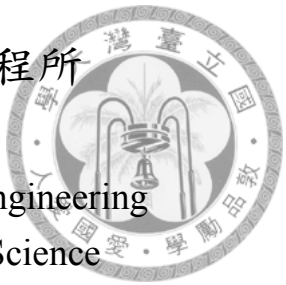


國立臺灣大學電機資訊學院資訊工程所

博士論文

Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Doctoral Dissertation

基於使用者生成多媒體內容之巨量資料分析

Human-Centric Data Analytics from User-Contributed Media Collections

陳殷盈
Yin-Ying Chen (A.K.A. Yan-Ying Chen)

指導教授：
徐宏民博士　廖弘源博士
Advisor:
Winston H. Hsu, Ph.D.　Hong Yuan Mark Liao, Ph.D.

中華民國 103 年 5 月
May, 2014


國立臺灣大學
教育學程・愛國愛人

Acknowledgements

Thank Prof. Mark Liao for his great support to my research career. Without his encouragement, I could not purchase my research dream without hesitation. Thank Prof. Shih-Fu Chang for his kindly help in my visiting research. I did learn a lot during the wonderful time in DVMM. Thank Prof. Winston Hsu for everything that happened in my Ph.D. life. He let me know courage and passion are the keys to reach one's potential. In addition, I would like to express the deepest appreciation to the committee of my dissertation. Their precious comments help to considerably improve the quality of this thesis. Finally, I am so grateful to have many dear friends in Academia Sinica, Columbia University and NTU, particularly my partners in MiRA. Because of their warm regards, happiness was always with me in the past five years.


國立臺灣大學
教育部・學術倫理・愛國愛人

摘要

數量持續成長的社群媒體用戶基於共享和社交的目的，貢獻大量人物照片。這些人物多媒體資料（如旅行照片和家人影片等）保有豐富的人群活動資料，對行動推薦系統，個人化，廣告和更多以人為中心的應用非常有利。有鑒於這些強烈需求，我們提出利用影像中自動偵測獲得的人物訊息（如人臉屬性，人物群體類別，視覺情感概念）來幫助社交多媒體分析。本計畫進一步結合計算社會學和認知心理學來了解社群使用者所提供的視覺資料中所挖掘出的知識訊息。最後，我們並展示利用百萬規模的社群影像及其周邊資訊（如地理位置，時間，標籤和評論）來幫住人物特徵分析，人口統計調查和社群情感運算。就我們所知，這是第一個利用大規模社交視覺資料來幫助分析使用者行為的研究工作。

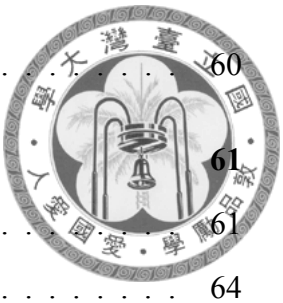

國立臺灣大學
教育學程品德愛國

Abstract

A growing population of the Internet users are contributing a huge amount of photos and videos to social media for the purpose of sharing and social communication. These big human-centric media collections such as travel photos and family videos retain abundant people activities inherently beneficial for mobile recommender system, personalization, advertisement and more people-related applications. Witnessing these strong needs, we propose to exploit the human-centric contexts automatically detected from visual content, e.g., people attributes, social group types and visual concepts, for social multimedia analytics. The proposed approach further incorporates computational sociology and cognitive psychology to understand the knowledge mined from the visual content contributed by real users. Finally, we demonstrate its effectiveness for user profiling, demographic investigation and social affective computing by using million-scale social images and the associated metadata (i.e., geo-locations, time stamps, tags and comments) crawled from social media. To the best of our knowledge, this is the first work addressing how large-scale visual contexts can help user profiling and improve user behavior analysis.


國立臺灣大學
教育學術・愛國愛人

Contents




國立臺灣大學
教學勵品愛國愛人

List of Figures

List of Tables


國立臺灣大學
教育學生愛國愛人

Chapter 1

Introduction

Social media, as a major platform for virtual interactions among the Internet users, provides a rich repository of human-centric information as well as people activities. Witnessing plentiful knowledge in such big media collections, recent studies have shown promising progress in social media analytics from varied research aspects, including demographic analysis [13, 60], spatial data mining [90, 22], sentiment classification [58] and more measuring as well as interpreting approaches for large-scale user-contributed data. These studies mainly target text-based media, e.g., blogs, tweets and reviews, but few of them address the knowledge mined from visual content such as images and videos.

Different from the previous studies, we aim at further considering the plentiful visual contexts in social multimedia for human-related data analytics. With the prosperity of multimedia-sharing websites, like Flickr and Youtube, the volume of community-contributed multimedia has increased drastically. Beyond text-based media, images and videos usually comprise more interactions and associations among users and their social communities (e.g., family, friends) in real life. In our study from more than 17 million photos collected from Flickr using the keyword "family", we found that around 60% of them contain at least one person. Such collections not only appear more visually interesting to human [38] but also visually reveal rich people activities. These phenomena motivate more attention on multimedia analytics and encourage visual content analysis towards more scalable to cope with the big media collections in social media.

Moreover, these publicly available images and videos are commonly associated with

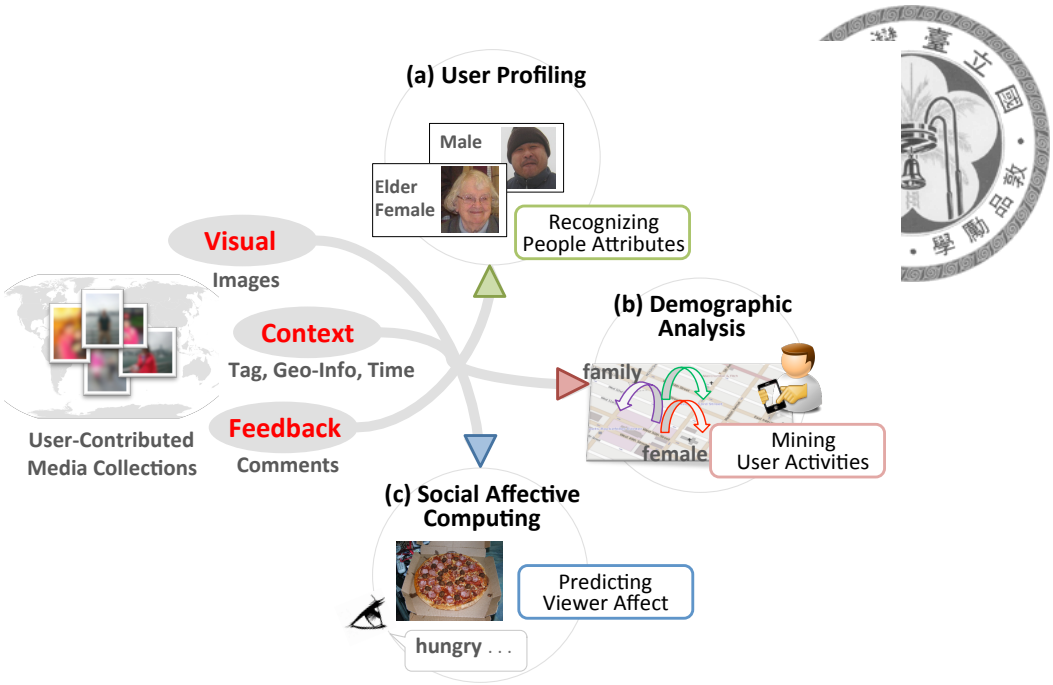
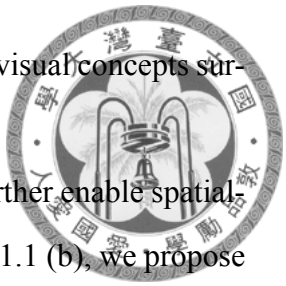


Figure 1.1: The proposed human-centric data analytics by leveraging the big user-contributed media collections.

rich metadata such as tags, time-stamps, geo-locations and comments. These overwhelming amounts of multi-relational contexts, though noisy, are tremendously essential for many multimedia applications including annotation, searching, marketing, advertising and recommendation. The studies [55, 29] in visual content analysis are moving forward the conception of "crowdsourcing" – obtaining needed content and annotations by soliciting contributions from online communities, rather than from costly manual labeling. Our previous work has demonstrated the effectiveness in learning people attributes by crowdsourcing metadata along with user-contributed images [18]. Most importantly, this mechanism shows the power of incrementally learning recognition models and thus benefits more precise social media analytics.

One of the significant social media analytics is user profiling by latent or explicit people attributes. User profiling based on the online reviews, tweets, blogs [13] has been shown promising and more cost-effective compared to traditional user investigation. We further address user profiling (cf. Figure 1.1 (a)) by involving the contexts detected from social visual content. These visual contexts possess several additional benefits, including (1) less bias – taking the more people shown in images into account rather than the photo



publisher him/herself and (2) more content – considering informative visual concepts surrounded with the people in image content.

Meanwhile, like millions of human-sensors, geo-tagged photos further enable spatial-temporal pattern mining aware of demographics. As shown in Figure 1.1 (b), we propose to exploit people attributes (e.g., gender), social group types (e.g., family) and semantic concepts (e.g., scene) detected from visual content for more preference-aware or topic-aware travel pattern mining. Furthermore, because these travel patterns categorize people activities by visually detected attributes rather than metadata only, they are less affected by data sparsity problem – little or even no labels at cold-start stage.

Finally, consumer photos usually contain human-scale scenes or objects which easily trigger stronger human affects in the audience (i.e., the social media users). Taking Fig. 1.1 (c) as an example, after viewing the visual content with "yummy food," the viewers are very likely to respond with a comment "hungry" (viewer affective concept). We target what viewer affective concepts will be evoked after the semantics expressed in visual content is viewed. The proposed affective data analytics capture the interactions among publishers and viewers via visual and text content, which is complementary to pure text-based sentiment and affect analysis in social media.

To sum up, we aim at human-centric data analytics from user-contributed media collections to achieve user profiling (Figure 1.1 (a)), large-scale demographic study (Figure 1.1 (b)) and social affect computing (Figure 1.1 (c)), which are still very challenging and important problems both in academic and industry research. In the remainder of this proposal, we will discuss the literature review (Chapter 2) and proposed methodologies for (a) learning people attributes by crowdsourcing (Chapter 3), (b) demographic anaylsis by using social visual content (Chapter 4), (c) predicting viewer affective concepts and comments (Chapter 5), and finally closing with a conclusion.


國立臺灣大學
愛國愛學・勵品教人

Chapter 2

Literature Review

This proposal targets the emerging research field – the enterprise of visual content analysis and social media analytics, especially, with the focus on human-centric media collections. We will discuss the promising supports from the two related research areas, respectively, followed by the review of our studies in this novel field of social multimedia analytics.

Analyzing people, human-scale objects and concepts in visual content has been an active research subject for decades because it is one of the enabling technologies for visually understanding and interpreting the environment, interactions and behavior of human. One of the significant field focuses people attribute analysis such as facial attribute detection (e.g., gender, age, race) [41] and clothing attribute classification [16] (e.g., dress, outerwear). Our preliminary studies [18, 20], further show the possibility of learning people attributes by crowdsourcing social media without tedious manual annotations for collecting training data. These people attributes greatly benefit profiling the attributes of social media users from the visual aspect. In addition, recent studies in visual content analysis has been moving forward to the psychological perspective; for example, analyzing image interestingness [38] and emotions [10]. Applying these human-centric measurements in social media augments the knowledge interpretation with plentiful visual contexts and human affect, which are essential but rarely addressed in the previous work.

Social communication motivates users to share their ideas, comments and photos with their social circles via multiple media format. Such publicly available data provide a

cost-effective way to obtain demographic information – the statistics for the specific user groups in certain events or locations such as restaurants, hotels, tourist attractions, etc. The previous studies [22, 90] have demonstrated that popular travel landmarks can be discovered by data-driven approaches in user-contributed travel logs. In addition, consumer groups (e.g. family, friends, couple) have quite different preferences when searching for travel accommodations [46]. However, they mainly focus on analyzing text-based metadata or logs without considering the visual contexts mined from the large-scale community-contributed travel photos.

Our previous studies [21, 17] have shown the potential for exploiting visual content analysis to improve social media analytics. In [21], we demonstrate that face attributes detected from travel photos are promising for more accurate travel recommendation because these attributes carry more personalization factors for travel-pattern mining. In addition, the social relationships shown in image content [19] are informative cues for mining group preferences [17] from big group photo collections. These preliminary results encourage us to incorporate more people attributes (e.g., actions, accessories, clothing styles), human-scale concepts (e.g., food, scenes, transportation), visual affective measurement (e.g., beautifulness, interestingness, emotions) in interpreting human-related analysis from social media. The detailed related works are discussed in each section.

Chapter 3

Learning facial attributes by weakly labeled images

3.1 Introduction

Beyond the low-level features commonly used for face recognition, the rich set of facial attributes such as gender, race, age, beard, smile, etc., have been shown to be very promising for characterizing designated persons [41] as well as for identity verification [42]. Moreover, facial attributes make photo management easier. Lei et al.[45] designed an efficient framework to retrieve photos of the target persons by graphically specifying the face icons with attributes on a query canvas. In addition, the statistics of (automatically detected) facial attributes from certain user groups (e.g., young girls) can approximate users' preferences. Cheng et al.[21] proposed a travel recommender by mining people attributes from community-contributed photos. Combining with specific time, location, etc., the plentiful facial attributes greatly benefit mining consumer activities from large-scale and less organized photos.

Prior research for facial attribute detection [54, 5, 41] solely relied on supervised learning with manually annotated training photos, which is very time-consuming and labor intensive. On average, manual annotation requires 5-6 seconds for tagging a photo or 15 seconds through gaming-based annotations [82]. Furthermore, manual annotation is sub-

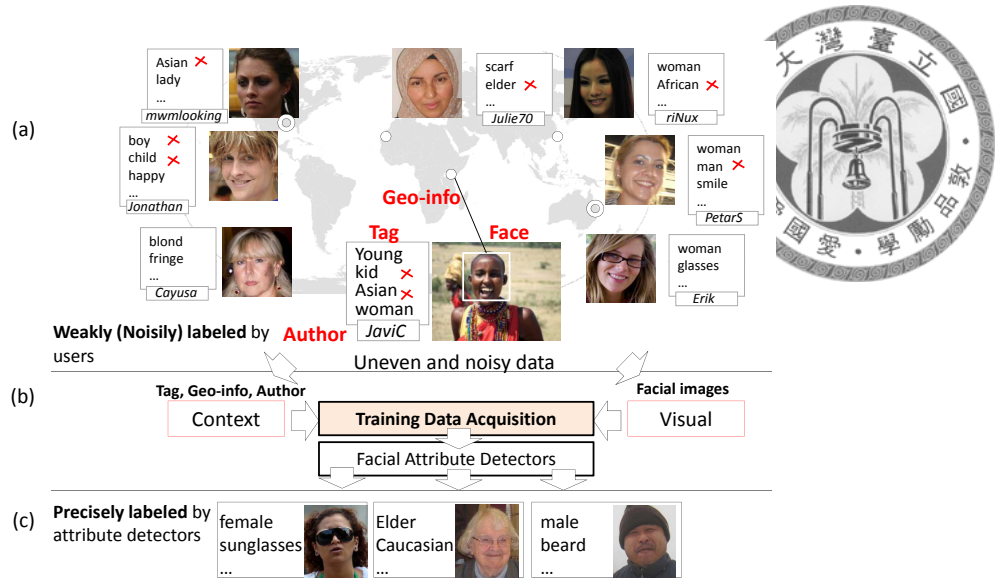


Figure 3.1: Goal – automatically acquiring training images for generic facial attribute detection by leveraging visual and contextual cues from publicly available community-contributed photos in an unsupervised manner. (a) Besides visual appearances in Internet images, the rich contextual cues such as tags, geo-locations are promising to ease the data bias problem in training facial attributes. Though contextual cues are noisy (e.g., the crossed tags), we aim to mine the effective training images from them. (b) Combining visual relevance and contexts to rank the effective and visually diverse training images. (c) Learning and detecting generic facial attributes from the automatically acquired training images for each attribute.

jective and biased; for example, being restricted to limited domains or locations. The problems get worse when preparing to analyze a large set of facial attributes as proposed in [41, 42].

With the prevalence of capturing devices and photo sharing services such as Flickr and Youtube, the volume of multi-media resources have been dramatically increased. There are reportedly more than four billion images in Flickr and even more than 70,000TB broadcast video data generated every year [51]. Such ultra-large-scale multimedia brings about profound social impact upon the society and has potential for easing the burden of large-scale training image acquisition [27, 63, 55] by means of freely available user-contributed data. In this work, we aim to acquire effective training images from community-contributed photos for facial attribute detection. It is promising since social media are full of user activities via the photos associated with tags, comments, locations, etc. However, simply acquiring training images by keywords (e.g., "beard") brings significant amount of false

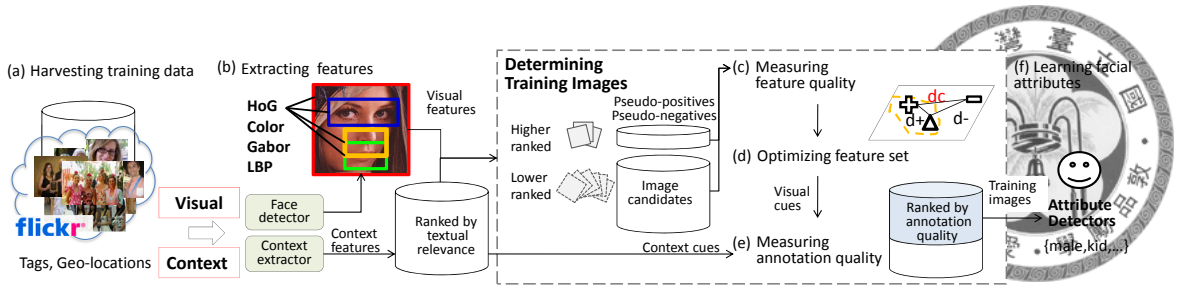


Figure 3.2: The framework to automatically acquire training images for learning generic facial attributes includes: (a) harvesting photos and the associated context information (e.g., tags, GPS) from the community-contributed photos by keyword queries as the initial candidates, (b) extracting the visual features from the detected (frontal) faces and the context features from the associated text as well as geo-locations, (c) measuring feature quality according to the discriminability voting results from image candidates over multiple visual feature spaces, (d) optimizing feature set of a designated attribute for measuring the visual relevance, (e) fusing the visual relevance (estimated in (d)) and the contextual cues extracted in (b) to estimate and rank the annotation quality, and (f) learning generic facial attributes by the automatically acquired training images.

positives due to an uncontrolled annotation quality; learning with such noisy data degrades the accuracy of facial attribute detectors.

With an effective feature representation (e.g., supervector [83]) to a designated facial attribute (e.g., age), examining visual relevance has been shown to be promising to reject certain false positives in the previous research [55]. In reality, users are not expected to predetermine well which features are important to a designated attribute, for example, edge features for detecting eyeglasses [80] and texture features for estimating age [34]. To enable automatic training image acquisition to be adaptive to various facial attributes, we propose to automatically select effective features from a rich set of visual features, which are potential feature candidates for different facial attributes. The proposed feature selection mechanism first measures the discriminant capability of each visual feature by *discriminability voting* – voting upon unlabeled images by pseudo-positives (negatives) retrieved by textual relevance – and then it selects effective features according to the estimated discriminant capability and the degree of mutual similarity. Discriminability voting can reduce the interference of noisy labels in the training images and does not require heuristic thresholds. Therefore, it has better generalization capability for multiple feature modalities. Another critical deficiency in prior research is rejection of false posi-

tives by the use of visual relevance only (e.g., [27, 55]) because that may cause the set of acquired training images to be dominated by color or other visual features (cf. Fig 3.8(b)). The above mentioned images do bring marginal improvement for learning facial attribute detectors. However, it may cause data skew at the same time. Therefore, we propose to exploit the rich context cues (e.g., tags, geo-locations, etc.) along with the community-contributed photos to increase the degree of diversity for the training images (Fig 3.8(c)). The proposed approach is conducted in an unsupervised manner and most importantly, it can be applied to different facial attributes.

For the proposed framework, as shown in Fig. 3.1, we first measure the quality of each visual feature given a noisy set of keyword-retrieved (e.g., "beard") training image candidates. Optimized by discriminability and mutual similarity, the selected features are then used to evaluate the annotation quality of the training image candidates from the visual aspect. Second, context information is further augmented to ensure the degree of diversity and the quality of automatically collected training images. Experiments show that the proposed method – balancing visual and context cues, outperformed two baseline approaches (1) measuring textual relevance (text-based) and (2) measuring visual reconstruction error via Principal Component Analysis (PCA-based); the error rates are reduced by up to 23.24% and up to 38.50% (relative improvement), respectively. More excitingly, we found that the facial attribute detectors trained by the proposed method are competitive with those trained by the use of manually annotated photos. Note that our work requires no manually collected training images but automatically mines semantically related training images from the initial candidate photos and their associated metadata retrieved by facial attribute keywords. The primary contributions of the work include:

- Devising a generic framework for learning numerous facial attributes by automatically acquiring training images from freely available and growing community-contributed photos without tedious manual annotations.

- Proposing a robust-to-noise feature selection approach by discriminability voting to measure visual relevance adaptive to different facial attributes (Sec. 3.4).

- Balancing visual relevance and contextual cues along with community-contributed photos to optimize automatic training image acquisition (Sec. 3.5).

- Experimenting on consumer photo benchmarks and showing great improvement in accuracy for facial attribute detection and superiority to its counterpart which requires costly manual annotations (Sec. 3.6).

3.2 Related Works

Facial attribute detection has made remarkable progress through decades of researches. Moghaddam and Yang [54] propose a gender classification approach using the Support Vector Machine and showed good results regarding the FERET data set [61]. Baluja and Rowley [5] further focuse on real-time gender classification based on the Adaboost to select a linear combination of weak classifiers. The above two works are carried out using small data sets and highly controlled environment settings. Recently, facial attribute detection has been shown to be important for retrieving specific people in large-scale media such as surveillance videos or photo sharing websites. Kumar et al. [41] propose an adaptive framework for learning multiple facial attributes based on the face images downloaded from the Internet with intensive manual annotations. Those user-contributed data are rich and diverse, thus providing better generalization capability for learning numerous facial attributes from large-scale media.

However, manual annotation is a burden and unrealistic for large-scale facial attributes and training images. To address this issue, user-generated content has been used to automatically identify the correct association between labels and images without any human intervention. Such new attempts have been shown promising in image classification [27, 63], object attribute learning [74, 7] and face identity retrieval [6, 52, 70]. Even though we can easily crawl images with keywords, the Internet images contain plenty of unmanageable noisy labels. Fergus et al. [27] use latent semantic models to learn object categories using the raw output of image search engines. They validate a designated category in the first 5 images returned from the queries of 7 different languages; however,

the learning performance is still highly dominated by the limited appearance of the small validation set. Taneva et al. [70] harness relational facts in the existing knowledge about named entities for gathering diverse photos of the entities and integrated image-similarity computations for improving the final ranking. Berg et al. [7] simply leverage text information associated with the crawled images and the MILboost [73] to identify the appropriate attribute type of general object attributes from noisy web images.

The aforementioned works generally focus on object image classification, which are not appropriate for learning subtle attributes of a face. For facial attribute detection, Ni et al. [55] used multiple-instance learning (MIL) to construct a universal age estimator. That work indicates the importance of learning facial attributes by automatically acquired training images. However, two critical problems are still unsolved in their framework. First, their work on age estimation is not generic for various facial attributes since it does not address how to discover the discriminative features for different facial attributes. Second, they only focus on removing false positives by visual relevance; therefore, the visual-based noise filtering process might lessen the diversity of the training data and result in data skew.

In [55, 5, 41], the authors use different combinations of facial features to represent a face. For finding an appropriate feature set to best represent a designated facial attribute, Kumar et al. [41] and Zhou and Wei [91] propose different feature selection techniques to achieve their goals. Multiple feature selection (e.g., boosting) and combination (e.g., late fusion, fusion via optimization scheme) are active research problems in multimedia analysis and have been investigated to improve learning-based video annotation [74] and image classification [77]. In this work, we aim to deal with one additional problem – the interference from the noisy labels along with training images, which pose great challenges for measuring the discriminative power of multiple features. To tackle the interference caused by noisy training images, we propose discriminability voting and leverage unsupervised (unlabeled) images for selecting effective features adaptive to different facial attributes.

Conventionally, people use a verification strategy based on potential visual features to filter out incorrectly labeled images. This process can resist the interference of noisy

labels. However, the above process also reduces the visual diversity of images which is essential for image retrieval [76] and classification [27]. For example, the visual appearance of female faces across locations are very diverse as shown in Fig. 3.1. If the female images for training are limited in certain groups, the detector might not be able to deal with general female images. Similar situation also happens to other facial attributes since the visual appearances of faces have substantial intra-class variation which requires more consideration for retaining diversity in training data. In order to balance the unfavorable effects caused by noisy labels and the diversity problem of facial attributes, we propose to introduce context cues, such as tags or geo-location, to ease the problem of solely relying on visual relevance. The proposed framework is scalable and can be generalized for adaptively learning numerous facial attributes. On the other hand, our approach requires no tedious manual annotations.

3.3 System Overview

As shown in Fig. 3.2 (a) and (b), we first harvest image candidates along with the context from the Internet by keyword queries (Sec. 3.4.1). Then we extract potential visual and context features (Sec. 3.4.2, 3.4.3). In Fig. 3.2 (c) and (d), we measure the feature quality (Sec. 3.4.4) and optimize the feature set representative to a designated attribute (Sec. 3.4.5). The features chosen in the previous steps are then fused for measuring visual relevance of the training image candidates (Sec. 3.5.1). In Fig. 3.2 (e), an optimization framework is proposed to combine both visual and context cues for estimating the annotation quality of each image candidate (Sec. 3.5.2). We further consider the cues from geo-locations in Sec. 3.5.3, since photos collected around the world are essential for training a generalized facial attribute. The candidates with higher annotation quality are superior training images for learning facial attributes (Fig. 3.2 (f)).

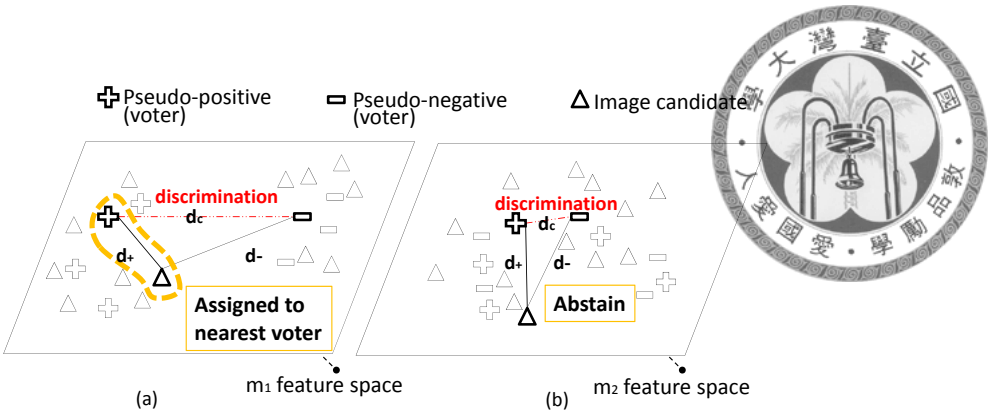


Figure 3.3: **Concept – measuring the discriminative capability of each visual feature by voting.** For a feature space m, a vote is operated by an image candidate ($\triangle$), a pair of pseudo-positive ($+$) and pseudo-negative ($-$), which are initially labled by textual relevance. If the image candidate is closer to the pseudo-positive (as (a)), it would get a positive vote, otherwise it would get a negative one. However, if the distance between the pseudo-positive and the pseudo-negative (d_c) is shorter than the distance from both of them to the candidate (d_+ and d_-), the vote would be abstained due to insufficient discrimination (as (b)). More abstained votes accumulated on a feature space indicate that the feature is less discriminative. By comparing the relative distance, the voting scheme measures the discriminative capability of different feature space without any heuristic thresholds (cf. Sec. 3.4.4).

3.4 Selecting Effective Features from Noisily Labeled Images

Initially, the images and the associated context from Flickr are acquired as the input. We try to automatically select effective features for measuring the annotation quality of candidates.

3.4.1 Harvesting Training Image Candidates

Instead of manually filtering noisily labeled images, our work only requires the users to define keywords positively correlated to the designated attribute as the input. Taking female attribute as an example, users provide the keywords, "woman" or "girl" as the input for retrieving positive training image candidates as well as the surrounding contexts such as tags and geo-locations (Fig. 2(a)). Similarly, the antonyms of female attribute such as "man" or "boy" are used as the negative keywords for retrieving negative training image candidates. If the antonyms of an attribute are ambiguous, the negative training

image candidates can be obtained through universal background data (UBD) that includes generalized background data crawled by neutral words (e.g., "people," "persons") without implying any specific facial attributes. UBD reflects the photo distribution across facial attributes in consumer photos so that it is less biased than defining ambiguous antonyms for retrieving negative training images. Note that UBD is more effective for the attributes (e.g., sunglasses) appearing in consumer photos less frequently, because they result in fewer false negatives. For the attributes (e.g., male) appearing frequently, they often have clear antonyms for collecting images rather than relying on UBD.

3.4.2 Extracting Multiple Visual Feature Combinations

For precisely analyzing the facial content, the facial regions of each photo are extracted by a face detection algorithm [72]. Then, the primary facial organs such as the eyes, philtrum and mouth are extracted. A feature combination, (l, r), represents an aggregate feature, where l is one element of the set L:{Gabor filter [28], HoG [11], Color, Local Binary Patterns [57]}, $l \in L$. Every element l in L is a varying low-level feature. The low-level features defined in L are extracted from one of the facial organs defined in R, where R:{whole face, eyes, philtrum, mouth}, $r \in R$. Different combinations of (l, r) pairs constitute M aggregate features ($\{(l, r)|\forall l \in L, \forall r \in R\}$). This set provides abundant options for effective visual feature set selection. In [41], the effectiveness of the above approach is proven.

3.4.3 Computing Textual Relevance

Given Q^+ positive keywords and Q^- negative keywords of a desired facial attribute and its associated text (e.g., tags, titles), the textual relevance t_k of the k-th facial image corresponding to that attribute can be estimated with the following equation,

$$t_k = (\sum_{q_1}^{Q^+} tf_{kq_1} \cdot idf_{q_1} - \sum_{q_2}^{Q^-} tf_{kq_2} \cdot idf_{q_2}) \times \frac{1}{A_k}, \tag{3.1}$$

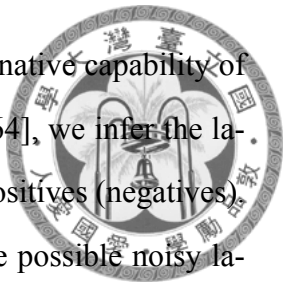
where tf_{kq} is the q-th term frequency associated with the k-th facial image, idf_q is the log inverse frequency of the facial images associated with the q-th term. A_k, the number of faces within the album containing the k-th face, is involved to prohibit top ranked images from being dominated by certain sources. For some attributes (e.g., beard, sunglasses) which have ambiguous antonyms (negative keywords), we do not assign any negative keywords. Under these circumstances, we neglect the influence of negative keywords in the second term in Eq. 3.1 and measure the text relevance via the use of positive keywords only. Using the top ranked images, we are able to perform better cold-start [1] and achieve better performance. We shall discuss this issue in the subsequent section.

3.4.4 Measuring Feature Quality by Discriminability Voting

Feature selection (e.g. Adaboost) is an essential technique for discovering a subset of relevant features for building robust learning models [41, 91]. The main idea is to estimate the weight of each feature according to its discriminative capability measured from the annotated training samples on each feature space. However, keyword-retrieved training images are very easy to be interfered by incorrect (unreliable) labels, therefore result in incorrect decision boundaries.

Knowing the limitation of a supervised feature selection process, we further take the unlabeled images and the *pseudo-positives (negatives)* into consideration. First, a small number of images that have the strongest textual relevance are selected as *pseudo-positives*, which means they are very likely to be positives. Similarly, the images with the weakest textual relevance are most likely to be negatives, therefore are selected as *pseudo-negatives*. The concept resembles *transductive learning* – reasoning the unlabeled image candidates by the very few labeled ones – but also tackles the interference from possible noisy labels in pseudo positives (negatives). We propose to measure the quality of potential visual features by *discriminability voting* – voting the unlabeled images by exploiting the discrimination between the pseudo-positives and the pseudo-negatives as an index of reliability (as d_c in Fig. 3.3). The absolute distance between pseudo positives and nega-

[1] Cold-start problem concerns the issues that the system lacks sufficient information for inferring or associating the non-annotated data.



tives on a feature space is an informative cue to evaluate the discriminative capability of that feature modality. Similar to *Kernel Density Estimation (KDE)* [64], we infer the label of the image candidate by the distance from them to the pseudo-positives (negatives). However, KDE estimates probability density without considering the possible noisy labels while our approach would confirm the discriminability d_c before voting the unlabled image candidates.

However, the scale of multiple feature space may be quite different; therefore, measuring the absolute distance between two examples may cause the evaluation to be dominated by certain feature modalities. In contrast, the proposed *discriminability voting* only compares the relative distance between examples such that it can be applied to the measurement of multiple feature modalities without concerning the variation of distance scale. As shown in Fig. 3.3, a vote is determined by an image candidate and a pair of pseudo-positive and pseudo-negative. The candidate would get a positive vote (denoted as label *1*) if it is closer to the pseudo-positive (as Fig. 3.3 (a)). Otherwise, the candidate would get a negative vote (denoted as label *-1*). Further, we abstain the votes as uncertainty occurs. Referring to Fig. 3.3 (b), if the distance between the pseudo-positive and the pseudo-negative (d_c) is shorter than the distance from both of them to the candidate (d_+, d_-), the vote would be abstained due to insufficient discrimination (denoted as label *0*). An uncertain vote may come from the weak discrimination of a feature itself as well as the disturbance of incorrect annotations. We separate those abstained votes from either positives or negatives to reduce the uncertainty caused by unreliable training images. The voting scheme requires no heuristic thresholds but only compares the relative distance. Therefore, it has higher generalization capability to multiple feature modalities without concerning the sensitivity of threshold setting. The statistics of the abstained votes helps estimate the discriminative capability of each visual feature. This issue will be discussed in Sec. 3.4.5.

The voting results are represented as the matrices shown in Fig. 3.4. K image candidates, K^+ pseudo-positives and K^- pseudo-negatives would induce $K \times K^+ \times K^-$ voting results for each visual feature, thus obtaining a feature-wise voting vector (E_m) of the m-

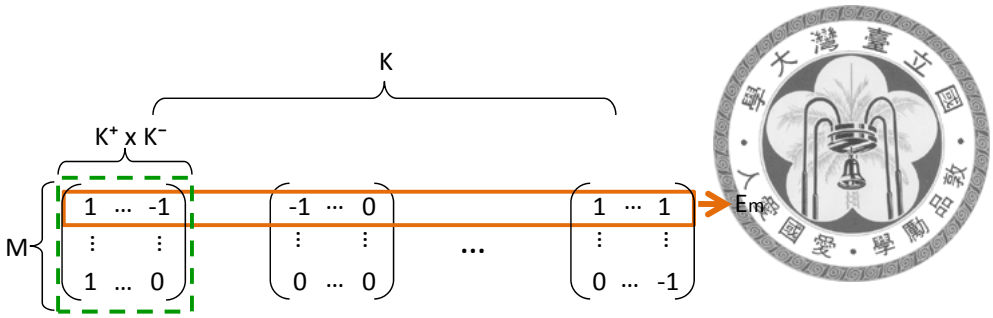


Figure 3.4: Voting results – each row (in the orange rectangle) is a feature-wise voting vector E_m, which represents the voting results of all the candidates conducted by the m-th visual feature. Each sub-matrix (in the green square) is voting results of the k-th candidate, which represents the labels of the k-th candidate assigned by $K^+ \times K^-$ pairs of pseudo-positives and pseudo-negatives across M visual features. See more details in Sec. 3.4.5.

th feature (shown in the orange rectangle in Fig. 3.4). Similarly, every candidate would be assigned $M \times K^+ \times K^-$ votes by all the pairs of pseudo-positives and pseudo-negatives across M visual features (shown in the green square in Fig. 3.4). K^+ and K^- are generally less than 50 (investigated in Sec.3.6.1) and K depends on the number of the involved candidates. Given $K = 2000$ candidates, the computation time of voting process for a feature is 2.5 seconds on average for a 2.40GHz Intel Xeon server. Keeping the whole voting matrices is not required, since the voting results of candidates on each feature space would be accumulated by the label type (1, -1 and 0), which are denoted as $v^+_{k,m}$, $v^-_{k,m}$ and $v^a_{k,m}$ (cf. Algorithm 1). Our approach only keeps the original voting results E_m of any two visual features for evaluating the mutual similarity (i.e., requiring 80 megabytes when K^+, $K^- = 50$ and $K = 2000$). In the next section, we will use the voting results as the index of feature quality to select the optimal feature set.

3.4.5 Optimizing Feature Set

Generally, importance and complementarity are two kernels of feature selection approaches [33]. Accordingly, our work exploits the voting results of each feature to evaluate the importance of the feature itself and the complementarity between any two of them. First, more abstained votes in the voting vector E_m of the m-th feature, more uncertainty the feature would induce. Second, more similarity between the two voting vectors implies less complementarity between the two features. Motivated by the intuitions, we estimate

Algorithm 1 Voting on m-th visual feature space

Input: the features (F,F^+,F^-) of candidates, pseudo-positives and pseudo-negatives

Output: feature-wise voting vector (E_m), accumulated positive, negative and abstained votes ($v^+_{k,m}$, $v^-_{k,m}$, $v^a_{k,m}$)

```
cnt := 0;
for all f_k in F, f_i^+ in F^+, f_j^- in F^- do
    if distance(f_i^+, f_k) < distance(f_j^-, f_k) then
        vote := 1;
        shortest := distance(f_i^+, f_k);
    else
        vote := -1;
        shortest := distance(f_j^-, f_k);
    end if
    if distance(f_i^+, f_j^-) < shortest then
        vote := 0;
        v^a_{k,m} := v^a_{k,m} + 1;
    else
        if vote = 1 then
            v^+_{k,m} := v^+_{k,m} + 1;
        else
            v^-_{k,m} := v^-_{k,m} + 1;
        end if
    end if
    E_m(cnt) := vote;
    cnt := cnt + 1;
end for
```

the importance u_m of a visual feature by the number of abstained votes in its voting vector E_m.

$$u_m = 1 - (1 + e^{-count(E_m=0)})^{-1}. \tag{3.2}$$

The $count(E_m = 0)$ is the total number of abstained votes conducted by the m-th visual feature. Further, the similarity $s_{m,n}$ between the m-th and the n-th features can be evaluated by the Euclidean distance as follows,

$$s_{m,n} = 1 - (1 + e^{-\|E_m - E_n\|^2})^{-1}. \tag{3.3}$$

As aforementioned, we want to select those features with the largest importance and the largest complementarity (the least similarity). We can realize the physical meaning by

converting it into an optimization formulation [33] as follows,

$$\min_{x} \sum_{m}^{M} \sum_{n,n\neq m}^{M} s_{m,n} x_m x_n - \sum_{m}^{M} u_m x_m, \text{s.t. } x_m \in [0,1], \tag{3.4}$$

where $\boldsymbol{x} : \{x_m | m = 1, \ldots, M\}$ is a vector containing the optimized weights of all visual features. For a training candidate, the visual relevance is estimated by the voting results aggregated from different visual features weighted by $\boldsymbol{x}$ (cf. Sec. 3.5.1). The feature weights not only reduce the interference of less-effective features but integrate the results from multiple visual modalities, which bring more diversity to the estimation of annotation quality. Eq. 3.4 can be further reformulated as the following equation,

$$\min_{x} x^\top s x - u^\top x, \text{s.t. } x_m \in [0,1], \tag{3.5}$$

s is an $M \times M$ matrix where the diagonal is assigned with 0, and u is an M-dimensional vector. The equation can be solved by gradient descent method with the consideration of box constraint.

3.5 Measuring Annotation Quality for Determining Effective Training Images

The annotation quality of a training image candidate means the possibility that the candidate is correctly labeled corresponding to the target facial attribute. The candidates with higher annotation quality have higher priority to be chosen as the training images. To begin with, we measure the annotation quality from the degree of visual relevance. Furthermore, the annotation quality would be optimized by both the visual relevance and the textual relevance (estimated by Eq. 3.1) to prevent the training images from being dominated by some special visual appearances. For training generalized facial attributes, we further measure relative visual relevance in a specific geographic location to include facial images more uniformly around the world.

3.5.1 Measuring Visual Relevance

A facial image carries essential information for evaluating annotation quality for a designated attribute, hence we measure the visual relevance v_k of an facial image to represent the likelihood of belonging to an attribute category through visual modality only. After executing the voting process described in Sec. 3.4.4, each image candidate would be assigned $M \times K^+ \times K^-$ votes with labels. According to the assigned labels, the visual relevance v_k of the k-th candidate can be measured as follows,

$$v_k = \sum_{m}^{M} x_m \times v_{k,m}^{a}(v_{k,m}^{+} - v_{k,m}^{-}), \tag{3.6}$$

where $\boldsymbol{x}$ are the feature weights measured by Eq. 3.4, which indicate the effectiveness of each feature. $v_{k,m}^{+}$, $v_{k,m}^{-}$ and $v_{k,m}^{a}$ are the accumulated positive votes (label *1*), negative votes (label *-1*) and abstained votes (label *0*) for the k-th candidate on the m-th feature space, respectively (cf. Algorithm 1). All the accumulated votes are normalized to $[0, 1]$. $(v_{k,m}^{+} - v_{k,m}^{-})$ is considered to favor the images which have more tendency to be positives than to be negatives. However, it is important to choose the most informative example for learning a function. One interpretation of this is to choose the examples with high uncertainty such as the strategy of uncertainty sampling in active learning [71]. Considering the trade-off, we use $v_{k,m}^{a}$, the uncertainty of classifying a candidate, for moderately encouraging the candidates carrying informative cues for classification. In the following processes, the visual relevance v_k would be used to rank the annotation quality of the faces from the visual aspect.

3.5.2 Combining Textual and Visual Relevance

Examining the visual relevance of candidates can suppress the false positives, but sacrifices the diversity in visual appearances, which is essential for collecting training images of a generalized facial attribute. Balancing visual relevance v_k and the textual relevance t_k (semantic relevance), we refine the annotation quality score p_k for each candidate face

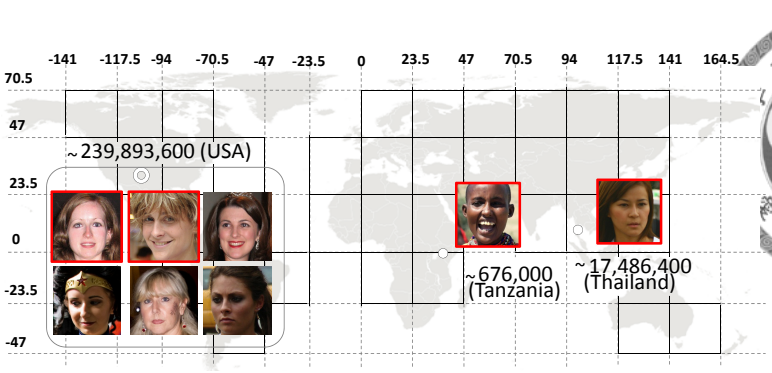


Figure 3.5: Inherently uneven distribution in user-contributed photos: the data skew in the web images due to the huge gap of Internet usage, e.g., USA (239 million users) vs. Tanzania (0.6 million users) [2].

by the following optimization criterion:

$$\min_{p} \sum_{k}^{K} [(1-\alpha)(p_k - t_k)^2 - \alpha v_k p_k + \gamma \|p_k\|^2]. \tag{3.7}$$

The first term is to measure the error between the estimated annotation quality and the textual relevance. The second term is to refine the possible error annotations by the visual relevance, and the last term is for regularization. The equation favors the candidates with higher visual relevance v_k. α, γ are the parameters used to control the effect of visual relevance and to prevent the overfitting effect. These parameters will be further investigated in Sec. 3.6.2 for maximizing the system performance. The equation can be solved by gradient descent which iteratively updates the annotation quality p_k starting from an arbitrary vector. p_k is the annotation quality of the k-th candidate image. Annotation quality means the correlation between the designated facial attribute and the facial attribute in the candidate image. The higher the p_k is, the better annotation quality the k-th candidate image has. The candidates with higher annotation quality would be chosen as the training images.

3.5.3 Considering Geo-locations

The statistics of global Internet users [2] reveals that there is a big gap of Internet usage across countries; for example, 239 million users in USA and 0.6 million in Tanzania.

So do the numbers for community-contributed photos across countries. Learning with those biased face distributions neglects the generality of facial attributes, since the visual appearances of people from the same area are probably more similar (e.g., Europeans) than those from other areas (e.g., Asians). Though many applications only concern specific groups (usually the majority), the proposed approach aims to deal with more general cases in real life. Thus reducing the geographic bias is critical for the purpose of enhancing generalization. To tackle the problem, we divide the world into equal grids, where the grids containing the continents (solid-line rectangles in Fig. 3.5) are preserved as individual location groups (totally 34 groups) and the other grids (dashed-line rectangles in Fig. 3.5) are aggregated to the same location group. We evaluate the *relative visual relevance*, which is the Borda rank [9] of a training face candidate within a location group assigned by the location contexts (e.g., GPS) along with the photos containing that face. Relative visual relevance will favor the training image candidates with higher visual relevance within each location group, therefore prevents the appearances of training candidates from being dominated by the same places.

Given a training face candidate with visual relevance v_k in a location grid G, the relative visual relevance g_k in its location group is measured by the following equation.

$$g_k = 1 - \frac{B_G(v_k)}{|G|}. \tag{3.8}$$

$B_G(v_k)$ is the number of image candidates in G which have the visual relevance value larger than v_k, where $B_G(v_k) \in \{0, 1, 2, ..., |G|-1\}$. $|G|$ is the number of photos collected in a location grid. To prevent the training images from including too many photos of the same location grid, we limit the value of $|G|$ to be the number of required training images divided by that of total location grids. Through the arrangement, the faces with higher visual relevance within a group are given higher relative visual relevance according to their location context, hence only a few photos in a location group get the opportunities to be chosen as the training data. The relative visual relevance g_k is further integrated to the annotation quality measurement for introducing more locational diversity into the

acquired training images. The optimization formulation in Eq. 3.7 is refined as follows,

$$\min_{p} \sum_{k}^{K} [(1-\beta)(p_k - t_k)^2 - \beta g_k p_k + \gamma \|p_k\|^2]. \quad (3.9)$$

The sample with higher relative visual relevance g_k has higher priority to be selected as the training samples in that area. Moreover, β and γ are the parameters used to control the influence of the geo-context and the regularization process, respectively. These parameters will be further investigated in Sec. 3.6.3 for analyzing the merits and the limitations of the location contexts.

3.6 Experiments

Our algorithm was conducted on up to 0.2 million face images from community-contributed photos (from Flickr) and the associated context (e.g., tag, title, GPS) as the candidates for training image acquisition. All of the photos were preprocessed by face detector and eye detector [72] for more precisely locating the facial components on a face. Rather than manually labeling the faces with correct annotations, we only input relative keywords for the designated facial attribute. We will first describe the implementation detail regarding the threshold setting in the feature selection process (Sec. 3.6.1) and then show the experiments based on several evaluation metrics (Sec. 3.6.2) to demonstrate the quality (correctness and diversity) of the acquired images and the superiority of learning from those freely available images.

3.6.1 Threshold Selection

Our work exploits the few top ranked images as the pseudo-positives and the pseudo-negatives in the voting approach (Sec. 3.4) and it is important to adaptively determine a proper number of initial seeds to facilitate effective voting. Typically, the precision of text-based retrieval would decay when the number of retrieved images increases. More images retrieved brings more incorrect labels. On the other hand, more images are more representative to carry out confident voting. To balance the trade-off between the precision

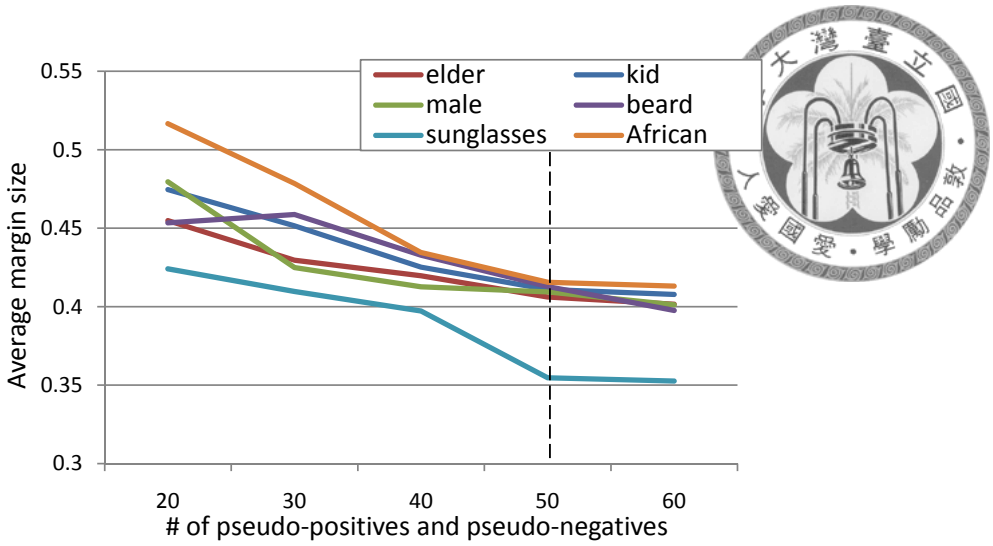
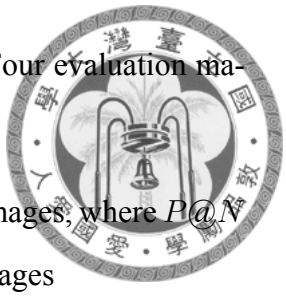


Figure 3.6: With more pseudo-positives and pseudo-negatives included as the seeds for evaluating the training image quality, the average margin size (Sec. 3.6.1) decreases because the rich data tends to mix together and decreases discriminability. To balance the diversity (more training images) and discriminability, our system progressively increases the number of pseudo-positives (negatives) until the average margin size nearly flattens. The approach ensures acceptable representativeness without sacrificing much precision in the included images.

and the representativeness of the selected images, we compute the *average margin size* by averaging the Euclidean distance $d(f_i^+, f_j^-)$ between two data groups, i.e., the pseudo-positives F^+ ($f_i^+ \in F^+$) and the pseudo-negatives F^- ($f_j^- \in F^-$), to measure the discriminability. With more pseudo-positives (negatives), the average margin size decreases rapidly as shown in Fig. 3.6 because the rich data tends to mix together and decreases discriminability. To encourage the representativeness along with more training images, our system progressively increases the number of pseudo-positives (negatives) until the average margin size nearly flattens. The approach ensures acceptable representativeness without sacrificing much precision.

3.6.2 Evaluation Metrics

Our work aimed at automatic data acquisition in weakly labeled web images for learning facial attributes. In the experiments, we show that our approach is able (1) to acquire correctly labeled images, (2) to enhance classification capability, (3) to select effective features from noisy image collections and (4) to bring about more classification gain by



increasing the number of (automatically) acquired training images. Four evaluation matrices were leveraged to investigate the following challenges:

1. *Quality of acquired images* – measuring $1-P@N$ in acquired images, where $P@N$ means the precision at different number (N) of the retrieved images

2. *Effectiveness in Classification* – examining the classification error rate on benchmark data [41] using facial attribute detectors learned by the acquired images

3. *Robust-to-noise feature selection* – verifying the overlaps between the visual features selected by manually labeled images and those selected by noisily labeled images

4. *Classification gains as increasing the number of training images* – evaluating classification accuracy of the attribute detector trained with the growing number of acquired images

Without loss of generality, the proposed approach was experimented on six different types of facial attributes and compared with (a) manually labeled [41], (b) text-based and (c) PCA-based methods, respectively. (a) represents the oracle by costly human annotation. (b) indicates the effect of simply including required images based on textual relevance (measured by Eq. 3.1). (c) has been commonly used in previous works [55] for measuring visual relevance to eliminate outliers within the same category. For (c), we collected images by given keywords for the target attribute and performed PCA to filter out the faces with large reconstruction errors according to the reduced eigenvectors.

Quality of Acquired Images

We evaluate the correctness in the acquired images by $1-P@100$ and $1-P@200$, where lower $1-P@N$ indicates higher correctness. As shown in Fig. 3.7, our approach can effectively acquire images with higher correctness across most of attributes. Comparing with the text-based method, our approach could reduce by 24% the incorrect labels on average in top 100 expanded images. For sunglasses attribute, the reduction rate reaches 32% by the use of our method. Even the least improved attribute, "kid," our approach

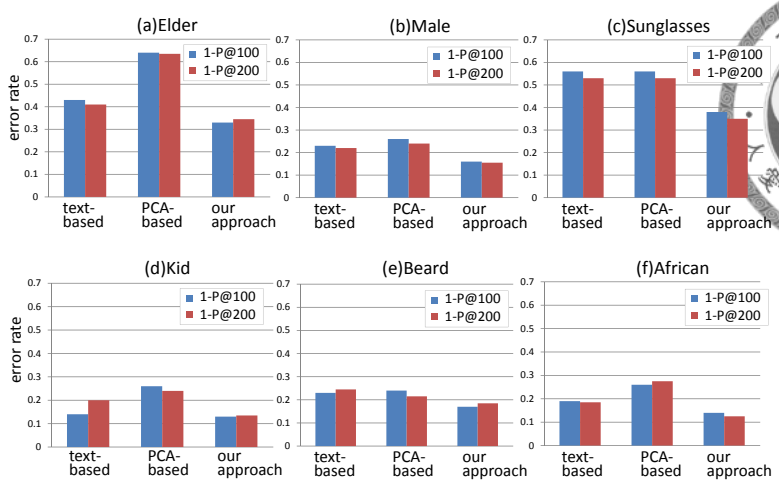


Figure 3.7: The error rate in the acquired images: for all the six types of attributes, our approach can successfully acquire more correctly labeled images comparing with text-based [27] and PCA-based approach [55], thus ensuring the correctness in the training images for learning facial attributes.

could still reduce 7% error labeling. In addition, when the number of acquired images was increased from 100 to 200, the precision of the text-based approach dropped significantly while our approach still maintained good precision (Fig. 3.7 (d)). One thing to be noted is that the PCA-based approach performed the worst among the three comparing methods in terms of precision. Its error rate maintained the highest in the six sets of experiments shown in Fig. 3.7. PCA is effective in extracting the majority of visual contents from the collected faces. However, when the extracted majority is misled by incorrect labels, the degradation of performance is even worse than applying other methods. Our approach considered the visual relevance as well but exploited the proposed verification strategy (Sec. 3.4.4) to select effective features at first, therefore survived the aforementioned challenges.

The diversity of acquired images is a critical problem if we intend to learn attribute detectors by these images. Since textual features possess high semantic meaning by nature, the text-based approach (T) acquires visually different images ($\alpha = 0$ in Eq. 3.7, cf. Fig. 3.8 (a)) but brings much more incorrect labeling at the same time. On the other hand, acquiring images solely by visual content eliminates incorrect labeling but loses diversity ($\alpha = 1$ cf. Fig. 3.8 (b)). Witnessing the drawbacks of the two types of image acquisition approaches, our work (F+T+V) considered both textual relevance and visual content ($\alpha =$

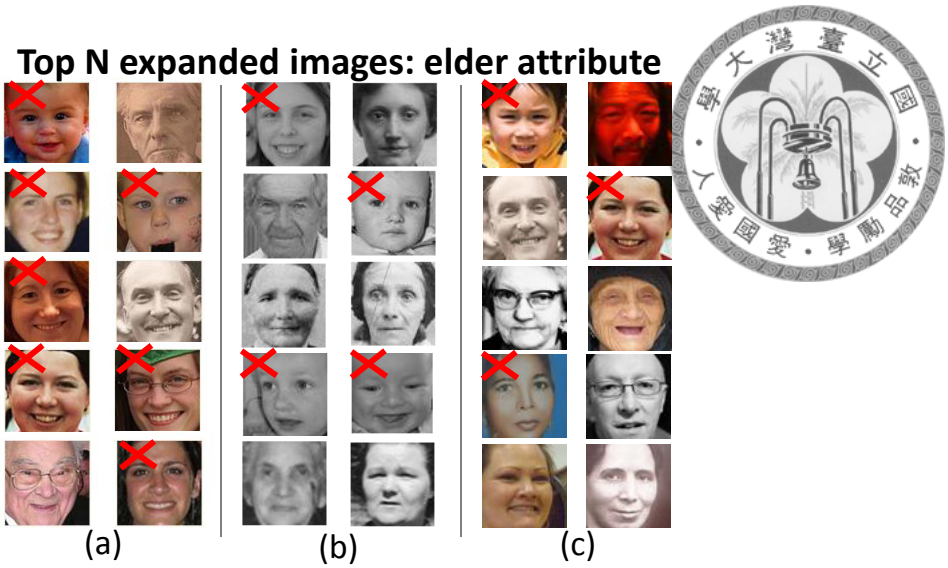


Figure 3.8: The images acquired by textual relevance (a), which contains many incorrect annotations (marked by the crosses); by visual relevance (b), which contains dominated appearances; by annotation quality considering context and visual relevance (c), which contains both diverse appearances and correct annotations.

Table 3.1: The number of test images of 6 facial attributes from FaceTracer [41], totally around 1500 images.

attribute	elder	kid	male	beard	sunglasses	African
# of test data	322	315	200	200	322	313

0.45). As shown is Fig. 3.8 (c), the expanded images are semantically consistent but visually different. In other words, it covers diverse appearances of a designated attribute. We will demonstrate that the diverse images acquired by our approach greatly advance the classification capability in Sec. 3.6.2.

Effectiveness in Classification

To demonstrate that our approach can discover effective training images for learning facial attributes, we adopt an adaptive learning method based on a boosted set of SVM classifiers [14]. Two key tasks are required, (1) organizing potential visual features, (2) finding the best feature combination for a designated facial attribute [41]. As described in Sec. 3.4.2, a rich set of visual features provides higher flexibility for describing various facial attributes. Here, we train a weak classifier by SVM using a specific visual feature extracted from the training images. For a facial attribute, the optimal set of visual

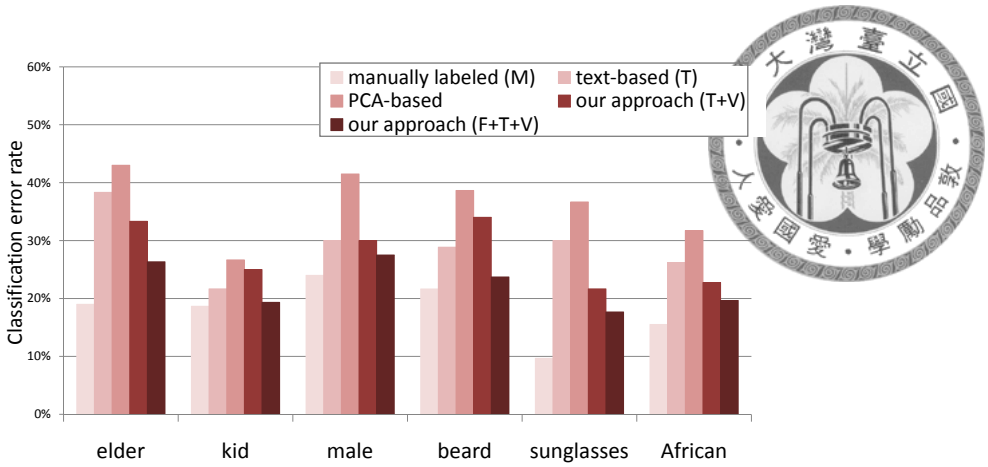


Figure 3.9: The classification error rate (%) of 6 facial attributes using 5 training data acquisition approaches, (a) manually labeled (M), (b) retrieved by textual relevance (T), (c) ranked the keyword-retrieved training images by PCA reconstruction error, (d) balancing context and visual relevance (T+V) and (e) selecting effective features for measuring visual relevance before combining context and visual relevance (F+T+V). Apparently, augmenting the freely available photos from the Internet by visual and contextual cues yields the best automatically acquired training data. Preliminarily selecting effective features further improves the annotation quality. In certain attributes, the proposed method is comparable with manual labeling.

feature classifiers is selected by Adaboost. The combined strong classifier represents the most important parts of that attribute, for example, (philtrum, color) and (mouth, LBP) are most important for beard attribute. Through the training process, six facial attributes were learned by our approach and other data acquisition approaches for comparing the classification capability.

In the experiments, we use 400 positive images and 400 negative images for training to evaluate the classification accuracy via the different approaches of training data acquisition. The text-based approach (T), the PCA-based approach and manual labeling (M) are adopted as the baselines of collecting the required training images. In the testing phase, we evaluate the attribute detectors by measuring classification error rate in the publicly released facial attribute dataset, FaceTracer [41], which is composed of facial images grabbed from the web across various facial-attribute annotations. For fair comparison with the related works, more than 1500 facial images in this benchmark collection and their attribute labels are included as the test data. Table 3.1 shows the exact number of test images for each attribute in detail.

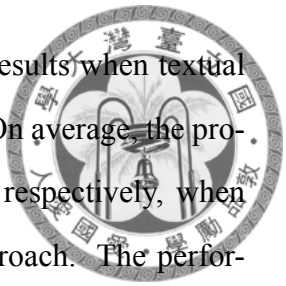


Fig. 3.9 reveals that all attributes show improved classification results when textual and visual relevance were considered for collecting training images. On average, the proposed approach (F+T+V) cut down 23.35% and 38.50% error rate, respectively, when compared with the text-based (T) approach and the PCA-based approach. The performance of our approach was comparative to that of learning by manually labeled images for most attributes. The exceptions were the elder and the sunglasses attributes, where the incorrect labeling rate in top retrieved images were much higher (cf. Fig. 3.7). Although the precision in the raw images (initial input) would affect the quality of training data acquisition, comparing with the text-based approach, our approach still cut down 12.0% and 12.3% error rate, respectively, for the two attributes. The PCA-based approach may make the majority appearance dominate the training images, thus having little effect upon the classification model. The same phenomenon happened in the approaches purely relied on visual content.

It is worthy mentioning that feature selection in noisily labeled images also helps the classification performance. We simply omitted the feature selection process in Sec. 3.4.4 and used equal weight for each feature when measuring the visual relevance (T+V). The classification result in Fig. 3.9 (T+V) shows that the overall error rate increased by 5.6% than that by (F+T+V), the approach with robust-to-noise feature selection. The results evidence the importance of feature selection for image acquisition. In the next section, we shall examine whether the feature set optimized by our approach matches the one chosen by manually training images (oracle) using the Adaboost-based approach [41].

Robust-to-noise Feature Selection

Feature selection aims to identify a set of optimal features for describing designated facial attributes. The existing approaches can already work on correctly annotated images. In this work, we extend the capability of our approach to handle noisily labeled images. We demonstrate the reliability of an optimized feature set by estimating the overlaps between the selected features (based on images collected by different approaches) and the top features highly weighted by Adaboost conducted with manually labeled images. Since

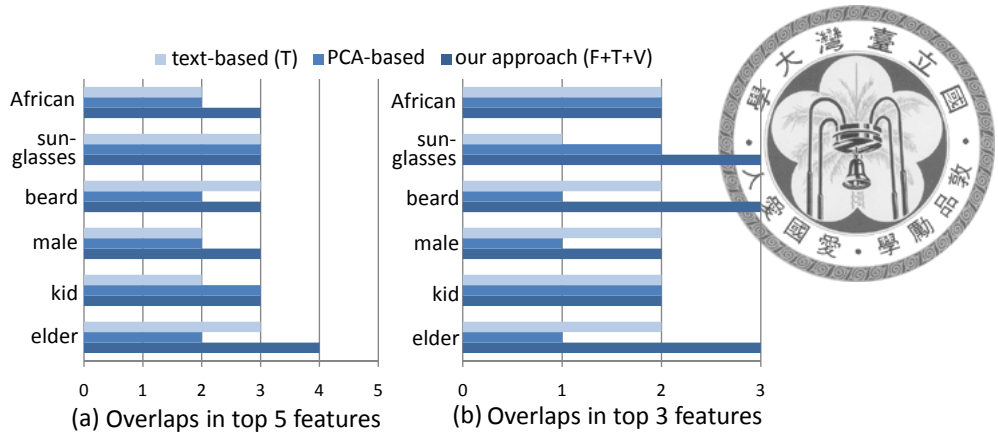


Figure 3.10: The overlaps with the most important features: we compare the number of overlapped features between those selected by manually annotated images (most effective) and those selected by the images collected with text-based, PCA-based and our approach. Text-based approach and PCA-based approach averagely overlap less than 50% ($< 3/5$) features, while our approach overlaps more than 50% ($>= 3/5$) features for all the six facial attributes. The phenomenon becomes more apparent in top 3 features, where the overlap ratio by our approach achieves 100% ($= 3/3$) for half of the attributes.

the top boosted features are the most vital for discriminating the positive and negative data, a higher feature overlap ratio leads to better classification capability. Likewise, the boosting process was conducted on the images collected by the text-based approach and the PCA-based approach for comparison. In Fig. 3.10 (a), we can see that the text-based approach and the PCA-based approach overlapped less than 50% ($< 3/5$) (top 5) features on average. The loss was due to incorrectly labeled images which misled the feature weighting process. On the other hand, the overlap ratio of our approach was more than 50% ($>= 3/5$) for all the six facial attributes. The phenomenon is very obvious if the top 3 features were checked (cf. Fig. 3.10 (b)). When we conducted this check, our approach achieved 100% ($= 3/3$) for half of the six attributes. This result indicates that the selected features would be affected by the incorrectly labeled images from text-based retrieval. The mis-selected features may degrade the classification accuracy as shown in Fig. 3.9 (M) and (T). Our approach, on the other hand, could survive the challenges caused by noisy annotations and select similar features as those selected by manually labeled images.

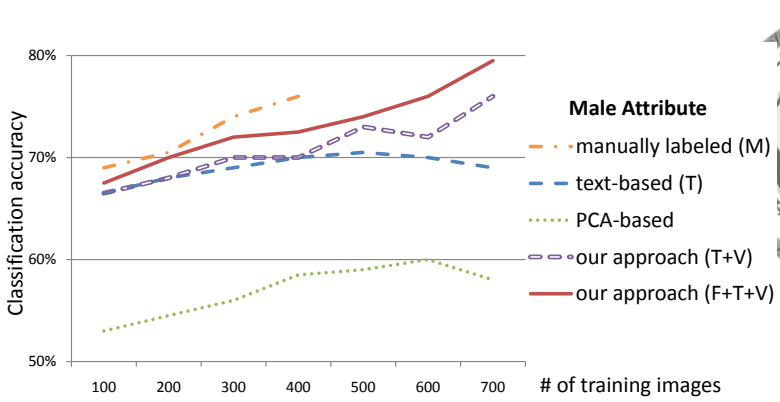


Figure 3.11: The trend of classification accuracy with more training images (female attribute): manual annotation would confront the bottleneck of annotation effort. We can conquer the problem through automatic training image acquisition methods. The accuracy of our results with 700 training images yielded 79.5%, which outperformed 76.0% by 400 manually annotated training images. Note again that our approach does not incur any extra manual effort to expand more images.

Classification Gains as Increasing the Number of Training Images

In this section, we study the relation between the classification gain and the number of training images. In general, increasing the number of effective training images should benefit the classification accuracy, but the growth will become slow after the number of training images reaches a threshold. For facial attribute detection such as gender, age and race, the accuracy can reach 80% accuracy on average with around 400 to 800 (manually collected) positive (negative) training images [41, 54, 34]. In our work, we also investigate the classification gains with similar number of training images. Fig. 3.11 (M) shows the classification accuracy did increase from 69% to 76% if we introduced more manually labeled images. However, to use manually labeled training images is not a good choice once the number of images is too huge. The above mentioned problem can be solved by prior or baseline methods; however, they are not consistently successful. In Fig. 3.11, learning with more acquired images by the text-based approach and the PCA-based approach both received good performance improvement when the number of training images was less than 600. However, when the number of training images was over 600, the classification accuracy dropped. When our approach (F+T+V) was applied, the classification accuracy consistently increased from the beginning (100 images) to the end (700 images) according to Fig. 3.11. A notable phenomenon is that when the number of training images

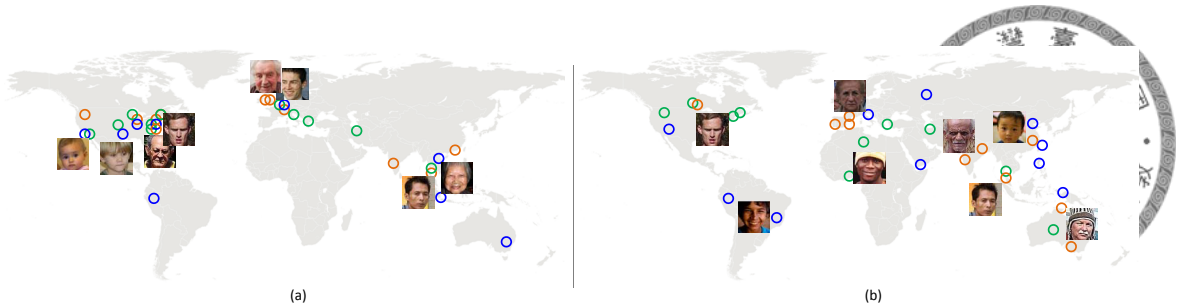


Figure 3.12: The geo-distribution of top 10 acquired training images for three facial attributes, which are denoted by orange (elder), blue (kid) , green (male) circles. (a) contains the results by considering visual and textual relevance (F+T+V). (b) contains the results by the method (F+T+V) and further considering relative visual relevance of each location group (F+T+G). The images near the relatively intensive circles represent certain training faces located at that area. Obviously, the training faces in (b) are more widely distributed around the world and much diverse in appearance comparing with those in (a). Therefore, using (b) as training images benefits the attributes which require diverse facial appearance from different locations (e.g., gender, age).

reached 700, the classification accuracy was 79.5%, which outperformed the 76% classification accuracy obtained by using 400 manually annotated training images. Note again that our approach does not incur extra manual effort to expand more images. The nature demonstrates its superiority to its counterpart, generally requiring expensive manual annotations. Moreover, our approach can continuously learn from the updated images and metadata from the Internet without additional human intervention.

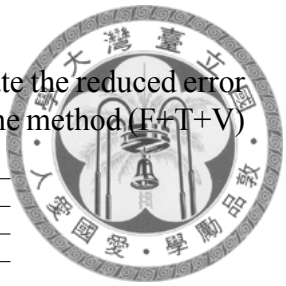
3.6.3 Effects of Geo-Context

Finally, we investigate the effects of geo-context. The geo-context enhanced approach is experimented on the three facial attributes, i.e., elder, kid and female, where the facial appearance of these attributes are strongly correlated to locations; for example, the Asians and the Africans have very different facial appearances. The geo-locations help the proposed method collect more generalized training images for those facial attributes as shown in Fig. 3.12 [2]. Meanwhile, the correctness is not sacrificed because the relative visual relevance within each location grid is kept in measuring geo-context Eq. (3.9). For the kid and the male attribute (cf. Table 3.2), the error rate is considerably reduced and very close to that by costly manual annotations (oracle). For the elder attribute, the classification er-

[2]For privacy concerns, we only present the facial photos under Creative Commom Liscense.

Table 3.2: Classification error rate (%): the numbers in brackets indicate the reduced error rate by the geo-context enhanced approach (F+T+G) comparing with the method (F+T+V) ignoring geo-context.

attribute	elder	kid	male
our approach (F+T+G)	26.33 (-0.00)	18.66 (-0.67)	24.50 (-3.00)
manually labeled (M)	19.00	18.66	24.00



ror rate stays the same as that without considering geo-context. The possible reasons are: (1) rather than diversity, the lower precision in the training images of the elder attribute might be the bottleneck of classification performance; (2) the crawled elder-face images are less extensively distributed around the world (comparing with the male and the kid attributes), so that the geo-context does not bring much improvement.

3.7 Remarks

Leveraging the freely available community-contributed photos and their plentiful and informative contexts such as tags, geo-locations, we propose a generic framework to automatically acquire effective training photos for facial attribute detection. Through the experiments, we confirm the effectiveness of the proposed feature selection approach, which successfully escapes from the interference of noisy labels (tags) and ensures the correctness in the acquired training images. We also show that the contextual cues can boost the diversity and keep accuracy in the automatically acquired training images, comparing with prior methods by limited keyword queries or considering visual relevance only. As the first work for mining effective training images for facial attributes, we investigate different configurations of parameters to optimize the proposed method. Meanwhile, generally increasing diversity (by contexts) are considerably helpful for the six attributes. We also show that the proposed framework is promising to alleviate costly human annotations for learning facial attributes. For the future work, we are to increase the number of facial attributes and further devise effective methods to adaptively optimize the balance of visual and contextual cues for acquiring training faces for numerous facial attributes.

3.8 Extensive Applications: Retrieving Images by Facial Attributes

3.8.1 Face Image Retrieval using Attribute-Enhanced Sparse Codewords

Facial attributes also benefit face verification [42] and face retrieval [15]. In [42], their experiments also show that human can achieve salient verification performance using facial attributes and even only the surrounding context of face images, where the surrounding contexts are strongly correlated to people attributes such as hair style, hair color and accessories. Meanwhile, our preliminary study [15] has also shown that the semantic information provided by automatically detected facial attributes can substantially complement the missing contexts. In [15], we propose to utilize automatically detected facial attributes that contain semantic cues of the face photos to improve content-based face retrieval by constructing semantic codewords for efficient large-scale face retrieval. We also investigate the effectiveness of different attributes and vital factors essential for face retrieval. Experimenting on two public datasets, the results show that the proposed methods can achieve up to 43.5% relative improvement in MAP compared to the existing methods. These studies evidence that people attributes (e.g., gender, race, hair style) are high-level semantic descriptions about a person and are effective for people-centric multimedia analysis. Therefore, we propose to further investigate effective features for learning frontal and profile facial attributes and conduct consumer photo retrieval by aggregating multiple facial attributes.

3.8.2 Face Image Retrieval by Facial Attributes and Canvas Layout

Facial attributes are important characters for describing a human face. Nowadays, because of the prevalence of camera devices, people are growing accustomed to preserving important moments in life by photos. With an increasing number of personal photos, it is difficult and inefficient for users to indicate the exact file location in the storage even

though they are well categorized by time stamps or geo-locations. Some photo sharing websites employ crowd-sourcing to obtain free tags semantically associated to images, but the mechanism cannot be duplicated to personal photo management because users are not expected to actively annotate their photos. Recently, certain commercial software began to exploit technologies of face recognition and face clustering; such solutions still lack the capability of searching for scenes with faces deployed in a specific layout. In light of this observation, we attempt to make consumer photo retrieval faster and easier by facial attributes, face similarity and overall layout. We are (1) to analyze "wild photos" with no tag information at all by automatic facial attribute detection and face similarity estimation [41], (2) to advance search pattern from query by single face instance to query by multiple attributed faces allocated on a canvas and (3) to enable rapid search response by block-based indexing approach. The framework has been realized in a touch-based user interface which allows interactively refining the query canvas [45].

Chapter 4

Mining facial attributes and social relationships

4.1 Introduction

The freely available media provide a cost-effective way to obtain demographic information – the statistics for the user preferences in certain events or locations such as restaurants, hotels, landmarks, etc., which is essential for marketing, advertising, and recommendation systems. Such rich information collected from the huge user-contributed photos reveals diverse activities and preferences and can be treated as multimedia life "logs." To deal with the big data, many studies focus on exploiting facial content analysis such as facial attribute detection (e.g., gender, age, race, etc.) to support large-scale demographic research. For example, our preliminary study [21] adopted the associated contexts (e.g., time, location) and the people attributes mined from community-contributed photos to facilitate profiling consumer activities for mobile recommendations.

In fact, consumer activities and user intentions are not limited in individuals. Group recommendation, which recommends to a group of people instead of individuals, is vital for daily life. In Li et al.'s work [46], they analyzed the transaction logs and discovered that different types of consumer groups (e.g. family, friends, couple) have quite different preferences when searching for travel accommodations. For example, family groups

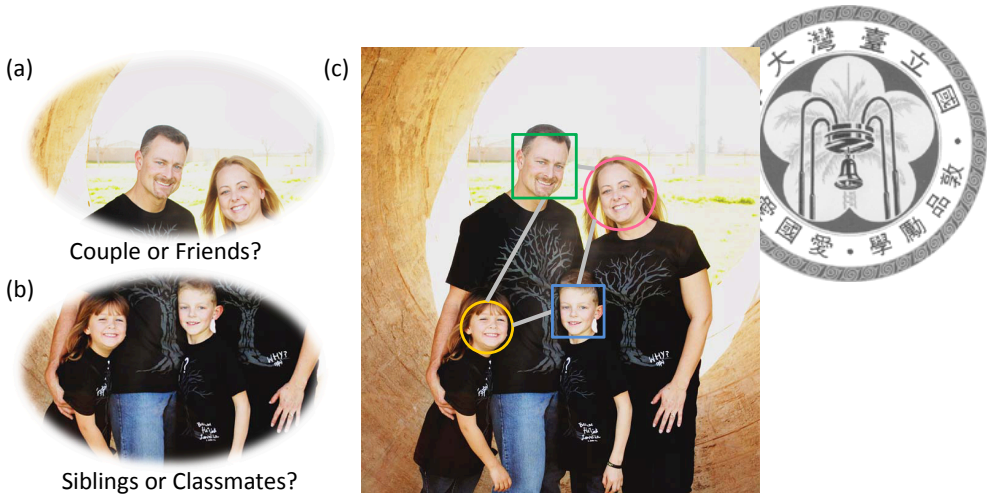


Figure 4.1: It is difficult to determine the pairwise relationships, e.g., couple or friends in (a) and siblings or classmates in (b), if the observations are limited to the pairs only. Interestingly, the ambiguity greatly decreases when all the faces are considered simultaneously as shown in (c). The contribution comes from the contextual cues from all the other faces. The social links resemble a graph parameterized by facial attributes and topological information. Therefore, we propose a novel graph representation to model the potential social subgroups among a group of people and to predict pairwise relationships by leveraging atomic subgroups in the group photos. (Photo courtesy of Spencer Finnley [1].)

prefer the hotels in downtown areas while friend groups are more concerned about transportation convenience. The discoveries evidence the importance in profiling consumer groups. However, transaction logs are not easily accessible due to complicated privacy and commercial issues. As a substitute for transaction logs, group activities can be observed from the growing and freely available sources – social media. As aforementioned, the large-scale user-contributed media possess a huge number of group photos and the associated metadata. Besides, mining from the rich media not only improves the accessibility but also escapes from the huge language gaps and culture differences (cf. Fig. 4.8 (a)).

It has been evidenced that the social interactions and relationships can be observed from the social contexts in a photo [67, 32, 75]; for example, a mother stands close to her child(ren) and they naturally form a subgroup in the group photo. For group analysis, it has been shown that the cohesive subgroups represent an important construct to study a group and individuals [30]. For example, the basic properties of a social group (e.g., a family as Fig. 4.1 (c)), are organized by the **social subgroups** (e.g., a couple as (a) and

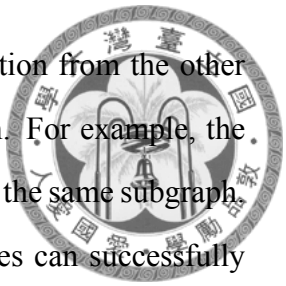
siblings as (b)). In other words, the social subgroups provide meaningful features to infer the overall look of a group.

In addition, social subgroups also play a critical role in understanding individuals, because individuals are influenced the most by the members of their tight subgroup than others [31]. For example, if we have identified a social subgroup as a "couple" relation (as Fig. 4.1 (a)) and have also known the identity of a member (e.g., the wife's name), the identity of the other (e.g., the husband) can be intuitively inferred. Because social subgroups act as the crucial link to holistic group and individuals, we argue to automatically discover informative social subgroups embedded in community-contributed group photos. The mined subgroups would strongly benefit (1) classifying the holistic group types and (2) predicting the pairwise relationships in a (dense) group photo [1].

Intuitively, the correlation of a social group and its social subgroups resembles that of a graph and its subgraphs. Using a graph to link faces in a group photo preserves the social connections among the whole group (e.g., Fig. 4.1 (c)) and does not limit the social contexts to one or two individuals (e.g., Fig. 4.1 (a)(b)). Therefore, we represent the faces in a photo by their gender and age attributes [2], and further consider the spatial proximity among them to form a **face graph**. Also, we enumerate the subgraphs of a face graph to automatically discover the potential subgroups in a group photo. Applying on a large number of consumer photos, we can extract the informative subgroups in the communities, i.e., a vocabulary of face subgraphs. The mined subgraphs are informative to represent a group photo by **a bag of face subgraphs (BoFG)**, which records the occurrence pattern of meaningful social subgroups appearing in certain group types or events. Taking family-type image classification as an example, we demonstrate that learning by BoFG achieved 30.5% relative improvement comparing to the state-of-the-art low-level features for image classification. The proposed framework can excel on photos of more group types (e.g., nuclear family, friends of different ages, etc.) and further enables investigating comprehensive demographics in group photos.

[1] Note that, in this work, we target at group photos with more people since they contain richer social relationships and are more challenging for the existing technologies.

[2] Though we only involve gender and age attributes in this work, there is a potential to extend to dozens of attributes with reasonable detection accuracies (>80%) [42].



Moreover, the mined subgraphs bring the co-occurrence information from the other faces, which benefit predicting pairwise relationships in a face graph. For example, the pairwise relationship "husband-wife" usually co-occurs with a child in the same subgraph. We demonstrate that using the co-occurrence in subgraphs as features can successfully predict four typical pairwise relationships in a family photo. Because labeling names in a photo is very tedious, predicting pairwise relationships is precious to help the association of faces and names for automatic name annotation. In summary, the primary contributions of this work include:

- Proposing a novel graph representation to model a group of people in a photo.
- Devising a methodology to automatically discover informative subgraphs, which resemble the meaningful social subgroups in communities.
- Introducing a novel feature, BoFG, for representing a group of people and demonstrating its effectiveness in recognizing family-type photos.
- Investigating the various factors, i.e., subgraph selection, learning with kernels and sensitivity to normalization, which affect the performance of BoFG.
- Arguing to predict pairwise relationships by the co-occurrence information in the mined subgraphs.

4.2 Related Works

Facial attribute detection is an important technique in facial photo analysis. Dozens of works demonstrate that the detected attributes are quite helpful for image retrieval [45], personalized recommendation [21], and face verification [42]. Facial attributes have been broadly exploited as additional knowledge to categorize or recognize the person of interest. Since consumer photos usually contain more than one person, the coming challenge is how to represent a group of persons. In those cases, simply aggregating or averaging attributes from individuals may lead to information loss. The phenomenon is getting obvious when the group becomes larger and more diverse in attributes.

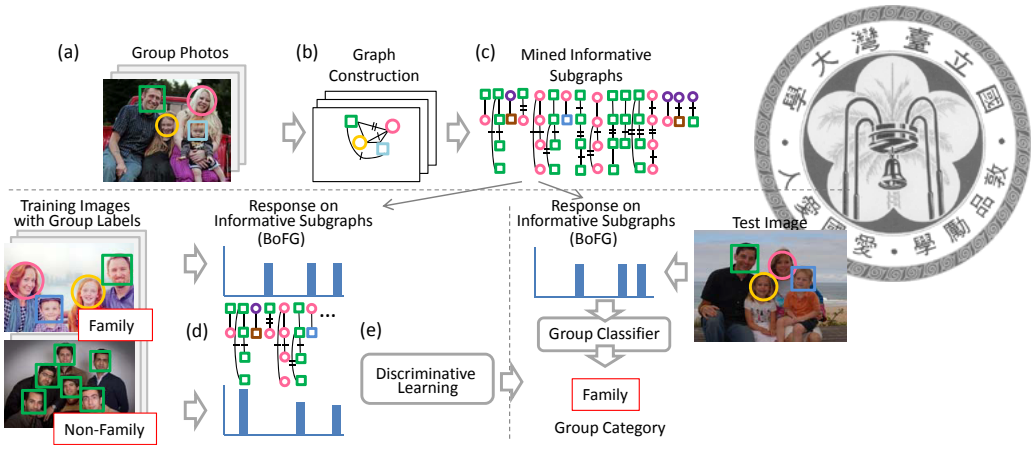
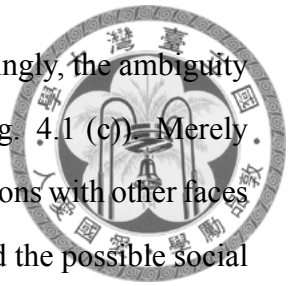


Figure 4.2: Framework – The inputs (a) of our approach are consumer photos containing faces with automatically estimated gender and age attributes (extendable to other attributes as well). The faces in a photo are modeled as a face graph (b) by the proposed graph construction method. From the face graphs, we can automatically discover the informative subgraphs (c) which resemble the social subgroups commonly appearing in communities. We propose to represent a photo by a bag-of-face-subgraphs (BoFG) (d). BoFG preserves the occurrence patterns of social subgroups among a group of faces and acts as effective features for classifying family-type photos by supervised learning (e). (Best seen in color.)

The early studies tend to predefine several typical pairwise relationships (e.g., mother-child, sibling) between people to compensate the lack. Singla et al. [67] used rule-based approach to identify pairwise relationships in photos by a predefined knowledge base. Afterwards, Gallagher [32] gathered real statistics of facial attributes, positions, face size to correlate the social contexts with certain pairwise relationships in consumer photos. Wang et al.[75] further proposed to involve pairwise relationships as cues for learning the correspondence between facial appearances and their names. Pairwise relationships were also adopted as an index for personal photo management [89, 81] and aesthetic assessment [47] when it comes to group photos. The aforementioned works have evidenced that pairwise relationships concern the arrangement of face positions in a photo; however, they only focused on a small set of pairwise relations and limited the social contexts to one or two individuals.

In fact, the social contexts between two persons are only partial factors in inferring their relationship. In a number of cases, the pairwise relationship is ambiguous when only two persons are exposed. For example, it is very difficult to identify whether the two persons in Fig. 4.1 (a) are a couple or just friends. Similarly, we have not enough cues to



identify the relationship between the two kids in Fig. 4.1 (b). Interestingly, the ambiguity extremely drops when we observe the holistic faces in the photo (Fig. 4.1 (c)). Merely relying on the social contexts from a pair of faces neglects the connections with other faces in the social group. On the other hand, if we consider all the faces and the possible social links among the faces as a graph, each of them can propagate its contextual cues to the others. Seeing the potential cues, we propose to exploit the holistic relations in a photo by a face graph. Graph representation has been adopted for modeling co-occurrences and geometrical relations among a set of visual words in image categorization [56]. Due to the large variations in scene and object images, the graph representations are much complicated and very possible to be interfered by cluttered backgrounds. As for face graph, it is relatively easy to filter out unintended points of interest by face detection [72].

Resembling to mining the subcomponents in chemical compound [25], we enumerate all the substructures in consumer photos by subgraph mining [84] to preserve pattern regarding both the facial attributes and the topological proximity. Furthermore, subgraph selection is introduced to reduce the representative dimensionality [53], and thus ensures the scalability for the proposed framework. In the rest of this paper, we will depict how to transform a group photo (Fig. 4.2 (a)) to a face graph (Fig. 4.2 (b)) in Sec. 4.3 and how to discover informative face subgraphs (Fig. 4.2 (c)) as a vocabulary over a set of face graphs. We will further represent every photo as a bag of face subgraphs (Fig. 4.2 (d)) for profiling group types or events in Sec. 4.4 and predict pairwise relationships (Fig. 4.6) in Sec. 4.5. Finally, we demonstrate the effectiveness of BoFG for recognizing family-type photos (as Fig. 4.2 (e)) and the superior performance in predicting pairwise relationships in Sec. 4.6.

4.3 Building a Vocabulary of Facial Subgraphs

A rich amount of social subgroups are embedded in a group photo and also shown effective for understanding group activities and pairwise relationships [30, 31]. We argue to automatically discover the meaningful subgroups from community-contributed photos. In our approach, a social subgroup resembles a subgraph in a face graph constructed from

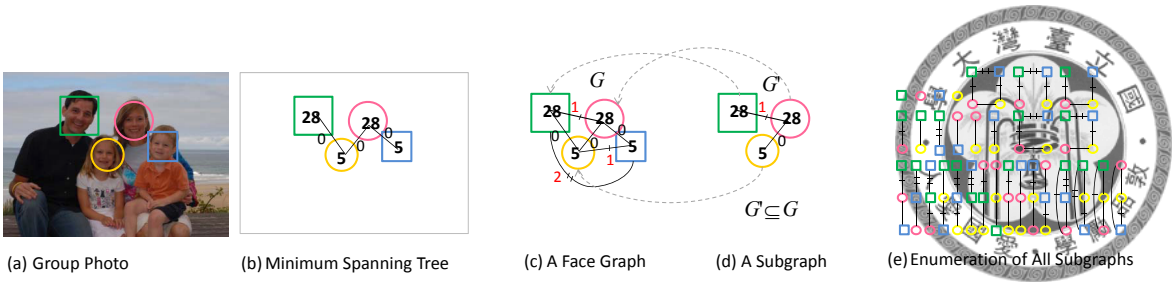


Figure 4.3: Once the faces in a photo are detected as (a), we depict the basic skeleton of a group as a minimum spanning tree (MST) (b) weighted by pixel distance of any two faces. The face vertices are then fully connected as (c), where an edge of two vertices are labeled by the order distance (numbers on the edges) – the length of the shortest path from one vertex to the other in the MST, which represents the social order to other members. To discover potential subgraphs (e.g., (d)) of the face graph, we enumerate all the subgraphs as (e) by subgraph mining. Each of the subgraphs resembles a certain social subgroup. (Best seen in color. Photo courtesy of Steve Polyak [1].)

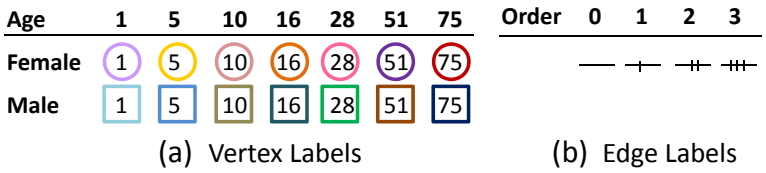


Figure 4.4: For describing a face vertex, the ages are quantized into seven clusters coupled with gender attribute, thus resulting in fourteen vertex labels as (a). The label of an edge between two face vertices is the order distance between them. (b) denotes the edge labels with order distance equal to $0, 1, 2, 3$.

a group photo. We first establish a face graph to model a group of faces as shown in Fig. 4.3 (c). Then, we enumerate the potential subgraphs (as Fig. 4.3 (e)) in a face graph. Applying graph construction and enumeration to all the collected photos, we discover and select a small set of informative subgraphs, which are analogous to the subgroups commonly appearing in consumer photos, as a vocabulary for semantic representations.

4.3.1 Graph Construction

We establish a face graph by all the faces in a photo (as Fig 4.3 (a)), where each face is regarded as a vertex. All of the vertices are categorized by their (automatically detected) facial attributes. For example, the pink circle means a female who is around 28 years old and the blue square means a 5-year-old boy. The ages are quantized into seven clusters

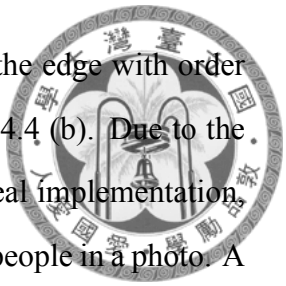
[3] coupled with the gender attribute, thus resulting in fourteen vertex labels (cf. Fig. 4.4 (a)). The spatial distance between any two faces is used as the edge label to represent the closeness of two persons.

The spatial distance between two people is strongly correlated with their interactions and relationship [35]. Therefore, pixel distance is adopted as an informative cues to measure the interpersonal relation in a photo [75, 89, 81]. Unfortunately, pixel distance is sensitive to environment factors like obstacles, atypical poses and culture differences [4]. Another critical problem is how to normalize the pixel distance under various image resolutions and discretize continuous distance into separate degrees of closeness. These concerns make pixel distance lose its superiority (also confirmed in our experiments in Sec. 4.6.4). Actually, for a group of people, "**order distance**" can be another index to evaluate the closeness between any two people. *Order distance between two faces means how the group people intervene the space between them.* The concept originates from that people who do not want to interact would seldom arrange themselves with the other side-by-side [69]. That is, order distance also approximates the tendency to interact in a social group.

The following challenge is how to estimate the order distance of any two faces. Measuring pixel distance will not suffice because people arrange themselves in a free organization rather than in a strict line. We have to shape the basic skeleton of a group at first. Here, we propose to use a minimum spanning tree (MST) to find the basic structure as shown in Fig 4.3 (b). We first leverage the pixel distance of two face vertices as the weights to find a unique MST. This way, we preserve the influence of pixel distance in estimating order distance. Once the MST of a group is obtained, the order distance of two faces can be estimated by the shortest path starting from one vertex to the other on the MST. As shown in Fig. 4.3 (b), the order is counted from 0, which means no face intervenes in between, and steps up progressively as the number of intermediate faces increases. For example, the order distance between the green and blue squares is 2.

In the face graph construction (Fig 4.3 (c)), all the faces are fully connected using

[3] The age categories are decided by the social status of a person, including infant, kid, school-age child, teenager, youth, middle-aged adult and elder, totally seven clusters as shown in Fig. 4.4. Note that the framework can be extended to other attributes such as race, etc.



the order distances as edge labels. For example, the edge labels for the edge with order distance equal to 0, 1, 2, 3 would be denoted as the symbols in Fig. 4.4 (b). Due to the space limitation, we only show four edge labels in the notation. In real implementation, the number of edge labels depends on the number and the structure of people in a photo. A larger group may require more edge labels to denote the growing order distance. Due to the nature of group photos, the range of order distance is bounded [4]. After graph construction, a group photo would be translated into a face graph (as Fig. 4.3 (c)) represented by a 4-tuple $G = (V, E, L, l)$. V is a set of vertices. $E \subseteq V \times V$ is a set of edges. L is a set of labels. l is a mapping for assigning labels to V and E, where $l : V \cup E \rightarrow L$.

4.3.2 Enumeration of Subgraphs

In real life, a group of people comprises many smaller subgroups, which are important characteristics of the group itself [30]. The subgroups resemble the subgraphs in the face graphs constructed from the numerous consumer photos. For example, Fig. 4.3 (a) is a family, and the faces of the family form a face graph G in Fig. 4.3 (c). A subgraph $G' = (V', E', L', l')$ of G should satisfy the criteria, $V' \subseteq V$, $E' \subseteq E$, $L' \subseteq L$ and $l' = l$. By definition, G' in Fig. 4.3 (d) is a subgraph of G. Semantically speaking, G' is a subgroup of parents-child and G is the whole family. In this way, we further enumerate all the subgraphs of a face graph G. After subgraph enumeration, a face graph would be decomposed into a set of subgraphs as shown in Fig. 4.3 (e). An enumerated subgraph indicates a social subgroup, which is not limited in two or three people. The subgraph G' in a face graph G contains $|V'|$ people, where $0 < |V'| \leq |V|$.

To gather various types of social subgroups, we propose to extract the informative social subgraphs from consumer photos. We categorize the subgraphs which preserve the same structure and correspondences in terms of facial attributes (the labels of vertices) and order distance (the labels of edges). To examine the mapping between two subgraphs, we exploit graph isomorphism which allows us to identify identical subgraph representations

[4] In our investigation, the informative subgraphs discovered from consumer photos seldom contain the edges with order distance larger than 4. Therefore, removing the edges with order distance > 4 only has little effects on mining results.

among face graphs (photos). In graph theory, an **isomorphism** of graphs G and H is a bijection f between the vertex sets of G and H, where $f : V(G) \rightarrow V(H)$. That means any two vertices v_α and v_β of G are adjacent in G if and only if $f(v_\alpha)$ and $f(v_\beta)$ are adjacent in H. We write $G \cong H$. For example, the subgraph G_1 in Fig. 4.5 (a) and the subgraph G_2 in Fig. 4.5 (b) are isomorphic ($G_1 \cong G_2$) and are categorized as the same type of subgraph in a vocabulary. The subgraphs G_3 and G_4 in Fig. 4.5 (b) are isomorphic as well ($G_3 \cong G_4$). Similar to calculating text terms in a document, we can count subgraphs of the same type in an image. To accelerate the mining process, we adopt the subgraph mining algorithm [84] which combines enumerating and checking into one procedure. The algorithm transfers graphs to tree-based codes and apply depth first search to speed up the mining process. Finally, the face graphs of a set of consumer photos M would generate a subgraph-image matrix T of $|M| \times |S|$, where S is a subgraph vocabulary mined from M, $\forall s_i, s_j \subseteq S, \nexists s_i \cong s_j$. The m-th row in T contains the frequency of occurrence of subgraph appearing in the m-th image. The i-th column in T comprises the frequency of occurrence of i-th subgraph appearing in each image.

Actually, enumerating subgraphs is time-consuming when the number of vertices in a graph is huge. The computation load is relatively light in our approach since the number of people in a group photo is not as many as the vertices in complicated networks. Besides, the process would be done in the training phase and the mined subgraphs are general for different learning tasks. However, the subgraph matching in the test phase is inevitable. The effort increases along with the size of subgraph vocabulary (S). To ensure scalability, we further introduce the subgraph selection and representation in the next section.

4.4 Bag-of-Face-Subgraphs

The subgraph vocabulary enables interpreting a group photo by a bag-of-subgraphs; for example, the m-th photo in M can be represented by the m-th row in subgraph-image matrix T. Extending the proposed bag-of-face-subgraphs (BoFG) as features for classification tasks would confront two challenges: (1) how to reduce costly graph matching in the test phase, (2) how to translate bag-of-facial-subgraphs into an effective feature

representation.

4.4.1 Subgraph Selection

We conduct feature (subgraph) selection for reducing the substantially large subgraph vocabulary generated in Sec. 4.3.2. The huge amount of subgraphs would be a big problem for scalability in learning models. Besides, it may incur intensive computation for graph matching in the classification (test) phase, and thus makes it infeasible to analyze the large-scale social media. Seeing the requirements, we investigate two approaches for subgraph selection, (1) document frequency and (2) sequential covering, to reduce the size of subgraph vocabulary.

Document Frequency (DF)

Document frequency (df), is a manner of feature selection commonly used in text categorization [87] and visual-words based image classification [85]. df_i is the number of photos that contain the i-th facial subgraph. According to df, the subgraphs are selected by how common they are in the whole training data set without considering the class labels. The approach does not require class labels, and therefore saves the effort to re-select subgraphs for different classification tasks.

Sequential Covering (SC)

In addition to document frequency, we introduce a feature selection approach, sequential covering [53], by taking into account the class labels. Sequential covering algorithm proceeds by iteratively selecting the most discriminative subgraph from the candidates, by measuring its individual classification capability as provided the class labels. Here we treat a subgraph s as a feature (and classifier quality measure $C(s)$) and iteratively select a subgraph s^* which has maximum discriminative capability (classification accuracy) in

the remaining training images compared with the other candidate subgraphs in S.

$$s^* \leftarrow \max_s \frac{\sum_m^{|M|} C_m(s)}{|M|},$$
$$S \leftarrow S \backslash s^*,$$
$$W \leftarrow W \cup s^*, \tag{4.1}$$

where W is the selected subgraphs, $C_m(s)$ is the result of the m-th training image classified by s. $C_m(s) = 1$, if the m-th image is correctly classified, otherwise $C_m(s) = 0$. The process would repeat iteratively until the designated number of subgraphs is selected. To speed up the selection process, we first take document frequency in the training images as the initial ranking. The subgraphs are initially ordered by the confidence scores (i.e., DFs) [48]. The prefiltering step greatly reduces the number of checking processes on the training images.

4.4.2 Feature Representation of Group Photos

Image categorization and retrieval are research problems of great interest; therefore, dozens of image features are proposed for solving different challenges. For example, Histograms of Oriented Gradient (HoG) descriptor [23] shows its superiority to extract subtle edge features for human detection. Pyramid HoG (PHoG) [11] further preserves the traits of spatial layout in the image representation. The aforementioned works have demonstrated that local shape patterns and spatial information are effective for scene classification. As for understanding human activities or group types of a photo, the occurrences of social subgroups should be more critical than the visual shape patterns. Our experiments also confirmed that in Sec. 4.7.

Our approach, BoFG, stands as better representation when considering the facial attributes, the social links, and the spatial proximity for a group of people. Motivated by visual words [68] that extract the local patterns of a image, face subgraphs represent local relation approximated by the people attributes. The feature representation of bag-of-face-

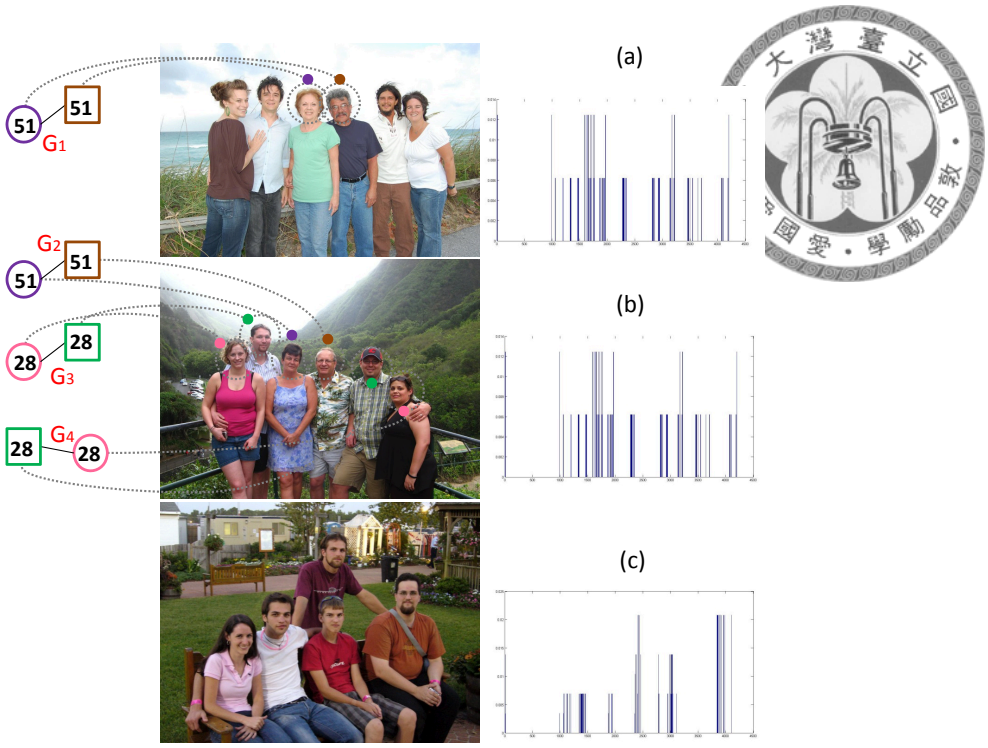


Figure 4.5: Representativeness of BoFG for different social groups (e.g., family vs. non-family). The first and second photos are with the same group type (e.g., family), thus generating very similar BoFG features ((a) and (b)). The third group photo contains much different social subgroups, therefore, the feature vector (c) generated from the photo is quite different.

subgraphs is analogous to that of the bag-of-visual-words [85] and is applicable for group photo classification. The bag-of-face-subgraphs of a group photo are represented by a feature vector f_j,

$$f_j = (t_1, ..., t_i, ...t_{|W|})^T, \tag{4.2}$$

$$t_i = \frac{n_{ij}}{n_j}, \tag{4.3}$$

where W is the selected subgraphs in Eq. 4.1. n_{ij} is the frequency of occurrence of the i-th subgraph appearing in image j. n_j is the number of subgraphs in the image j.

The feature vector f_j contains the histogram information of each subgraph, and is normalized by the total number of subgraphs in image j. Subgraph frequency t_i resembles term frequency (tf) in text domains and likewise each face subgraph is a term and each image is a document. The feature representation is visualized in Fig. 4.5 (a)(b)(c). The first and second photos are of the same group type (i.e., family) and possess similar social

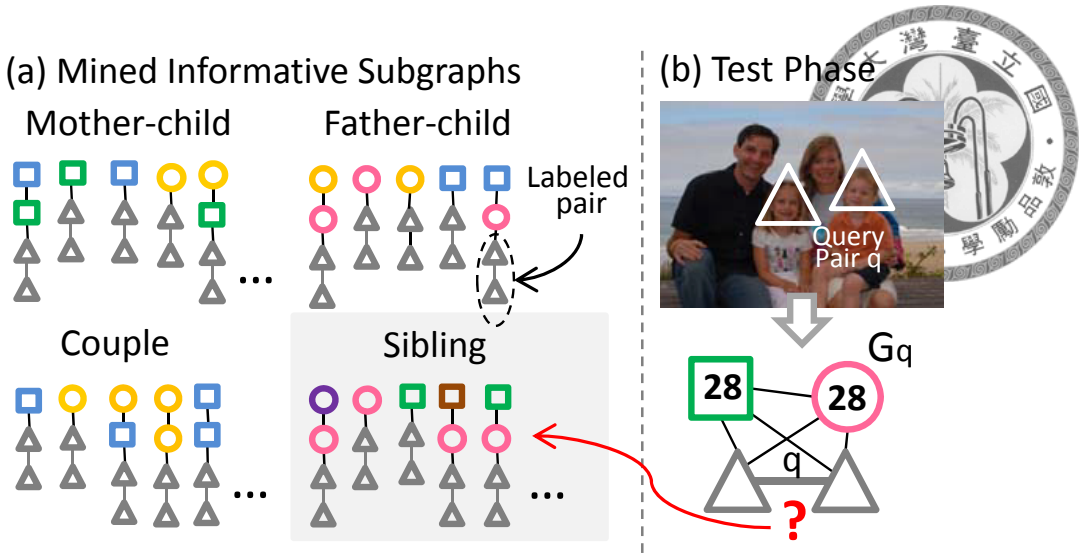


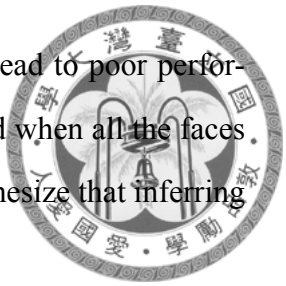
Figure 4.6: (a) shows the mined informative subgraphs (from supervised learning) containing different pairwise relationships including mother-child, father-child, couple and sibling (denoted by gray triangles and their connected line). For "sibling" relation, the informative subgraphs often contain a woman or a man, which are possibly their mother or father. When a query pair of faces (b) arrives, we predict its relationship by checking the presence of the informative subgraphs belonged to each pairwise relationship. (Best seen in color.)

subgroups, thus generating very similar feature vectors ((a) and (b)). On the other hand, the third group photo contains much different social subgroups. Therefore, the feature vector (c) is quite different from (a) and (b). Accordingly, BoFG can capture the informative cues of social subgroups in a group of faces.

4.5 Predicting Pairwise Relationships

Through the studies, users are reluctant to annotate photos and even the faces in photos. The phenomenon makes automatically predicting pairwise relationship (e.g., mother-child, father-child) by image content more important. Besides annotation by face recognition, which is still very challenging for (wild) consumer photos, once the pairwise relationships are identified, the unknown identities are potential to be automatically inferred by partial name labels and their existing social relationships. Traditionally, predicting pairwise relationships relied on the social contexts between the two people, such as relative distance, face size, gender and age attributes [67, 75]. As mentioned in Fig. 4.1 (a)(b),

the social contexts between two people are really limited, and thus lead to poor performance in recognition. However, more contextual cues can be inferred when all the faces are considered in a holistic way as shown in (c). Therefore, we hypothesize that inferring the pairwise relationships by the proposed face graph is promising.



The face graph of a group photo may contain many faces which might inevitably confuse co-occurrence measurement. On the other hand, informative subgraphs are potential to filter out unintended information, and also preserve the co-occurring relationships. Therefore, we exploit the subgraphs co-occurring with the designated pairwise relationship as the features. In the training phase, we manually label pairwise relationships on a face graph according to their social relationships in the photo. By subgraph mining (as the process in Sec. 4.3.2) from the labeled face graphs, we discover the informative subgraphs containing the edges labeled with the designated relationship. As shown in Fig. 4.6 (a), the mined informative subgraphs are different for different designated pairwise relationships (denoted by gray triangles and their connected lines). Taking "sibling" as an example, the informative subgraphs often contain a woman (circle) or a man (rectangle), which are possibly their mother or father.

When predicting a pair q (as shown in 4.6 (b)), we first construct the face graph G_q as the process in Sec. 4.3.1. In G_q, we use graph matching to check the presence of informative subgraph s_i, mined from the training images. Finally, the pairwise relationship r^* is predicted by Naive Bayesian classifier by taking the image frequency $P(s_i|r_l)$ of the informative subgraph s_i in the image collections containing r_l pairwise relationship:

$$r^* = argmax_{r_l} \prod_i P(s_i|r_l), \tag{4.4}$$

Because the subgraphs in G_q is relatively few, appropriately smoothing $P(s_i|r_l)$ is required. In the experiments, we will demonstrate its superiority against prior work in predicting four typical pairwise relationships.

4.6 Experiments

In this section, we will (1) evaluate the effectiveness of BoFG for classifying family-type photos and then (2) evaluate the capability of informative subgraphs for predicting pairwise relationships (in Sec. 4.6.6). The techniques of face detection and facial attribute detection have been developed for years either in academic studies or commercial products. The previous work [42] has shown that the classification accuracy of facial attributes can achieve more than 80% on average. However, to prevent the evaluation from the error caused by face attributes, we experiment on the public data set [32], which provides group photos and the associated attributes of the faces. The data set is collected from social media (Flickr) with specific keywords, and categorized to family images, group images and wedding images. We leverage the keywords as the soft ground truth to obtain family-type images. Totally, 1,167 family images and 1,263 non-family images are retained for experiments which are conducted with 5-fold cross-validation. Note that, we evaluate the proposed approach by the photos containing at least three faces because those groups are more complex and very challenging for analysis and prediction. For groups containing less than three people, the prediction can be intuitively conducted by their attributes and distance directly [75]. Moreover, the proposed approach involves facial attributes rather than face identities; therefore, the discovered informative subgraphs are general and cross-family. In other words, our method operates on a per photo basis rather than a per family basis. We further investigate vital factors such as (1) different learning approaches, (2) the mined informative subgraphs , (3) sensitivity to normalization and (4) subgraph selection to evaluate classifying family photo by BoFG.

4.6.1 Classification

The analysis from text categorization [39] has concluded that Support Vector Machines (SVMs) is excellent in classification for BoW-like representations. The proposed bag-of-facial-subgraphs is in the similar paradigm, therefore we adopt SVMs as the learning method for family photo classification. To maximize the performance, we evaluate three

common SVM kernels for group classification.

$$Linear : K(x, y) = x^T y,$$
$$RBF : K(x, y) = e^{-\gamma \|x-y\|^2},$$
$$RBF - \chi^2 : K(x, y) = e^{-\sum \gamma \frac{(x_k - y_k)^2}{\frac{1}{2}(x_k + y_k)}},$$

where x, y are BoFG feature vectors and $\gamma > 0$. RBF kernel can map the training data to high dimensional space non-linearly, therefore can handle the case when the mapping between class label and feature vector is nonlinear. RBF-χ^2 kernel is another type of non-linear kernel, which are commonly used in image classification.

Although SVMs is a very powerful algorithm for learning high-dimensional features, it is deficient in feature selection and can only work on fixed (provided) features (subgraphs). Due to the high computation cost from subgraph enumeration, Kudo et al. [40] proposed a boosting-based algorithm to couple the subgraph mining and classification, which avoids wasting time to enumerate non-discriminative subgraphs. In the experiments, the aforementioned kernel-based and boosting-based approaches are both applied to compare the effects from different learning methods on the proposed feature representation.

4.6.2 Effects from Learning Approaches

As shown in Fig. 4.7, linear kernel results in the worse accuracy by BoFG features, partially due to the number of training data is relatively few comparing with the adopted high-dimensional features. On the other hand, RBF kernel can non-linearly map training data to the high-dimensional space, therefore leads to better classification results. In our experiments, Chi-square kernel shows its superiority to both linear and RBF kernels, because the proposed features are basically organized by histograms of informative subgraphs. Actually there is no big difference in accuracy generated by linear and non-linear kernels, because the proposed feature representations are sparse and discriminative. Therefore, similar to the cases in document vector or visual word vector, they are more linearly separable [86]. The classification accuracy of the boosting-based approach also

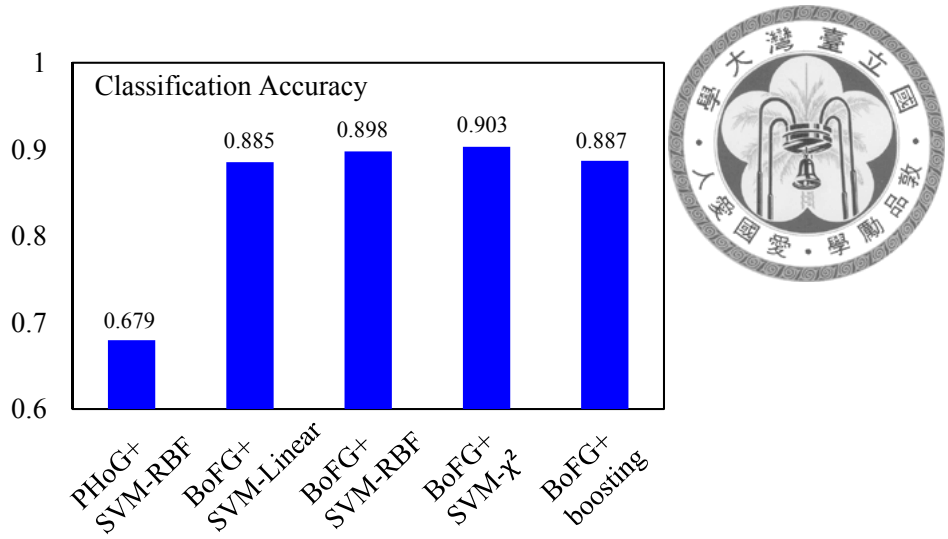


Figure 4.7: Performance comparisons for social group type classification (family vs. non-family) by different features. Chi-square kernel shows its superiority over both linear and RBF kernels as it has been found excellent in histogram representations (e.g., BoW [85], BoFG). Note that, the accuracy for using low-level feature PHoG is only 67.94 %.

achieve 88.67%, which is on par with SVMs with linear kernel. We also train a family photo classifier by SVMs using low-level (and competitive) PHoG feature. The classification accuracy only achieved 67.94%, mainly due to the lack of (semantic) social cues addressed by BoFG.

4.6.3 Mined Informative Subgraphs for Family

In Fig. 4.8, we display the mined informative subgraphs for the two different classes organized by the number of vertices ($|V'|$) in them. Block (a) is the most informative subgraphs in family photos and block (b) holds the counterparts. Obviously, the informative subgraphs in family photos contain faces with larger age gaps (e.g., Fig. 4.8 a-2, a-3, a-4). Besides, the order distance between two faces are much smaller (most are equal to 0). That is, the families tend to stand closer to each other. Also, the couple-like subgroups frequently co-occur with kids in family photos (e.g., a-2). The seniors tend to stand in the center of a family group (e.g., a-4) such that have smaller order distance and usually link to the others. On the other hand, the informative subgraphs in non-family groups are mostly comprised of young people with smaller age gaps (due to the collected dataset photos). People of the same gender stand together (e.g., b-4) more frequently than that in family

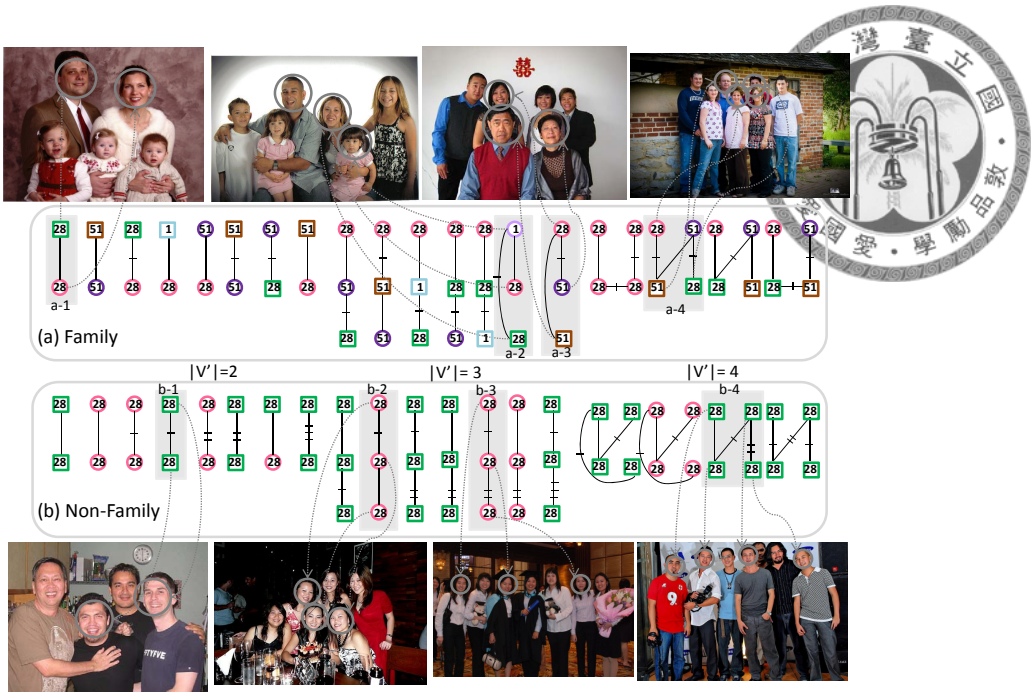


Figure 4.8: Block (a) is the most informative subgraphs (G') in family photos and block (b) holds the counterparts. Both of them are grouped by the number of vertices ($|V'|$). Obviously, the informative subgraphs in family photos contain faces with larger age gaps (e.g., a-2, a-3, a-4). Besides, the order distance between two faces are much smaller; that is, the families tend to stand closer to each other. Also, the couple-like subgroups frequently co-occur with kids in family photos (e.g., a-2). On the other hand, the informative subgraphs in non-family groups are mostly comprised of young people with smaller age gaps. People of the same gender stand together more frequently than that in family photos. They might like to arrange themselves in a row (e.g., b-3, b-4); therefore, the order distance is relatively larger. (Best seen in color.)

photos. They might like to arrange themselves in a row; therefore, the order distance is relatively larger (e.g., b-3, b-4).

4.6.4 Sensitivity in Pixel vs. Order Distance

BoFG adopts order distance as the edge labels and are free of different photo variations (e.g., size, face number, etc.). As for pixel distance, the sensitivity to normalization scale is relatively high. In the experiments, we reveal that pixel distance normalized by different scales results in unstable classification performance. We quantized the pixel distance into different scale ranged from 5 to 15 degrees. The normalized distance degrees are then used as the edge labels. Fig. 4.9 shows the classification accuracy using BoFG constructed by pixel distance and constructed by order distance. All of them are learned by

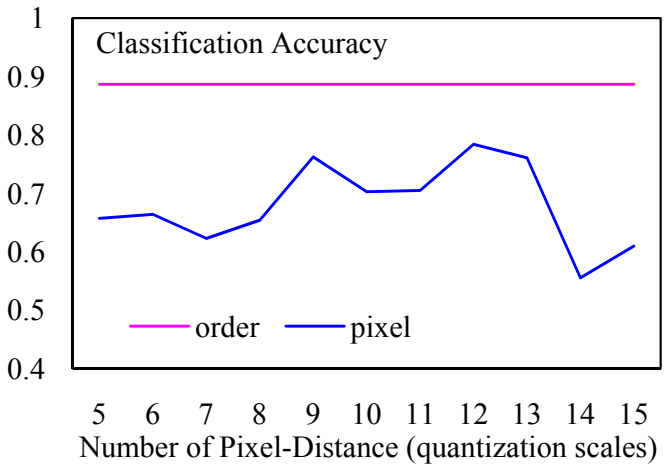


Figure 4.9: The pixel distance adopted in prior work suffers from the high variations in photo sizes, face scales, number of people, etc. The proposed order distance is more robust to the variances.

the boosting-based approach. As it shows, the results of pixel distance fluctuate by varying normalization scales and somehow are affected by the test photos. The proposed order distance can escape from the instability and perform robustly across consumer photos.

4.6.5 Effects of Subgraph Selection

The large number of features (subgraphs) would inevitably incur heavy computation cost in learning models and on-line classification. This problem is especially critical for social media, where the data are growing exponentially. To reduce the size of subgraph vocabulary, we further select the informative subgraphs by document frequency and sequential covering (Sec. 4.4.1). As Fig. 4.10 shows, both subgraph selection methods can effectively retain only 10% subgraphs but still ensure the same classification accuracy (89.75% with 4,315 subgraphs), therefore make the proposed framework more scalable. The performance of sequential covering (Fig. 4.10, DF+SC) is slightly better than document frequency (Fig. 4.10, DF). The difference may come from the utilities of the given class labels, which are provided in sequential covering only. Interestingly, increasing the number of subgraphs is not always a gain for learning. As the experiment shows, the classification accuracy notably degrades while the number of features is larger than 30,000. The drops should be attributed to the overfitting problem in learning from high dimen-

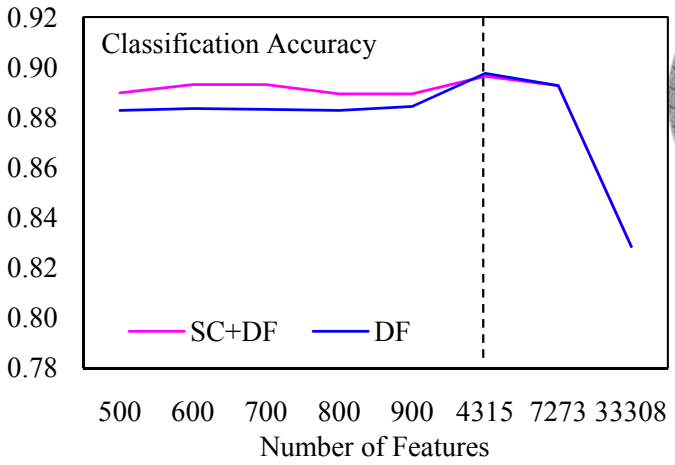


Figure 4.10: Both the subgraph selection methods, document frequency (DF) and sequential covering (SC), can effectively retain only 10% subgraphs but still ensure the classification accuracy and therefore make the proposed framework more scalable. Notably, besides efficiency, subgraph selection is vital since avoiding the overfitting problem commonly observed in learning from high-dimensional features.

sional features.

4.6.6 Performance of Predicting Pairwise Relationships

We use the family photos in [32] for experiments and predict the four pairwise relationships, including couple, mother-child, father-child, sibling. Totally 1,332 pairwise relationships are labeled in 772 photos (at least 250 labels for each pairwise relationship). We use one half of the labeled data for training and one half for testing. To verify the supports from the informative subgraphs, we remove the attributes of the two people involved in a pairwise relationship. That is, the social contexts between the two people are blind both in the training and testing phases. The confusion matrix in Fig. 4.11 shows that solely relying on the information from the subgroups on the face graph can successfully infer the pairwise social relationships and achieve very impressive accuracy. The results also support that the additional information augmented by face graph can compensate errors in estimating social contexts between the pair of faces. We also derive superior performance (36% relative improvement on the average) as comparing with the confusion matrix of classification in [75] which are experimented on the same database [32]. For example,

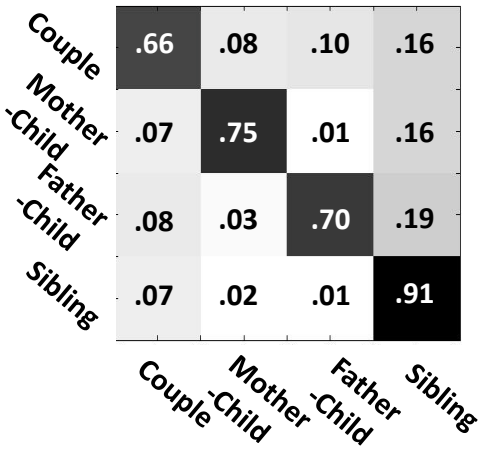


Figure 4.11: The confusion matrix for predicting pairwise relationships. The results outperform those reported in [24] since the informative subgroups provide supplemental supports for determining the pairwise relationship. For example, the most gain is in "sibling" since the co-occurring parent-like subgroups bring more supports.

the recognition of "sibling" relationship in [75] is less accurate and is probably due to the social contexts (relative distance, gender, etc.) between sibling is very ambiguous; as for our work, the co-occurred subgraphs, which frequently have the links to their parents, can provide further supports in recognizing pairwise relationships.

4.7 Remarks

We saw the sheer amount of consumer photos, which mostly contain groups of people. In this paper, we propose a novel graph feature, bag-of-face-subgraphs for describing the social subgroups in a group photo. The informative subgraphs are automatically discovered from community-contributed photos, which reflect the social subgroups commonly appearing in the communities. BoFG preserves the occurrence pattern of social subgroups that are effective for analyzing human-related activities and group types. We demonstrate the capability to classify family-type photos and achieved great improvement (30.5% relatively) against prior works using state-of-the-art low-level visual features. The proposed framework considers subgraph selection for ensuring the scalability as well. Furthermore, the co-occurrence cues in the informative subgraphs can also help predicting pairwise relationships, which benefit inferring unknown identities in group photos and show salient

improvement over the prior work (36% relatively). In the near future, we will investigate more social contexts (e.g., face angles) and people attributes (e.g., race) to enrich the potential social interactions in the emerging group photos. Moreover, we will extend the social groups discovered from the user-contributed photos to inferring implicit interactions in social networks.

4.8 Extensive Applications: Personalized and Group Recommendation for Tourism

4.8.1 Personalized Travel Recommendation

By intuition, we know that some landmarks are female-favored, and some are male-favored. So are by other attributes such as race, age. To examine the correlation between travel behavior and facial attributes, we measure the entropy and the mutual information in predicting next travel location by facial attributes. Taking the correlation between gender attribute and the travel route from Madison Square in Manhattan as an example, the mutual information gained from the facial attribute is 0.5329 (bits), about 25% reduction of the entropy. The result can be illustrated like this, if there are 4 random choices for the next destination, after knowing the facial attribute (e.g., for the male only), the number of choice is down to 3. We can see that the preferences can be partially observed by facial attributes; therefore, the proposed approach involves facial attributes for improving the recommendation performance. At first, in order to mine the travel information within each city, we crawl the photos from the on-line photo-sharing websites (i.e., Flickr). We then use a mean-shift based method on geo-locations of these photos to generate the important locations in each city for the following user trip mining process. We can further identify the demographic information (via automatically detected facial attributes) within travel paths by analyzing the associated photos. By mining the travel patterns users' day trips, we further propose two personalized travel recommendation applications – mobile travel recommendation and route planning, which are entailed by a probability Bayesian

model and dynamic programming technology [21].

4.8.2 Group Recommendation

In fact, consumer activities and user intentions are not limited to only individuals. Group recommendations are essential for daily life, for example, recommending a family-friendly travel path for family group. Beyond the preferences of an individual traveler, the preferences of a travel group, which may comprise people of very diverse attributes, have significant impacts on travel planning as well. Taking a family as an example, recommending a restaurant preferred by an aged family member (e.g., the father) may not satisfy the youngers (e.g., the child). Meanwhile, simply averaging attribute scores of each member in a group may lead to information loss. In Li et al.'s work [46], they analyzed specific transaction logs and found that different types of consumer groups (e.g. family, friends, couple) have quite different preferences when searching for travel accommodations. As a substitute for commercial transaction logs, group activities can be observed from growing and freely available sources ¡V social media. We will demonstrate that how social contexts, e.g., travel group types, can effectively improve recommendation services such as travel recommendation. We will seek the opportunities to leverage these automatically detected people attributes and social contexts mined from the large-scale photos in social media and uncover the differences in user behaviors across demographics.

Chapter 5

Predicting Affective Comments for Images in Social Media

5.1 Introduction

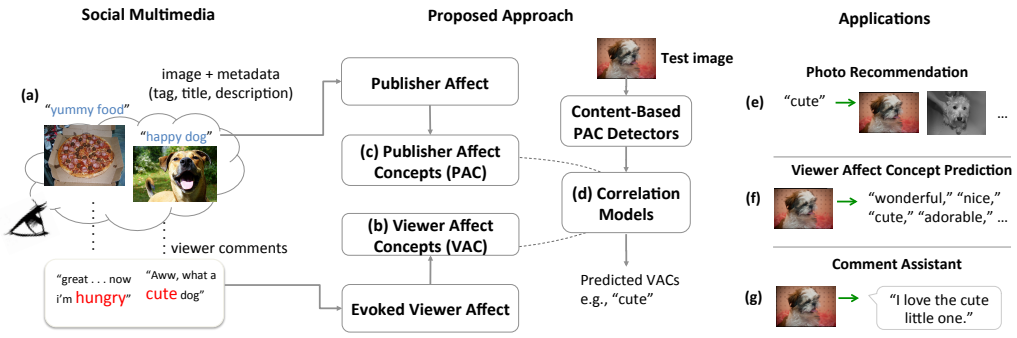


Figure 5.1: System overview of predicting Viewer Affective Concepts (VAC).

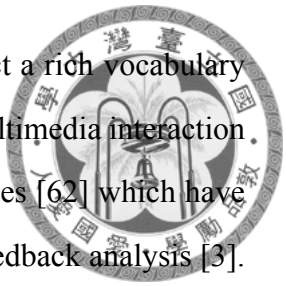
Visual content is becoming a major medium for social interaction on the Internet, including the extremely popular platforms, Youtube, Flickr, etc. As indicated in the saying "a picture is worth one thousand words," images and videos can be used to express strong affects or emotions of users. To understand the opinions and sentiment in such online interactions, visual content based sentiment analysis in social multimedia has been proposed in recent research and has been shown to achieve promising results in predicting sentiments expressed in multimedia tweets with photo content [10, 88]. However, these studies but usually do not differentiate *publisher affect* – emotions revealed in visual con-

tent from the publishers' perspectives, and *viewer affect* – emotions evoked on the part the audience after viewing the visual content.

Different from the previous work [10, 88], we specifically target what viewer affect concepts will be evoked after the publisher affect concepts expressed in images are viewed. Taking Figure 5.1 (a) as an example, after viewing the visual content with "yummy food" as the publisher affect concept, the viewers are very likely to respond with a comment "hungry" (viewer affect concept). Understanding the relation between the publisher affect concepts and the evoked viewer affect concepts is very useful for developing new user-centric applications such as affect-adaptive user interfaces, target advertisement, sentiment monitoring, etc. For example, as shown in Figure 5.1 (f), given an image posting, we may try to predict the likely evoked emotions of the audience even when there are no textual tags assigned to the image (namely visual content based prediction). The results can also be used to develop advanced software agents to interact in the virtual world and generate plausible comments including content relevant affect concepts in response to multimedia content.

The link between image content and subjective emotions it evokes has been addressed in some research on affect [43] and affective content analysis [36]. Meanwhile, from the statistics of the image sharing website Flickr, around 0.2% user comments associated with general images comprise the word "hungry" but the percentage will surge to 14% if we only consider comments associated with images containing visual content "yummy meat." In addition, users are more likely to comment "envious" on the image showing "pretty scene" and "sad" on the image showing "terrible tragedy." The above observations clearly confirm the strong correlation between the publisher affect concepts expressed in the image and the affect concepts evoked in the viewer part.

Visual affect has not been addressed much in terms of the relationships between publisher affect and viewer affect. To the best of our knowledge, this paper presents the first work explicitly addressing publisher affect concepts and viewer affect concepts of images, and aiming at understanding their correlations. Furthermore, we propose to predict viewer affect concepts evoked by the publisher affect concepts intended in image content.



Two challenges arise in this new framework; firstly, how to construct a rich vocabulary suitable for describing the affect concepts seen in the online social multimedia interaction (Figure 5.1 (a)). One option is to adopt the existing emotion categories [62] which have also been used for emotional image analysis [78, 49] and affective feedback analysis [3]. However, the affect concept ontology seen in online social interactions, e.g., "cute" and "dirty" in viewer comments may be different from those used in affect concepts intended by the image publishers. In this paper, we expand the basic emotions to a much more comprehensive vocabulary of concepts, called *viewer affect concepts (VAC)*. We propose to discover a large number of VACs (about 400) directly from million-scale real user comments associated with images on Flickr to represent the evoked affect concepts in viewer comments as shown in Figure 5.1 (b). Specifically, we focus on VACs defined as adjectives that occur frequently in viewer comments and reveal strong sentiment values.

The second challenge is how to model the correlations between publisher affect concepts and viewer affect concepts. We propose to measure such statistical correlations by mining from surrounding metadata of images (i.e., descriptions, title, tags) and their associated viewer feedback (i.e., comments). We develop a Bayes probabilistic model to estimate the conditional probabilities of seeing a VAC given the presence of publisher affect concepts in an image, as shown in Figure 5.1 (d). Furthermore, the mined correlations are used to predict the VACs by automatically detecting publisher affect concepts from image content (Figure 5.1 (c)) without needing the metadata tags of an image.

To demonstrate the effectiveness of the proposed approach, we design several interesting applications – recommend best images for each target VAC (Figure 5.1 (e)), and predict the VACs given a new image (Figure 5.1 (f)). In addition, we show how VACs may lead to designs of novel agent software that is able to select high quality comments for virtual social interaction (Figure 5.1 (g)). The results also suggest the potential of using VAC modeling in influencing audience opinions; for example, the automatically selected comments, when perceived as plausible and relevant, may help elicit more favorable responses from the targeted audiences.

The novel contributions of this paper include,

- hundreds of VACs automatically discovered from millions of comments associated with images of strong affective values.
- a novel affect concepts analysis model that explicitly separates the publisher and viewer affect concepts and characterize their probabilistic correlations.
- a higher than 20% accuracy gain in content-based viewer affect concept prediction compared to the baseline by using publisher affect concepts only.
- novel applications enabled by the proposed affect concept correlation model including image recommendation for targeted affect concepts and social agent software with the automated commenting ability.

5.2 Related Work

Making machine behave like human – not only at the perception level but also the affective level – is of great interest to researchers. Similar motivations have driven recent research in high-level analysis of visual aesthetics [24], interestingness [38] and emotion [49, 36, 78, 65]. These studies attempted to map low level visual features to high-level affect classes. Despite the promising results, the direct mapping from low level features is quite limited due to the well-known semantic gap and the emotional gap as discussed in [78]. Facing such challenges, recently a new approach advocates the use of mid-level representations, built upon Visual Sentiment Ontology and SentiBank classifiers [10]. It discovers about 3,000 visual concepts related to 8 primary emotions defined at multiple levels in [62]. Each visual sentiment concept is defined as an adjective-noun pair (e.g., "beautiful flower," "cute dog"), which is specifically chosen to combine the detectability of the noun and the strong sentiment value conveyed in adjectives. The notion of mid-level representation was also studied in [88], in which attributes (e.g., metal, rusty) were detected in order to detect high-level affect classes.

However, the aforementioned work on visual sentiment analysis only focuses on the affect concepts expressed by the content publishers, rather than the evoked emotions in

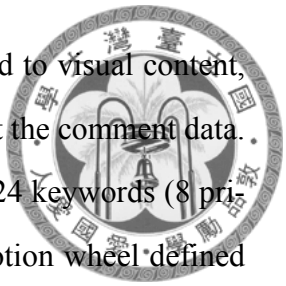
the viewer part. For example, a publisher affect concept "yummy food" expressed in the image often triggers VACs like "hungry" and "jealous." Analysis of review comments has been addressed in a broad spectrum of research, including mining opinion features in customer reviews [37], predicting comment ratings [66] and summarizing movie reviews [92]. Most of these studies focus the structures, topics and personalization factors in the viewer comments without analyzing the content of the media being shared. In this paper, we advocate that viewer responses are strongly correlated with the content stimuli themselves, especially for the visual content shared in social media. Thus, a robust VAC prediction system will need to take into account the publisher affect concepts being revealed in the visual content. Analogous to the large concept ontology constructed for the visual sentiment in [10], we believe a large affect concept pool can be mined from the viewer comments. Such viewer affect concepts offer an excellent mid-level abstraction of the viewer emotions and can be used as a suitable platform for mining the correlations between publisher and viewer affects (e.g., "yummy" evokes "hungry," "disastrous" evokes "sad").

In the remainder of this paper, we will discuss viewer affect concept discovery in Section 5.3 and further introduce the publisher-viewer affect concept correlation model in Section 5.4. The experiments for three applications, image recommendation, viewer affect concept prediction and automatic commenting assistant, will be shown in Section 5.5, with conclusions in Section 5.6.

5.3 Viewer Affect Concept Discovery

This section presents how and what VACs are mined from viewer comments. We introduce the strategy for crawling observation data, then a post-processing pipeline for cleaning noisy comments and finally the criteria for selecting VACs.

Online user comments represent an excellent resource for mining viewer affect concepts. It offers several advantages: (1) the comments are unfiltered and thus preserving the authentic views, (2) there are often a large volume of comments available for major social media, and (3) the comments are continuously updated and thus useful for investigating



trending opinions. Since we are primarily interested in affects related to visual content, we adopt the semi-professional social media platform Flickr to collect the comment data. To ensure we can get data of rich affects, we first search Flickr with 24 keywords (8 primary dimensions plus 3 varying strengths) defined in Plutchik's emotion wheel defined in psychology theories [62]. Search results include images from Flickr that contain metadata (tags, titles, or descriptions) matching the emotion keywords. We then crawl the comments associated with these emotional images as the observation data. The number of comments for each emotion keyword is reported in Table 5.1, totally around 2 million comments associated with 140,614 images. To balance the impact from each emotion on the mining results, we sample 14,000 comments from each emotion, resulted in 336,000 comments for mining VACs.

The crawled photo comments usually contain rich but noisy text with a small portion of subjective terms. According to the prior study of text subjectivity [79, 12], adjectives usually reveal higher subjectivity which are informative indicators about user opinions and emotions. Following this finding, we apply part-of-speech tagging [8] to extract adjectives. To avoid the confusing sentiment orientation, we exclude the adjectives within a certain neighborhood of negation terms like "not" and "no." Additionally, to reduce the influence by spams, we also remove the hyperlinks and HTML tags contained in the comments.

We focus on sentimental and popular terms which are often used to indicate viewer affective responses. Per the first criterion, we measure the sentiment value of each adjective by SentiWordNet [26]. The sentiment value ranges from -1 (negative sentiment) to $+1$ (positive sentiment). We take the absolute value to represent the sentiment strength of a given adjective. To this end, we only keep the adjectives with high sentiment strength (at least 0.125) and high occurrence frequency (at least 20 occurrences). Totally 400 adjectives are selected as viewer affect concepts (VACs). Table 5.2 presents the example VACs of positive and negative sentiment polarities, respectively.

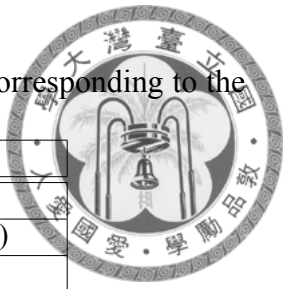


Table 5.1: Flickr training corpus for mining viewer affect concepts corresponding to the 24 emotions defined in psychology.

emotion keywords (# comments)
ecstasy (30,809), joy (97,467), serenity (123,533)
admiration (53,502), trust (78,435), acceptance (97,987)
terror (44,518), fear (103,998), apprehension (14,389)
amazement (153,365), surprise (131,032), distraction (134,154)
grief (73,746), sadness (222,990), pensiveness (25,379)
loathing (35,860), disgust (83,847), boredom (106,120)
rage (64,128), anger (69,077), annoyance (106,254)
vigilance (60,064), anticipation (105,653), interest (222,990)

Table 5.2: The example VACs of positive and negative sentiment mined from viewer comments.

sentiment polarity	**viewer affect concepts (VACs)**
positive	beautiful, wonderful, nice, lovely, awesome, amazing, fantastic, cute, excellent, interesting delicious, lucky, attractive, happy, adorable
negative	sad, bad, sorry, scary, dark, angry, creepy, difficult, poor, sick stupid, dangerous, freaky, ugly, disturbing

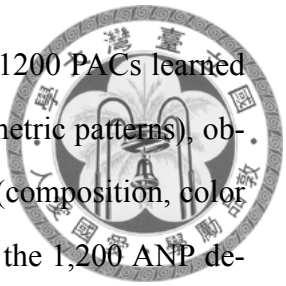
5.4 Publisher-Viewer Affect Correlation

Given an image, we propose to predict the evoked VACs by (1) detecting publisher affect concepts (PACs) in the image content and (2) utilizing the mined co-occurrences between PACs and VACs. This process considers the PACs as the stimuli and aims at exploring the relationships between the stimuli and evoked VACs.

5.4.1 Publisher Affect Concepts

We adopt 1,200 sentiment concepts defined in SentiBank [10] as the PACs in image content (Figure 5.1 (c)). As mentioned earlier, these concepts are explicitly selected based on the typical emotion categories and data mining from images in social media. Each concept combines a sentimental adjective concept and a more detectable noun concept, e.g., "beautiful flower," "stormy clouds." The advantage of adjective-noun pairs is its capability to turn a neutral noun like "dog" into a concept with strong sentiment like "dangerous dog" and make the concept more visually detectable, compared to adjectives only.

The concept ontology spreads over 24 different emotions [62] which capture diverse



publisher affects to represent the affect content. SentiBank includes 1200 PACs learned by low-level visual features (color, texture, local interest points, geometric patterns), object detection features (face, car, etc.), and aesthetics-driven features (composition, color smoothness, etc.). According to the experiment results in [10], all of the 1,200 ANP detectors have F-score greater than 0.6 over a controlled testset.

As shown in Figure 5.1 (c), given an image d_i, we apply SentiBank detectors to estimate the probability of the presence of each publisher affect concept p_k, denoted as $P(p_k|d_i)$. Such detected scores will be used to perform automatic prediction of affect concepts to be described in details later.

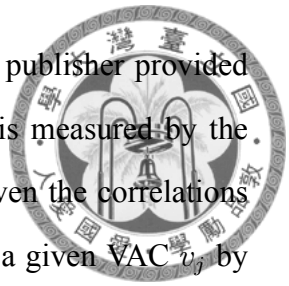
Another version of the PAC data use the "ground truth" labels found in the image metadata for the 1,200 PACs. In other words, we detect the presence of each PAC in the title, tags, or description of each image. Such ground truth PAC data will be used in the next section to mine the correlation between PACs and VACs. One potential issue with using such metadata is the false miss error - images without explicit labels of a PAC may still contain content of the PAC. We will address this issue by a smoothing mechanism discussed in Section 5.4.3.

5.4.2 Bayes Probabilistic Correlation Model

We apply Bayes probabilistic models and the co-occurrence statistics over a training corpus from Flickr to estimate the correlations between PACs and VACs. Specially, we used the 3 million comments associated with 0.3 million images containing rich PAC keywords crawled from Flickr [1] as the training data. Given a VAC v_j, we compute its occurrences in the training data and its co-occurrences with each PAC p_k over the training data θ. The conditional probability $P(p_k|v_j)$ can then be determined by,

$$P(p_k|v_j;\theta) = \frac{\sum_{i=1}^{|D|} B_{ik} P(v_j|d_i)}{\sum_{i=1}^{|D|} P(v_j|d_i)}, \tag{5.1}$$

[1] The training corpus [10] containing the Flickr images and their metadata are downloaded from http://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html



where B_{ik} is a variable indicating the presence/absence of p_k in the publisher provided metadata of image d_i and $|D|$ is the number of images. $P(v_j|d_i)$ is measured by the occurrence counting of v_j in comments associated with images. Given the correlations $P(p_k|v_j;\theta)$, we can measure the likelihood of a given image d_i and a given VAC v_j by multivariate Bernoulli formulation [50].

$$P(d_i|v_j;\theta) = \prod_{k=1}^{|A|}(P(p_k|d_i)P(p_k|v_j;\theta) \\ +(1-P(p_k|d_i))(1-P(p_k|v_j;\theta))). \tag{5.2}$$

A is the set of PACs in SentiBank. $P(p_k|d_i)$ can be measured by using the scores of SentiBank detectors (cf. Section 5.4.1), which approximate the probability of PAC p_k appearing in image d_i. Here, PACs act as shared attributes between images and VACs, resembling the probabilistic model [50] for content-based recommendation [59].

Based on the above probabilistic model, we can answer the question – what is the possibility that an image will evoke a specific VAC. This is very useful for the application of target advertisement applications - selecting the most possible images that will stimulate the given VAC.

Conversely, we can measure the posterior probability of VACs given a test image d_i by Bayes' rule,

$$P(v_j|d_i;\theta) = \frac{P(v_j|\theta)P(d_i|v_j;\theta)}{P(d_i|\theta)}. \tag{5.3}$$

$P(v_j|\theta)$ can be determined by the frequency of VAC v_j appearing in the training data and $P(d_i|\theta)$ is assumed equal over images. The above equation is useful for another interesting application – given an image, we can predict the most possible VACs by the posterior probability in Eq. 5.3. We will demonstrate the performance of these two applications in Section 5.5.2 and 5.5.3, respectively.

5.4.3 Smoothing

In this subsection, we address the issue of the missing associations – unobserved correlations between PACs and VACs. For example, a PAC "muddy dog" will likely trigger the VAC "dirty," but there are no viewer comments comprising this VAC in our data. To deal with such unobserved associations, we propose to add a smoothing factor in the probabilistic model.

Intuitively, some publisher affect concepts share similar semantic or sentimental meaning; for example, "muddy dog" and "dirty dog." More examples can be found in the 1200 publisher affect concepts in SentiBank [10], e.g., "weird cloud" and "strange cloud," "delicious food" and "delicious meat." To this end, we propose to apply collaborative filtering techniques to fill the potential missing associations. The idea is to use matrix factorization to discover the latent factors of the conditional probability ($P(p_k|v_j)$ defined in Eq. 5.1) and use the optimal factor vectors t_j, s_k for smoothing missing associations between PAC p_k and VAC v_j. The matrix factorization formulation can be expressed as follows,

$$\min_{t,s} \sum_{k,j} (P(p_k|v_j) - {t_j}^T s_k)^2, \tag{5.4}$$

Note that, we specifically use non-negative matrix factorization [44] to guarantee the smoothed associations are all non-negatives which can fit the calculation in the probabilistic model. The approximated associations between PAC p_k and VAC v_j can then be smoothed as follows,

$$\hat{P}(p_k|v_j) = {t_j}^T s_k. \tag{5.5}$$

With the smoothed correlations $\hat{P}(p_k|v_j)$, given a VAC v_j, the likelihood with an image d_i is reformulated as,

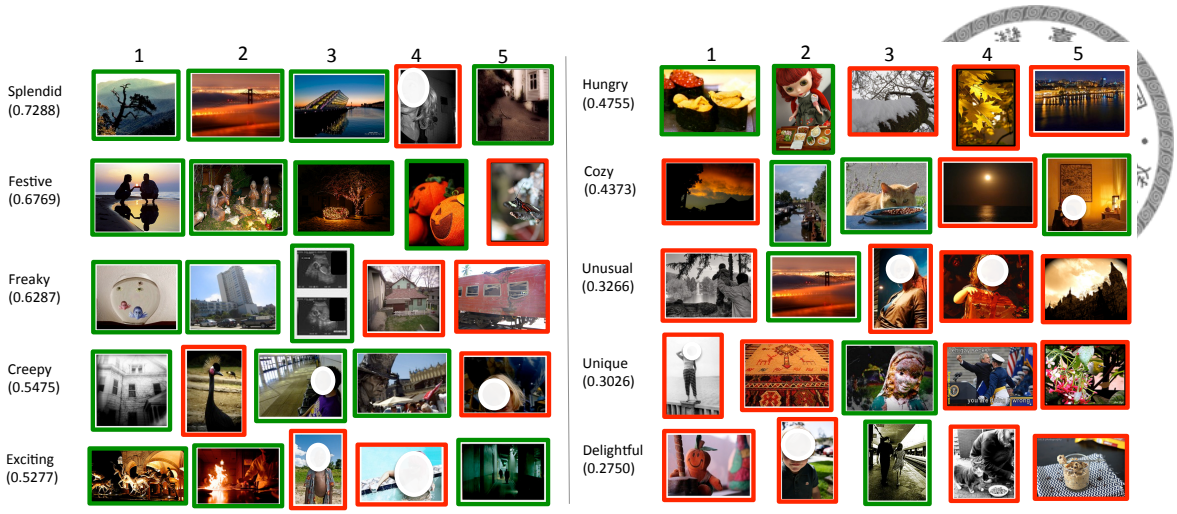


Figure 5.2: Examples of recommended images for each target view affect concept.

$$P(d_i|v_j;\theta) = \prod_{k=1}^{|A|}(P(p_k|d_i)\hat{P}(p_k|v_j) \tag{5.6}$$
$$+(1-P(p_k|d_i))(1-\hat{P}(p_k|v_j))).$$

To avoid floating-point underflow when calculating products of probabilities, all of the computations are conducted in the log-space.

5.5 Applications and Experiments

5.5.1 Dataset for Mining and Evaluation

This section introduces the dataset for mining PAC-VAC correlations and the additional dataset for evaluation. All the images, publisher provided metadata and comments are crawled from Flickr.

(a) **Dataset for mining correlations between PAC and VAC** comprises comments associated with the images (along with descriptions, tags and titles) of 1200 publisher affect concepts publicly released by SentiBank [10]. Totally, around 3 million comments associated with 0.3 million images are collected as the training data. On the average, an images is commented by 11 comments, and a comment comprises 15.4 words. All the

comments are further represented by 400 VACs for mining PAC-VAC correlations. Table 5.3 reports the example mined PAC-VAC correlations ranked by $P(p_k|v_j)$ (cf. Eq. 5.1) and filtered by statistical significance value (p-value). PAC and the evoked VACs may be related but not exactly the same, e.g., "hilarious" for "crazy cat," "delicate" for "pretty flower" and "hungry" for "sweet cake." In some cases, their sentiment are even extremely different, e.g., "cute" for "weird dog" and "scary" for "happy halloween." Because PAC may evoke varied VACs, further considering PAC-VAC correlations will benefit understanding viewer affect concepts. We will demonstrate how PAC-VAC correlations benefit viewer-centric applications in the following sections.

(b) **Test image dataset** contains 11,344 images from the public dataset [10] to conduct the experiments for the proposed three applications, image recommendation by viewer concepts (Section 5.5.2), viewer affect concept prediction (Section 5.5.3), and automatic commenting by viewer affect concepts (Section 5.5.4). Note that, the images from the databases (a) and (b) are not overlapped.

5.5.2 Image Recommendation for Target Affect Concepts

The first application is to recommend the images which are most likely to evoke a target VAC. Given a VAC v_j, the recommendation is conducted by ranking images over the likelihood $P(d_i|v_j)$ measured by Eq. 5.6. For each VAC, 10 positive images and 20 negative images are randomly selected from the test database (cf. Section 5.5.1 (b)) for evaluation. The ground truth of VAC for each image is determined by whether the VAC can be found in the comments associated with this image. For example, if the VACs "nice," "cute" and "poor" are found in the comments of an image, then this image will be a positive sample for "nice," "cute" and "poor" VAC image recommendation. The performance is evaluated by average precision (AP) over 400 mined VACs.

As shown in Table 5.4, the mean value of the average precision of the 100 most predictable VAC is around 0.5321. Mean AP exceeds 0.42 in the best 300 VACs and decreases to 0.3811 over the entire set of 400 VACs. Figure 5.2 shows the top five recommended images of 10 sampled VACs sorted by average precision from top to bottom. We found

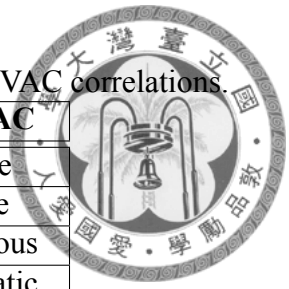


Table 5.3: The significant VACs for example PACs ranked by PAC-VAC correlations.

PAC	#1 VAC	#2 VAC	#3 VAC
tiny dog	cute	adorable	little
weird dog	weird	funny	cute
crazy cat	hysterical	crazy	hilarious
cloudy morning	ominous	serene	dramatic
dark woods	mysterious	spooky	moody
powerful waves	dynamic	powerful	sensational
wild water	dangerous	dynamic	wild
terrible accident	terrible	tragic	awful
broken wings	fragile	poignant	poor
bright autumn	bright	delightful	lovely
creepy shadow	creepy	spooky	dark
happy halloween	spooky	festive	scary
pretty flowers	delicate	joyful	lush
fresh leaves	fresh	green	vibrant
wild horse	wild	majestic	healthy
silly girls	sick	funny	cute
mad face	mad	funny	cute
beautiful eyes	expressive	intimate	confident
sweet cake	yummy	hungry	delicious
nutritious food	healthy	yummy	delicious
shiny dress	shiny	sexy	gorgeous
colorful building	colourful	vivid	vibrant
haunted castle	spooky	mysterious	scary

Table 5.4: Performance of image recommendation for target VACs.

top VACs	100	200	300	overall
MAP	0.5321	0.4713	0.4284	0.3811

that the most predictable VACs are usually of higher visual content and semantic consistency. For example, top recommended images for "splendid" affect concept are correlated with beautiful scenic views (e.g., rank #1, #2, #3 in Figure 5.2) while the "festive" images usually display warm color tones. That suggests the viewers usually have common evoked affect concepts for these types of visual content. Moreover, our approach can recommend images containing more diverse semantics in visual content (e.g., "freaky" and "creepy"), because it aims to learn PAC-VAC correlations from a large pool of image content with rich comments (millions).

As discussed in Section 5.5.1, the comments associated with images are naturally sparse (averagely 11 comments for each image and 15.4 words per comment in our training

Table 5.5: The performance of viewer affect concept prediction given a new image.

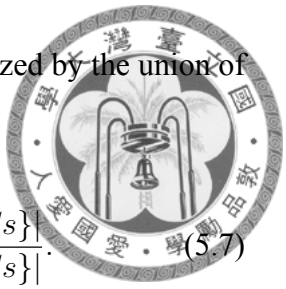
method	**PAC-only** [10]	**Corr**
overlap	0.2295	0.4306 (+20.1%)
hit rate	0.4333	0.6231 (+19.0%)
hit rate (3)	0.3106	0.5395 (+22.9%)

data) and leads to many missing associations. For example, the top 1 and 2 recommended images for "delightful" actually comprise smile, which likely evokes "delightful" affect concept. But because this term was never used in the comments of the images, it was treated as incorrect prediction even though the results should be right upon manual inspection. In general, the VACs without clear concensus among viewers (e.g., "unusual" and "unique") usually are less predictable by the proposed approach.

5.5.3 Evoked Viewer Affect Concept Prediction

The second application, viewer affect concept prediction, is opposite to the aforementioned image recommendation. Given an image d_i, we aim at predicting the most possible VACs stimulated by this image. We measure the posterior probability of each VAC v_j by the probabilistic model in Eq. 5.3. The higher posterior probability means the more likely that the VAC v_j will be evoked by the given image d_i. In addition, we compare our method (Corr) with the baseline using PACs [10] only. Given a test image, the baseline method (PAC-only) chooses all the VACs appearing in the comments associated with the training images which comprises the PACs with the highest detection scores in the test image. In contrast, our method (Corr) considers the soft detection scores of all PACs and use the PAC-VAC correlations described in Eq. 5.3 to rank VACs based on $P(v_j|d_i;\theta)$. The predicted VACs are the VACs with probabilities higher than a threshold. For fair comparisons without being affected by sensitivity of threshold setting, the threshold is set to include the same number of VACs predicted by the baseline method.

The test images are selected from database (b) described in Section 5.5.1 and each test image has comments comprising at least one VAC. Totally 2,571 test images are evaluated by the two performance metrics, overlap ratio and hit rate. Overlap ratio indicates how



many predicted VACs are covered by the ground truth VACs, normalized by the union of predicted VACs and ground truth VACs.

$$overlap = \frac{|\{groundtruthVACs\} \cap \{predictedVACs\}|}{|\{groundtruthVACs\} \cup \{predictedVACs\}|}. \quad (5.7)$$

As shown in Table 5.5, the overlap of our approach (Corr) outperforms the baseline approach by 20.1%. The higher overlap indicates higher consistency between the predicted VACs and the ground truth VACs given by real users.

Considering the sparsity in comments, the false positives in the predicted VACs may be simply missing but actually correct. To address such missing label issue, we further evaluate hit rate, that is, the percentage of the test images that have at least one predicted VAC hitting the ground truth VACs. Hit rate is similar to overlap ratio but deemphasizes the penalty of false positives in the predicted VACs. As shown in Table 5.7, our approach achieves 19.0% improvement in overall hit rate compared to the baseline. The gain is even higher (22.9%) if the hit rate is computed only for the top 3 predicted VACs (hit rate (3)). Some example prediction results are shown in Figure 5.3 (e.g., "gorgeous," "beautiful" for image (a) and "lovely," "moody," "peaceful" for image (b)). In the next section, we will introduce how to exploit the predicted VACs in generating comments for images, for which subjective evaluation will be used instead of the aforementioned overlap and hit ratios.

5.5.4 Automatic Commenting Assistant

We propose a novel application – given an image, automatically recommend comments containing the most likely VACs predicted based on image content. Automatic commenting is an emerging function in social media [2], aiming at generating comments for a given post, e.g., tweets or blogs, by observing the topics and opinions appearing in the content. However, commenting image has never been addressed because of the difficulty in understanding visual semantics and visual affects. Intuitively, commenting behavior is strongly influenced by viewer affect concepts. This motivates us to study automatically

[2] More details regarding commenting bot is introduced in http://en.wikipedia.org/wiki/Twitterbot

commenting images by the proposed viewer affect concept prediction.

The proposed method (Corr) considers the PACs detected from the visual content and the PAC-VAC correlations captured by the Bayesian probabilistic model described in Section 5.4.2. First, we detect the PACs in the test image and construct a candidate comment pool by extracting comments of images in the training set that contain similar PACs (the top 3 detected PACs with the highest $P(p_k|d_i)$) in the visual content. Each comment is represented by bag-of-viewer-affect-concepts as a vector C_l, indicating the presence of each VAC in that comment. Meanwhile, the test image is represented by a vector V_i consisting of the posterior probability $P(v_j|d_i)$ (cf. Eq. 5.3) of each VAC given the test image, d_i. The relevance between a comment and the test image is measured by their inner product $s_{li} = C_l \cdot V_i$. Finally, we select the comment with the highest relevance score s_{li} from the candidate comment pool for automatic commenting. Note that, the images, which are used to extract comments in the candidate pool, do not overlap with the test image set. We compare our method with the two baselines (1) PAC-only: selecting one of the comments associated with another image having the most similar PAC to that of the test image and (2) Random: randomly selecting a comment from the comments of training images.

We conduct user study to evaluate the automatic commenting quality in terms of (1) plausibility, (2) specificity to the image content and (3) whether it is liked by users. Totally, 30 users are involved in this experiment. Each automatic comment is evaluated by three different users to avoid potential user bias. Each user is asked to evaluate 40 automatic comment, each is generated for a test image. The users are asked to rate the comment in three different dimensions (score from 1 to 3 in each dimension), **Plausibility**: how plausible the comment given the specific image content; **Specificity**: how specific the comment is to the image content; **Like**: how much does the user like the comment. Totally, 400 image-comment pairs are included in this investigation.

As shown in Figure 5.4, the most gain appears in plausibility where our method significantly outperforms the other two baselines (PAC-only) and (Random) by 35% and 56% (relative improvement), respectively. Additionally, the proposed approach also clearly improves specificity of the generated comments to the visual content in the image. For

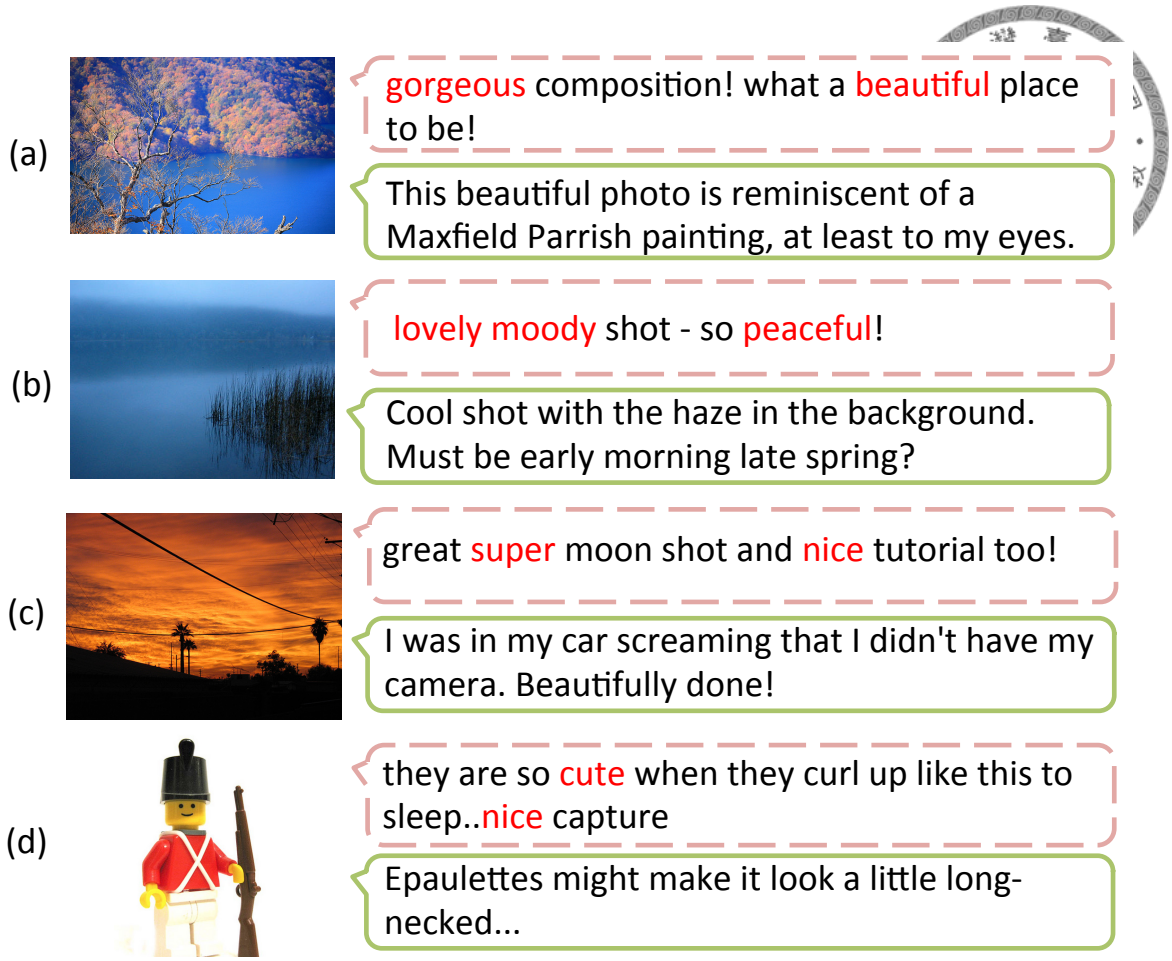


Figure 5.3: Example results of VAC prediction and automatic comment selection.

example, comments containing the affect concept "cute" are selected by our methods for images containing "dog," "kid." Our method (Corr) produces comments that are more liked by users. The potential reasons are, (1) our methods tend to include viewer affect concepts that comprise more emotional words and thus evoke stronger responses from the subjects; (2) our method uses the correlation model that tries to learn the popular commenting behavior discovered from real comments in social multimedia, as described in Section 5.4.2. Overall, commenting by our method has the quality closest to original real comment. Figure 5.3 (a) and (b) shows a few plausible and content relevant fake comments (dashed) automatically generated by the proposed commenting robot. One additional finding is if selected comments mention incorrect objects ("moon" in (c)) or actions ("sleep" in (d)) in the given image, users can easily distinguish them from the real ones. This points out interesting future refinement by incorporating object detection in the automatic commenting process.

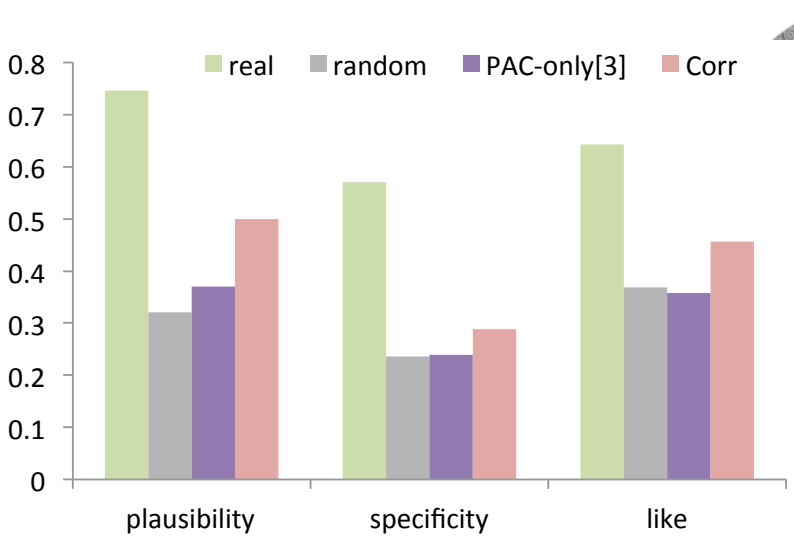


Figure 5.4: Subjective quality evaluation of automatic commenting for image content.

In another evaluation scheme, we focus on plausibility of the faked comments. Each test includes an image, one original comment and the fake comments selected by the proposed method and the baseline (Random). User is asked to decide which one of the four comments is most plausible given the specific image. Comments generated by content-aware method can confuse the users in 28% of times, while the real comment was considered to be most plausible in 61% of times. This is quite encouraging given the fact that our method is completely content-based, namely the prediction is purely based on analysis of the image content and the affect concept correlation model. No textual metadata of the image was used. It is also interesting that 11% of randomly selected comments are judged to be more plausible than the original real comment. However, as discussed earlier, such random comments tend to have poor quality in terms of content specificity.

5.6 Remarks

In this paper, we study visual affect concepts in the two explicit aspects, publisher affect concepts and viewer affect concepts, and aim at analyzing their correlations – what viewer affect concepts will be evoked when a specific publisher affect concept is expressed in the image content. For this purpose, we propose to discover hundreds of viewer affect concepts from a million-scale comment sets crawled from social multimedia. Furthermore, we predict the viewer affect concepts by detecting the publisher affect concepts in

image content and the probabilistic correlations between such affect concepts and viewer affect concepts mined from social multimedia. Extensive experiments confirm exciting utilities of our proposed methods in the three applications, image recommendation, viewer affect concept prediction and image commenting robot. Future directions include incorporation of the viewer profiles in predicting the likely response affects, and extension of the methods to other domains.


國立臺灣大學
教育部・學術倫理・愛國愛人

Chapter 6

Conclusions and Future Work

In summary, we address the human-centric data analytics from the three perspectives, (1) people-centric visual search, (2) demographic data mining and (3) viewer affective comment prediction. We propose a framework for learning facial attributes by crowdsourcing weakly labeled data in social multimedia. Based on these automatically detected attributes, we demonstrate the effectiveness in retrieving images and mining user preferences. Beyond profiling people in visual content, we further propose to analyze the viewer affective feedback elicited by social multimedia. The proposed methodologies are beneficial for cross-discipline research in computational sociology and cognitive psychology. Furthermore, the mined knowledge are essential for advertisement, personalization services and more human-centric applications, which always draw great industry attention in terms of mobile, search, cloud computing and online advertising technologies. We believe these strong links will encourage more opportunities in collaborations and developments between academia and industry.


國立臺灣大學
愛國愛學・勵品教人

Bibliography

[1] available at
http://www.flickr.com/photos/spencerfinnley/5377578656/,
http://www.flickr.com/photos/spolyak/1031569673/.

[2] Internet world stats: The latest internet indicators, usage, penetration rates, population, country size and iso 3316 symbol. http://www.internetworldstats.com/.

[3] I. Arapakis, J. M. Jose, and P. D. Gray. Affective feedback: An investigation into the role of emotions in the information seeking process. In *ACM SIGIR Conference*, 2008.

[4] M. Argyle and J. Dean. Eye-contact, distance and affliation. In *Sociometry*, 1965.

[5] S. Baluja and H. A. Rowley. Boosting sex identification performance. In *International Journal of Computer Vision*, 2007.

[6] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[7] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, 2010.

[8] Bird, Steven, E. Loper, and E. Klein. Natural language processing with python. 2009.

[9] D. Black. The theory of committees and elections. *Cambridge University Press, London, 1958, 2nd ed.*, 1963.

[10] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM International Conference on Multimedia*, 2013.

[11] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *International Conference on Image and Video Retrieval*, 2007.

[12] R. F. Bruce and J. M. Wiebe. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 1999.

[13] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In *International AAAI Conference on Weblogs and Social Media*, 2011.

[14] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15] B.-C. Chen, Y.-H. Kuo, Y.-Y. Chen, K.-Y. Chu, and W. H. Hsu. Semi-supervised face image retrieval using sparse coding with identity constraint. In *ACM International Conference on Multimedia*, 2011.

[16] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European Conference on Computer Vision*, 2012.

[17] Y.-Y. Chen, A.-J. Cheng, and W. H. Hsu. Personalized travel recommendation by mining people attributes and social group types from community-contributed photos. In *IEEE Transactions on Multimedia*, 2013.

[18] Y.-Y. Chen, W. H. Hsu, and H.-Y. M. Liao. Learning facial attributes by crowdsourcing in social media. In *International Conference on World Wide Web*, 2011.

[19] Y.-Y. Chen, W. H. Hsu, and H.-Y. M. Liao. Discovering informative social subgraphs and predicting pairwise relationships from group photos. In *ACM International Conference on Multimedia*, 2012.

[20] Y.-Y. Chen, W. H. Hsu, and H.-Y. M. Liao. Automatic training image acquisition and effective feature selection from community-contributed photos for facial attribute detection. In *IEEE Transactions on Multimedia*, 2013.

[21] A.-J. Cheng, Y.-Y. Chen, Y.-T. Huang, W. H. Hsu, and H.-Y. M. Liao. Personalized travel recommendation by mining people attributes from community-contributed photos. In *ACM International Conference on Multimedia*, 2011.

[22] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world' s photos. In *International Conference on World Wide Web*, 2009.

[23] N. Dalal and B. Trigg. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[24] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, 2006.

[25] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent sub-structure-based approaches for classifying chemical compounds. In *IEEE Transactions on Knowledge and Data Engineering*, 2005.

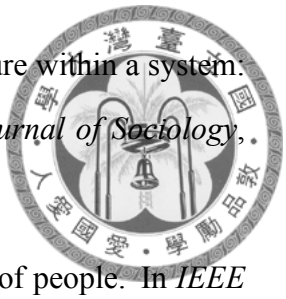
[26] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *International Conference on Language Resources and Evaluation*, 2006.

[27] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *IEEE International Conference on Computer Vision*, 2005.

[28] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 1987.

[29] A. C. Florian Schroff and A. Zisserman. Harvesting image databases from the web. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[30] K. A. Frank. Identifying cohesive subgroups. In *Social Networks*, 1995.



[31] K. A. Frank and J. Y. Yasumoto. Linking action to social structure within a system: Social capital within and between subgroups. In *American Journal of Sociology*, 1998.

[32] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[33] X. Geng, T.-Y. Liu, T. Qin, and H. Li. Feature selection for ranking. In *International ACM SIGIR conference on Research and development in information retrieval*, 2007.

[34] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[35] E. T. Hall. The hidden dimension. In *Culture*, 1966.

[36] A. Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 2006.

[37] M. Hu and B. Liu. Mining opinion features in customer reviews. In *AAAI Conference on Artificial Intelligence*, 2004.

[38] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[39] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Machine Learning and Data Mining Conference*, 1998.

[40] T. Kudo, E. Maeda, and Y. Matsumoto. An application of boosting to graph classification. In *Conference on Neural Information Processing Systems*, 2004.

[41] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *European Conference on Computer Vision*, 2008.

[42] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009.

[43] P. Lang, M. Bradley, and B. Cuthbert. International affective picture system (iaps): Affective ratings of pictures and instruction manual. *Technical Report A-8. University of Florida, Gainesville, FL*, 2008.

[44] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Conference on Neural Information Processing Systems*, 2001.

[45] Y.-H. Lei, Y.-Y. Chen, B.-C. Chen, L. Iida, and W. H. Hsu. Where is who: Large-scale photo retrieval by facial attributes and canvas layout. In *ACM SIGIR Conference*, 2012.

[46] B. Li, A. Ghose, and P. G. Ipeirotis. Towards a theory model for product search. In *International Conference on World Wide Web*, 2011.

[47] C. Li, A. Gallagher, A. C. Loui, and T. Chen1. Aesthetic quality assessment of consumer photos with faces. In *International Conference on Image Processing*, 2010.

[48] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1998.

[49] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*, 2010.

[50] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI Workshop on Learning for Text Categorization*, 1998.

[51] T. Mei, W. H. Hsu, and J. Luo. Knowledge discovery from community- contributed multimedia. *IEEE Multimedia Magazine*, 2010.

[52] T. Mensink and J. Verbeek. Improving people search using query expansions: how friends help to find people. In *European Conference on Computer Vision*, 2008.

[53] T. M. Mitchell. In *Machine Learning*, 1998.

[54] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[55] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *ACM International Conference on Multimedia*, 2009.

[56] S. Nowozin and K. Tsuda. Weighted substructure mining for image analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[57] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 1996.

[58] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Conference on Empirical Methods in Natural Language Processing*, 2002.

[59] M. J. Pazzani and D. Billsus. Content-based recommendation systems. In *The Adaptive Web: Methods and Strategies of Web Personalization. Volume 4321 of Lecture Notes in Computer Science*, 2007.

[60] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *International AAAI Conference on Weblogs and Social Media*, 2011.

[61] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[62] R. Plutchik. Emotion: A psychoevolutionary synthesis. *Harper & Row, Publishers*, 1980.

[63] F. Schroff, A. Criminisi, and A. Zissermann. Harvesting image databases from the web. In *IEEE International Conference on Computer Vision*, 2007.

[64] D. W. Scott. Multivariate density estimation: Theory, practice, and visualization. *John Wiley & Sons, Inc., Hoboken, NJ, USA. doi 10.1002/9780470316849.fmatter*, , 2008.

[65] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Emotion recognition based on joint visual and audio cues. In *International Conference on Pattern Recognition*, 2006.

[66] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro. How useful are your comments?: Analyzing and predicting youtube comments and comment ratings. In *International Conference on World Wide Web*, 2010.

[67] P. Singla, H. Kautz, A. Gallagher, and J. Luo. Discovery of social relationships in consumer photo collections using markov logic. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[68] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, 2003.

[69] R. Sommer. Further studies of small group ecology. In *Sociometry*, 1965.

[70] B. Taneva, M. Kacimi, and G. Weikum. Gathering and ranking photos of named entities with high precision, high recall, and diversity. In *ACM International Conference on Web Search and Data Mining*, 2010.

[71] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2002.

[72] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[73] P. Viola, J. C. Platt, and C. Zhang. Multiple instance boosting for object detection. In *Neural Information Processing Systems*, 2006.

[74] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *IEEE International Conference on Computer Vision*, 2009.

[75] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *European Conference on Computer Vision*, 2010.

[76] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia*, 2010.

[77] S.-Y. Wang, W.-S. Liao, L.-C. Hsieh, Y.-Y. Chen, and W. H. Hsu. Learning by expansion: Exploiting social media for image classification with few training examples. *Neurocomputing*, 2012.

[78] W. Wang and Q. He. A survey on emotional semantic image retrieval. In *IEEE International Conference on Image Processing*, 2008.

[79] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Conference of the Association for Computational Linguistics*, 1999.

[80] C. Wu, C. Liu, H.-Y. Shum, Y.-Q. Xu, and Z. Zhang. Automatic eyeglasses removal from face images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[81] P. Wu, W. Ding, Z. Mao, and D. Tretter. Close & closer: Discover social relationship from photo collections. In *IEEE International Conference on Multimedia and Expo*, 2009.

[82] R. Yan, A. Natsev, and M. Campbell. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[83] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang. Regression from patch-kernel. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[84] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *IEEE International Conference on Data Mining*, 2002.

[85] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *ACM International Conference on Multimedia Information Retrieval*, 2007.

[86] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[87] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*, 1997.

[88] J. Yuan, Q. You, S. McDonough, and J. Luo. Sentribute: Image sentiment analysis from a mid-level perspetive. In *Workshop on Sentiment Discovery and Opinion Mining*, 2013.

[89] T. Zhang, H. Chao, C. Willis, and D. Tretter. Consumer image retrieval by estimating relation tree from family photo collection. In *International Conference on Image and Video Retrieval*, 2010.

[90] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *International Conference on World Wide Web*, 2009.

[91] M. Zhou and H. Wei. Face verification using gaborwavelets and adaboost. In *International Conference on Pattern Recognition*, 2006.

[92] L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. In *ACM International Conference on Information and Knowledge Management*, 2006.