An Analytic System for User Gender Identification through User Shared Images

Ming Cheung, HKUST-NIE Social Media Lab
James She, HKUST-NIE Social Media Lab

Many social media applications, such as recommendation, virality prediction and marketing, make use of user gender, which may not be explicitly specified or kept privately. Meanwhile, advanced mobile devices have become part of our lives and a huge amount of content is being generated by users every day, especially user shared images shared by individuals in social networks. This particular form of user generated content is widely accessible to others due to the sharing nature. When user gender is only accessible to exclusive parties, these user shared images are proved to be an easier way to identify user gender. This work investigated 3,152,344 images by 7,450 users from Fotolog and Flickr, two image-oriented social networks. It is observed that users who share visually similar images are more likely to have the same gender. A multimedia big data system that utilizes this phenomenon is proposed for user gender identification with 79% accuracy. These findings are useful for information or services in any social network with intensive image sharing.

CCS Concepts: •**Networks** → **Online social networks;** •**Human-centered computing** → *Social networking sites; Social tagging systems;* Social networks;

General Terms: Big data analytic system, Images

Additional Key Words and Phrases: big data, mobile, user shared images, gender, recommendation, social network analysis

1. INTRODUCTION

User gender is important information for many personalized services or applications in online social networks, such as recommendation, virality prediction and connection discovery. However, this information may be hidden, or not specified and makes these applications inaccurate or not possible. Trending mobile social applications, such as Instagram (owned by Facebook from the US) and WeChat (owned by Tencent from China), do not provide this information for others. This is the trend of today's social networks for preserving user information due to privacy concerns. Nowadays, users tend to use different social networks for sharing different types of content, for example, sharing images on Flickr but sharing videos on Facebook. By combining information on heterogeneous networks, additional information about users can be obtained. Yet, this information on social networks could be missing, or difficult to obtain due to incomplete user information provided.

Meanwhile, a huge amount of content is being generated daily from our mobile devices as they have become part of our daily lives. The advances in devices, such as smartphones and wearables, as well as wireless technologies, make taking and sharing high quality images much easier than before, and these images can be processed by a cloud-assisted approach. Unlike traditional computer vision tasks to identify objects and understand the context of an image, this work attempts to identify user gender with the shared images of a user, with the help of non-user generated label[Cheung

This work is supported by the HKUST-NIE Social Media Lab.
Author's addresses: M. Cheung and J. She, Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

DOI: 0000001.0000001

et al. 2015b], and it is proven that any computer vision technique will work.

A non-user generated label is annotated on an image to represent the visual features, and two images with the same label are visually similar. Recently, researches have confirmed that two users with more similar non-user generated labels on their shared images are more likely to have online friendships, be of the same gender and be from the same origin [Cheung and She 2016]. An example of user generated images on Instagram is shown in Fig. 1, both users A and B shared images of cars and user C shared an image of a flower. As the features of cars are similar, the similarity between users A and B are higher than that between users A and C or B and C. User A and B are more likely to be of the same gender as they have a higher image similarity in their shared images. When more shared images from each of users A, B and C are accessible for evaluation, the gender identification should become reliably and accurately detectable. However, it is not clear how the gender of a user affects the image sharing behaviors, or how to make uses of these behaviors to identify user gender. Also, it is necessary to compute the pairwise similarity of all images shared by any 2 users, which requires $O(N_I^2)$ comparisons, where N_I^2 is the number of images. This becomes impossible when there are billions of images shared on social media.

With these motivations and challenges, this work has investigated over 3,152,344 user shared images shared by 7,450 users from 2 image-oriented social networks – Fotolog and Flickr– using a novel image processing technique, non-user generated labels [Cheung et al. 2015a][Cheung et al. 2015b]. Unlike traditional computer vision approaches that recognize objects and contexts in an image, non-user generated labels are generated as a non-biased representation to reflect the similarity of images. It is an unsupervised approach, which means there is no assumption about image, or predefined objects needed. User connections can be discovered based on the similarity of the occurrence of labels for gender identification. Non-user generated labels are not limited by the techniques used for image encoding. Different computer vision techniques, such as GIST, are proven to be able to annotate the images with non-user generated labels [Cheung et al. 2015b]. This paper applies a deep learning and a feature-based technique to demonstrate the effectiveness of using non-user generated labels for gender identification An interesting phenomenon between user genders and their shared images are observed from our intensive measurements, and this is nicely formulated with a proposed analytic system to identify user genders from user shared images on those social media. In summary, the contributions of this paper include the following:

- measured intensive and characterized user shared images from two social networks, Fotolog and Flickr, which proved the phenomenon that two users with a higher similarity of their shared images is likely to have the same genders;
- proposed a practical analytic system using non-user generated labels to identify user gender by their shared images;
- verified extensively the proposed formulation and analytic system with over 3 million images from 7,450 users in the two social networks to prove the effectiveness of using user shared images through non-user generated labels for user gender identification;

This paper is organized as follows: Section 2 is the related works. Section 3 introduces the image-based method for gender identification, while Section 4 shows the measurements of user shared images on the datasets. Section 5 proposes and formulates the gender identification analytic system, followed by the experimental results on the system in Section 6. Section 7 concludes the paper.

2. RELATED WORKS

User behaviors in online social networks have been recently studied in terms of user gender. Users with the names ”Jack”, ”Jason” and ”Alan” are identified as males, while

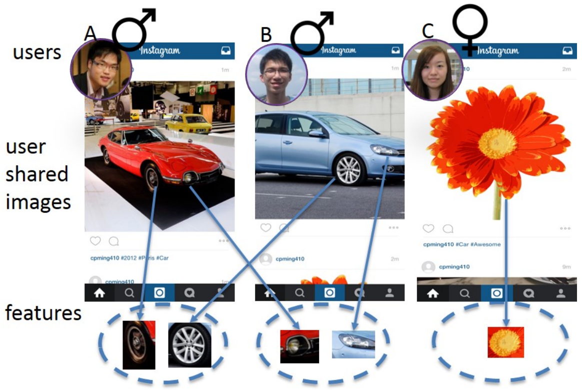

Fig. 1: Examples of the user shared image and their features

user with name ”Zoe” is identified as a female as these names are a good indicator of the user genders[Alowibdi et al. 2013a][Peersman et al. 2011][Liu and Ruths 2013]. An example can be found in Fig. 2 (a). However, as names can be fake or unusual, the use of such methods may not be applicable in some social networks such as Fotolog, where the real name is not available. Beside the user name, it is concluded the sharing behaviors on social media are different for males and females [Muscanell and Guadagno 2012][LOUGHEED 2012]. Text-related user content, such as tweets and chat messages, can also tell the gender of a user [Peersman et al. 2011][Liu and Ruths 2013][Burger et al. 2011][Schwartz et al. 2013][Goswami et al. 2009][Mukherjee and Liu 2010][Argamon et al. 2009][Rao et al. 2010]. Researches focus on building classification systems that make use of feature vectors generated by n-grams contained in user tweets, frequently use words, user generated tags and other information extracted from social media. The gender of users can be identified, as users with the same gender are likely to use similar wording in text. An example can be found in Fig. 2 (b). However, most of the works focus on text in English, and gender identification becomes challenging on global social networks with users from all over the world, such as Fotolog and Flickr. Another common method to identify user gender is to analyze user annotated tags on shared images[Li et al. 2008][Zhou et al. 2010], in which a tag in the

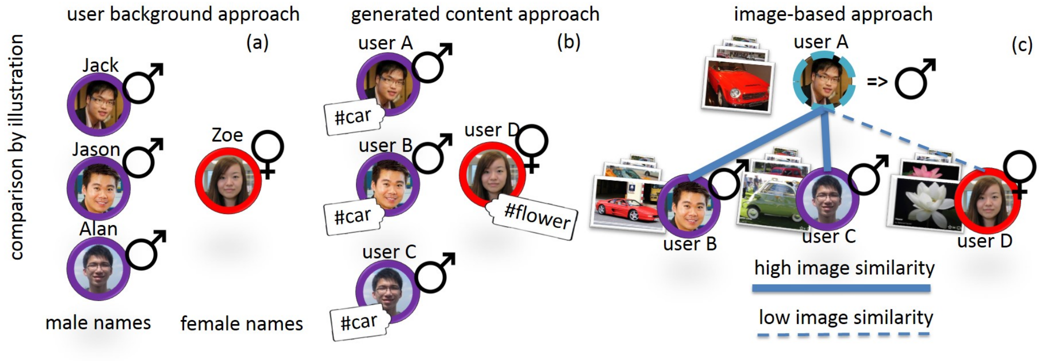

Fig. 2: Examples of gender identification with different approaches using, (a) user names, (b) tags, (c) image-based.

form of text is provided by a user as meta-data to describe the shared image. These tags can represent users, and the similarity between two users can therefore be calculated by the user annotated tags. An example can be found in Fig. 2. Users A, B and C are males, and they use the word, ”car”, a lot as they are interested in cars and are likely to share content related to cars. User D, who is a female, uses the word ”flower” a lot as she is interested in flowers. However, user generated tags are unreliable [Cheung et al. 2015a][Kennedy et al. 2007][Shepitsen et al. 2008][Zhang et al. 2012b] due to the use of different and inconsistent language, levels of detail and even inaccurate or missing words, which results in noisy or low performance. Collaborative filtering (CF) techniques [Zhang et al. 2012a][Sigurbjrnsson and Van Zwol 2008][Sang et al. 2012] are used to improve the tag accuracy for better gender identification, but only popular images are annotated by many users, while the rest are either not correctly annotated or missing annotation which leads to a poor gender identification performance.

Besides textual information, other information such as visual cues can also be used to identify gender, as users with different genders have different preferences. The choice of the layout of their personal pages, such as the color and profile picture, are also studied [Alowibdi et al. 2013b][Hum et al. 2011][Strano 2008], and it is concluded that users with different genders have different preferences on the visual appearance of their profile. The user shared images are also a good source for gender identification. Male users are more likely to share images reflecting the traits of activeness, dominance, and independence, while those shared by females reflect the traits of attractiveness and dependence[Rose et al. 2012]. An emerging image-based approach [Zhang et al. 2012b][Moxley et al. 2009] applies computer vision techniques to produce non-user generated labels that reflect the context of images. One of the techniques for tagging images with non-user generated labels is bag-of-features tagging, which applies a bag-of-feature framework, which has recently been proven to be an alternative to social graphs for connection discovery [Cheung et al. 2015a]. It is proven in [Cheung and She 2016] that using non-user generated labels is effective in de-anonymizing user identity, gender identification and origin inference. As the images shared by males and females are different, the visual content can also be used for gender identification [Cheung et al. 2015a]. In [You et al. 2014], user shared images are encoded using their group assigned (user annotated tags) by users (e.g., gadgets) and a probabilistic Latent Semantic Analysis (pLSA)-based approach. Users are represented by combining the distribution in pLSA, as the user profile. The user profile is then classified as a male or a female user profile. However, as discussed, the assigned groups of an image could be inaccurate, not available or limited by social media platform, for example, they are not available on Fotolog. Also, only vectors with all elements being positive values, such as the encoded vector from SIFT, can be used in pLSA. Those methods, such as CNN, that generate vectors with some elements being negative, cannot be used in pLSA. Fig. 2 (c) is an example of an image-based approach, in which the gender of user A is unknown. As user A, B and C are all interested in cars, they share a lot of images of cars, while user D shares images of flowers. As a result, the image similarity between user A and B, as well as user A and C are higher than that between user A and D. The gender of user A can be identified as male accordingly.

As there has previous been no study on how to utilize user characteristics and similarity distribution for gender identification, this paper is different from the previous works in the following: 1) scraped one more massive dataset, Flickr, in which there are 1.5 million user shared images, which has never been reported before; 2) measured the characteristic of user shared images and image similarity distribution to explain how they are related to gender identification and user gender; 3) studied how the choice of classifiers and other system parameters affect the performance of the gender identification, which has practical importances to system design; and 4) compared

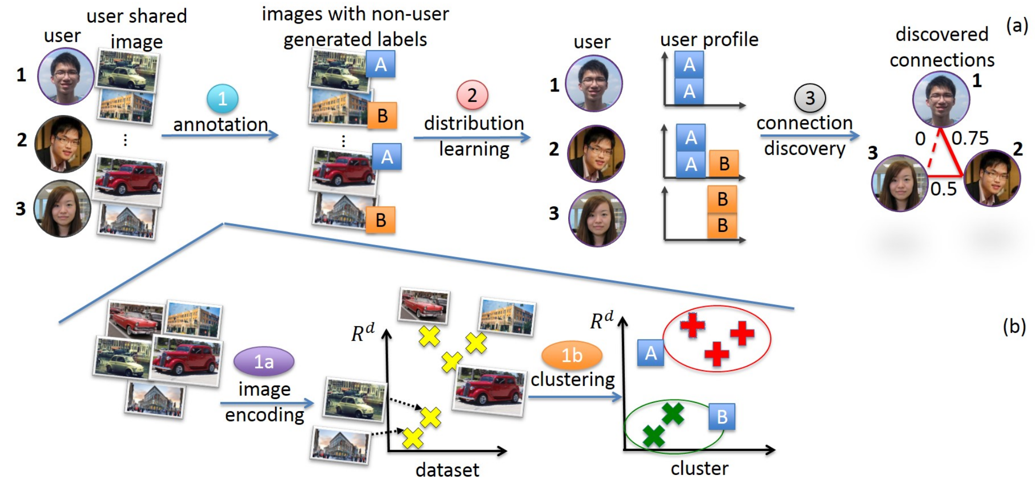

Fig. 3: Non-user generated labels for: (a) connection discovery, (b) gender identification from connections.

the proposed method with a pLSA-based algorithm[You et al. 2014] which requires user annotated tags.

3. NON-USER GENERATED LABELS FOR GENDER IDENTIFICATION

This section introduces the proposed method to identify user gender from the use of non-user generated labels. The images are first encoded as vectors, followed by a clustering process to group similar images, and the images among the same cluster are assigned with the same label.

3.1. Non-user Generated Annotation

The goal of non-user generated annotation is to represent each image with a non-user generated label to indicate the visual appearance of the image, as shown in step 1 of Fig. 3. The use of non-user generated labels is proven to be able to reveal a user online profile information, such as gender and origin. User shared images are first encoded as vectors, followed by an unsupervised clustering. Each cluster corresponding to a unique non-user generated label, indicates that they are visually similar. With these vectors obtained from the images, a clustering process is conducted using methods such as k-means. In this paper, scale-invariant feature transform (SIFT)[Wang et al. 2013][Lowe 2004] and CNN[Chatfield et al. 2014] are employed for the non-user generated annotation. Other encoding techniques, such as GIST and colour histograms, are also proven to be able to discover connections [Cheung et al. 2015b]. As the goal of tagging non-user generated labels is not object recognition, the clusters generated are not necessarily representing objects. They are clustered to represent the social signals hidden in the shared images. This is an advantage of the algorithm: there is no training needed for object recognition. For example. even if the social network is specific to certain kinds of images, such as food selfies, the proposed framework can still be used for connection discovery. This section introduces the 2 encoding techniques and the clustering processes.

3.1.1. Image Encoding using SIFT. SIFT-based features are detected first, and then the encoding is conducted. Unique local features are first obtained by technique such as the Harris Affine detector and KadirBrady saliency detector [Kadir and Brady 2001]. The extracted local features are relatively consistent across images taken under different viewing angles and lighting conditions. Those local features are then grouped and represented by visual words clustering techniques such as the Canopy clustering algorithm [McCallum et al. 2000] and LindeBuzoGray algorithm [Linde et al. 1980]. A K-means clustering with cosine similarity is used in our work. Instead of only assigning a feature to the nearest visual word (hard label), a soft label is used, in which features are represented by a vector. Soft labels have been used in many SIFT-based image problems [Kapoor et al. 2007]. The feature vector is obtained by summing the soft labels on the visual words of an image.

3.1.2. Image Encoding using CNN. CNN is motivated by the human' visual system, in which neurons are arranged to respond to small regions. The structure of CNN comprises of several layers of non-linear feature extractors, which are handcrafted with learnable weights and biases from data. Unlike the classification technique neural network, CNN is designed for images by adding a new convolution layer and out-perform all state-of-the-art techniques. In this work, the last layer of the CNN from [Chatfield et al. 2014] is used as the encoded vector. The network was trained based on ImageNet, which is a database containing millions of well-tagged images for object recognition. The CNN was trained for the recognition of 1000 objects. If the encoded vectors of two images are similar, it implies that they contain similar objects. However, the recognition results of CNN cannot be directly used for tagging, as images on social media are diverse and contain much more than 1000 objects.

3.1.3. Clustering and Non-user Generated Labels. Clustering groups' images that are visually similar through the similarity in their feature vectors is shown in step 1a of Fig. 3 (b). For example, when two images contain cars in the countryside, the feature vectors of the two images are similar in terms of the number of occurrences of each unique visual word. As a result, the two images will be assigned the same non-user generated label to indicate that they are visually similar. The framework applies a modified version of K-means, one of the most popular clustering algorithms, K-means++ [Arthur and Vassilvitskii 2007]. It applies the same procedure as K-means for the clustering, except for the cluster centroid initialization. Instead of randomly generating K cluster centroids as in K-means, they are picked using a weighted probability distribution, with probability proportional to the distance from the nearest cluster. It then iteratively assigns points to their nearest centroids, followed by a recomputing of the centroids until it converges. However, K-means does have its drawbacks in that the points lying far from any of the centers can significantly distort the position of the centroids and the number of centers must be known in advance. More discussion of this can be found in Section VI. The next step, labeling, assigns each cluster a unique non-user generated label so that those images with the same label are visually similar. The set of non-user generated labels of user shared images of user i, L_i, is obtained. L_i is a vector, with each element being the set of occurrences of a non-user generated label in the shared images of user i. The step is an unsupervised operation that analyzes user shared images without any manual input or process.

3.2. User Profile and Non-user Generated Labels

This section introduces how connections between two users can be discovered through non-user generated labels.

3.2.1. User Profile. The distribution of non-user generated labels, which reflects the content of a user shared images, is key in the discovery. The proposed method uses the number of occurrences of the non-user generated labels of the shared images of a user as his/her user profile, as in step 2 of Fig. 3 (a). A user i is represented by his/her user profile, L_i, and the distribution of the non-user generated labels that the user has is defined as:

$$L_i = \{l_1, ...l_k, ...l_K\} \tag{1}$$

where l_k is the number of occurrences of the k-th label among the shared images of user i, and K is the total number of labels, which is set to 500. Unlike comparing pairwise image similarity, which requires $O({N_I}^2)$, using non-user generated label only requires $O({N_u}^2)$, where N_u is the number of users. As the number of images is much higher than the number of users, using a user profile can reduce the runtime.

3.2.2. Connection Discovery and User Profiles. When the user profile of each user is established, the next step is to identify user genders based on the connections discovered from similarity, $S_{i,j}$, of users i and j, in which users who share highly similar images will have a high similarity. This requires a pairwise similarity comparison among user profile, the number of occurrences of non-user generated labels, and this is calculated using the following formula:

$$S_{i,j} = S(L_i, L_j) = \frac{L_i \cdot L_j}{||L_i|| \cdot ||L_j||} \tag{2}$$

where L_i and L_j are the set of non-user generated labels of the shared image in the user profiles of users i and j respectively. The similarity, $S_{i,j}$, among all users is then calculated as the discovered connections in step 3 of Fig. 3 (a).

3.3. Gender Identification using Discovered Connections

Based on the known genders and the discovered connections from non-user generated labels, the user gender can be identified The user profile of users with an unknown gender are first learned from the distribution of non-user generated labels, and their connections with users with a known gender are discovered. The identification can be based on voting approaches such as K-NN, in which users with high similarities vote for the gender of an unknown user. Standard machine learning techniques such as support vector machine (SVM), can also be used for gender identification, by using the known genders for training. The details of the gender identification are discussed in the coming sections.

4. USER SHARED IMAGES AND USER GENDER

This section first describes the dataset, followed by the characteristics of the user shared images and user gender. The third part of this section analyzes the user similarity distribution.

4.1. The Datasets

Fotolog and Flickr are two image-oriented social networks, which originate from the West. As image-oriented social networks, they allow users to share images and messaging, in which images are the only or the primary form of sharing. The two social networks are global networks, in which the users are from different parts of the world. As the user base on the two social networks is diverse, it is interesting to observe the user behaviors in these social networks. Fig. 4 shows the user interfaces of the two social networks. On Fotolog, as shown in Fig 4 (a) and (b), users can share images via the user interface of the web page and mobile applications. On Flickr, users can also

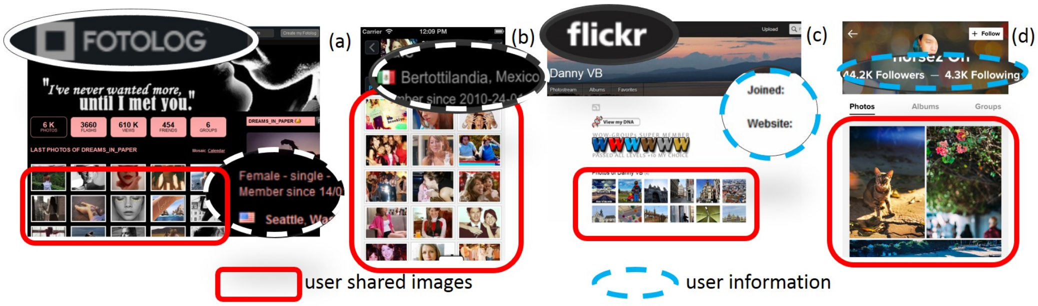

Fig. 4: The user interface of: (a) Fotolog web page, (b) Fotolog app, (c) Flickr web page, (d) Flickr app.

share images via both the user interface of the web page and the mobile application, as shown in Fig 4 (c) and (d), respectively. On the two social networks, a user can decide if they want to share information about their background, such as their name, gender, relationship status and other information, as indicated by the blue broken circle in Fig 4. The experiments in this paper involve 1,598,769 images shared by 6,036 users from Fotolog and 1,553,575 images shared by 1,414 users from Flickr, which were collected by Ruby-based scrapers during mid-2015. All the users, who have shared their gender, were selected randomly from a large set of users collected from their follower/followee relationships. The gender of users was also collected as the ground truth in the experiment.

4.2. Characteristics of User Shared Images

This section describes the characteristics of the user shared images and user genders. Fig. 5 (a) and Fig. 5 (b) show the distribution of the number of user shared images a user has, and the frequency of this number, on Fotolog and Flickr, respectively. It is observed that a few users share a large number of images, while most of the users share a few images only, and the same trend can be observed on both social networks, and is the same for most social networks. It is concluded that the selected users are a good representation of the users in the two social networks. There are 42% and 72% male users on Fotolog and Flickr, respectively.

4.3. User Similarity Distribution

Any two users can be considered as a user pair, and there are two types of pairs. The first type is user pairs with the same gender, that is, the two users both are males,

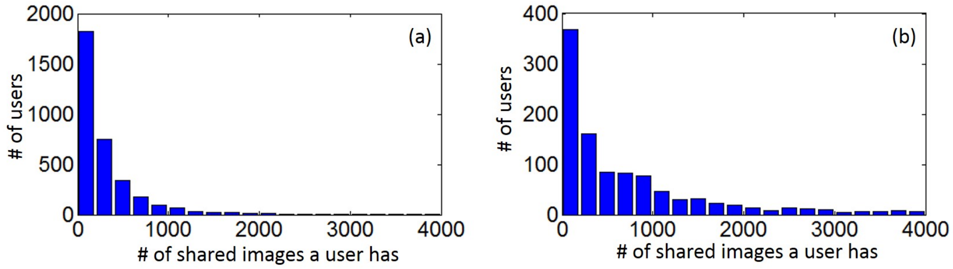

Fig. 5: Distribution of the number of shared images a user has: (a) Fotolog, (b) Flickr.

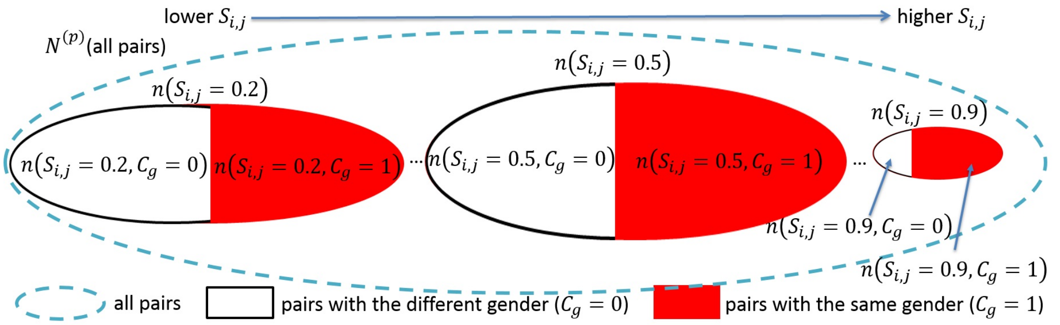

Fig. 6: The set of all pairs: the solid line ellipses are user pairs with a given $S_{i,j}$, with the white area being user pairs with different genders and the coloured areas are pairs with the same gender. The size of the ellipses represents the number of user pairs.

or both are females. The second type of user pairs is those with different genders, in which one of them is male and the other is female. The two types of user pairs can be considered as two classes, and the class of each pair, C_g, can be defined as:

$$C_g = \begin{cases} 1 & \text{if two users have the same gender} \\ 0 & \text{if otherwise,} \end{cases} \tag{3}$$

where $C_g = 1$ is the class in which the two users of the pair have the same gender, and $C_g = 0$ is the class in which the two users of the pair have different genders. Fig. 6 illustrates the idea. The ellipse in the broken line is the set of all pairs, with total $N^{(p)}$ pairs. It contains pairs with the same and different genders, as well as pairs with different $S_{i,j}$s. The solid line ellipses are those pairs with different ranges of $S_{i,j}$. The pairs in these ellipses can be classified into 2 types of pairs: pairs with different genders and pairs with the same gender. The coloured areas represent pairs with the same gender, while the other areas in white represent pairs with different genders. Hence, a white area, for a given $S_{i,j}$, contains $n(S_{i,j}, C_g = 0)$ pairs. $n(S_{i,j}, C_g)$ is a function of $S_{i,j}$ and C_g that gives the number of pairs with different genders when $C_g = 0$. The coloured area is the set of pairs with the same gender for a given $S_{i,j}$, with $n(S_{i,j}, C_g = 1)$ pairs.

It is interesting to investigate the distribution of $S_{i,j}$ among users. Fig. 7 shows an example of distribution using SIFT. Fig. 7 (a) and (b) show the distribution of the number of pairs with the same gender, given a $S_{i,j}$, $n(S_{i,j}, C_g = 1)$, on Fotolog and Flickr, respectively. Fig. 7 (c) and (d) show the distribution of the number of all pairs, given a $S_{i,j}$, $n(S_{i,j})$, on Fotolog and Flickr, respectively. It is observed that the distributions for $n(S_{i,j}, C_g = 1)$ and $n(S_{i,j})$ are similar on both social networks. They reach a peak value and decrease gradually, and there are only a few pairs have a high $S_{i,j}$. The probability that a pair of users have the same gender for a given $S_{i,j}$, $P(C_g = 1|S_{i,j})$, can be calculated as:

$$P(C_g = 1|S_{i,j}) = \frac{n(S_{i,j}, C_g = 1)}{n(S_{i,j})} \tag{4}$$

where $n(S_{i,j}, C_g = 1)$ and $n(S_{i,j})$ are numbers of pairs with the same gender and the total number of pairs, given a similarity $S_{i,j}$, as obtained in Fig. 7.

Fig. 8 shows $P(C_g = 1|S_{i,j})$ of the two social networks. It is observed that when a pair has a low $S_{i,j}$, they are less likely to be of the same gender than a high $S_{i,j}$. However, when $S_{i,j}$ is higher than a value, $P(C_g = 1|S_{i,j})$ increases with $S_{i,j}$, which indicates that two users with a higher $S_{i,j}$ are more likely to have the same gender. This idea can be

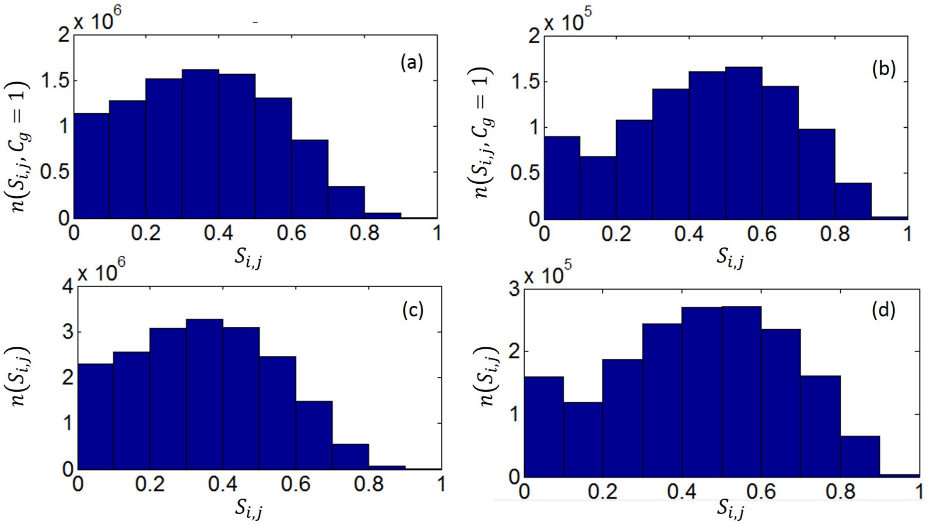

Fig. 7: Distribution of $S_{i,j}$ from SIFT among pairs of: (a) same gender on Fotolog, (b) same gender on Flickr, (c) different genders on Fotolog, (d) different genders on Flickr.

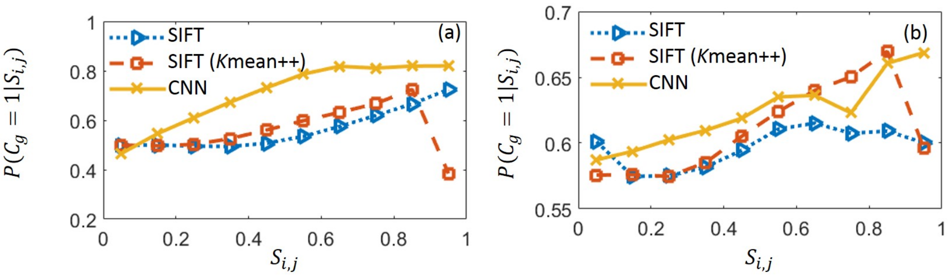

Fig. 8: Distribution of probability of being the same gender for a given similarity: (a) Fotolog, (b) Flickr.

illustrated by the area of the colored ellipses in Fig. 6: although the area (number of pairs) is smaller when $S_{i,j}$ is higher, a larger area of the coloured part (a higher portion of pairs with the same gender) can be observed. The same observation can be obtained from both social networks. Motivated by the above observations, an analytic system that utilizes these observations to identify user gender is proposed and the details are discussed in the next section.

5. PROPOSED ANALYTIC SYSTEM FOR GENDER IDENTIFICATION

This section introduces the analytic system flow and formulation of how gender can be identified based on the observations on the last section. This is a 3-stage (stages A to C) system as shown in Fig. 9. The first part is image collection, followed by similarity calculation using non-user generated labels. The third part is how to make use of the K-nearest neighbors approach based on known genders to identify the gender of a user. The fourth part introduces the implementation that annotates the 3 million user shared images with non-user generated labels.

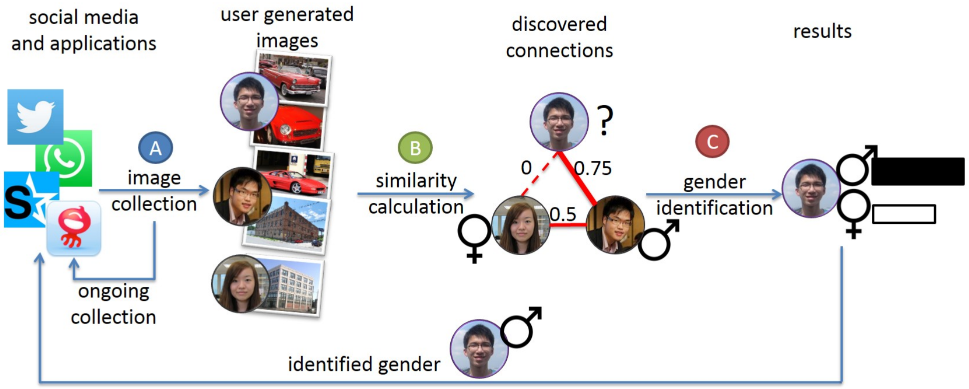

Fig. 9: System flow of the proposed analytic system, (a) image collection from social media; (b) connection discovery by collected images; (c) gender identification by connection discovery.

5.1. Image Collection

The proposed analytic system carries out data collection as shown in step A of Fig. 9, which shows the process to collect user generated images from social media applications such as Fotolog and Flickr. The images can be provided by the operators of the social media and mobile applications or collected through the API of the social networks. The user generated images can be shared in various forms, such as posted images on social media or images shared through instant messaging applications. On social networks such as Fotolog and Flickr, user generated images are those images shared by users. This process is ongoing, which means that user shared images are collected continuously.

5.2. Connection Discovery using Non-user Generated Labels

The objective of the image understanding is to annotate user generated images with non-user generated labels, as shown in step B of Fig. 9. The proposed system applies a computer vision approach to give a label to a user shared image, which is not affected by the language, culture or other characteristics of the user who shares the image, but is only based on the image's visual appearance. The accuracy of the user annotated tags is unreliable, sometimes even unavailable and the accuracy of discovery is affected. The proposed analytic system annotates user generated images with non-user generated labels. The set of user shared images of user i is processed by the proposed method, and a set of labels, L_i, is generated to represent user i. With L_i, the user profiles of users, can be calculated by Eq. 2

5.3. Gender Identification

Gender identification can be considered as a binary classification problem with 2 classes, male and female. The two classes can be defined as:

$$G_i = \begin{cases} 1 & \text{if user } i \text{ is a male} \\ 0 & \text{if user } i \text{ is a female} \end{cases} \tag{5}$$

Based on the observations that users with a high $S_{i,j}$ are likely to have the same gender, there is a motivation to apply the K-Nearest Neighbor (K-NN) to calculate $P(G_i|L_i)$, the probability that user i is male or female, given the distribution of labels of user i, $L,$1. The list of top J users with the highest $S_{i,j}$ with user i, $U_{i,J}$ is obtained for

K-NN. The probability that user i, given the distribution of non-user generated labels is a male, $P(G_i = 1|L_i)$, can be calculated as:

$$P(G_i = 1|L_i) = \sum_{j \subset U_{i,J}} \frac{G_j}{J} \tag{6}$$

where G_j equals 1 when user j is a male. Similarly, the probability that user i is a female, $P(G_i = 0|L_i)$, can be calculated as:

$$P(G_i = 0|L_i) = \sum_{j \subset U_{i,J}} \frac{(1 - G_j)}{J} \tag{7}$$

Hence, the binary classification of identifying the gender of a user i. The identified gender of user i, $\tilde{G}_i$, can be defined as:

$$\tilde{G}_i = \begin{cases} 1 & \text{if } P(G_i = 1|L_i) > P(G_i = 0|L_i), \\ 0 & \text{if otherwise,} \end{cases} \tag{8}$$

where $P(G_i = 1|L_i)$ and $P(G_i = 0|L_i)$ are the probability that user i is male or female given the distribution of the non-user generated labels, respectively. If $P(G_i = 1|L_i)$ is higher than $P(G_i = 0|L_i)$, G_i will be 1, that is, user i is classified as male, and if $P(G_i = 0|L_i)$ is higher than $P(G_i = 1|L_i)$, the user will be classified as female. It is also interesting to compare the result with a weighted K-NN using $S_{i,j}$ as the weight. Eq. 6 becomes:

$$P(G_i = 1|L_i) = \sum_{j \subset U_{i,J}} \frac{G_j S_{i,j}}{J} \tag{9}$$

Similarly, Eq. 7 becomes:

$$P(G_i = 1|L_i) = \sum_{j \subset U_{i,J}} \frac{(1 - G_j) S_{i,j}}{J} \tag{10}$$

$\tilde{G}_i$ can be obtained by Eq. 8. The identified genders are back to the social media and applications, and other applications, such as personalized recommendation, virality prediction and connection discovery, become possible with the identified gender.

5.4. Implementation and Design Requirements

This section introduces the implementation of the analytics that processes the image big data collected from Fotolog and Flickr. To annotate an image, processes such as extracting features from user shared images, become a challenge with the 3 million shared images. The system consists of a master and a list of slaves. The slaves are running on cloud-based platforms, such as Amazon EC2 and Microsoft Azure, and waiting for the master to connect to. Once the master starts, it checks the availability of each slave to obtain a list of available slaves, and the master is ready to get a task list as step 2. For example, when the task is to obtain the image vector of the 3 million images, the task is the link to access the images on an image server. These tasks are sent to the slaves for processing, and the master keeps checking the status of each slave periodically until at least one of the slaves has finished the task. When a slave receives the link from the master, it downloads the images and encodes the downloaded image into a feature vector. The feature vector is then sent to the master. Once the master receives the data from a slave, i.e., the feature vector. Steps 2 to 4 are repeated until all tasks in the list are completed. In this work, slaves are running on an Amazon EC2, with Ubuntu Server 14.04 LTS (HVM) using Compute optimized

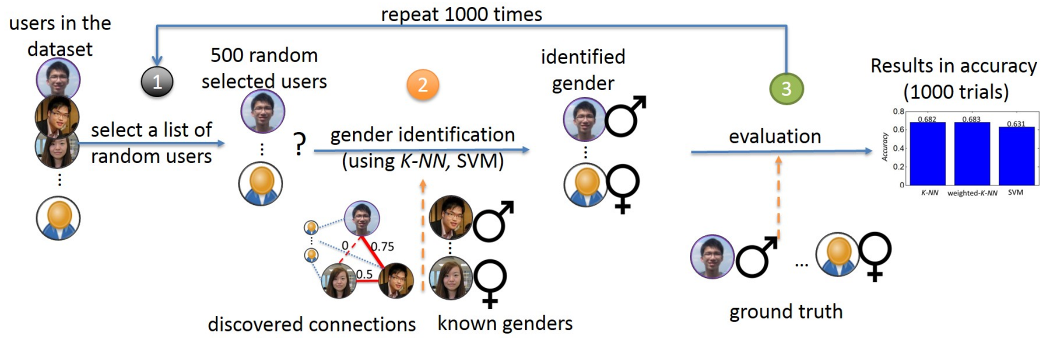

Fig. 10: Settings of the experiment. Steps 1) select 500/150 users as the testing set while the rest are for training; step 2) identify gender using machine learning approaches; Step 3) evaluate the identified gender with the ground truth. The result is the mean of 1000 trials

instances (c3.xlarge). Each machine consists of 4 virtual CPUs (vCPUs), and 7.5 GB of memory. Unlike general purpose virtual machines, such as m4.xlarge on Amazon EC2, compute optimized instances have a higher ratio of vCPUs to memory and are suitable for compute intensive tasks such as feature extraction. The processing of the 3 million user shared images involve six slaves, and takes more than a week with each slave having a 100% CPU utilization rate during the processing.

To identify the gender of a user, there are different considerations that may affect the performance of the identification and the system design, and the following discusses three important considerations. The first one is the number of images a user has. When a user shares more images, the system has a better understanding of the users, but it requires more computational resources to process the images. Hence, it is interesting to investigate how the number of images a user has affects the accuracy of the system. The second one is the choice of the classifier. As motivated by the observation, the K-NN is one the best approaches for this task. The third one is the number of K, the number of neighbors to be used in the K-NN. These parameters are investigated in the coming section.

6. EXPERIMENTAL RESULTS

This section shows the experimental results based on the 3 million images from the 2 image-oriented social networks, Fotolog and Flickr. The first part describes the setting of the experiments, while the second part is the results of the experiments. The third part discusses some important considerations for the experiments.

6.1. Experimental Settings

Fig. 10 shows the setting of the experiment, are there are 3 steps. In the first step, 500/150 random users are selected from the available users on Fotolog/Flickr as the testing set, while the rest are the training set. The discovered connections based on the non-user generated labels from user shared images and the gender of the users in the training set are the input of the gender identification. The gender of the users in the testing is obtained from gender identification, as shown in step 2 of Fig. 10. The identified genders are then evaluated by the ground truth, the gender of the users in the testing set, and then the accuracy, the number of correct identification divided by the total number of users identified, is recorded. These steps are repeated 1000 times

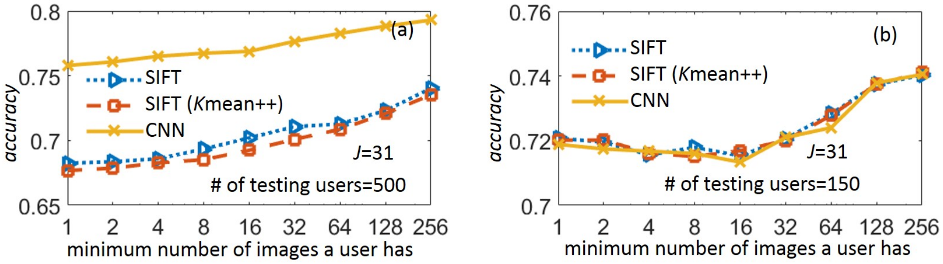

Fig. 11: Accuracy with different minimum numbers of user shared images on: (a) Fotolog, (b) Flickr.

and the averaged accuracy is recorded. Three methods are implemented and tested to show the efficiency of the proposed analytic system, J nearest neighbor (K-NN), SVM and random (Rand). Note that J is used here to distinguish another parameter of the system, K, the number of unique non-user generated labels. K-NN locates the top J users in the training data with the highest $S_{i,j}$ with the users in the testing data. A similar approach is also implemented, weighted K-NN (wK-NN), in which $S_(i,j)$ is used as the weight for the identification. The second approach is SVM, which is a standard machine learning technique that a classifier is trained from the training data using L_i, and the gender is identified based on the classifier. The third approach is a random approach, in which the gender of users is identified randomly. This is the baseline for the evaluation. The clustering process was conducted by K-means and K-means++ for comparison.

6.2. Gender Identification Results

This section shows the results of the three methods in the experiment. Tab. I shows the comparison of the averaged accuracy of the different methods on the two social networks, in which Tab. I (a) shows the results for Fotolog, while Tab. I (b) shows the results for Flickr. In the experiment, the number of unique non-user generated labels, K, is set to 500, other values of K will also work. More discussion on the values of K can be found in the next subsection. The users in the test have shared at least 16 images. As there are only 2 classes, by random guessing (Rand), there is a 50% chance to be correct. It is observed that K-NN and wK-NN give better results than SVM on both social networks, in all encoding methods. In terms of the method, CNN gives the best performance, achieving 77% accuracy. Note that J is set to 31, and more results can be found in the later part of this section. One of the reasons behind this may be that the two classes are not linearly separable. Interestingly, the wK-NN does not improve the accuracy much. The reason behind this is that when there are a number of users in the training data, the top few users with the highest $S_{i,j}$ have a similar value of $S_{i,j}$ and a similar result is obtained.

SNs	Fotolog			Flickr		
	CNN	SIFT(Kmeans++)	SIFT	CNN	SIFT(Kmeans++)	SIFT
K-NN	0.77	0.70	0.70	0.72	0.72	0.72
WK-NN	0.77	0.69	0.70	0.71	0.72	0.72
SVM	0.74	0.65	0.66	0.56	0.59	0.5
Rand	0.5			0.5		

Table I: Accuracy with different methods.

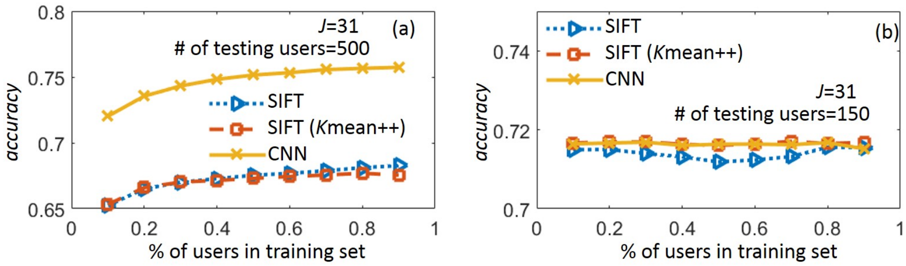

Fig. 12: Accuracy with the percentage of users in testing set on: (a) Fotolog, (b) Flickr.

As shown in Fig. 5, most of the users share only a few images, while a few users share many images. It is interesting to investigate how the number of user shared images affects the performance of the proposed analytic system. An experiment is conducted to investigate this effect by repeating the experiment in Fig. 10, but excluding users who have their number of share images smaller than a threshold. The averaged accuracy of 1000 trials with different minimum numbers of shared images is shown in Fig. 11. In each trial, 500/150 users are randomly selected as the testing set on Fotolog/Flickr, while the others are training set. Fig. 11 (a) shows the results on Fotolog, while Fig. 11 (b) shows the results on Flickr. It is observed that the accuracy increases with the minimum number of shared images on both social networks, for all image encoding methods. One of the possible reasons is that when there are more shared images, the system has a better understanding of the user and can obtain a more accurate calculation of $S_(i, j)$. Note that the experiment stops at 256, as the number of users in the training set will be smaller than the testing set, which gives a bad identification performance when the minimum number of images a user has is greater than 256.

Another experiment is conducted to show the effect of size of training set, that is, the number of users whose gender is known. As discussed in the last section, the wK-NN produces similar results to the K-NN method as the top few users with the highest $S_{i,j}$ have a similar value of $S_{i,j}$. It is interesting to discover the number of users in the database that is needed for the system to be accurate. Fig. 12 shows the accuracy of the percentage of users in the testing set of 1000 trials. For example, if the percentage is 10%, then 90% of the data is in the training set. Fig. 12 (a) shows the result on Fotolog, while Fig. 12 (b) shows the result on Flickr. In each trial, 500/150 users are selected randomly from Fotolog/Flickr with a different size of the training set of randomly selected users. It is observed that the accuracy becomes stable when the number of users is large on Fotolog and Flickr. This proves the robustness of the proposed analytic system when there are many users with known genders.

6.3. Showcase: pLSA-based Approach

In order to compare the effectiveness of the proposed system, the approach used in [You et al. 2014] is also implemented for comparison. Each image is encoded using pLSA [1] with 10 visual topics, as in [You et al. 2014]. The user profile is the mean of the images shared by the user, and then the same procedures is applied to the user profiles for gender identification. Unlike in [You et al. 2014], the datasets do not contain group information of images, so all images are encoded as one group.

Fig. 13 shows the comparison of the averaged accuracy of different image encod-

[1] code is downloaded from http://www.robots.ox.ac.uk/ vgg/software/, and modified to process the huge number of images

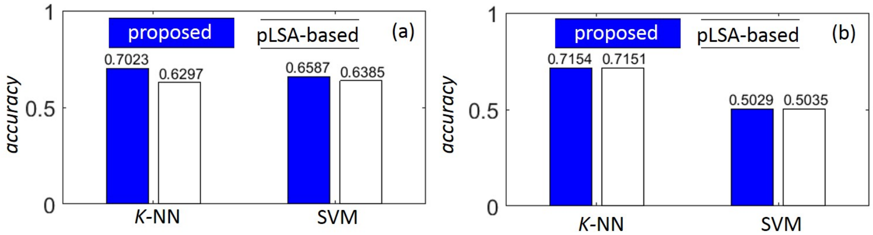

Fig. 13: Averaged accuracy with the proposed approach and the pLSA-based approach[You et al. 2014] for image encoding: (a) Fotolog, (b) Flickr.

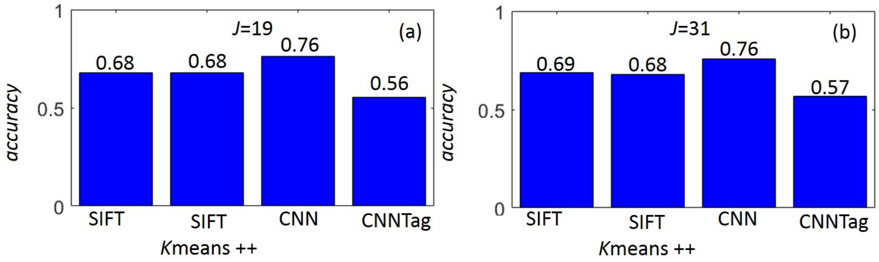

Fig. 14: Averaged accuracy with different approaches and J: (a) 19, (b) 31.

ing approaches on the two social networks, in which Fig. 13 (a) shows the results for Fotolog, while Fig. 13 (b) shows the results for Flickr, with K-NN and SVM for classification. It is observed that the proposed approach performs better than the pLSA-based approach in Fotolog, on both K-NN and SVM for classification using SIFT-based encoding, as proposed in [You et al. 2014]. The proposed approach is 11.5% better than the pLSA-based approach, while on Flickr, it achieves similar performance to the the pLAS-based approach. The results prove that the proposed approach can achieve better results than the pLSA-based approach.

6.4. Showcase: Object Recognition for Gender Identification

As CNN performs very well on Fotolog, it is interesting to investigate whether the 1,000 objects that can be recognized in [Chatfield et al. 2014] are useful for gender identification. The user shared images in the Fotolog dataset are classified by CNN, and the tag of an image is the one with the highest score (CNNTag). There are 1,000 classes, where they are trained using ImageNet. The histogram of the occurrence of the tags is used as the user profile and the similarity is computed accordingly. The result is shown in Fig. 14, in which Fig. (a) and (b) show J=19 and 31, respectively. The results prove that the proposed approach gives better performance than CNNTag. One of the reasons that CNNTag cannot give a good performance is that there are images with a large variety of objects shared on social media. The proposed approach has an advantage that it can be applied to domain-specified social media unsupervised. For example, on a social network that focuses on food, a classifier cannot provide a good result.

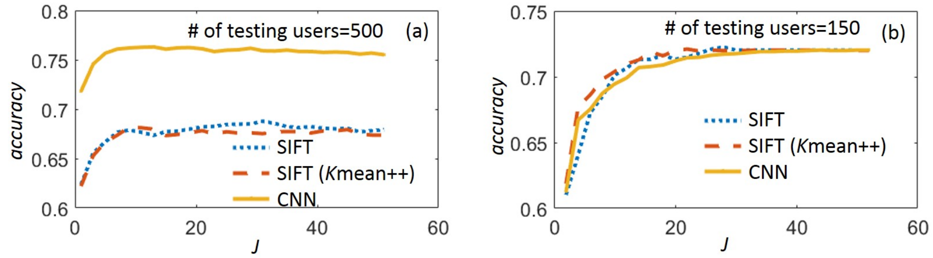

Fig. 15: Accuracy with different J on: (a) Fotolog, (b) Flickr.

6.5. Discussion

This paper has successfully proved and characterized the phenomenon that two users are likely to have the same gender if their shared images are similar. It then formulated and developed the results into a practical analytic system to identify user genders by mass user shared images from real-world social networks. There are two parameters that need to be pre-defined for the proposed analytic system: the first one is the number of non-user generated labels. A small value could make two images that are not similar be annotated with the same label, while a large value will annotate two similar images with different labels. In this work, a fixed K of 500, is used. Strategies, such as that in [Jurie and Triggs 2005], combine the advantages of on-line clustering [Meyerson 2001] and mean-shift [Comaniciu and Meer 2002] in an under sampling framework [Estabrooks et al. 2004]. The method does not require knowing K in advance, and performs better than K-means in image categorization [Jurie and Triggs 2005]. More investigations are needed on how to select K.

The second parameter, J, the top J user to be used in K-NN, also needs to be predefined. An experiment is conducted to check the results with different J and the results are shown in Fig. 15, in which Fig. (a) and (b) are the results of Fotolog and Flickr, respectively. In the experiment, the gender of user i is identified with the list of user J with the highest $S_{i,j}$ with user i, and the test is repeated with all users in the dataset. It is observed that the accuracy increases with J, and become stable when J is large. The results indicate that J should be set to greater than 10.

To obtain the list of similar user from millions of users requires computing of computationally intensive tasks. Millions of images are generated every day, so an analytic system that can process big data with scalable storage design is needed for collecting and processing these user shared images, such as a cloud-assisted system to handle profile learning and similarity calculation [Jie et al.] for a scalable system. The feature vectors are first split into multiple blocks in the Hadoop Distributed File System (HDFS) and distributed to virtual machines (VMs) for the K-means clustering process. Each VM is in charge of computing the distribution of different labels for several users, and the user similarity is also calculated in a distributed way. Other possible extensions include selecting gender-sensitive features and labels instead of directly applying all labels and features obtained.

7. CONCLUSIONS

This work investigated 1,598,769 user shared images by 6,036 users from Fotolog and 1,553,575 user shared images by over 1,414 users from Flickr, two image-oriented social networks. Based on the intensive measurements and characterizations on the user shared images, this work proved a phenomenon that two users with a higher similarity of their shared images are likely to have the same genders. From this phenomenon, a practical analytic system using non-user generated labels to identify user genders by their shared images is proposed and verified by the 3 million shared images. It is proven that the accuracy of the analytic system is up to 80% and 11.5% better in Fotolog. This analytic system solves the problem in many social media applications, that user gender may be kept private or not explicitly specified. With advanced mobile devices such as smartphones and wearables, billions of user shared images are generated by individuals in many social networks today. These findings are useful for information or services recommendations in any social network with intensive image sharing.

ACKNOWLEDGMENTS

The authors would like to thank Allan Jie from HKUST-NIE Social Media Lab for his support.

REFERENCES

Jalal S. Alowibdi, Ugo Buy, and Paul Yu. 2013a. Empirical evaluation of profile characteristics for gender classification on twitter. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, Vol. 1. IEEE, 365–369.

Jalal S. Alowibdi, Ugo Buy, and Paul Yu. 2013b. Language independent gender classification on Twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*. IEEE, 739–743.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52, 2 (2009), 119–123.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1301–1309.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).

Ming Cheung and James She. 2016. Evaluating the Privacy Risk of User Shared Images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 12, 58 (2016).

Ming Cheung, James She, and Zhanming Jie. 2015a. Connection discovery using big data of user-shared images in social media. *Multimedia, IEEE Transactions on* 17, 9 (2015), 1417–1428.

Ming Cheung, James She, and Li Xiaopeng. 2015b. Non-user Generated Annotation on User Shared Images for Connection Discovery. In *Green Computing and Communications (GreenCom), 2015 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*. IEEE.

Dorin Comaniciu and Peter Meer. 2002. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24, 5 (2002), 603–619.

Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence* 20, 1 (2004), 18–36.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers age and gender. In *Third International AAAI Conference on Weblogs and Social Media*.

Noelle J. Hum, Perrin E. Chamberlin, Brittany L. Hambright, Anne C. Portwood, Amanda C. Schat, and Jennifer L. Bevan. 2011. A picture is worth a thousand words: A content analysis of Facebook profile photographs. *Computers in Human Behavior* 27, 5 (2011), 1828–1833.

Zhanming Jie, Ming Cheung, and James She. A cloud-assisted framework for bag-of-features tagging in social networks. In *Network Cloud Computing and Applications. Proceedings. 4th IEEE Symposium on*. IEEE.

Frederic Jurie and Bill Triggs. 2005. Creating efficient codebooks for visual recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, Vol. 1. IEEE, 604–610.

Timor Kadir and Michael Brady. 2001. Saliency, scale and image description. *International Journal of Computer Vision* 45, 2 (2001), 83–105.

Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. 2007. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. 2007. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the 15th international conference on Multimedia*. ACM, 631–640.

Xin Li, Lei Guo, and Yihong E. Zhao. 2008. Tag-based social interest discovery. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 675–684.

Yoseph Linde, Andres Buzo, and Robert M. Gray. 1980. An algorithm for vector quantizer design. *Communications, IEEE Transactions on* 28, 1 (1980), 84–95.

Wendy Liu and Derek Ruths. 2013. What's in a name? Using first names as features for gender inference in Twitter.. In *AAAI Spring Symposium: Analyzing Microtext*.

ERIC LOUGHEED. 2012. Frazzled by Facebook? An exploratory study of gender differences in social network communication among undergraduate men and women. *College Student Journal* (2012), 88–99.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 169–178.

Adam Meyerson. 2001. Online facility location. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*. IEEE, 426–431.

Emily Moxley, Jim Kleban, Jiejun Xu, and BS Manjunath. 2009. Not all tags are created equal: Learning Flickr tag semantics for global annotation. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 1452–1455.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*. Association for Computational Linguistics, 207–217.

Nicole L. Muscanell and Rosanna E. Guadagno. 2012. Make new friends or keep the old: Gender and personality differences in social networking use. *Computers in Human Behavior* 28, 1 (2012), 107–112.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 37–44.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, 37–44.

Jessica Rose, Susan Mackey-Kallis, Len Shyles, Kelly Barry, Danielle Biagini, Colleen Hart, and Lauren Jack. 2012. Face it: The impact of gender on social media images. *Communication Quarterly* 60, 5 (2012), 588–607.

Jitao Sang, Changsheng Xu, and Jing Liu. 2012. User-aware image tag refinement via ternary semantic analysis. *Multimedia, IEEE Transactions on* 14, 3 (2012), 883–895.

H. A. Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin E. Seligman. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8, 9 (2013), e73791.

Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. 2008. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems (RecSys '08)*. ACM, New York, NY, USA, 259–266. http://doi.acm.org/10.1145/1454008.1454048

Brkur Sigurbjrnsson and Roelof Van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 327–336.

Michele M. Strano. 2008. User descriptions and interpretations of self-presentation through Facebook profile images. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 2, 2 (2008), 5.

Zhi Wang, Lifeng Sun, Wenwu Zhu, Shiqiang Yang, Hongzhi Li, and Dapeng Wu. 2013. Joint social and content recommendation for user-generated videos in online social network. (2013).

Quanzeng You, Sumit Bhatia, Tong Sun, and Jiebo Luo. 2014. The eyes of the beholder: Gender prediction using images posted in Online Social Networks. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE, 1026–1030.

Xiaoming Zhang, Zhoujun Li, and Wenhan Chao. 2012a. Tagging image by merging multiple features in a integrated manner. *Journal of Intelligent Information Systems* 39, 1 (2012), 87–107.

Xiaoming Zhang, Xiaojian Zhao, Zhoujun Li, Jiali Xia, Ramesh Jain, and Wenhan Chao. 2012b. Social image tagging using graph-based reinforcement on multi-type interrelated objects. *Signal Processing* (2012).

Tom C. Zhou, Hao Ma, Michael R. Lyu, and Irwin King. 2010. UserRec: A User Recommendation Framework in Social Tagging Systems.. In *AAAI*.

Received May 2016; revised Oct 2016; accepted TBC