

Artistic Image Analysis using the Composition of Human Figures

Qian Chen and Gustavo Carneiro

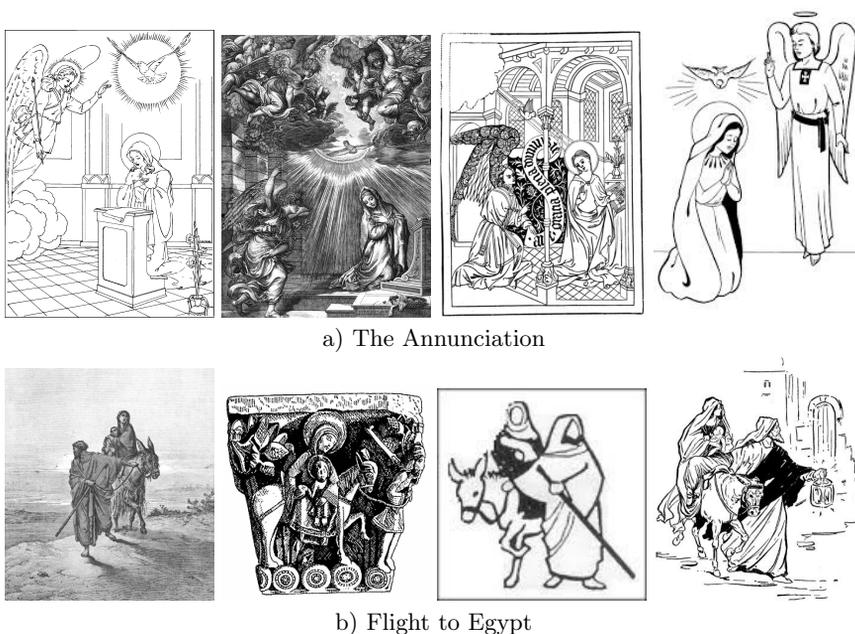
Australian Centre for Visual Technologies
The University of Adelaide

Abstract. Artistic image understanding is an interdisciplinary research field of increasing importance for the computer vision and art history communities. One of the goals of this field is the implementation of a system that can automatically retrieve and annotate artistic images. The best approach in the field explores the artistic influence among different artistic images using graph-based learning methodologies that take into consideration appearance and label similarities, but the current state-of-the-art results indicate that there seems to be lots of room for improvements in terms of retrieval and annotation accuracy. In order to improve those results, we introduce novel human figure composition features that can compute the similarity between artistic images based on the location and number (i.e., composition) of human figures. Our main motivation for developing such features lies in the importance that composition (particularly the composition of human figures) has in the analysis of artistic images when defining the visual classes present in those images. We show that the introduction of such features in the current dominant methodology of the field improves significantly the state-of-the-art retrieval and annotation accuracies on the PRINTART database, which is a public database exclusively composed of artistic images.

Keywords: Artistic Image Analysis, Image Feature, Image Annotation and Retrieval

1 Introduction

Artistic image understanding is a research area gaining increasing importance in the field of computer vision, as evidenced by the recent two editions of the workshop VISART, which have been held in conjunction with the European Conference on Computer Vision (ECCV) in 2012 and 2014, and also the two editions of the conference SPIE Computer Vision and Image Analysis of Art, held in 2010 and 2011. There are a large number of problems involved in artistic image understanding, but this paper is focused on the tasks of retrieving and annotating artistic images. These problems can be defined as follows [?]: 1) given an artistic keyword, retrieve un-annotated test images that are related to that keyword, and 2) given a test image, automatically produce relevant annotations represented by artistic keywords. The current dominant method in the field [?,?]



a) The Annunciation

b) Flight to Egypt

Fig. 1. Different artistic images depicting the theme "The Annunciation" (a) and "Flight to Egypt" (b). Notice how the composition of human figures are quite similar among images of the same theme, where the "Annunciation" images show Angel Gabriel slightly elevated and Virgin Mary kneeling down at a relatively lower height; while the "Flight to Egypt" images show St. Joseph pulling a donkey, on which the Virgin Mary and baby Jesus are sitting. It is also worth noticing the low appearance similarity among artistic images of the same theme (even though they have very similar annotations).

solves the two sub-problems above by exploring the artistic influence among different artistic prints with graph-based learning methodologies that take into consideration appearance and label similarities. However, the annotation and retrieval results reported show that there is quite a lot of room for improvement. For instance, the current retrieval and annotation results of this dominant method on the PRINTART database [?] show a mean average precision (MAP) of 0.18, an F1 score of 0.26 for class-based annotation, and an F1 score of 0.38 for example-based annotation.

In this paper we improve the retrieval and annotation accuracies of the methodology proposed by Carneiro et al. [?,?] with the use of a new type of feature that estimates the composition of human figures in the image. Our main motivation for exploring such feature lies in the fact that the composition of human figures of a scene depicted in an artistic image represents a powerful feature to be explored in its characterization. For instance, Figure ?? shows different prints of the same artistic themes ("The Annunciation" in (a) and "Flight

to Egypt” in (b)), and although the texture of the images are quite different from each other (which makes the use of appearance similarity not quite useful), the composition of human figures are strikingly similar. The features that we propose explore different ways of representing the geometric distribution of human figures in the image, and we incorporate such features in the methodology by Carneiro et al. [?,?] when computing the similarity between images. We show that this new feature improves the retrieval and annotation accuracies of that methodology [?,?] on the PRINTART database [?]. Specifically, we show that the retrieval accuracy is improved by 44%, achieving an MAP of 0.26. In terms of class-based annotation, the results are improved by 46% for the average F1 score, which reaches the value of 0.38. Finally, the example-based annotation is improved by 32% in terms of the average F1 score, reaching 0.50.

2 Literature Review

The majority of works being published in computer vision on the topic of artistic image analysis annotation and retrieval is focused on the artistic identification problem with the goal of determining if a test image is an original work of art by a famous painter or is a fake [?,?,?]. Other works have also handled the problem of identifying stylistic patterns in paintings [?,?,?]. The automatic classification of Chinese paintings [?] and brushwork [?] also represent important contributions in this field. Nevertheless, most of these works can be regarded as successful adaptations of the content-based image retrieval systems [?] to these somewhat limited artistic image analysis problems, but a more deep interpretation of an artistic image, with the goal of annotating the image with global and local keywords [?], is still far from being possible.

One of the main reasons hindering the development in this area lies in the use of the aforementioned adaptations of systems developed for analyzing photos, but it is important to note that the analysis of photos and artistic image are intrinsically different problems. For example, the photo of a face has arguably fewer degrees of freedom when compared to an artistic representation of a face (drawing, printing), as exemplified in Figure ???. Specifically, notice how the face proportions and geometry are different among the artistic faces. Another compelling example presented by Carneiro [?] and reproduced here in Figure ?? is the variation in the representation of the visual class "sea" among different paintings. Simply building a classifier that identifies if a pixel (or an image patch) belongs to the visual class "sea" (similarly to what is done in typical computer vision analysis of photos [?]) cannot work in these examples due to lack of a consistent representation of the visual class "sea" among different works of art.

The examples in Figure ??-?? indicate that the analysis of artistic images must follow a different development path compared to the usual photo analysis. In particular, we believe that the analysis of artistic images must explore the influence between artistic works, which suffer small variations over the course of time. This idea has been explored by Carneiro et al. [?,?], by Abe et al. [?] and by Graham et al. [?]. In the work by Carneiro et al. [?,?], they explore link analysis algorithms (a.k.a. graph-based learning methods) that compute the similarity

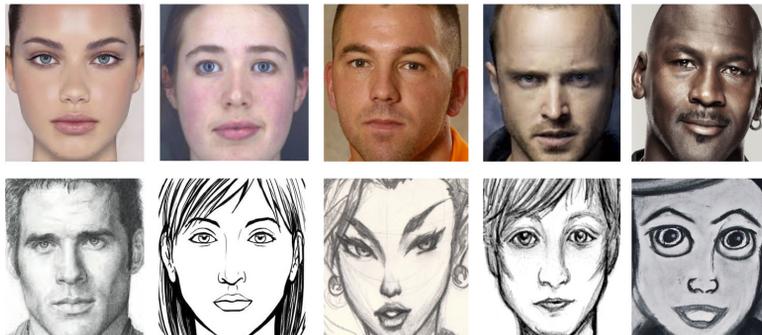


Fig. 2. Comparison between images of real faces (top row) and artistic representation of faces (bottom row). Notice that artistic faces have an arguably larger number of degrees of freedom in the sense that the face proportions and geometry can vary substantially among the artistic faces.



Fig. 3. Different paintings showing the visual class "sea" with different patterns of color and texture. In (a), we show Pieter Brueghel il Giovane's *Christ on the Storm of Galilee*; in (b) we have Claude Monet's *Shadows on the Sea*";(c) shows August Renoir's *Oarsmen Chatou*"; and (d) displays John Marin's *Sea Piece*. Figure from [?].

between images using global image representation (based on bag of features) and expert annotation. On the other hand, Abe et al. [?] explores high and low-level image features and Graham et al. [?] proposes the use of several image statistics. Nevertheless, we also believe that the analysis of artistic images must also explore the composition of human figures in the image, which is a feature that has not been explored in the field, to the best of our knowledge.

3 Methodology

We first explain the database PRINTART [?] and the original image representation available with the database, then we present the new human figure composition features and how they are incorporated into the methodology developed by Carneiro et al. [?,?]. It is important to note that the detection of faces from

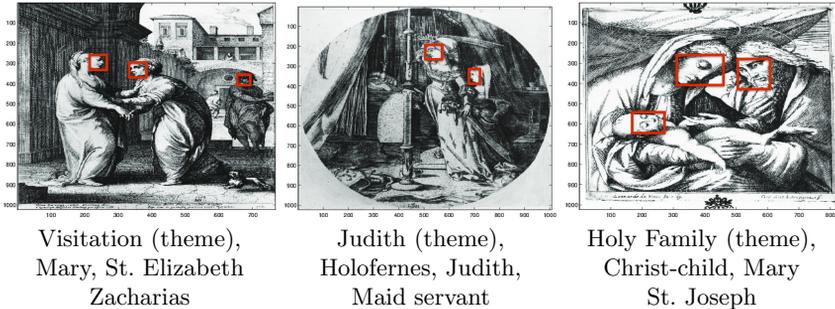


Fig. 4. Example of global (text below the image) and face (red bounding box) annotations of the PRINTART images (from Artstor [?]) produced by an art historian. Note that the classes identified with '(theme)' in brackets represent the multi-class problems defined in Sec. ??.

artistic images is a relatively easier problem compared to the detection of human bodies, so hereafter we use faces as a proxy for the presence and location of human figures in an image.

3.1 Database PRINTART

The PRINTART database [?] contains 988 artistic images with global and face annotations (see Fig. ??). These images have been acquired from the Artstor digital image library [?] and annotated by art historians with global annotation keywords, representing one multi-class problem (theme with 27 classes) and 48 binary problems, and face annotations, representing the main characters relevant for the theme of the image.

The training set is defined with $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i, \mathcal{P}_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathbb{R}^X$ represents the feature extracted from the image I_i , \mathbf{y}_i denotes the global annotation of that image representing M multi-class and binary problems, and \mathcal{P}_i represents the annotated faces in the image. Hence, $\mathbf{y}_i = [\mathbf{y}_i(1), \dots, \mathbf{y}_i(M)] \in \{0, 1\}^Y$, with each problem being denoted by $\mathbf{y}_i(k) \in \{0, 1\}^{|\mathbf{y}_i(k)|}$ and $|\mathbf{y}_i(k)|$ representing the dimensionality of $\mathbf{y}_i(k)$ (i.e., $|\mathbf{y}_i(k)| = 1$ for binary problems, $|\mathbf{y}_i(k)| > 1$ with $\|\mathbf{y}_i\|_1 = 1$ for multi-class problems, $Y = 75$, and $M = 49$ with one multi-class problem (with 27 classes) and 48 binary problems). In other words, the binary problems are about the detection of a visual class (i.e., their presence or absence), and the multi-class problem (representing the image theme) considers the identification of which class is relevant in the image (assuming that one of the classes must be present). Finally, the set $\mathcal{P}_i = \{\mathbf{b}_{i,j}\}_{j=1}^{|\mathcal{P}_i|}$ denotes the face annotation in image I_i , where $\mathbf{b}_{i,j} \in \mathbb{R}^4$ denotes the bounding box (with two 2-D image coordinates) of the face j and $|\mathcal{P}_i|$ denotes the number of faces in image I_i . The test set is represented by $\mathcal{T} = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i, \tilde{\mathcal{P}}_i)\}_{i=1}^{|\mathcal{T}|}$, with $\tilde{\mathbf{x}}_i$ denoting the features of test image \tilde{I}_i , $\tilde{\mathbf{y}}_i$ representing the global annotation of the test image (available in order to compute the retrieval and annotation accuracies), and $\tilde{\mathcal{P}}_i$ denoting

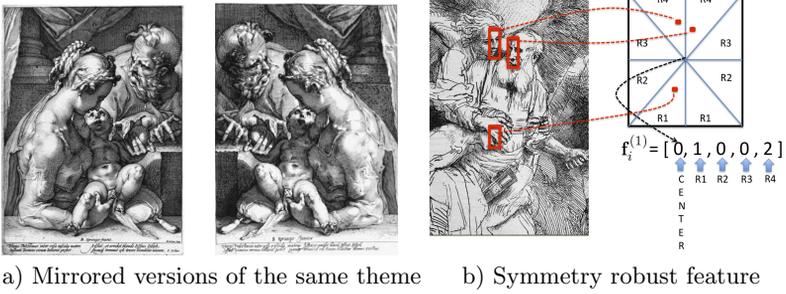
the estimated faces. It is also important to note that there is no overlap between the training and test sets, which means that $\mathcal{D} \cap \mathcal{T} = \emptyset$. The union of \mathcal{D} and \mathcal{T} produces the full PRINTART dataset with 988 images (i.e., $|\mathcal{D} \cup \mathcal{T}| = 988$). The label cardinality of the database, computed as $LC = \frac{1}{|\mathcal{D}|+|\mathcal{T}|} \sum_{i=1}^{|\mathcal{D}|+|\mathcal{T}|} \|\mathbf{y}_i\|_1$, is 4.22, while the label density $LD = \frac{1}{(|\mathcal{D}|+|\mathcal{T}|)^Y} \sum_{i=1}^{|\mathcal{D}|+|\mathcal{T}|} \|\mathbf{y}_i\|_1$, is 0.05.

The images in the PRINTART database use the standard bag of features (BoF) representation, which is publicly available from the PRINTART web site [?]. More specifically, 10000 SIFT descriptors [?] are extracted from a uniform grid over the image and scale spaces and a spatial pyramid [?] is used to form the image representation. The spatial pyramid representation is achieved by tiling the image in the following three levels, [?]: 1) the first level comprising the whole image, 2) the second level dividing the image into 2×2 regions, and 3) the third level breaking the image into 3×1 regions. Another important part of the BoF representation is the implementation of the visual vocabulary, which is built with the hierarchical clustering algorithm with three levels, where each node in the hierarchy has 10 descendants [?]. This means that the resulting directed tree has $1 + 10 + 100 + 1000 = 1111$ vertices, 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 vertices). Each SIFT descriptor from the image is then represented by four votes (weighted by the node entropy) in a histogram of the visited vertices (containing 1111 bins). Using this hierarchical tree (with a total of 1111 vertices) and the tiling described above (with 8 tiles), an image is represented with 8 histograms as in $\mathbf{x} \in \mathbb{R}^X$, where $X = 8 \times 1111$.

3.2 Human Figure Composition Features

In this section, we propose three different types of features that represent the human figure composition in an image. Note that to build these features, we assume that the faces detected from image I_i are represented with $\mathcal{P}_i = \{\mathbf{b}_{i,j}\}_{j=1}^{|\mathcal{P}_i|}$, as defined above.

Symmetry Robust Feature It is important to notice that images annotated with the same theme (i.e., the multi-class problem in Sec. ??) present similar distribution of faces in the image (see Fig. ??), but it is also common to see a mirrored representation of the faces, such as the ones in Figure ??-(a). This is the motivation for the design of this feature, which is based on first, splitting the image into two halves (using the vertical line that divides the image into two halves) then sub-dividing each half into four regions, as shown in Fig. ??-(b). Finally, the feature is based on a histogram of five bins that represent the number of faces lying in each region and at the center of the image, but in order to make the feature robust to symmetric transformations, the symmetrically equivalent regions in both (vertical) halves have the same label, and are consequently represented by the same bin in the histogram (see vector $\mathbf{f}_i^{(1)}$ in Fig. ??-(b) that takes 0 faces from the centre, 1 face from R1 and 2 faces from R4). Assume that



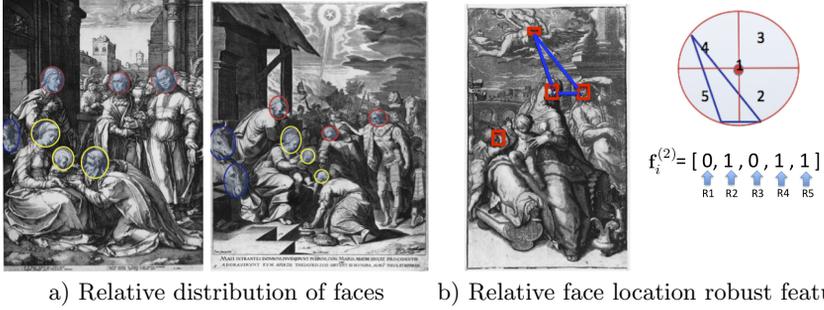
a) Mirrored versions of the same theme b) Symmetry robust feature

Fig. 5. Mirrored distribution of human figures on images from the same theme (in this case, the theme is 'Holy Family') in (a) and the depiction of the symmetry robust feature representation (b).

the histogram from this feature is represented by the vector $\mathbf{f}_i^{(1)} \in \mathbb{R}^5$, extracted with $\mathbf{f}_i^{(1)} = \phi_1(I_i, \mathcal{P}_i)$.

Relative Face Location Robust Feature The number and relative location of faces with respect to one another is also a powerful representation of the composition of human figures in an image. For instance, consider the two images in Fig. ??, which shows two images from theme 'Magi'. Notice that even though the absolute location of faces are different in the two images, the number of human faces are the same and their relative locations are quite similar. Specifically in Fig. ??-(a), notice that there are always three faces at a relatively higher position, almost lying in a horizontal line (see the red-contour faces), and three faces at a lower position (see the yellow-contour faces). We propose the following algorithm to build a representation for this relative face location robust feature (Fig. ??-(b)): 1) select three faces in the image and draw a circle that pass through the center of the bounding box of all faces; 2) divide the circle into five regions and count the number of faces lying in each region (not only the three faces used to draw the circle); 3) repeat steps (1) and (2) for all sets of three faces and accumulate all results. In case only two faces are available, just draw a circle that passes through the center of the two faces, divide it in five regions and record the bins where the two faces reside (similarly to what is done for three faces - Fig. ??-(b)). Finally, for the case when only one face is available just build a vector with bin 1 equal to 1. Assume that the histogram from this feature is represented by the vector $\mathbf{f}_i^{(2)} \in \mathbb{R}^5$, extracted with $\mathbf{f}_i^{(2)} = \phi_2(I_i, \mathcal{P}_i)$.

Rule of Thirds Feature The rule of thirds is a well-used composition technique that can also be applied to estimate the distribution of human figures in an artistic image. The technique is essentially about dividing the image into a 3×3 tiling and placing the important objects of the scene along these divisions. Our proposed representation consists of first dividing the image into a 3×3 tiling



a) Relative distribution of faces b) Relative face location robust feature

Fig. 6. Relative distribution of faces in images of the same theme ('Magi') in (a) and an example of one step of the algorithm to build the relative face location robust feature (b). Note that in this image, our proposed algorithm will run for four steps to get all sets of three faces in the image (i.e., 4 choose 3).

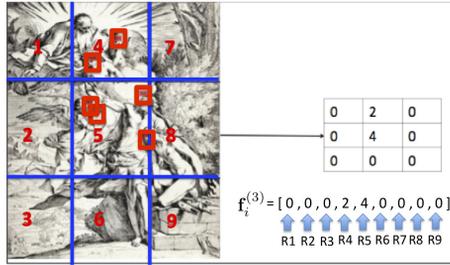


Fig. 7. Rule of thirds applied to the distribution of faces.

and then counting the number of faces lying in each one of these regions, as shown in Figure ?? . Similarly to the features presented above, this feature is also represented by a histogram, where each bin contains the number of faces in each bin. Assume that the histogram from this feature is represented by the vector $\mathbf{f}_i^{(3)} \in \mathbb{R}^9$, extracted with $\mathbf{f}_i^{(3)} = \phi_3(I_i, \mathcal{P}_i)$.

3.3 Similarity Between Human Figure Distribution Features

The similarity between images I_i and I_k (with detected faces represented by the sets \mathcal{P}_i and \mathcal{P}_k , respectively) with respect to the three human figure composition features defined above in Sec. ?? is computed with the following equation:

$$s_p(i, k) = \exp \left\{ -\frac{1}{\lambda} \left(\alpha \|\mathbf{f}_i^{(1)} - \mathbf{f}_k^{(1)}\|_2^2 + \beta \|\mathbf{f}_i^{(2)} - \mathbf{f}_k^{(2)}\|_2^2 + \gamma \|\mathbf{f}_i^{(3)} - \mathbf{f}_k^{(3)}\|_2^2 \right) \right\}, \quad (1)$$

where α, β, γ are the weights controlling the importance of each one of the features, λ is a weight controlling the slope of the function $s(\cdot, \cdot)$, and $\mathbf{f}_i^{(1)}, \mathbf{f}_i^{(2)}, \mathbf{f}_i^{(3)}$

are defined in Section ???. Also, assume that the similarity between a training image I_i , with face bounding boxes represented by \mathcal{P}_i , and a test image \tilde{I} , with face bounding boxes $\tilde{\mathcal{P}}$, is denoted by $s_p(i, \sim)$.

3.4 Incorporation of Human Figure Distribution Similarity

Following the notation by Carneiro [?], the annotation of a test image \tilde{I} , represented by $(\tilde{\mathbf{x}}, \tilde{\mathcal{P}})$ from the test set \mathcal{T} is achieved by finding $\tilde{\mathbf{y}}^*$ that solves the following optimization problem:

$$\begin{aligned} & \text{maximize } p(\mathbf{y}|\tilde{\mathbf{x}}, \tilde{\mathcal{P}}) \\ & \text{subject to } \mathbf{y} = [\mathbf{y}(1), \dots, \mathbf{y}(M)] \in \{0, 1\}^Y, \\ & \quad \|\mathbf{y}(k)\|_1 = 1 \text{ for } \{k \in \{1, \dots, M\} | \mathbf{y}(k) > 1\}, \end{aligned} \quad (2)$$

where $p(\mathbf{y}|\tilde{\mathbf{x}}, \tilde{\mathcal{P}})$ is a probability function that computes the confidence of annotating the test image with a vector $\mathbf{y} \in \mathcal{Y}$ (with \mathcal{Y} denoting the set of all possible annotations \mathbf{y} in the training set). The retrieval problem is solved by building a set of test images that are relevant to a query $\mathbf{q} \in \{0, 1\}^Y$, as follows:

$$\mathcal{Q}(\mathbf{q}, \tau) = \{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}^*, \tilde{\mathcal{P}}) | (\tilde{\mathbf{x}}, \tilde{\mathcal{P}}) \in \mathcal{T}, (\mathbf{q}^\top \tilde{\mathbf{y}}^*) > 0, p(\tilde{\mathbf{y}}^*|\tilde{\mathbf{x}}, \tilde{\mathcal{P}}) > \tau\}, \quad (3)$$

where $\tilde{\mathbf{y}}^*$ is obtained from (??), $\tau \in [0, 1]$ is a threshold, and \mathcal{T} is the set of test images defined in Sec. ???.

We incorporate the similarity between human figure composition features (??) into the inverse label propagation framework [?], which basically takes a test image represented by $(\tilde{\mathbf{x}}, \tilde{\mathcal{P}})$ and ranks the most relevant training images $(\mathbf{x}, \mathbf{y}, \mathcal{P}) \in \mathcal{D}$ via a random walk process, which uses a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ built with the training set \mathcal{D} , where the nodes \mathcal{V} represent the images and the weights of each edge in \mathcal{E} are computed based on the appearance, label and human figure composition similarities between training images. Then given a test image in \mathcal{T} , represented by $(\tilde{\mathbf{x}}, \tilde{\mathcal{P}})$, we start a random walk process in this graph by taking into account the appearance and human figure composition similarities between the test image and training images. Carneiro [?] proposes three different ways to solve this process, and we follow the *combinatorial harmonics*, which shows the best performance among the proposed methods.

The random walk process based on the combinatory harmonics estimates the probability of first reaching each of the database samples $(\mathbf{x}_i, \mathbf{y}_i, \mathcal{P}_i) \in \mathcal{D}$ starting from the test image $(\tilde{\mathbf{x}}, \tilde{\mathcal{P}})$ [?], using the following adjacency matrix: $\tilde{\mathbf{W}} = \begin{bmatrix} \mathbf{W} & \tilde{\mathbf{w}} \\ \tilde{\mathbf{w}}^\top & 0 \end{bmatrix}$, where $\tilde{\mathbf{w}} = [s_x(\mathbf{x}_1, \tilde{\mathbf{x}})s_p(1, \sim), \dots, s_x(\mathbf{x}_{|\mathcal{D}|}, \tilde{\mathbf{x}})s_p(|\mathcal{D}|, \sim)]^\top$ (note that $s_p(1, \sim)$ is defined in (??)), and

$$\mathbf{W}(i, j) = s_y(\mathbf{y}_i, \mathbf{y}_j)s_x(\mathbf{x}_i, \mathbf{x}_j)s_p(i, j)s_x(\mathbf{x}_i, \tilde{\mathbf{x}})s_p(i, \sim). \quad (4)$$

where the label similarity function is the Jaccard index defined by $s_y(\mathbf{y}_i, \mathbf{y}_j) = \frac{\mathbf{y}_i^\top \mathbf{y}_j}{\|\mathbf{y}_i\|^2 + \|\mathbf{y}_j\|^2 - \mathbf{y}_i^\top \mathbf{y}_j}$, and the feature similarity function is the histogram intersection defined as $s_x(\mathbf{x}_i, \mathbf{x}_j) = \sum_{d=1}^X \min(\mathbf{x}_i(d), \mathbf{x}_j(d))$ (i.e., this is the histogram

intersection kernel over the spatial pyramid, where $\|\mathbf{x}\|_1 = 1$). The goal is then to find the distribution $\mathbf{g}^* \in \mathbb{R}^{|\mathcal{D}|}$ ($\|\mathbf{g}^*\|_1 = 1$), representing the probability of first reaching each of the training images in a random walk procedure, where the labeling matrix $\mathbf{G} = \mathbf{I}$ (i.e., an $|\mathcal{D}| \times |\mathcal{D}|$ identity matrix) denotes a problem with $|\mathcal{D}|$ classes, with each training image representing a separate class. The estimation of \mathbf{g}^* is based on the minimization of the following function:

$$E([\mathbf{G}, \mathbf{g}]) = \frac{1}{2} \left\| [\mathbf{G}, \mathbf{g}] \tilde{\mathbf{L}} \begin{bmatrix} \mathbf{G}^T \\ \mathbf{g}^T \end{bmatrix} \right\|_2^2, \quad (5)$$

where $\tilde{\mathbf{L}} = \tilde{\mathbf{D}} - \tilde{\mathbf{W}}$ is the Laplacian matrix computed from the the adjacency matrix $\tilde{\mathbf{W}}$, where $\tilde{\mathbf{D}}$ is a matrix that has the sum of the rows in the diagonal (i.e., it is a diagonal matrix). This Laplacian matrix can be divided into blocks of the same sizes as in $\tilde{\mathbf{W}}$, that is $\tilde{\mathbf{L}} = \begin{bmatrix} \mathbf{L}_1 & \mathbf{B} \\ \mathbf{B}^T & \mathbf{L}_2 \end{bmatrix}$. Solving the following optimization problem produces \mathbf{g}^* [?]:

$$\begin{aligned} & \text{minimize } E([\mathbf{G}, \mathbf{g}]) \\ & \text{subject to } \mathbf{G} = \mathbf{I}, \end{aligned} \quad (6)$$

which has the closed form solution [?]: $\mathbf{g}^* = (-\mathbf{L}_2^{-1} \mathbf{B}^T \mathbf{I})^\top$. Note that $\mathbf{g}^* \in [0, 1]^{|\mathcal{D}|}$ and $\|\mathbf{g}^*\|_1 = 1$.

The probability of annotation \mathbf{y} is then computed from the test image with:

$$p(\mathbf{y} | \tilde{\mathbf{x}}, \tilde{\mathcal{P}}) = Z \sum_{i=1}^{|\mathcal{D}|} \mathbf{g}^*(i) p(\mathbf{y} | (\mathbf{x}_i, \mathcal{P}_i)) \quad (7)$$

where $\mathbf{g}^*(i)$ is the i^{th} component of the solution vector from (??), $p(\mathbf{y} | (\mathbf{x}_i, \mathcal{P}_i)) = \frac{\delta(\|\mathbf{y} - \mathbf{y}_i\|_1)}{\sum_{j=1}^{|\mathcal{D}|} \delta(\|\mathbf{y} - \mathbf{y}_j\|_1)}$ [?] ($\delta(\cdot)$ is the delta function), and Z is a normalization factor. In the experiments presented below in Sec. ??, this method is called 'FACGT' in case we use the manual face annotations and 'FAC++' or 'FACVJ' if we use the automated face detection Face++ [?] or Viola and Jones [?], respectively.

4 Experiments

Using the PRINTART database [?] and the setup described below, we measure the performance of the annotation methodologies with one type of retrieval and two types of annotation evaluations. The *retrieval* evaluation is based on the following precision and recall measures [?] computed with the first $Q \leq |\mathcal{T}|$ images retrieved with respect to query \mathbf{q} (sorted by $p(\tilde{\mathbf{y}}^* | \tilde{\mathbf{x}}, \tilde{\mathcal{P}})$ in (??) in descending order, where $\mathbf{q}^\top \tilde{\mathbf{y}}^* > 0$):

$$p_r(\mathbf{q}, Q) = \frac{\sum_{i=1}^Q \delta(\tilde{\mathbf{y}}_i^\top \mathbf{q} - 1^\top \mathbf{q})}{Q}, \text{ and } r_r(\mathbf{q}, Q) = \frac{\sum_{i=1}^Q \delta(\tilde{\mathbf{y}}_i^\top \mathbf{q} - 1^\top \mathbf{q})}{\sum_{j=1}^{|\mathcal{T}|} \delta(\tilde{\mathbf{y}}_j^\top \mathbf{q} - 1^\top \mathbf{q})}, \quad (8)$$

where $\delta(\cdot)$ is the delta function. The retrieval task is assessed with the mean average precision (MAP), which is defined as the average precision over all queries, at the ranks where recall changes.

The annotation of a test image, represented by $(\tilde{\mathbf{x}}, \tilde{\mathcal{P}})$, is computed by finding $\tilde{\mathbf{y}}^*$ that solves the optimization problem (??). The first type of annotation evaluation, called *label-based* annotation [?], is computed for each class $y \in \{1, \dots, Y\}$ with the respective precision, recall and F1 measures:

$$p_c(y) = \frac{\sum_{i=1}^{|\mathcal{T}|} (\pi_y \odot \tilde{\mathbf{y}}_i^*)^\top \tilde{\mathbf{y}}_i}{\sum_{i=1}^{|\mathcal{T}|} \pi_y^\top \tilde{\mathbf{y}}_i^*}, r_c(y) = \frac{\sum_{i=1}^{|\mathcal{T}|} (\pi_y \odot \tilde{\mathbf{y}}_i^*)^\top \tilde{\mathbf{y}}_i}{\sum_{i=1}^{|\mathcal{T}|} \pi_y^\top \tilde{\mathbf{y}}_i}, f_c(y) = \frac{2p_ga(y)r_ga(y)}{p_ga(y)+r_ga(y)}, \quad (9)$$

where $\pi_y \in \{0, 1\}^Y$ is one at the y^{th} position and zero elsewhere, and \odot represents the element-wise multiplication operator. The values of $p_ga(y)$, $r_ga(y)$ and $f_ga(y)$ are then averaged over the visual classes. The measure in (??) is called *label-based* because the result is assessed class by class, independently. On the other hand, the *example-based* annotation computes an image-based performance considering all labels jointly (i.e., a multi-label evaluation). Example-based annotation evaluation is computed for each test image and then averaged over the test set with respect to precision, recall, F1 and accuracy measures, which are respectively defined by [?]:

$$p_e = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \frac{(\tilde{\mathbf{y}}_i^*)^\top \tilde{\mathbf{y}}_i}{\|\tilde{\mathbf{y}}_i^*\|_1}, \quad r_e = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \frac{(\tilde{\mathbf{y}}_i^*)^\top \tilde{\mathbf{y}}_i}{\|\tilde{\mathbf{y}}_i\|_1}, \quad (10)$$

$$f_e = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \frac{2(\tilde{\mathbf{y}}_i^*)^\top \tilde{\mathbf{y}}_i}{\|\tilde{\mathbf{y}}_i^*\|_1 + \|\tilde{\mathbf{y}}_i\|_1}, \quad a_e = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \frac{(\tilde{\mathbf{y}}_i^*)^\top \tilde{\mathbf{y}}_i}{\|\min(1, \tilde{\mathbf{y}}_i^* + \tilde{\mathbf{y}}_i)\|_1}.$$

We follow the same experimental setup proposed by Carneiro [?], with a 10-fold cross validation, where the PRINTART database is divided into a training set \mathcal{D} with 90% of the database (i.e., $|\mathcal{D}| = 889$), and a test set \mathcal{T} with the remaining 10% (i.e., $|\mathcal{T}| = 99$). We compute the retrieval and annotation evaluation measures (??)-(??) and display the results using the average and standard deviation in this 10-fold cross validation experiment. We use the same acronyms for the methodologies defined in [?], as follows: inverse label propagation (ILP), combinatorial harmonics (CH), stationary solution (SS), random walk (RW), bag of features (BoF), label propagation (LP), class label correlation (CC), matrix completion (MC), structural learning (SL), random (RND), and nearest neighbor (NN). In addition, the methods proposed in this paper are labeled 'FACGT' for the ones that uses the manual face annotations, and 'FAC++' or 'FACVJ' for the ones using automated face detection. Note that for the FAC++, we use the face detection 'Face++' [?] (available from <http://www.faceplusplus.com/>), which has produced state of the art face detection results in recent challenges. Essentially, Face++ is based on a multi-scale convolutional neural network methodology, and on the PRINTART database it produces an average precision of 0.75 and recall 0.11 for the face detection problem (i.e., quite high precision, but relatively low recall - see Fig. ??-(a) for an example produced by the Face++ detector). For 'FACVJ', we use the classic Viola and Jones face detector [?] (available from OpenCV - <http://opencv.org/>), which has a relatively poor performance on PRINTART with a precision of 0.17 and recall also of 0.17 for the

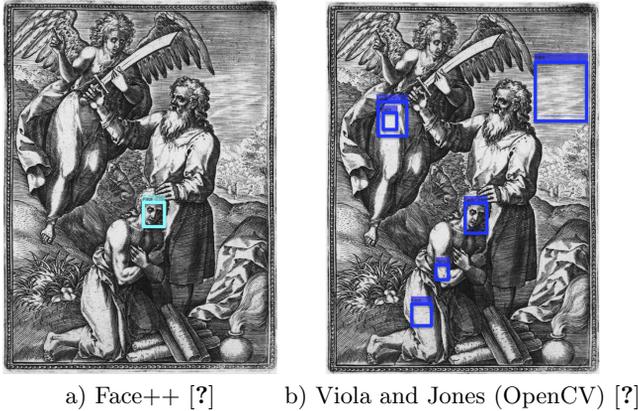


Fig. 8. Example of face detection results from Face++ [?] (a) and Viola and Jones (OpenCV) [?] (b).

face detection problem (see Fig. ??-(b) for an example produced by the Viola and Jones detector).

Finally, the values of α, β, γ in (??) are estimated via the 10-fold cross validation experiment explained above, with $\alpha, \beta, \gamma \in [0, 1]$. Note that different values of α, β, γ are estimated for each one of the proposed methods FACGT, FAC++ and FACVJ. The value of λ in (??) is estimated as follows: $\lambda = \kappa(\alpha\sqrt{5} + \beta\sqrt{5} + \gamma\sqrt{9})$, where $\kappa \in [0, 1]$ is also estimated via cross validation and the constants $\sqrt{5}$ and $\sqrt{9}$ are related to the number of dimensions of the respective human figure composition features, defined in Sec. ??.

4.1 Results

Table ?? shows the retrieval and annotation results (??)-(??) for all methodologies from [?] and the methodologies FACGT, FAC++ and FACVJ proposed in this paper. Note that the methodology ILP-CH currently holds the state-of-the-art results for the PRINTART database. Figure ?? compares the annotation results between the proposed FACGT and ILP-CH, for cases where FACGT improves the results produced by the ILP-CH.

5 Discussion

According to the results in Sec. ??, we can conclude that for the database PRINTART used in this paper, the method that uses the manually annotated faces, FACGT, improves substantially the retrieval and annotation results produced by the previously best method ILP-CH. Thus we can conclude that there is enough evidence to accept the main hypothesis being tested by this paper, which is that the use of human figure composition improves the classification

Table 1. Retrieval, label-based and example-based results using the mean and standard deviation of the measures described in Sec. ???. The highlighted value in each column indicates the highest for each measure.

	Retrieval	Label-based annotation			Example-based annotation			
Models	Label MAP	Average Precision	Average Recall	Average F1	Average Precision	Average Recall	Average F1	Average Accuracy
FACGT	0.26 ± .03	0.39 ± .05	0.37 ± .06	0.38 ± .05	0.51 ± .04	0.50 ± .05	0.50 ± .04	0.45 ± .04
FAC++	0.18 ± .03	0.25 ± .05	0.23 ± .03	0.24 ± .04	0.37 ± .03	0.37 ± .04	0.36 ± .03	0.31 ± .03
FACVJ	0.13 ± .02	0.16 ± .03	0.16 ± .04	0.16 ± .03	0.31 ± .03	0.31 ± .03	0.30 ± .03	0.25 ± .03
ILP-CH	0.18 ± .04	0.26 ± .05	0.26 ± .05	0.26 ± .05	0.39 ± .03	0.39 ± .04	0.38 ± .03	0.33 ± .03
ILP-SS	0.12 ± .01	0.15 ± .02	0.16 ± .05	0.15 ± .04	0.24 ± .04	0.24 ± .04	0.23 ± .04	0.20 ± .04
ILP-RW	0.10 ± .01	0.10 ± .03	0.13 ± .02	0.11 ± .03	0.33 ± .03	0.36 ± .03	0.34 ± .03	0.26 ± .03
BoF	0.12 ± .05	0.14 ± .11	0.10 ± .06	0.11 ± .08	0.47 ± .05	0.26 ± .08	0.30 ± .05	0.23 ± .05
LP	0.11 ± .01	0.12 ± .02	0.12 ± .02	0.12 ± .02	0.32 ± .03	0.28 ± .02	0.26 ± .02	0.19 ± .01
LP-CC	0.11 ± .01	0.13 ± .02	0.14 ± .02	0.13 ± .02	0.27 ± .03	0.26 ± .03	0.25 ± .03	0.18 ± .02
MC	0.17 ± .01	0.24 ± .03	0.11 ± .02	0.15 ± .02	0.47 ± .03	0.28 ± .02	0.32 ± .02	0.25 ± .02
SL	0.14 ± .01	0.20 ± .04	0.15 ± .03	0.17 ± .03	0.37 ± .04	0.32 ± .03	0.34 ± .03	0.28 ± .03
RND	0.08 ± .06	0.06 ± .01	0.07 ± .01	0.06 ± .01	0.26 ± .02	0.21 ± .01	0.22 ± .01	0.15 ± .01
NN	0.13 ± .01	0.17 ± .02	0.17 ± .04	0.17 ± .03	0.32 ± .04	0.32 ± .03	0.31 ± .03	0.26 ± .03

of artistic images. Nevertheless, when using actual face detection systems, the results are either comparable to ILP-CH (see the results for FAC++) or worse (see FAVJ). The main issue with these detectors lies in its quite low recall (below 0.2 for both detectors) and low precision (for the case of FACVJ). It is worth mentioning that we have tried most of the state-of-the-art face detectors that are publicly available, and these two are the ones with the best results on the PRINTART database. That relatively low performance can be explained by the fact that faces in artistic images are less constrained mainly in terms of geometric distribution of facial features (see Fig. ??), and consequently the generalization of current face detectors to this new domain is not straightforward.

6 Conclusion

In this paper we show that by exploring the composition of human figures in artistic prints, we can produce the currently best results on the PRINTART database. This shows empirically that this composition is in fact decisive for estimating the visual classes present in artistic images. Furthermore, we show that the current best face detectors in the field are not reliable enough to be used in this new domain of artistic image analysis. Therefore, there is plenty of room for improvement in this area of face detection from artistic images, where faces like the ones displayed in Fig. ?? can be reliably detected. We plan to adapt current face detectors by re-training them using artistic images. Finally, given the positive results shown by the proposed composition of human figures, we also plan to explore other types of composition techniques, such as the rule of odds, composition of the eye gazing of human figures, and depth and pictorial cues (depth, illumination, depth, etc.) [?].



GT: magi
christ-child
mary
melchior
st.joseph
wise men

ILP-CH: calvary
christ
cross

FACGT: magi
christ-child
mary
melchior
st.joseph
wise men



GT: holy family
anthony abbot
christ-child
mary
st. john baptist
st. joseph

ILP-CH: christ carrying cross

FACGT: holy family
anthony abbot
christ-child
mary
st. john baptist
st. joseph



GT: tobit

ILP-CH: flight egypt
christ-child
donkey
mary
st. joseph

FACGT: tobit



GT: holy family
christ-child
mary
st. john baptist
st. joseph



GT: washing of the feet

ILP-CH: the ascension
apostle
christ

FACGT: washing of the feet



GT: st. anthony of padua

ILP-CH: circumcision of christ
christ-child
mary
st. joseph
priest

FACGT: holy family
christ-child
mary
st. john baptist
st. joseph

Fig. 9. Sample results on PRINTART showing the annotations from the ground truth (GT), the previously best method ILP-CH [?], and our proposed approach FACGT.

References

1. Carneiro, G., da Silva, N.P., Bue, A.D., Costeira, J.P.: Artistic image classification: An analysis on the printart database. In: ECCV (4). (2012) 143–157
2. Carneiro, G.: Artistic image analysis using graph-based learning approaches. *Image Processing, IEEE Transactions on* **22**(8) (2013) 3168–3178
3. Berezhnoy, I., Postma, E., van den Herik, H.: Computerized visual analysis of paintings. In: *Int. Conf. Association for History and Computing*. (2005) 28–32
4. Lyu, S., Rockmore, D., Farid, H.: A digital technique for art authentication. *Proceedings of the National Academy of Sciences USA* **101**(49) (2004) 17006–17010
5. Polatkan, G., Jafarpour, S., Brasoveanu, A., Hughes, S., Daubechies, I.: Detection of forgery in paintings using supervised learning. In: *International Conference on Image Processing*. (2009)
6. Graham, D., Friedenber, J., Rockmore, D., Field, D.: Mapping the similarity space of paintings: image statistics and visual perception. *Visual Cognition* **18**(4) (2010) 559–573
7. Hughes, J., Graham, D., Rockmore, D.: Stylometrics of artwork: uses and limitations. In: *Proceedings of SPIE: Computer Vision and Image Analysis of Art*. (2010)
8. Jafarpour, S., Polatkan, G., Daubechies, I., Hughes, S., Brasoveanu, A.: Stylistic analysis of paintings using wavelets and machine learning. In: *European Signal Processing Conference*. (2009)
9. Li, J., Wang, J.: Studying digital imagery of ancient paintings by mixtures of stochastic models. *IEEE Trans. Image Processing* **13**(3) (2004) 340–353
10. Yelizaveta, M., Tat-Seng, C., , Jain, R.: Semi-supervised annotation of brushwork in paintings domain using serial combinations of multiple experts. In: *ACM Multimedia*. (2006) 529–538
11. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* **40**(2) (2008)
12. Abe, K., Saleh, B., Elgammal, A.: An early framework for determining artistic influence. In: *New Trends in Image Analysis and Processing–ICIAP 2013*. Springer (2013) 198–207
13. <http://www.artstor.org>
14. Lowe, D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* (2004) paper accepted for publication.
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features, spatial pyramid matching for recognizing natural scene categories. In: *CVPR*. (2006)
16. <http://pascallin.ecs.soton.ac.uk/challenges/VOC/>
17. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: *CVPR*. (2006) 2161–2168
18. Grady, L.: Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(11) (2006) 1768–1783
19. Fan, H., Cao, Z., Jiang, Y., Yin, Q., Doudou, C.: Learning deep face representation. *arXiv preprint arXiv:1403.2802* (2014)
20. Viola, P., Jones, M.J.: Robust real-time face detection. *International journal of computer vision* **57**(2) (2004) 137–154
21. Nowak, S., Lukashevich, H., Dunker, P., Rügner, S.: Performance measures for multilabel evaluation: a case study in the area of image classification. In: *Multimedia Information Retrieval*. (2010) 35–44
22. Arnheim, R.: *Art and visual perception: A psychology of the creative eye*. Univ of California Press (1954)