image Processing*, 13(3):340âĂŞ–353, 2004.

[30] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

[31] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. paper accepted for publication.

[32] S. Lyu, D. Rockmore, and H. Farid. A digital technique for art authentication. *Proceedings of the National Academy of Sciences USA*, 101(49):17006–âĂŞ17010, 2004.

[33] H. Maitre, F. Schmitt, and C. Lahanier. 15 years of image processing and the fine arts. In *IEEE Int. Conf. Image Processing*, pages 557âĂŞ–561, 2001.

[34] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV (3)*, pages 316–329, 2008.

[35] A. Ng, A. Zheng, and M. Jordan. Link analysis, eigenvectors and stability. In *IJCAI*, pages 903–910, 2001.

[36] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[37] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, pages 2161–2168, 2006.

[38] S. Nowak, H. Lukashevich, P. Dunker, and S. Rüger. Performance measures for multilabel evaluation: a case study in the area of image classification. In *Multimedia Information Retrieval*, pages 35–44, 2010.

[39] G. Polatkan, S. Jafarpour, A. Brasoveanu, S. Hughes, and I. Daubechies. Detection of forgery in paintings using supervised learning. In *International Conference on Image Processing*, 2009.

[40] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.

[41] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

[42] D. Stork. Computer image analysis of paintings and drawings: An introduction to the literature. In *Proceedings of the Image Processing for Artist Identification Workshop*, 2008.

[43] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *NIPS*, pages 945–952, 2001.

[44] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: a large margin approach. In *ICML*, pages 896–903, 2005.

[45] V. N. Vapnik. *Statistical Learning Theory.* Wiley, 1998.

[46] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Image annotation with tagprop on the mirflickr set. In *Multimedia Information Retrieval*, pages 537–546, 2010.

[47] X. Xue, H. Luo, and J. Fan. Structured max-margin learning for multi-label image annotation. In *CIVR*, pages 82–88, 2010.

[48] M. Yelizaveta, C. Tat-Seng, , and R. Jain. Semi-supervised annotation of brushwork in paintings domain using serial combinations of multiple experts. In *ACM Multimedia*, pages 529âĂŞ–538, 2006.

[49] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *NIPS*, pages 321–328, 2004.

[50] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[51] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.