IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

A Study on Gender Prediction using Online Social Images

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: Nowadays identifying user attributes from their social network activities has been a common research topic. 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 propose an algorithm that predicts the user gender on different networking sites.

Keywords: Social Network, Social Images, Gender Prediction, Demographics.

INTRODUCTION

Social networking sites are now a very widely used communication medium, especially among young people. Facebook and Twitter, etc have become very popular this days. Many people on a very large scale have adopted this 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 users of Pinterest. 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 from their collections of pins in a few different categories, such as art, travel and technology. For posting behaviors, we focus on the users' own labeled distributions 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.

LITERATURE REVIEW

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 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[5].

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

DOI 10.17148/IJARCCE.2016.5284

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[6].

METHODOLOGY

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.

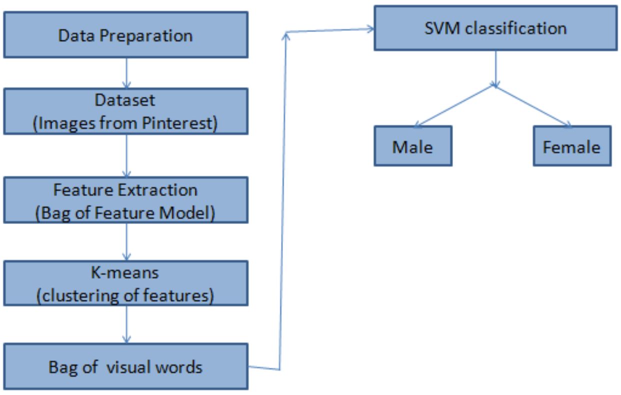

Fig 1: Sample Design Architecture

In Visual Content of posted Images we will first construct a bag of visual words representation(visual vocabulary). Then Scale-invariant feature transform(SIFT) will be used to discover local features for each image in the dataset. Visual words will be discovered by clustering all the SIFT features. We will use K-means for clustering. Bag of Visual words will be created in which we include different categories of feature and each category will be known as a visual word. We will use SVM algorithm for classification of user as male or female.

A. Data Preparation

Dataset is been prepared using different categories of image posted by users. These images are collected from Pinterestwebsites(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:

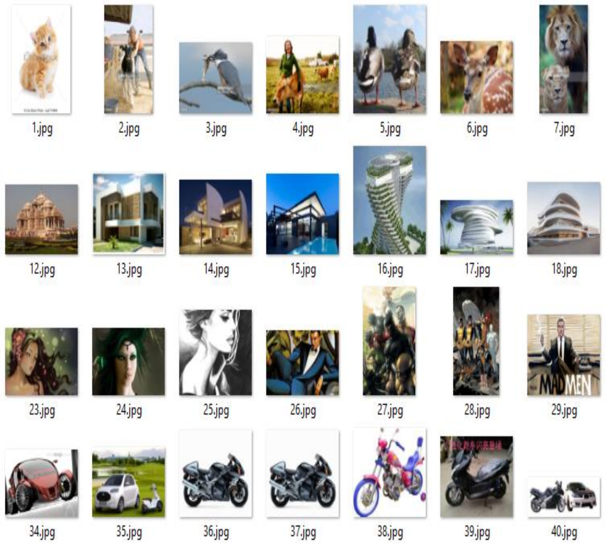

Fig 2: Sample dataset of Pinterest

B. Bag of Feature Model: Various features of images will be extracted here(fig 3) and all the SIFT features will be collected in the bag of feature.

SIFT: Scale-invariant feature transform is an algorithm in computer vision to detect and describe local features in images. SIFT can robustly identify objects even among cluster and under partial occlusion, because SIFT feature descriptor is invariant to uniform scaling, orientation and partially invariant to affine distortion and illumination changes.

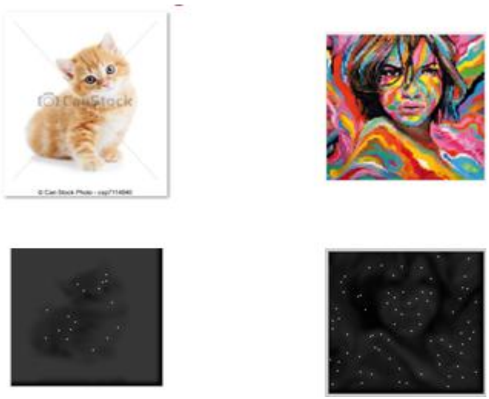

Fig 3: Sample images after applying SIFT

C.Clustering of features:
collection of all the similar features will be done under one category using K-means. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

DOI 10.17148/IJARCCE.2016.5284

D.Bag of visual words: These different categories of features will be different visual words. All this visual words will be collected in one place which will be known as bag of visual words.

E.SVM classification: 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 will predict the gender of the user from the posting behavior and the visual content of the images and then the performance will be measured 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] L. Fei-Fei and P. Perona, A Bayesian hierarchical model for learning natural scene categories, in Computer Vision and Pattern Recognition, 2005.CVPR 2005. IEEEComputer Society Conference on, vol. 2. IEEE, 2005, pp. 524-531.

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

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

BIOGRAPHIES

Minal Gadiya has 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.

DOI 10.17148/IJARCCE.2016.5284