International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 06 | June-2016 www.irjet.net p-ISSN: 2395-0072

Gender Prediction using Images Posted on Online Social Networks

Minal Gadiya[1], S. V. Jain[2]

Student, Computer Science and Engg, Shri Ramdeobaba College of Engineering and Management, Nagpur, India[1]

Assistant Professor, Computer Science and Engg., Shri Ramdeobaba College of Engg. andManagement,Nagpur, India[2]

Abstract:Identifying user attributes from their social network activities has been a common research topic nowadays. Age, gender and interest can be common user attributes which can be predicted and are essential for personalization and recommender systems. Most of the researches are based on the textual content created by user, whereas recently multimedia has gained popularity in social networks. In this paper we use feature extraction algorithm that predicts the user gender on different networking sites.

Keywords – Social Network, Social Images, Gender Prediction, Demographics

INTRODUCTION

Online Social Networking sites such as Facebook and Twitter are widely used communication medium, especially among young people. These social sites have become very popular these days and therefore it has become a research topic for studying relationship between users' digital behavior and their demographic attributes such as age, gender, relationship status, etc.

Instagram and Pinterst are mainly image based social sites. Images posted by users on online social network may be useful to learn various personal and social attributes of users.

We mainly extract the features from the images posted by users using their posting behavior and posted content. We use the images from Pinterestimage dataset. There is a difference between male and female preferences. For male users, they are mostly interested in electronics, buildings, men clothes and so on. On the other hand, female users are mainly interested in jewelry, women clothes, gardening and so on. For each user, we extract features like color, texture and shape from their collections of pins in a few different categories, such as art, cars & motorcycles and food & drinks. For posting behaviors, we focus on the users' own labeled distribution of their collections of pins over the limited number of categories provided by Pinterest. Our results suggest that both posting behavior and posted content are beneficial for gender prediction.

Our contribution includes predicting the gender of the user based on the type of images posted by him/her and increasing the accuracy of the system. We frame gender classification as a binary classification problem (male and female categories) and evaluate the use of a variety of image based features.

RELATED WORK

In 2014, Quanzeng You and JieboLuo and Sumit Bhatia presented a paper "A Picture Tells a Thousand Words-About You! User Interest Profiling from User Generated Visual Content," in which they analyze the content of individual images and then aggregate the image-level knowledge to infer user-level interest distribution. They employ image-level similarity to propagate the label information between images, as well as utilize the image category information derived from the user created organization structure to further propagate the category-level knowledge for all images. A real social network dataset created from Pinterest is used for evaluation[2].

In 2013, J. S. Alowibdi, U. A. Buy, and P. Yu presentd a paper "Empirical evaluation of profile characteristics for gender classification on Twitter,"in which they explore profile characteristics for gender classification on Twitter. Unlike existing approaches to gender classification that depend heavily on posted text such as tweets, here they study the relative strengths of different characteristics extracted from Twitter profiles (e.g. first name and background color in a user's profile page). Their goal is to evaluate profile characteristics with respect to their predictive accuracy and computational complexity. In addition, they provide a novel technique to reduce the number of features of text-based profile characteristics from the order of millions to a few thousands and in some cases, to only 40 features. They prove the validity of their approach by examining different classifiers over a large dataset of Twitter profiles[3].

In 2012, Sridhar, Gowri, presented a paper "Color and Texture Based Image Retrieval," in which they build an interactive image recommendation system, which firstly uses color histogram feature and GCLM texture feature to express image contents, then a kernel based K-means is utilized to cluster images into multiple classes by their visual features, finally based on a feature vectors stored in the database the similar images are retrieved. The HSV color histogram is calculated and the joint histogram is derived based on the combination of the hue and saturation in the hue and saturation histogram. The color feature is extracted from the joint histogram. The chi-square is used to find the similarity between the two images. Thus global feature is calculated using the joint histogram. The regional feature is extracted using the GCLM technique in which the neighbor pixels is considered into account. The evaluation results demonstrate the accuracy of the retrieval based on the precision and recall false positive and negative ratio. The ROC curve is used to compare the efficiency of the color, texture and the combination of both the color and the texture[4].

In 2010, Dr. H.B. Kekre, SudeepThepade, Priyadarshini Mukherjee, MitiKakaiya,, ShobhitWadhwa and Satyajit Singh presented a paper "Image Retrieval with Shape Features Extracted usingGradient Operators and Slope Magnitude Technique withBTC," in which novel image retrieval methods based on shape features extracted using gradient operators and slope magnitude technique with Block Truncation Coding (BTC). Four variations of proposed „Mask-Shape-BTC" image retrieval techniques are proposed using gradient masks like Robert, Sobel, Prewitt and Canny. The proposed image retrieval techniques are tested on generic image database with 1000 images spread across 11 categories. In all 55 queries (5 from each category) are fired on the image database. The average precision and recall of all queries are computed and considered for performance analysis. In all

the considered gradient operators for shape extraction, „Mask-Shape- BTC" CBIR techniques outperform the „Mask-Shape" CBIR techniques. The performance ranking of the masks for proposed image retrieval methods can be listed as Robert (best performance), Prewitt, Sobel and lastly the Canny[5].

In 2000, B. Moghaddam and M. H. Yang, presented a paper "Gender classification with support vector machines," in Automatic Face and Gesture Recognition in which they addressed the problem of classifying gender from thumbnail faces in which only the main facial regions appear (without hair information). In their study, they demonstrate that SVM classifiers are able to learn and classify gender from a large set of hairless low resolution images with very high accuracy[6].

Michael Fairhurst and M'arjory Da Costa-Abreu, presented a paper "Using keystroke dynamics for gender identification in social network environment." In which they introduce an approach to addressing risks such as risk of transactions with individuals who deliberately conceal their identity or, importantly, can easily misrepresent their personal characteristics. They use a form of biometric data accessible from routine interaction mechanisms to predict important user characteristics, thereby directly increasing trust and reliability with respect to the claims made tomessage receivers by those who communicate with them [7].

PROPOSED APPROACH

We frame the task of predicting users' gender from their posted images as a binary classification task (fig 1). Given a set of images posted by a user on a social networking site, we predict whether the user is male or female. We suggest that males and females differ in terms of their image posting behavior as well as in the content of posted images. We extract features to capture visual content of images as well as users' posting behavior.

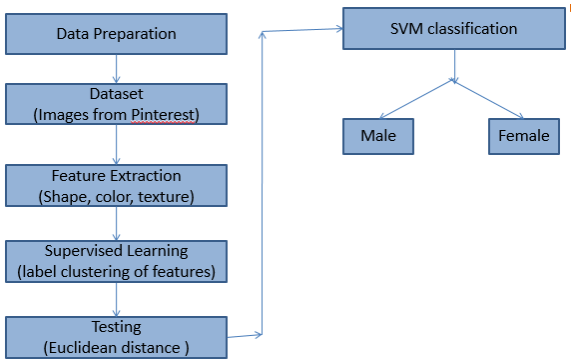

Fig 1: Sample Design Architecture

In feature extraction process we extract the color, texture and shape feature of the images. The color features are mainly based on the HSV histogram, color autocorrelogram and color moments. The texture feature are mainly besed on the contrast, correlation, energy and homogeneity properties. Then we used supervised learning approach for label clustering of the feature extracted images. And then we finally apply the SVM classifier to predict the gender of the user whether male or female.

IMPLEMANTATION DETAILS

A. *Data Preparation*

Dataset is been prepared using different categories of image posted by users. These images are collected from Pinterest websites (fig 2).

Pinterest: Pinterest is a free website that requires registration to use. Users can upload, save, sort and manage images known as pins and other media contents (e.g. videos and images) through collections known as pinboards. Pinterest acts as a personalized media platform.

User data: Like Facebook and Twitter, Pinterest now let marketers access the data collected on its users. By granting access to users' data, Pinterest lets marketers investigate how people respond to products. If a product has a high number of repins, this tells the producer of the product that it is liked by many members of the Pinterest community. Now that Pinterest lets marketers access the data, companies can view user comments on the product to learn how people like or dislike it. People use social media sites like Pinterest to direct or guide their choices in products. Sample dataset which is been collected is shown below:

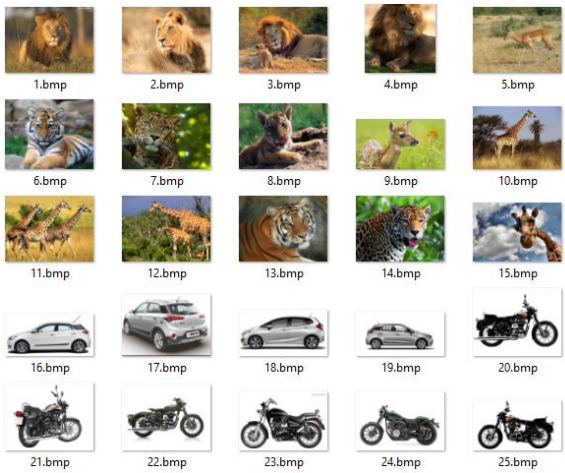

Fig 2: Sample dataset of Pinterest

B. *Feature Extraction*:

We are using color, texture and shape feature extraction techniques for extracting the features from the images.We are extracting 102 color features, 4 texture features and 124 shape features. These 230 features are the most important features of an image.

Color Feature: Color is an important feature for image representation which is widely used in image retrieval. This is due to the fact that color is invariance with respect to image scaling, translation, and rotation. The key items in color feature extraction consist of color space, color quantization, and the kind of similarity measurements. Color Feature can be extracted using color moment, color histogram, and Color Coherence Vector (CCV). Color Histogram is commonly based on the intensity of three channels. It represents the number of pixels that have colors in each of a fixed list of color ranges. Color Moment is used to overcome quantization effect in color histogram. It calculates the color similarity by weighted Euclidean distance. Color set is used for fast search over large collection of image. It is based on the selection of color from quantized color space.

A histogram is the distribution of the number of pixels for an image. The color histogram represents the color content of an image. It is robust to translation and rotation. Color histogram is a global property of an image. The number of elements in a histogram depends on the number of bits in each pixel in an image. For example, if we suppose a pixel depth of n bit, the pixel values will be between 0 and 2n-1, and the histogram will have 2n elements. The HSV space color histogram is calculated and the joint histogram is calculated by using Hue and Saturation Histogram by calculating the total number of pixel in both the Hue and Saturation Histogram. The joint histogram is calculated using

$$p(hi, sj) = N(hi, sj) / Ntotal$$

where, N (hi, sj) is the total number of pixel in both the hue and saturation histogram, Ntotal is the total number of pixel in the image. The joint histogramcan be used to efficiently calculate the mean,standard deviation, entropy, skewness and kurtosis of very large data sets. This is especially important for images, which can contain millions of pixels. The sum of all elements in the histogram must be equal to the number of pixels in the image. For evaluation some sample images in order to evaluate

different extracted features. In evaluation, a retrieved image is considered a match if and only if it is in the same category as the query image. In addition, the effectiveness of the extracted features has been measured by precision and recall parameters. Precision is the ratio of relevant retrieved images to the total number of retrieved images. Recall is the ratio of retrieved relevant images to the total number of relevant images in the database.

Texture Feature: Texture [6] refers to visual patterns with properties of homogeneity that do not result from the presence of only a single color such as clouds and water. Texture features typically consist of contrast, uniformity, coarseness, and density. There are two basic classes of texture descriptors, namely, statistical model-based and transform-based. The former one explores the grey-level spatial dependence of textures and then extracts some statistical features as texture representation. One example of this group is co-occurrence matrix representation. The latter approach is based on some transform such as DWT.

Gray-level co-occurrence approach uses Gray-Level Co-occurrence Matrices (GLCM) whose elements are the relative frequencies of occurrence of grey level combinations among pairs of image pixels. The GLCM can consider the relationship of image pixels in different directions such as horizontal, vertical, diagonal, and antidiagonal. The co-occurrence matrix includes second-order grey-level information, which is mostly related to human perception and the discrimination of textures. Four statistical features of the GLCMs are computed. The features are energy, entropy, contrast, and homogeneity. G *G GLCM Pdfor a displacement vector d = (dx; dy) is defined as follows. The (i; j) of Pd is the number of occurrences of the pair of gray-level i and j which are a distanced apart.

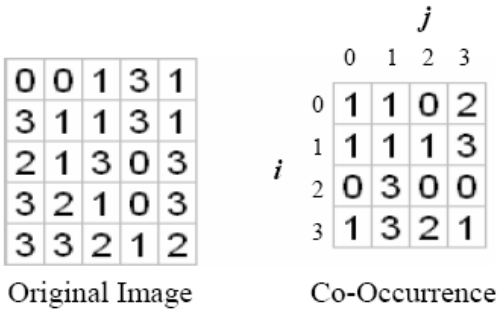

Figure 3: describes the Co-Occurrence Matrix Functionality of the gray level co-occurrence matrix in detail.

Thus the co-occurrence matrix is used to find the pixel level similarity in the image. So, the extractedfeature is based on the neighbor pixel value. Thedifferent angle images are not retrieved using theGCLM technique. So, the orientation of the imagecan be calculated using the Gabor texture in whichthe orientation and the frequency are taken intoconsiderations. A set of 113 images is considered forextracting the feature. The images are JPEG imageswith standard resolution of 640x480. The images andfeature are stored in MySQL database. The featuresare extracted using the MATLAB.

Shape Feature: Efficient shape features must present some essential properties such as:

- Identifiability: shapes which are found perceptually similar by human have the same features that are different from the others.
- Translation, rotation and scale invariance: the location, the rotation and the scaling changing of the shape must not affect the extracted features.
- Affine invariance: the affine transform performs a linear mapping from coordinates system to other coordinates system that preserves the "straightness" and "parallelism" of lines. Affine transform can be constructed using sequences of translations, scales, flips, rotations and

shears. The extracted features must be as invariant as possible with affine transforms.

• Noise resistance: features must be as robust as possible against noise, i.e., they must be the same whichever be the strength of the noise in a given range that affects the pattern.

• Occultation invariance: when some parts of a shape are occulted by other objects, the feature of the remaining part must not change compared to the original shape.

• Statistically independent: two features must be statistically independent. This represents compactness of the representation.

• Reliability: as long as one deals with the same pattern, the extracted features must remain the same.

In general, shape descriptor is a set of numbers that are produced to represent a given shape feature. A descriptor attempts to quantify the shape in ways that agree with human intuition (or task-specific requirements). Good retrieval accuracy requires a shape descriptor to be able to effectively find perceptually similar shapes from a database. Usually, the descriptors are in the form of a vector. Shape descriptors should meet the following requirements:

• The descriptors should be as complete as possible to represent the content of the information items.

• The descriptors should be represented and stored compactly. The size of a descriptor vector must not be too large.

• The computation of the similarity or the distance between descriptors should be simple; otherwise the execution time would be too long.

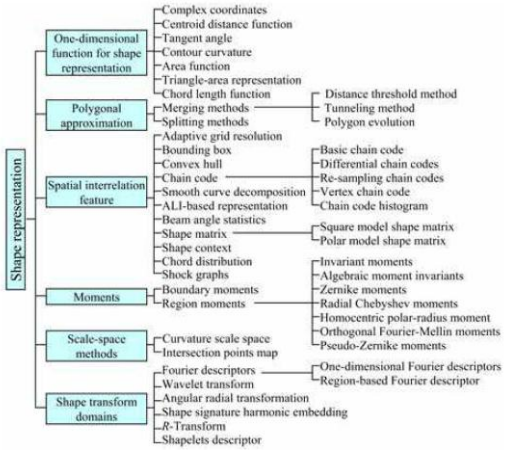

Figure 4: An overview of shape description techniques

C. *Supervised Learning*

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

D. *Testing*

In testing we apply Euclidian distance to the input image and all the images in the dataset to measure the least minimum distance between them and to specify to which category they belong either animal, architecture, cars or any other.

We use the following Euclidian distance:

Euclidian = sum ((dataset- input image) .^ 2

We consider the first five least minimum Euclidian distance results for further classification.

E. *SVM Classification*

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

SVM classification will be used to classify whether the user is male or female. In machine learning, support vector machine (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.

CONCLUSION

Initially we predict the category of the image to which in belongs and then the gender of the user from the posting behavior and the visual content of the images. We will then measure the performance in terms of accuracy, precision, recall and F-measure.

ACKNOWLEDGEMENT

We hereby thanks the authors listed in the references for the valuable information and survey statistics.

REFERENCES

[1]Quanzeng You, Sumit Bhatia, Tong Sun, JieboLuo, The eye of the beholder: Genderprediction using images posted in Online Social Networks, 2014 IEEE InternationalConference on Data Mining Workshop.

[2]Quanzeng You, JieboLuo and Sumit Bhatia, A Picture Tells a Thousand Words- AboutYou! User Interest Profiling from User Generated Visual Content, 2014 IEEE InternationalConference.

[3] J. S. Alowibdi, U. A. Buy and P. Yu, Empirical evaluation of profile characteristics forgender classification on Twitter, in Machine Learning and Applications(ICMLA), 201312th International Conference on, vol. 1. IEEE, pp. 365-369.

[4] Sridha and Gowri, Color and Texture Based Image Retrieval, ARPN Journal of Systems and Software, VOL. 2, NO. 1, January 2012.

[5]Dr. H.B. Kekre, SudeepThepade, Priyadarshini Mukherjee, MitiKakaiya,, ShobhitWadhwa and Satyajit Singh,"Image Retrieval with Shape Features Extracted usingGradient Operators and Slope Magnitude Technique withBTC," International Journal of Computer Applications (0975 – 8887), Volume 6– No.8, September 2010.

[6] L. Fei-Fei and P. Perona, A Bayesian hierarchical model for learning natural scenecategories, in Computer Vision and Pattern Recognition, 2005.CVPR 2005. IEEEComputer Society Conference on, vol. 2. IEEE, 2005, pp. 524-531.

[7] B. Moghaddam and M. H. Yang, Gender classification with support vector machines, inAutomatic Face and Gesture Recognition, 2000, pp. 306-311.

[8] Michael Fairhurst and M'arjory Da Costa-Abreu, Using keystroke dynamics for genderidentification in social network environment.

BIOGRAPHIES

MinalGadiyahas received her B.E. degree in Information Technology in 2014. She is pursuing Masters in Technology in Computer Science and Engineering from Shri Ramdeobaba College of Engineering and Management, Nagpur-440013. Her areas of interest include Image Processing and Network Security.

Professor Sweta Jain received the Masters in Technology from Nagpur University in 2009 as a first merit holder. She is currently Assistant professor in Computer Science and Engineering department at Shri Ramdeobaba college of Engineering and Management, Nagpur. She has a total teaching experience of around 13 years. Her research interest include Pattern Recognition, Digital Image Processing and Machine Learning.