TADA: Toolkit for Analysis of DeviantART

Bart Buter Nick Dijkshoorn Davide Modolo Quang Nguyen Sander van Noort Bart van de Poel

***Abstract*—abstract**

***Index Terms*—image analysis, image features, classification, deviantART, art, online social network**

I. Introduction

A. deviantART

B. Why interesting as a data source

C. Research questions

D. Solution - framework

E. Overview of the paper

deviantART[1] (commonly abbreviated as *dA*) is the largest online community [?] showcasing various forms of user-made artwork. The website was launched in 2000 and has over 13 million registered members. The platform allows emerging and established artists to exhibit, promote, and share their works within a peer community dedicated to the arts. Newly added art is submitted to the constantly-changing *Newest* listing, where it is viewable by the general public. Members are able to create profiles, galleries of their own work, and to choose *Favorites* from among other submissions. Through the publicity process, some pieces become *Popular*, and are added to a ranked list of *Most Popular* in the last 8 hours, 1 day, 3 days, 1 week, 1 month, and *All time.* During their browsing, members are also allowed to comment on one anothers art and on their profiles, making for a highly interactive and dynamic community. All artwork is organized in a comprehensive category structure, including traditional media, such as painting and sculpture, to digital art, pixel art, films and anime.

[social aspects/networks] Bart B.

[why relevant, problem with current research] Research questions

- Can we visualize important aspects of deviantART?
- Can artists and/or styles be distinguished?
- Are artists influencing each other?
- Do art styles change over time?
- Are there none-artists interesting for deviantART?

[goals - framework]

II. Previous work

A. Online social network (art - pictures)

small world network

B. Image features

C. Classification of image features

D. Visualization of image features

III. Implementation

A. Toolkit introduction

4 components, what should it do (online/offline), image2image, gal2gal, cat2cat, albert drawing

B. Data collection

C. Feature extraction

D. Classification

E. Visualization

F. Data collection

[Sander, Bart B.]

DeviantArt provides users, also called deviants, galleries of their images. When someone visits such a gallery a featured page will be shown. Furthermore the user is provided with options to browse the gallery or visit a sub gallery defined by the user. There are various types of members: normal members, premium members, banned members, staff members, etc. Premium members have extra benefits like no ads, gallery customization, beta-test new site features and more.

DeviantArt does not provide a web api to download images. This makes it more difficult to download images. On top of that changes to the website can possible break down the downloading application.

DeviantArt does provide rss, which allows us to download xml files containing information about the users galleries. RSS xml files are more easy to parse than the html gallery pages.

For each image the full sized image and two different sized thumbnails are available. DeviantArt supports png, jpeg, bmp and gif image formats.

In order to create our dataset we needed to download the images from Deviantart. The *Gallery Scraper* is written in Python and retrieves all images for a given list of deviants. In our initial approach we retrieved the backend link from *http://deviant.deviantart.com/gallery* using the SGML(Standard Generalized Mark-up Language) parser included in the sgmllib Python module. The backend link points to a xml file containing the first 60 images of the deviant. TO retrieve the other xml files we needed to change the offset in the url.

We parsed the xml files using the xml.sax module included with Python. For each item in the xml file a link to the full sized image, a big thumbnail and a small thumbnail are provided. The big thumbnails are usually around 300 width and the small thumbnails around 150 height. For

[1] `http://www.deviantart.com`

each deviant we created a seperate folder, which contains three subfolders for the different sized images.

For each image of the deviant a xml file is written containing the filename, category, deviantart link and title. An example of such a xml file looks like this:

(Insert example here)

Although this initial approach seems to work we did encountered a few problems. First it seems the backend link on the gallery page only contains the images listed under *featured*, which does not necessarily contains all images of the deviant. After some more investigation it turns out we can retrieve the xml file containing all images directly from the url $http : //backend.deviantart.com/rss.xml?q = gallery : \$deviant\$$.

The second problem is that downloading an image might fail due problems at the server side of DeviantArt. We solved this by removing corrupt downloads and not generating a xml file for the image. Then at a later point we run the scraper and try to download the missing images again.

To retrieve statistics about the dataset we made a quick analysis tool. It reads in all image xml files using the xml.minidom Python module and counts the categories. It has an option to this per user or for the whole dataset. It also has an option to only print the top category statistics or also print per sub category.

An analysis including the sub categories looks like this: (Insert example here)

G. *Feature extraction*

When working with images, it is usually not possible to work with the raw image data (the pixel values). The reason for this is the high dimensionality of images, which can easily exist in a space of more than a million dimensions. By extracting features from images, they can be represented in a lower dimensional feature-space. This feature extraction process has several advantages:

- The data becomes computationally easier to work with due to the smaller number of dimensions
- By using the right features, the data becomes more suitable for generalization across images
- Reducing the dimensionality makes it easier to visualize sets of images
- Features can have an intuitive basis, which makes it easier for non-computer-scientists to analyze (sets of) images

Here something general about different kinds of image features

In the extraction of image features, a distinction was made between low-level statistical features and higher level cognitive-based features....

G.1 Statistical features

As statistical features, many relatively simple low-level features were extracted from the images. Below is a list of all features and their meaning:

- Edge ratio: description
- Corner ratio: description
- etc...

G.2 Cognitively-inspired features

One of the more recent trends in computer vision research in the pursuit of human-like capability is the coupling of cognition and vision into cognitive computer vision. The term cognitive computer vision has been introduced with the aim of achieving more robust, resilient, and adaptable computer vision systems by endowing them with a cognitive faculty: the ability to learn, adapt, weigh alternative solutions, and develop new strategies for analysis and interpretation.

In our project we focus on computational models of focal visual attention. Attention allows us to break down the problem of understanding a visual scene into a rapid series of computationally less demanding, localized visual analysis problems. "Visually salient" are those location in the visual wolrd that automatically attract attention.

H. *Classification*

[Quang]

I. *Visualization*

[Nick]

IV. Experiments

A. *Image experiments*

dataset, unbalenced

A.1 1 vs all artists - distinct artists

result: ranking of artists (some good, some bad)

A.2 1 vs 1 artists - classification performance

B. *Network experiments*

dataset

small world, describe network, identify core

V. Future work

VI. Conclusion

given our dataset. We have made a toolkit. Capable to anaylize, visualize... this is the initial step, still many possibilities, experiments show it is interesting. artists that can be identified using simple features people pop out

References

[1] Jan-Mark Geusebroek and Arnold W. M. Smeulders. A physical explanation for natural image statistics. In *International Workshop on Texture Analysis and Synthesis*, 2002. Observations on 45,000 stock photos of natural images, covering a wide variety of topics, revealed that 60Weibull-shaped contrast distributions. The remaining percentage of images has a distribution close to Weibull or is highly regular. These images are typically composed of two (or more) parts in the image, each having a distinct Weibull contrast distribution.

[2] Arnold W. M. Smeulders Victor A. F. Lamme H. Steven Scholte Sennay Ghebreab, Lourens Waldorp. Brain responses strongly correlate with weibull image statistics when processing natural images. *Journal of Vision*, 2009. UvA research, the brain is capable of approximating the beta and gamma underlying the contrast distribution in natural images.

[3] Arnold W. M. Smeulders Jan-Mark Geusebroek. A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, 2005. Apparently, the Weibull parameters form a six-stimulus basis for stochastic texture description. The results indicate that texture perception can be approached like the experimental science of colorimetry.We have experimentally verified that the Weibull distribution characterizes spatial statistics for ergodic stochastic textures.

[4] C.J. Veenman C.G.M. Snoek A.W.M.Smeulders J.C. van Gemert, J. Geusebroek. Robust scene categorization by learning image statistics in context. *Unknown Journal*, 2010. Learning image categories using Weibull features and color invariant edges to represent texture and othe r natural image statistics.

[5] J.C. van Gemert, J. Geusebroek, C.J. Veenman, C.G.M. Snoek, and A.W.M. Smeulders. Robust scene categorization by learning image statistics in context. In *2006 Conference on Computer Vision and Pattern Recognition Workshop*, pages 105–105, 2006. Observations on 45,000 stock photos of natural images, covering a wide variety of topics, revealed that 60Weibull-shaped contrast distributions. The remaining percentage of images has a distribution close to Weibull or is highly regular. These images are typically composed of two (or more) parts in the image, each having a distinct Weibull contrast distribution.

Contents