The Automatic Annotation and Retrieval of Digital Images of Prints and Tile Panels using Network Link Analysis Algorithms

Gustavo Carneiro [*,†] and João P. Costeira[*]

Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisbon, Portugal

ABSTRACT

The study of the visual art of printmaking is fundamental for art history. Printmaking methods have been used for centuries to replicate visual art works, which have influenced generations of artists. Particularly in this work, we are interested in the influence of prints on artistic tile panel painters, who have produced an impressive body of work in Portugal. The study of such panels has gained interest by art historians, who try to understand the influence of prints on tile panels artists in order to understand the evolution of this type of visual arts. Several databases of digitized art images have been used for such end, but the use of these databases relies on manual image annotations, an effective internal organization, and an ability of the art historian to visually recognize relevant prints. We propose an automation of these tasks using statistical pattern recognition techniques that takes into account not only the manual annotations available, but also the visual characteristics of the images. Specifically, we introduce a new network link-analysis method for the automatic annotation and retrieval of digital images of prints. Using a database of 307 annotated images of prints, we show that the annotation and retrieval results produced by our approach are better than the results of state-of-the-art content-based image retrieval methods.

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

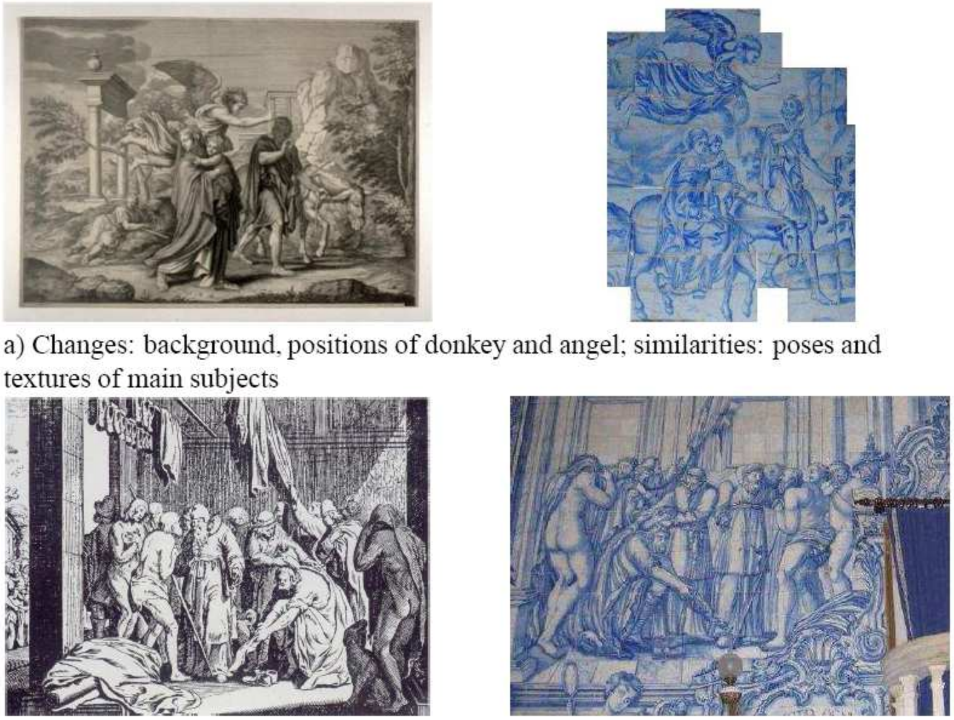
1. INTRODUCTION

Printmaking is a method for replicating paintings (usually on paper) based on intaglio printing (e.g., etching), relief printing (e.g., engraving) or planographic printing (e.g., lithography).[1] During the 15^{th} century, the fast and cheap production of paper and advancements in graphical arts made printmaking one of the main forms of reproduction of European masters' paintings. As a result, reproductions of the paintings reached a large number of people. Consequently it is worth understanding the importance of printmaking in the 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 printmaking productions constitute a quite important activity for artists and art historians in order to understand the art produced in the last five centuries.

In Portugal, printmaking images influenced generations of artistic tile panel painters, as evidenced by the large number of panels found in Churches, public buildings and palaces.[2,3] As a consequence, the study of printmaking is extremely important not only for classification and annotation purposes, but also for understanding how the Portuguese tile panel artists have been influenced, which has the potential to furnish relevant information to art historians in Portugal. It is important to mention that printmaking images were often used as references to produce the panels, but the artists often did not make an exact copy of the image. For example, original prints have generally suffered several modifications, such as background changes, arbitrary displacements of the position of subjects, mirror transformations, etc. (Fig. 1). For this reason, the task of discovering the influence of one or several prints in the composition of a tile panel is rather complex. This task requires an expert with a specialized visual knowledge of the current databases of prints together with peculiar abilities in to relate tile panel compositions and prints. A system which can automatically retrieve a set of printmaking images related to a tile panel image can dramatically help art historians in this task.

Compared to photographic digital image, the images from printmaking and artistic tile panels are quite poor in terms of color and texture. For instance, Fig. 2 shows three images of the theme *the Crucifixion of Jesus Christ*, which displays the large sensorial gap[4] between the original scene and the different forms of art productions. This means that much of the information (texture, color, depth, etc.) present in the original scene is lost. The loss of such information reduces

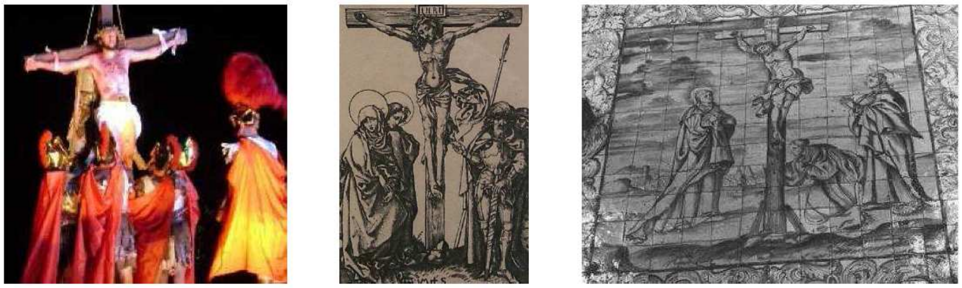
[*]This work was supported by the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds and Project PRINTART (PTDC/EEA-CRO/098822/2008). [†]This project is also funded by the European Commission; Marie Curie IIF, contract number PIIF-GA-2009-236173 (IMASEG3D).



a) Changes: background, positions of donkey and angel; similarities: poses and textures of main subjects

b) Changes: background, the whole scene suffered a mirror transform; similarities: poses and textures of main subjects

Figure 1. Examples of how printmaking images are altered in the process of becoming a tile panel.



a) Photograph

b) Printmaking Image

c) Artistic Tile Panel Image

Figure 2. Differences between a photographic image in (a), printmaking image in (b), and tile panel image in (c).

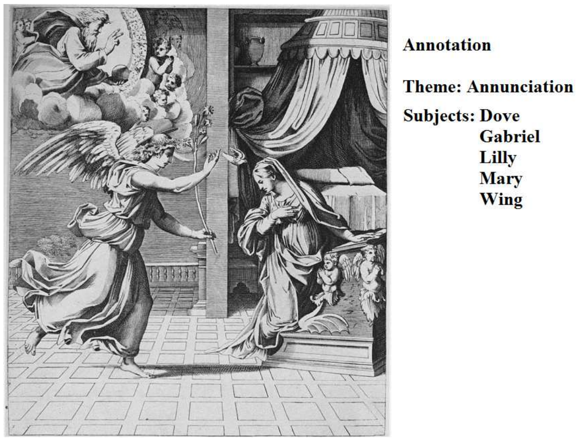


Figure 3. Example of annotation of printmaking images. The image (from Artstor[9]) is shown with its the manual annotation produced by an art historian.

the effectiveness of current image analysis techniques, which usually work with photographic digital images.[5] Art image analysis can also be used in this work,[6,7] but the great majority of these techniques have been developed for the analysis of digitized images of paintings, which still contain texture and color. The most similar type of images compared to printings and tile panels are the ancient Chinese paintings used in the image analysis method proposed by Li and Wang.[8]

In this work, we present a new method for the analysis of images from printmaking processes. This analysis of prints represents the first step in the analysis of tile panel images, which will is our long-term objective. The goal of the work is to automatically produce global annotations to new test images of prints using a statistical model whose parameters are estimated using a database of manually annotated training images. The prints 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 because a great number of tile panel productions are restricted in the same way. The manual annotation (Fig. 3) has been produced by art historians, who label the image theme and the relevant subjects present. The method presented in the paper is based on network link analysis (we use the terms network link analysis and graph-based learning algorithms interchangeably), which has the assumption that the visually similar images are likely to share the same annotations. Preliminary annotation and retrieval results are shown in a database containing 307 images, and we compare the results with bag of visual words methods using the following classifiers: support vector machines (SVM)[10] and random forests (RF).[11] The results show that our method provide several advantages in terms of accuracy of annotation and retrieval.

2. LITERATURE REVIEW

In this section, we provide a brief review of papers in the areas of retrieval and annotation of photographic images and art images. We also review graph-based learning methods that are relevant to our technique.

Currently in photographic image annotation and retrieval, the most successful methods are based on the bag of visual words framework using a multiple kernel learning (MKL) classifier,[12] which is an extension of the SVM classifier that allows the combination of several kernels. This methodology is quite effective when used in a retrieval setting where the number visual classes is relatively small (with a large number of training images per class), and these visual classes are fixed. Unfortunately, both constraints do not apply in our case where the number of visual classes can be quite large (with each visual class containing relatively small number of training images), and the introduction of new images to the database may happen often. 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 of compressed sensing,[13] which finds a sub-space of smaller dimensionality for classification. However, the dynamic nature of this learning 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 annotation and retrieval has attracted the attention of researchers in the fields of computer vision and machine learning.[6,7,14] 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[15–17] or to produce stylistic analysis of paintings.[18–20] Most of the methods above can be regarded as adaptations from the content-based image retrieval systems,[5] 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[8] 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.[21] is also similar to ours given that the authors are dealing with multi-class classification of brush strokes. Nevertheless, as we shall explain below, our problem involves not only a multi-class, but also a multi-label problem.[22]

The problem of graph-based learning (or network link analysis) has been thoroughly studied by the information retrieval community to solve the problem of ranking web pages from the World Wide Web using hyperlink structure analysis.[23–25] Essentially, a graph is built where the nodes are represented by the web pages and the edge weights denote existence of hyper-links, and several analysis algorithms based on random walks in this graph have been designed to rank the nodes (i.e., web pages). As suggested by the explanation above, graph-based approaches makes an implicit assumption of label smoothness over the graph.[26] These graph-based techniques have received considerable attention from the machine learning community for the problem of semi-supervised learning,[26] where the training set contains labeled and unlabeled samples, and the goal is to annotate the unlabeled samples using the similarity graph structure and the labeled samples. Note that the similarity graph is built based on the distances between the samples in some feature space, and one of the forms to explore the structure of this graph is with the use of random walk algorithms. The random walk algorithm has also been studied in the domains of unsupervised image segmentation[27] and multi-class classification.[28] Finally, random walk algorithms have also been explored in the area of image retrieval,[29–31] where the main contribution is the use of visual and non-visual cues in the random walk procedure, which is also explored in our work.

3. METHODOLOGY

Assume that a training set of annotated images is available, and is represented as follows: $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1..N}$ with $\mathbf{x}_i$ representing the feature vector of image I_i and $\mathbf{y}_i$ denoting the annotation of the same image. An annotated test set is also available and is represented by $\mathcal{T} = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)\}_{i=1..P}$, but note that the annotation in the test set is used only for the purpose of methodology evaluation. It is important to mention that 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 $\mathbf{y}_j \in \{0,1\}^{|\mathbf{y}_j|}$, where $|\mathbf{y}_j|$ represents the dimensionality of $\mathbf{y}_j$. Binary annotations are denoted by $\mathbf{y}_j \in \{0,1\}$. Multi-class problems have $1 < |\mathbf{y}_j| \leq M$, and $\sum_{k=1}^{|\mathbf{y}_j|} \mathbf{y}_j(k) \in \{0,1\}$, with each class in the problem represented by the dimension $\mathbf{y}_j(k) \in \{0,1\}$. In general, binary annotation involve problems that indicate either the presence or absence of a class, while multi-class annotation regards problems that at most one of the possible classes must be present (e.g., the theme of a print). Note that the sum of the dimensionalities of the L problems must be equal to the dimensionality of the annotation vector, so $\sum_{j=1}^{L} |\mathbf{y}_j| = M$.

Following the notation introduced by Estrada et al.,[32] 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 $\mathcal{D}$, 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:

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

In (1), $\mathcal{Z}_T$ is a normalization factor, K represents the total number of random walks,R denotes the length of one random walk, $p(\mathbf{y}|\mathbf{x}^{(r,k)}) = \delta(\mathbf{y} - \mathbf{y}^{(r,k)})$ (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,

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

with the derivation made assuming a Markov process and that the training labels and features are independent given the test image, $p(\mathbf{y}^{(r,j)}|\mathbf{y}^{(r,j-1)}) = I_y(\mathbf{y}^{(r,j)}, \mathbf{y}^{(r,j-1)})$ with $I_y(\mathbf{y}^{(r,j)}, \mathbf{y}^{(r,j-1)}) = \frac{1}{\mathcal{Z}_y} \sum_{l=1}^{M} \lambda_l \times y_l^{(r,j)} \times y_l^{(r,j-1)}$ (λ_l is the weight associated with the label l 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. 3.2.

We propose the use of class mass normalization[33] to determine the annotation for image $\tilde{\mathbf{x}}$. 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

$$\widehat{\mathbf{y}} = \sum_{i=1}^{N} \mathbf{y}_i \max(p(\mathbf{y}_i|\widetilde{\mathbf{x}}) - p(\mathbf{y}), 0), \tag{3}$$

where $p(\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i$. The use of this proportion of class annotation 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 $\widehat{\mathbf{y}}$ represents the probability that the image represented by $\tilde{\mathbf{x}}$ is annotated with each of the labels $[\hat{y}_1, ..., \hat{y}_M] = \widehat{\mathbf{y}}$. However, recall that some of these annotations belong to multi-class problems, so we process $\widehat{\mathbf{y}}$ as in

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

for multi-class problems, and for binary problems, we define:

$$\begin{array}{l} \forall j \in \{1, ..., L\}, \text{with } |\widehat{\mathbf{y}}_j| = 1 \\ \mathbf{y}^*_j = \begin{cases} 1, \text{if } \widehat{\mathbf{y}}_j > 0.5 \\ 0, \text{otherwise} \end{cases} . \end{array} \tag{5}$$

As a result, the final annotation for image $\widetilde{x}$ 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 a visual class $y_t \in \{0, 1\}$ for $t \in \{1, ..., M\}$ (i.e., one of the annotations in $\mathbf{y} \in \{0, 1\}^M$), as in:

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

where $p(\widetilde{\mathbf{x}}|y_t) = p(y_t|\widetilde{\mathbf{x}})p(\widetilde{\mathbf{x}})/p(y_t)$ with $p(\widetilde{\mathbf{x}}) =$ constant and $p(y_t)$ being irrelevant for (6). Furthermore, $p(y_t|\widetilde{\mathbf{x}})$ is defined as in (1) substituting the last term for $p(y_t|\mathbf{x}^{r,k}) = \delta(y_t - y_t^{(r,k)})$.

3.1 Data Set

The data set is formed by 307 annotated images with one multi-class problem (theme with seven classes) and 21 binary problems [†]. All images have been collected from the Artstor digital image library.[9] Fig. 3 shows an example of a manual annotation produced by an art historian. For the experiments in Sec. 4, 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).

3.2 Image Features

The images are represented with the bag of visual words model,[34] where each visual word is formed with a collection of scale invariant feature transform (SIFT) local descriptors.[35] The visual vocabulary is built using a 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), scale (patch size), and dominant orientation are randomly determined,[37] and we generate 1000 descriptors per image. Then, the vocabulary is

[†] The classes considered are: theme (annunciation, flight into Egypt, magi, rest on the flight into Egypt, shepherds, the baptism of Christ, visitation), angel, angels, angels floating into the air, Christ, Christ-child, donkey, dove, Gabriel, Lilly, Mary, Melchior, miracle of the bending palm tree, shepherd, st. Elisabeth, st. Frances, st. Joseph, vase, wing, wings, wise men, and Zacharias

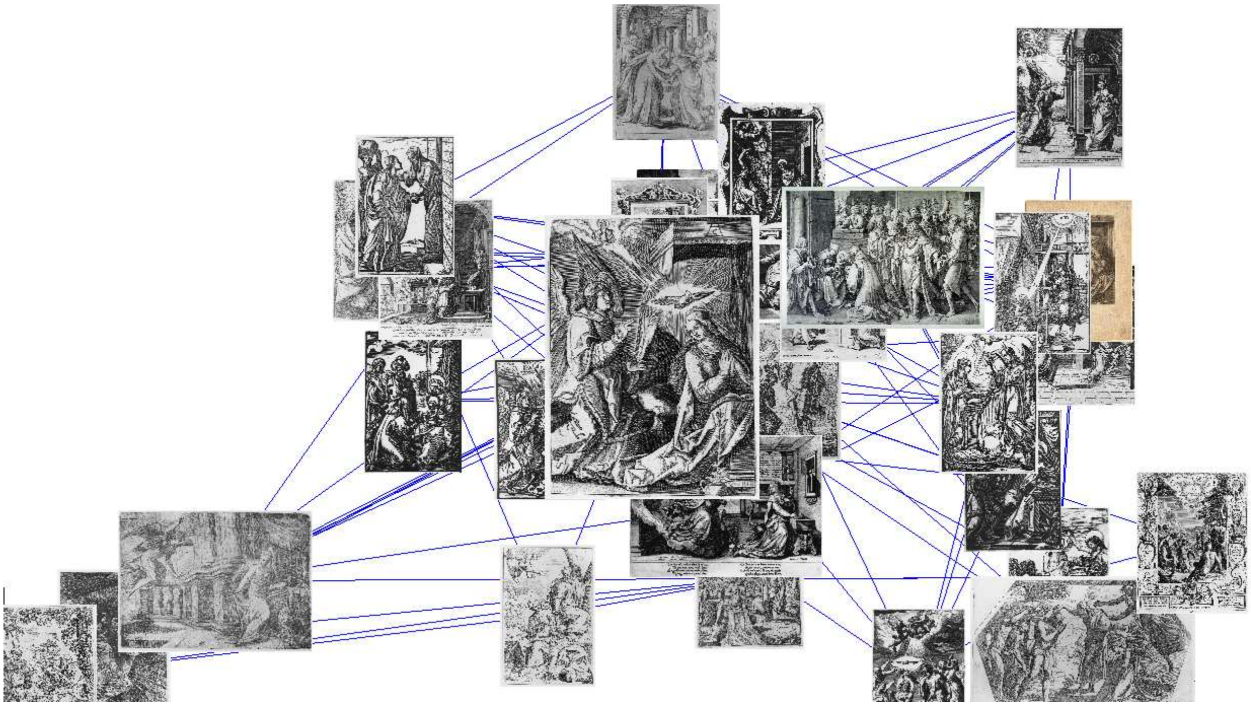


Figure 4. Network structure built using the adjacency matrix of (9) and shown using a variant of the MDS algorithm.[38] The large image in the center denotes the test image and the most visually similar images to it appear closer in the graph.

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 $1 + 10 + 100 + 1000 = 1111$ nodes, 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 nodes). The histogram of visited nodes is weighted by the node entropy (i.e., nodes that are visited more often receive smaller weights). As a result, an image I is represented by a vector $\mathbf{x} \in \Re^{1111}$, representing the histogram above.

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:

$$p(\mathbf{x}^{(r,j)}|\mathbf{x}^{(r,j-1)}, \widetilde{\mathbf{x}}) = I_x(\mathbf{x}^{(r,j)}, \mathbf{x}^{(r,j-1)}) I_x(\mathbf{x}^{(r,j)}, \widetilde{\mathbf{x}}), \tag{7}$$

and the transition probability between two feature vectors is given by

$$p(\mathbf{x}^{(r,1)}|\widetilde{\mathbf{x}}) = I_x(\mathbf{x}^{(r,1)}, \widetilde{\mathbf{x}}), \tag{8}$$

with $I_x(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\mathcal{Z}_x} \sum_{d=1}^{1111} \min(x_i(d), x_j(d))$ where $\mathcal{Z}_x$ is a normalization factor.

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 following adjacency matrix:

$$\mathbf{W}(j,i) = I_y(\mathbf{y}_i, \mathbf{y}_j) \times I_x(\mathbf{x}_i, \mathbf{x}_j) \times I_x(\mathbf{x}_j, \widetilde{\mathbf{x}}), \tag{9}$$

with $\mathbf{W}(i,i) = 0$ for all $i \in \{1, ..., N\}$. In this part of the graph, we take a test image represented by $\widetilde{\mathbf{x}}$ shown at the center (note the enlarged image in the figure), and display the graph structure from the training database around it. Notice that the neighboring images in the graph tend to be similar visually or with respect to the annotation. In order to show this graph, we use a variant of the multidimensional scaling algorithm for visualization (MDS).[38]

3.3 Random Walk

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

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}, \tag{10}$$

Table 1. Retrieval performance.

Models	Random Walk	RF	SVM
MAP	0.31	0.30	0.22

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 $\mathbf{u} = [I_x(\mathbf{x}_1, \widetilde{\mathbf{x}}), ..., I_x(\mathbf{x}_N, \widetilde{\mathbf{x}})]^T$, with $\|\mathbf{u}\|_1 = 1$. The random walk starts by selecting a training image (say i^{th} training image) by sampling the distribution $\mathbf{u}$. Then, use the distribution denoted by $\mathbf{A}\pi_i$ (with π_i a vector of zeros with a one at the i^{th} position) for selecting the next training image in the random walk procedure. After the random walk is finished, form the list of visited training images $T_{r,k}$ from Sec. 3, and produce the annotation as defined in (4) and (5) with $K = 100$ random walks and the length of each random walk is denoted by $R = 10$. The retrieval is produced as described in (6).

4. EXPERIMENTS

In this experiment we compare the model presented in (3) to models based on SVM[10] and on RF.[11] 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. 3.1), where each classifier is defined as $p(\mathbf{y}_l|\widetilde{\mathbf{x}}, \theta_{RF}(l))$, with the $\theta_{RF}(l)$ representing the parameters of the random forests classifier for the l^{th} classification problem (recall that $l = 1, ..., L$). Using the same notation as in (3), we have $\widehat{\mathbf{y}} = \max([p(\mathbf{y}_l|\widetilde{\mathbf{x}}, \theta_{RF}(l))]_{l=1..L} - p(\mathbf{y}), 0) = [\hat{y}_1, ..., \hat{y}_M]$. The main parameters of the random forests, which are the number and height of tress, 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. Then, the multi-class decisions are performed with (4) and binary problems with (5). For the SVM, we train $M = 28$ classifiers using the one-versus-all training method, where for the multi-class theme classification, we adopt the winner-takes-all strategy. Specifically, we train the following classifiers $p(y_t|\widetilde{\mathbf{x}}, \theta_{SVM}(t))$, for $t = \{1, ..., M\}$, and the annotation probability is produced by $\widehat{\mathbf{y}} = \max([p(y_t|\widetilde{\mathbf{x}}, \theta_{SVM}(t))]_{t=1..M} - p(\mathbf{y}), 0) = [\hat{y}_1, ..., \hat{y}_M]$. The main parameter of the support vector machine, which are penalty factor for the slack variables, is also determined with cross validation, where the training set $\mathcal{D}$ is again divided into a training and validation sets of 90% and 10% of $\mathcal{D}$, respectively. Also, we perform the multi-class decisions with (4) and binary problems with (5). Note that these two models roughly represent the state-of-the-art approaches for image annotation and retrieval problems explained in Sec. 2.

4.1 Retrieval

We measure the performance of the system in terms of retrieval using the precision and recall measures.[39] For each annotation class y_t belonging to the set of classes in the test set $\mathcal{T}$, find the n test images that produce the maximum values for (6). Out of those n images, let the set $\mathcal{A}$ be the images for which $y_t = 1$ (note that $|\mathcal{A}| \leq n$). Also, let $\mathcal{B} \subset \mathcal{T}$ be set of all test images that have $y_t = 1$. Then, the precision and recall are computed as follows:

$$precision_R = \frac{|\mathcal{A}|}{n}, \text{ and } recall_R = \frac{|\mathcal{A}|}{|\mathcal{B}|}. \tag{11}$$

The performance is computed with the mean average precision[39](MAP), which is defined as the average precision over all queries, at the ranks that the recall changes. The results in Table 1 show the average MAP for a 10-fold cross validation experiment with different sets $\mathcal{D}$ and $\mathcal{T}$, as explained in Sec. 3.1. Figure 6 shows the the top five retrieval results for four annotation classes. Also, Fig. 5 displays the MAP results for each visual class (top-left), the number of training image per visual class (bottom-left) and the MAP performance in terms of the number of training images (center graph). Notice that it is quite clear that the retrieval performance is positively correlated with the number of training images.

4.2 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).[39] For each annotation

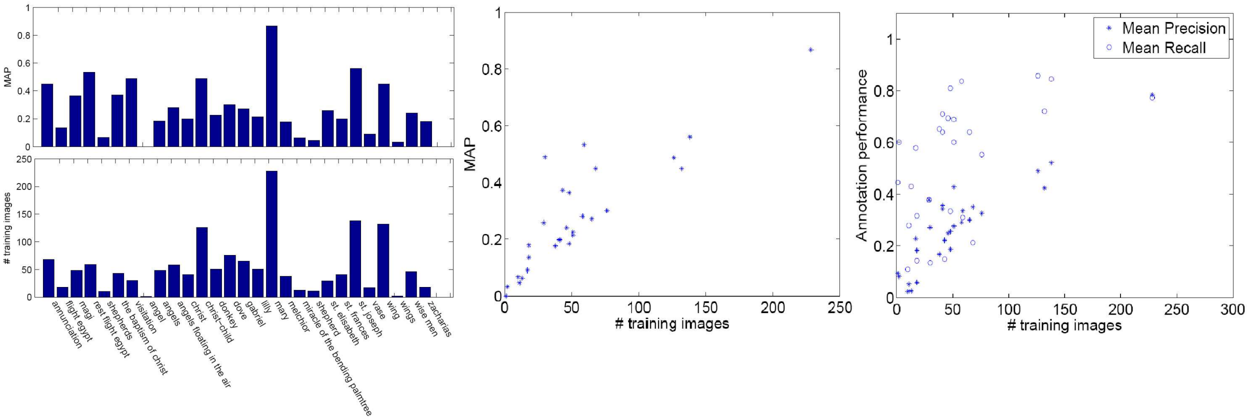


Figure 5. MAP results per class (left-top), number of training images per visual class (left-bottom), MAP as a function of the number of training images (center), and the annotation performance as a function of the number of training images (right).

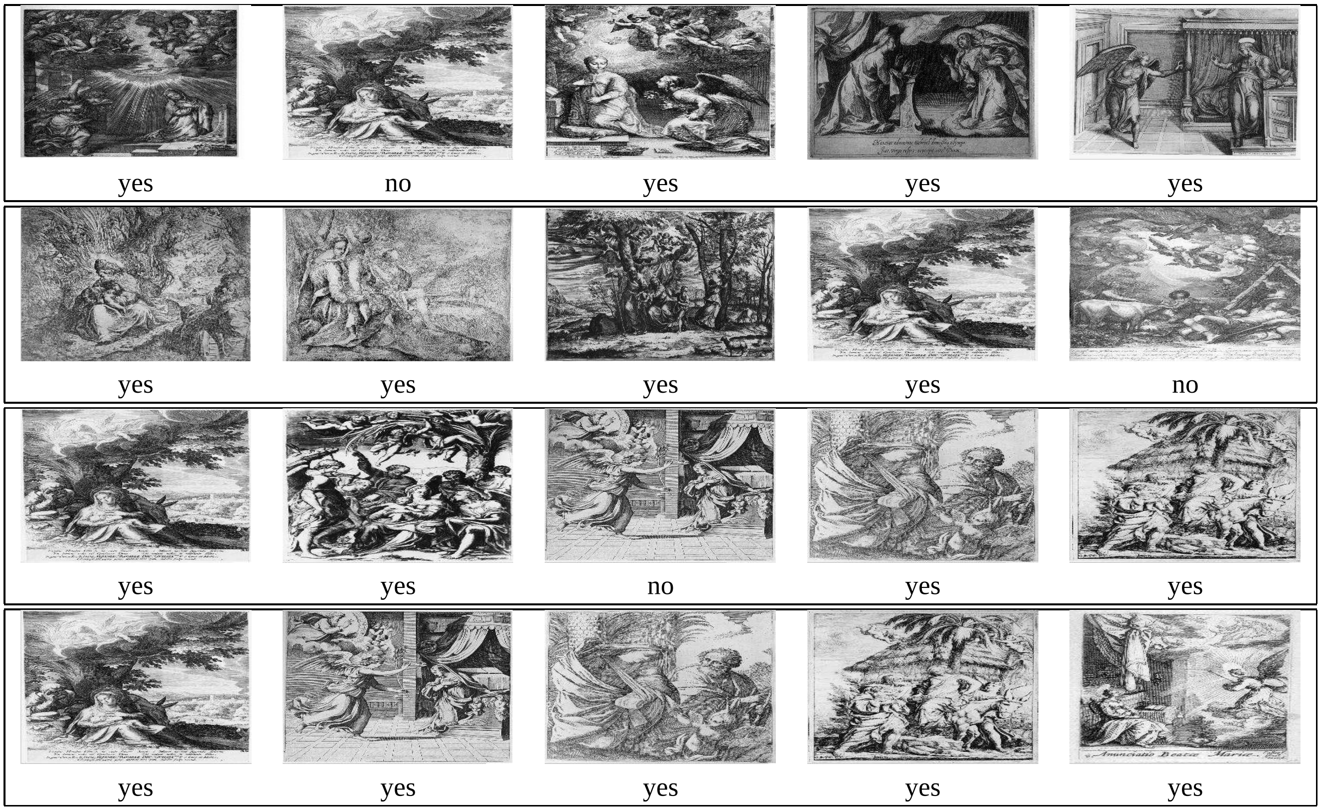


Figure 6. Retrieval results. Each row shows the top five matches to the following queries (from top to bottom): *'annunciation'*, *'rest on the flight into Egypt'*, *'Christ child'*, and *'Mary'*. Below each image (from Artstor[9]), it is indicated whether the image is annotated with the class.

problem (binary or multi-class) indexed by $j \in 1, ..., L$, assume that there are w_H manually annotated images in $\mathcal{T}$, and the system annotates w_{auto}, of which w_C are correct. We compute precision and recall as follows:

$$precision_A = \frac{w_C}{w_{auto}}, \text{ and } recall_A = \frac{w_C}{w_H}. \quad (12)$$

Then, the values of $precision_A$ and $recall_A$ are averaged over the set of binary and multi-class problems. The results in Table 2 show the average mean per-word precision and recall for a 10-fold cross validation experiment with different sets

Table 2. Annotation performance.

Models	Random Walk	RF	SVM
Mean Per-word Recall	0.51	0.40	0.12
Mean Per-word Precision	0.30	0.23	0.52

$\mathcal{D}$ and $\mathcal{T}$, as explained in Sec. 3.1. Figure 7 shows the annotation produced by our system in five test images. Finally, Fig. 5 displays the annotation performance in terms of the number of training images (right graph). Notice the correlation between precision and recall in terms of the number of training images.

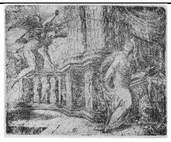
					
Human Annotation	Theme: Rest flight Egypt angels floating, Christ child, Mary, st. Joseph, wing	Theme: Magi Christ child, Mary, st. Joseph, wise men	Theme: Annunciation Gabriel, Lilly Mary, wing	Theme: Baptism of Christ angels floating, Christ dove, st. Frances, wing	Theme: Flight into Egypt angels floating,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

Figure 7. Comparison of Random Walk annotations with those of a human subject on images (from Artstor[9]).

5. 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 our model based on random walks produces slightly better results than current state-of-the-art approaches based on SVM and RF. The annotation results in Tab. 2 shows that our approach produces the best performance in terms of recall, but SVM appears 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, it is correct more often than with random walk and RF. We believe that this happens due to the limited number of training images per class to estimate the parameters of these models. We plan to improve our random walk model and use closed form solutions.[26,27] The dependencies between labels should also be encoded into the method using, for example, structural learning methods.[40,41] This would prevent the following two issues observed in Fig. 7: 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 *Melchior* in prints of theme *Annunciation* should not be allowed). The incorporation of structural learning in the methodology should be evaluated by more effective retrieval measures, such as the one described by Nowak et al.[42] We also intend to investigate other image features for image representation, such as wavelets and curvelets. In the near future, we shall make the training database available with several benchmark results.

ACKNOWLEDGMENTS

The authors would like to thank Duarte Lázaro and Rosário Carvalho for their help with the art history issues. We would like to acknowledge the use of SVMLight by Thorsten Joachims. Finally, we would like to thank David Lowe for valuable suggestions on the development of this work.

REFERENCES

[1] http://www.visual-arts cork.com/printmaking.htm.

[2] Campos, T., “Application des regles iconographiques aux azulejos portugais du xviieme siecle,” in [*Europalia*], 37–40 (1991).

[3] Carvalho, R., "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).

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

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

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

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

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

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

[10] Vapnik, V. N., [*Statistical Learning Theory.*], Wiley (1998).

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

[12] Vedaldi, A., Gulshan, V., Varma, M., and Zisserman, A., "Multiple kernels for object detection," in [*International Conference on Computer Vision*], (2009).

[13] Hsu, D., Kakade, S., Langford, J., and Zhang, T., "Multi-label prediction via compressed sensing," in [*Neural Information Processing Systems*], (2009).

[14] Maitre, H., Schmitt, F., and Lahanier, C., "15 years of image processing and the fine arts," in [*IEEE Int. Conf. Image Processing*], 557–561 (2001).

[15] Berezhnoy, I., Postma, E., and van den Herik, H., "Computerized visual analysis of paintings," in [*Int. Conf. Association for History and Computing*], 28–32 (2005).

[16] Lyu, S., Rockmore, D., and Farid, H., "A digital technique for art authentication," *Proceedings of the National Academy of Sciences USA* **101**(49), 17006–17010 (2004).

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

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

[19] Hughes, J., Graham, D., and Rockmore, D., "Stylometrics of artwork: uses and limitations," in [*Proceedings of SPIE: Computer Vision and Image Analysis of Art*], (2010).

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

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

[22] Dekel, O. and Shamir, O., "Multiclass-multilabel classification with more classes than examples," in [*International Conference on Artificial Intelligence and Statistics*], (2010).

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

[24] Brin, S. and Page, L., "The anatomy of a large-scale hypertextualweb search engine," in [*Computer Networks and ISDN Systems*], 107–117 (1998).

[25] Ng, A., Zheng, A., and Jordan, M., "Link analysis, eigenvectors and stability," in [*IJCAI*], 903–910 (2001).

[26] Zhu, X., "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison (2005).

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

[28] Szummer, M. and Jaakkola, T., "Partially labeled classification with markov random walks," in [*NIPS*], 945–952 (2001).

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

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

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

[32] Estrada, F., Fleet, D., , and Jepson, A., “Stochastic image denoising,” in [*BMVC*], (2009).

[33] Zhu, X., Ghahramani, Z., and Lafferty, J., “Semi-supervised learning using gaussian fields and harmonic functions,” in [*ICML*], 912–919 (2003).

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

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

[36] Nistér, D. and Stewénius, H., “Scalable recognition with a vocabulary tree,” in [*CVPR*], 2161–2168 (2006).

[37] Nowak, E., Jurie, F., and Triggs, B., “Sampling strategies for bag-of-features image classification,” in [*ECCV*], 2161–2168 (2006).

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

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

[40] Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C., “Learning structured prediction models: a large margin approach,” in [*ICML*], 896–903 (2005).

[41] Xue, X., Luo, H., and Fan, J., “Structured max-margin learning for multi-label image annotation,” in [*CIVR*], 82–88 (2010).

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