Explaining scene composition using kinematic chains of humans: application to Portuguese tiles history*

Nuno Pinho da Silva and Manuel Marques and Gustavo Carneiro†and João P. Costeira

ISR – Instituto Superior Técnico, Univ. Técnica de Lisboa
Av. Rovisco Pais 1, Torre Norte – 7^o piso, Lisboa, Portugal;

ABSTRACT

Painted tile panels (Azulejos) are one of the most representative Portuguese forms of art. Most of these panels are inspired on, and sometimes are literal copies of, famous paintings, or prints of those paintings. In order to study the Azulejos, art historians need to trace these roots. To do that they manually search art image databases, looking for images similar to the representation on the tile panel. This is an overwhelming task that should be automated as much as possible. Among several cues, the pose of humans and the general composition of people in a scene is quite discriminative. We build an image descriptor, combining the kinematic chain of each character, and contextual information about their composition, in the scene. Given a query image, our system computes its similarity profile over the database. Using nearest neighbors in the space of the descriptors, the proposed system retrieves the prints that most likely inspired the tiles' work.

Keywords: Painted Tile Panels, Art Image Retrieval, Kinematic Chain, Nearest Neighbors

1. INTRODUCTION

Painted tile panels, also called "Azulejos", are one of the most representative Portuguese forms of art. They can be found throughout the country, in public buildings and fountains, churches and palaces,[1,2] and there is a

*This work was supported by the FCT (IS/IST plurianual funding) through the PIDDAC Program funds and Project PRINTART (PTDC/EEA-CRO/098822/2008

†This work is also founded by the European Commission; Marie Courier IIF, contract number PIIF-GA-2009-236173 (IMAGE-SEG3D)

E-mail: { nmps, manuel, gcarneiro, jpc }@isr.ist.utl.pt

Figure 1. Azulejos. The image on the left is a painted tile panel, also called "Azulejos", one of the most representative Portuguese art forms. Typically, the tile panel's artists were inspired by famous paintings (or prints of those paintings). The author of a given panel may have been inspired by more than one painting (the images on the right). We rely on the art historian for choosing portions of the panel that are likely to match with a single image, thus reducing the problem to image retrieval.

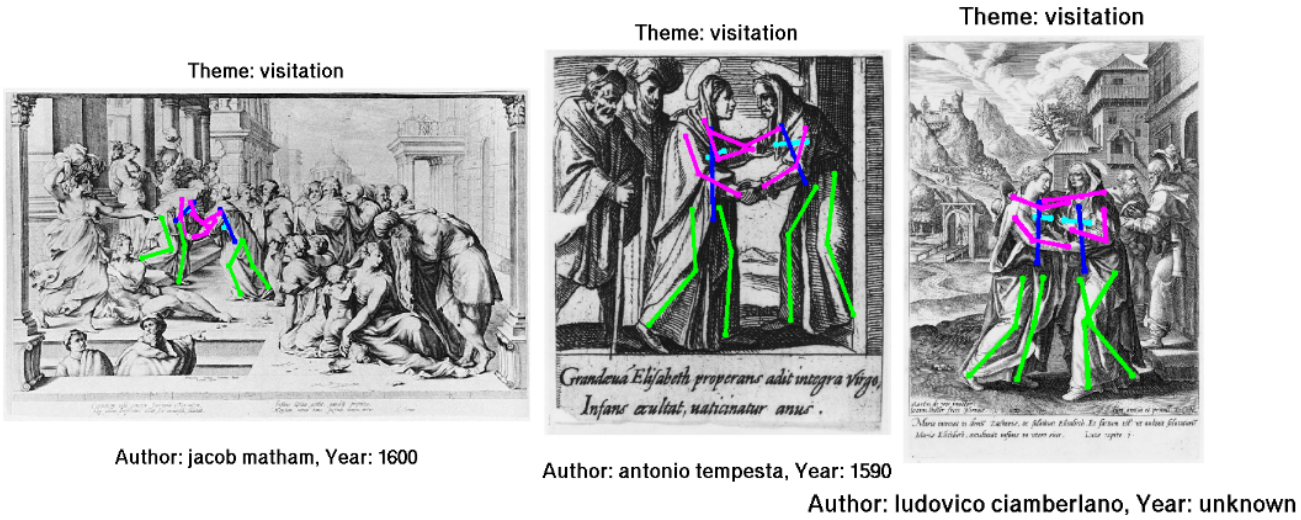

Figure 2. We propose exploiting the kinematic chain and composition description of the main characters in a scene, thus filtering the stylistic framing surrounding the central motive (theme). This is useful for searching large image databases, thus helping art historians tracing the roots of the images and explaining the scenes. In this example, we were able to match three globally different images (the query is on the left, followed by its first and second nearest neighbors, respectively).

national museum dedicated to it (Museu Nacional do Azulejo[3]).

The artists of the Azulejos were inspired by, and sometimes copied, paintings of famous authors, or prints of those paintings. In order to study the influences of these paintings and prints on the panel's artist, art historians invert the process: they manually search databases of prints, retrieving images similar to the representations painted on the panel, for tracing its roots.

Since the author of a painted tile panel may have been inspired by more than one painting, we rely on the criterion of the art historian studying the panel for giving an image of a patch that is likely to be explained by a single print (Fig. 1). Hence, explaining tile panels from images boils down to the problem of retrieving similar images from a given database.

In the past few years, computer vision and machine learning researchers gained interest in the retrieval and annotation of art images.[4,5] The bulk of this work concerns with studying the techniques employed by the artistic, thus enlightening the process of creation[6] or modeling the artists' distinctive features for authentication.[7,8]

In this work, we propose using the kinematic chain (pose) of the subjects, and their relationships (composition) in the scene, for retrieving similar images from a database. As shown in Fig. 2, our approach focus on the central theme of the scene, allowing to match images that are otherwise completely different.

More specifically, we combine the kinematic chain of each character, and contextual information about their composition, in the scene, for describing an image. Given a query image, we compare its descriptor with the image descriptors in the database. Each image in the database is ranked between zero and one, according to its similarity with respect to the query image. In other words, given a query image, we find its similarity profile over the database.

The main purpose of our system is helping art historians parsing image databases. Its output may also be used in conjunction with mechanisms leveraging semantic information for image retrieval[9] or image annotation.[5] Here, we apply it for retrieving images from the same semantic class (the theme represented in the images), using nearest neighbors[10] in the space of the descriptors. In particular, we use it for tracing the roots of a particular image (Sec. 4.2).

Figure 3. Mathching Images with Azulejos using Pose. The print and the right panel suggest a female *personage* at the bottom of the figure, while the left panel presents it as male character. Also, the characters' facial expression and hair style is different, but their individual poses and their relationship in the scene is the same. This suggest the usage of these salient features for image retrieval from a database of art images.

2. DESCRIBING IMAGES BY POSE AND COMPOSITION

Among several other cues for art image retrieval,[4,5] the pose and the spatial configuration of the subjects in the scene (or image) are discriminative. Consider Fig. 3: note the female breasts of the personage at the bottom of the original print (middle image) and the right panel, while the left panel presents it as a male subject. Also, the subjects' facial expression and hair style is different, but their individual poses and their spatial relationship in the scene is the same. To the spatial relationship between subjects we called (scene) composition. Hence, we search for images containing similar individual kinematic chains and similar compositions.

Each image in our database was annotated by an art historian (see Section. 4 for a sample of the database). The annotations contain semantic information, such as the theme of the scene or the year of the engraving, as well as lower level data like the characters' contour or a set of image points representing it. These were manually annotated and correspond to keypoints, such as eyes, elbows or knees. Fig. 4 (left) depicts the pose annotation. We discard the head points, as well as the hands and feet points (dashed red), because these body parts vary substantially even within the same classes of subjects and themes, thus making the keypoint annotation very noisy.

We build the image descriptors from the keypoints of each personage in the scene. Each image f is characterized by the individual pose of the scene's subjects and their relationship in the image plane. For an image with P subjects we have

$$\mathbf{v}_f = \begin{bmatrix} \mathbf{v}_{f1} \\ \vdots \\ \mathbf{v}_{fP} \end{bmatrix} \tag{1}$$

where $\mathbf{v}_{fk}$ $(k = \{1, \ldots, P\})$ describes the k^{th} subject in the scene.

This vector is further divided into individual information, describing the pose, and contextual information, describing the relationships between subjects in the scene, i.e.

$$\mathbf{v}_{fk} = \begin{bmatrix} \mathbf{p}_{fk} \\ \mathbf{s}_{fk} \end{bmatrix}. \tag{2}$$

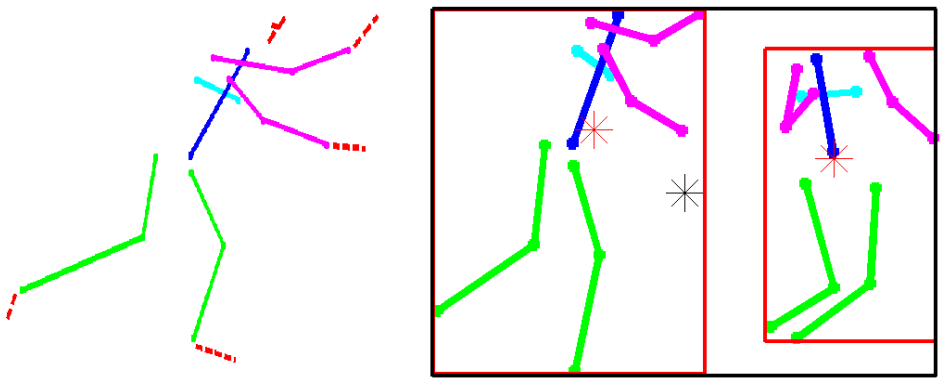

Figure 4. Pose Annotation and Descriptors. The left image shows the keypoint (elbows, knees, etc.) annotation underlying our image descriptor. The individual pose part of the descriptor contains the orientation of the line segments in the image. We discard the head, as well as the hands and the feet (dashed red), because these body parts vary substantially even within the same classes of subjects and themes, thus making the keypoint annotation very noisy. The right image depicts the composition part of the descriptor. It includes the relative areas occupied by each character (red boxes), as well as the relative centers of mass (red stars), with respect to the scene's area (black box) and center of mass (black star).

The first part of the subject's description $\mathbf{p}_{fk}$ was inspired in the pictorial models for articulated bodies[11] and is related to pose. In our case, pose must be scale invariant (Fig. 2). Hence, the individual pose descriptor only contains the orientations of the body parts, i.e. the solid line segments in Fig. 4 (left).

The cost of scale invariance is a misperception between depth and size: a character lying down and aligned with the scene depth is percieved as standing. Consequently, when only one subject is completely visible in the scene, we lose its relative scale and may get some curious results. In Fig. 5, the left image, modeled by the pose of the baby Christ alone, has as "closest" image (nearest neighbor) the right image, which only has the pose of the adult Christ fully annotated. If more subjects were completely visible, we would be able to leverage that contextual information for distinguishing the baby from the adult. Dealing with partial kinematic chains (missing data in general) and perspective effects in this setup are challenging tasks, requiring further research.

The second part of the subject's descriptor $\mathbf{s}_{fk}$ represents the spatial relationships between the subjects in the scene (see the right image in Fig 4). It contains the relative area of the personage (red boxes) with respect to the scene's bounding box (black box), as well as the image vector from the character's mean point (red stars) to the mean of the bounding box (black start), i.e. the relative center of mass.

When there is more than one character fully annotated in the scene, the relative scale produces a clear distinction between subjects like the child Christ and the adult Christ. Moreover, the relative centers of mass allow to organizing the descriptor vector from top to bottom, according to the left to right position of the characters on the image plane, and encode the dispersion of the characters within the scene's bounding box.

It is not yet clear if different descriptors for mirrored images benefit the retrieval process. Ultimately, this depends on the application. For example, it leverages retrieving the most similar image, but for semantic matching, such as theme matching, a scene, and its mirror, should be the same point.

3. COMPARING IMAGES BY POSE AND COMPOSITION

Given the image descriptors $\mathbf{v}_f$ from the previous section, we need a similarity function for comparing them. We use the radial basis function of the error d_{ij} between the descriptors of image i and j.

The length of our image descriptors is a function of the number of subjects. In fact, finding image representations that can handle and compare scenes with arbitrary number of subjects is one of the greatest challenges. As a consequence, two images described by a different number of subjects have infinite distance.

When the feature vectors have the same length, the error has two distinct contributions, one from the pose angles $\mathbf{p}_{fk}$ and the other from the composition data $\mathbf{s}_{fk}$.

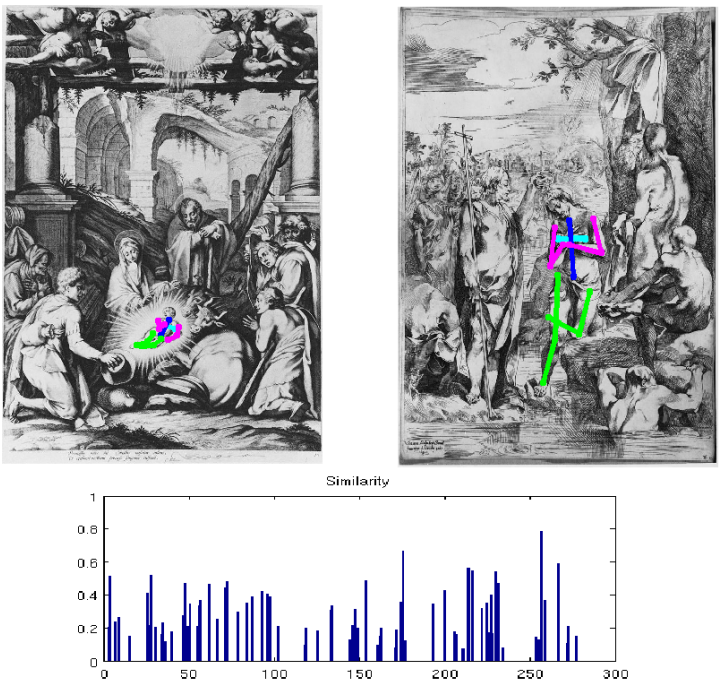

Figure 5. Scale Invariance of the pose descriptor. The query image (on the left) has only one fully described character – the child-Christ. The same happens with its nearest neighbor (on the right) which is only described by the adult Christ. Since the pose descriptor is scale invariant, the relative scale, and the depth, of the characters is lost when no more than one personage is considered, thus explaining this curious match between the child and the adult Christ.

Consider the l^{th} body part. The error between the body part l from characters k in images i and j is

$$d_{ij}(\mathbf{p}_{ik}(l), \mathbf{p}_{jk}(l)) = \min\{|\mathbf{p}_{ik}(l) - \mathbf{p}_{jk}(l)|, |2\pi - (\mathbf{p}_{ik}(l) - \mathbf{p}_{jk}(l))|\}, \tag{3}$$

where $\mathbf{p}_{ik}(l)$ is the orientation of the body part l from the k^{th} personage in image i. For the composition data $\mathbf{s}_{ik}$ we use euclidean distance. Hence, the squared error between two images with P subjects is

$$d_{ij}^2 = \frac{1}{n} \sum_{k=1}^{P} \sum_{l \in \text{limbs}} d_{ij}^2(\mathbf{p}_{ik}(l), \mathbf{p}_{jk}(l)) + \gamma \|\mathbf{s}_{ik} - \mathbf{s}_{jk}\|_2^2, \tag{4}$$

where n is the length of the feature vector and γ is an empirically adjusted parameter.

4. EXPLORING THE DATABASE

Our database contains images representing seven different themes of religious art. These are the *Annunciation* (ANN), the *Baptism-of-Christ* (BC), the *Flight-to-Egypt* (FE), the *Magi* (MAGI), the *Rest-during-the-flight-to-Egypt* (RFE), the *Shepherds* (SHEP) and the *Visitation* (VIS). Fig. 6 shows examples of each theme.

4.1 Theme Retrieval

The theme of a given scene is a semantic concept. Its unequivocal apprehension relies on features that cannot be fully explained by the character's pose and their composition in the scene. For example, all representations of the *Visitation* must have a particular subject named St. Elizabeth. However, the individual pose of the characters and their relationships may also shed some light into the theme of the scene, most specially when the main characters are surrounded by a complex framing, as in Fig. 2.

For retireving images from the same theme, we use nearest neighbors (k–NN) classification[10] in the space of the descriptors, according to Equation (4). To evaluate the system's performance, we compute the average

Figure 6. Religious Art Image Database. Our database contains images representing seven motives. From left to right and top to bottom: the *Annunciation* (ANN), the *Baptism-of-Christ* (BC), the *Flight-to-Egypt* (FE), the *Magi* (MAGI), the *Rest-during-the-Flight-to-Egypt* (RFE), the *Shepherds* (SHEP) and the *Visitation* (VIS).

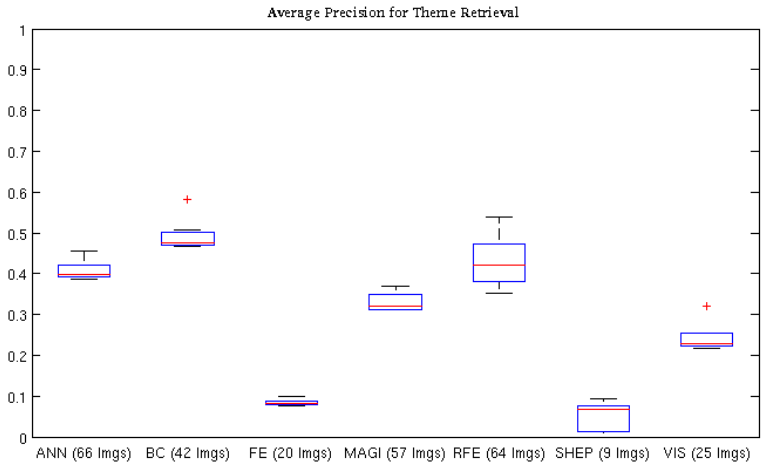

Figure 7. Theme retrieval. The boxlot shows the average precision of retrieving images from the theme. The themes are the annunciation (ANN), the baptism-of-Christ (BC), the flight-to-Egypt (FE), the magi (MAGI), the rest-during-the-flight-to-Egypt (RFE), the shepherds (SHEP) and the visitation (VIS).

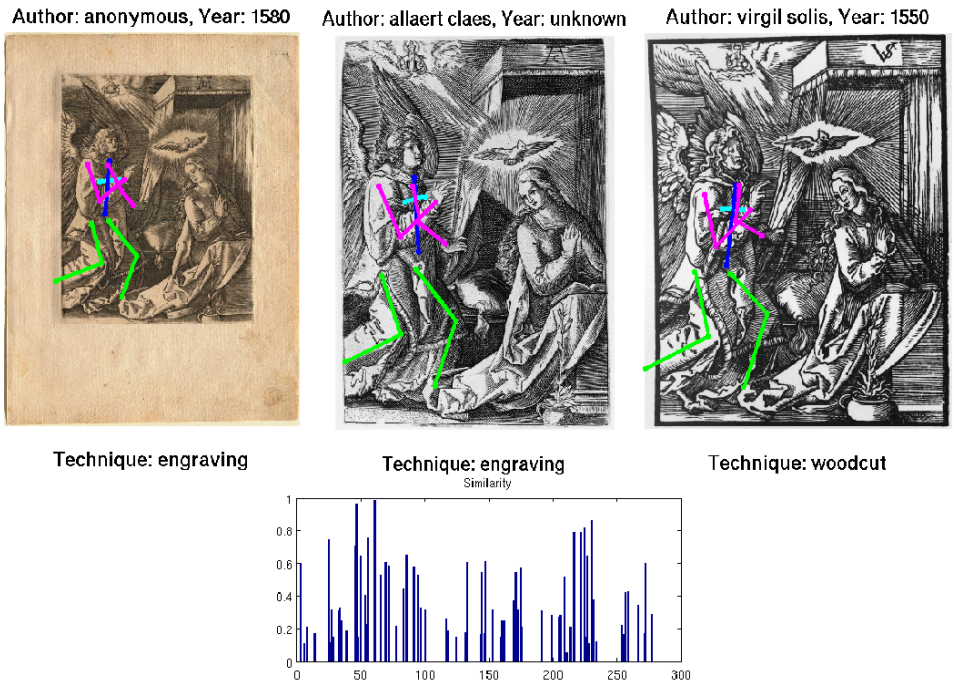

Figure 8. Tracing the roots of an image. We wish to provide a valid suggestion for the author and the roots of the query image (left). The right most images are the two nearest neighbors. From these results, we may suggest that in 1580 Allaert Claes made an engraving representing the *Annunciation*, depicted from the woodcut work of Virgil Solis, dated from 1550.

precision over all queries. For each query image, precision is defined as the ratio between the true positives and the number of nearest neighbors.[5] We perform eight sets of experiments, with $k \in \{2, \ldots, 9\}$ nearest neighbors. Nine is the number of images from the shepherds theme that have, at least, one character with its pose fully annotated, thus setting an upper bound for the possible nearest neighbors. Fig. 7 presents the boxplots of the average precision for these experiments.

The results rely heavily on the number of subjects fully annotated in the scene. The average precision for the *Shepherds* theme is influenced by the small number of fully annotated poses in its representations. The *Shepherds* theme contains a large number of characters, e.g. Mary, child-Christ and an arbitrary number of shepherds, which favors occlusions between characters. Hence, most of the times we are using no contextual information, as in the example of Fig. 5. On the contrary, the traditional composition of the images representing the *Rest-during-the-flight-to-Egypt* favours the visibility of 3–4 characters, increasing the intra-class similarity. For example, most of these images will have null similarity with the representations from the *Annunciation*, *Baptism-of-Christ* or *Visitation*, which typically contain one or two subjects.

Interestingly, the system achived the highest average precision for the theme *Baptism-of-Christ*, albeit its representations typically contain the same number of fully annotated poses as the *Annunciation* or the *Visitation*. In the *Baptism-of-Christ*, the two subjects are usually standing, very close to each other, and the arms of both subjects have a very distinctive pose, which is not the case for both the other themes (see Fig. 6). Notwithstanding, the precision can be increased, either by adopting other nonparametric classifiers, such as the naive-bayes nearest neighbors,[12] or by including contextual learning.[13–15]

4.2 Tracing the roots of an image

Here, we use our system to provide guidelines for filling the semantic gaps in the annotation of a particular image. In order to do that, we find the similarity profile of the left image in Fig. 8. The images on the right are the two nearest neighbors, corresponding to over 0.95 of similarity, as can be seen by the profile at the bottom of

Figure 9. The scope of applicability of exploring the pose of the subjects, and their composition, in scene goes to a scale that can involve all painted art. The left image shows a painting from the Madrilean painter Felix Castelo, "Los preparativos del viaje de Tobías desde Ecbatana a Ragues", Museo de Bellas Artes de Asturias. The right image show an engraving from Maarten van Heemskerck. Note that the pose of the subjects and their spatial relationships in the scene are identical in both images.

the figure. These three images are representations the *Annunciation*, and the annotated pose is from the angel Gabriel.

The query (left) image in Fig. 8 is from an engraving with unknown author and known date. The middle image is also an engraving, but has known author and unkown date. On the other hand, we have a full annotation from the right most image.

From these results, we may suggest that in 1580 Allaert Claes made an engraving representing the *Annunciation*, reproducing a woodcut from Virgil Solis, dated from 1550.

5. CONCLUSIONS AND PERSPECTIVES

Painted tile panels, representing religious motives, are a very representative Portuguese form of art. The panels' artists were usually inspired by famous paintings, or prints of those paintings. When studying a panel, art historians manually search image databases, looking for images similar to the representation in the panel, for tracing its roots.

Our system exploits the pose of the individual subjects, and their composition, in the scene, for helping art historians parsing art image databases. In particular, we use it here for tracing the roots of a given image. It may also be used in conjunction with mechanisms leveraging semantic information for image retrieval and annotation. Currently, the system relies on the annotated pose, whic should be done automatically, presenting challenges for future work.

Using pose to help retrieval of art images can be extended beyond the scope of this project. In some cases, finding influences or even clear copies among painters can be made possible only by the scene composition and relative pose of the subjects. A paradigmatic case is that of Madrilean painter Felix Castelo and the engravings of Maarten van Heemskerck shown in Fig 9. This case and a large piece of research in art history along this line (the Madrilean Baroque and the Flamish and Italian influences) is fully documented by Navarrete Prieto.[16] In conclusion, the scope of applicability goes to a scale that can involve all painted art.

ACKNOWLEDGMENTS

The authors would like to that Duarte Lázaro and Rosário Carvalho for their help with the art history issue, We would like th thank David Lowe for valuable suggestions on the development of this work.

REFERENCES

[1] Campos, T., "Application des regles iconographic aux azulejos portugais du xviieme siècle," *Europalia* , 31–40 (1991).

[2] Carvalho, R., "O programa artítico da ermida do rei salvador do mundo em castelo vide, no contexto da arte barroca," *Artis − Revista do Instituto de História da Arte da Faculdade de Lisboa* , 145–180 (2003).

[3] http://mnazulejo.imc−ip.pt.

[4] Stork, D. G., "Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature," *CAIP* (2009).

[5] Carneiro, G. and Costeira, J. P., "The automatic annotation and retrieval of digital images of prints and tile panels using network link analysis algorithms," *IS&T SPIE Computer Vision and Image Analysis of Art II* (2011).

[6] Stork, D. G., "Did jan van eyck build the first 'photocopier' in 1432?," in [*SPIE Electronic Imaging: Color Imaging IX: Processing, Hardcopy and Applications*], Eschbach, R. and Marcu, G. G., eds., 50–56, SPIE, Bellingham (2004).

[7] C. Richard Johnson, J., Hendriks, E., Berezhnoy, I. J., Brevdo, E., Hughes, S. M., Daubechies, I., Li, J., Postma, E., and Wang, J. Z., "Image processing for artist identification: Computarized analysis of vicent van gogh's painting brushstrokes," *IEEE Signal Processing Magazine* **37** (2008).

[8] Hughes, J. M., Graham, D. J., and Rockmore, D. N., "Stylometrics of artwork: uses and limitations," *IS&T SPIE Computer Vision and Image Analysis of Art* (2010).

[9] Marchenko, Y., Chua, T.-S., Aristarkhova, I., and Jain, R., "Representation and retrieval of paintings based on art history concepts," *IEEE International Conference on Multimedia and Expo* (2004).

[10] Friedman, J., Hastie, T., and Tibshirani, R., [*The Elements of Statistical Learning: Data Mining, Inference and Prediciton*], Springer (2001).

[11] Felzenszwalb, P. F. and Huttenlocher, D. P., "Pictorial structures for object recognition," *International Journal of Computer Vision* **61**(1) (2005).

[12] Boiman, O., Shechtman, E., and Irani, M., "In defense of nearest-neighbor based image classification," *IEEE Conference on Computer Vision and Patter Recognition* (2008).

[13] Desai, C., Ramanan, D., and Fowlkes, C., "Discriminative models for multi-class object layout," *IEEE International Conference on Computer Vision* (2009).

[14] Yao, B. and Fei-Fei, L., "Modeling mutual context of object and human pose in human-object interaction activities," *IEEE Conference on Computer Vision and Patter Recognition* (2010).

[15] Eichner, M. and Ferrari, V., "We are family: Joint pose estimation of multiple persons," *European Conference on Computer Vision* (2010).

[16] Prieto, B. N., "Flandes e italia en la pintura barroca madrilea," *Fuentes y Modelos de la Pintura Barroca Madrilea* (2008).