Discovering Informative Social Subgraphs and Predicting Pairwise Relationships from Group Photos

Yan-Ying Chen*†, Winston H. Hsu*, Hong-Yuan Mark Liao†
*National Taiwan University, Taipei, Taiwan
†Academia Sinica, Taipei, Taiwan
yanying@cmlab.csie.ntu.edu.tw, winston@csie.ntu.edu.tw, liao@iis.sinica.edu.tw

ABSTRACT

An increasing number of users are contributing the sheer amount of group photos (e.g., for family, classmates, colleagues, etc.) on social media for the purpose of photo sharing and social communication. There arise strong needs for automatically understanding the group types (e.g., family vs. classmates) for recommendation services (e.g., recommending a family-friendly restaurant) and even predicting the pairwise relationships (e.g., mother-child) between the people in the photo for mining implicit social connections. Interestingly, we observe that the group photos are composed of atomic subgroups corresponding to certain social relationships. For this work, we propose a novel framework to (1) connect faces of different attributes and positions as a face graph and (2) discover informative subgraphs to represent social subgroups in group photos. A group photo can be further represented by a bag-of-face-subgraphs (BoFG) – the occurring frequency of social subgroups, which is informative to categorize specific group types or events. We demonstrate the effectiveness of BoFG in recognizing family photos and achieve 30.5% relative improvement over the state-of-the-art low-level features. Moreover, we propose to predict the pairwise relationships (e.g., husband-wife) in a face graph by the co-occurrence information (e.g., co-occurring with a child) in the mined subgraphs. The experiments demonstrate that the informative social subgroups significantly outperform prior work (36% relatively) which considers merely facial attributes for determining pairwise relationships.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Feature Measurement

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Face graph, face subgraph mining, social relationship

MM'12, October 29–November 2, 2012, Nara, Japan.

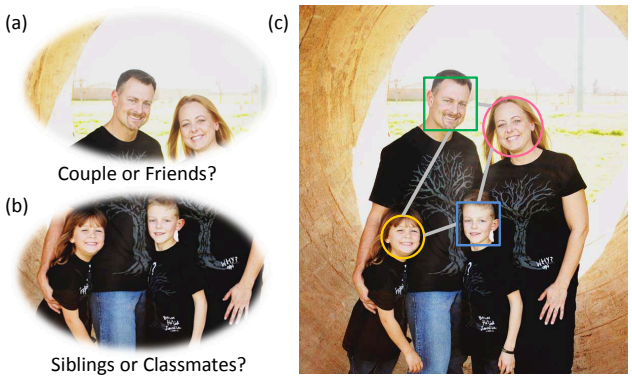

Figure 1: It is difficult to determine the pairwise relationships, e.g., couple or friends in (a) and siblings or classmates in (b), if the observations are limited to the pairs only. Interestingly, the ambiguity greatly decreases when all the faces are considered simultaneously as shown in (c). The 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].)

1. INTRODUCTION

With the prevalence of capturing devices and photo sharing services, the volume of community-contributed photos has increased exponentially. Isola et al. [11] pointed out that the photos containing people are the most memorable, especially those containing families or friends. The desire of social communication motivates users to share group photos on the social media to keep close relationships with others. In our study from more than 17 million photos collected from Flickr using the "family" keyword, we found that around 60% of them contain at least one face. [1]

Such 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, ad-

[1]All of the images presented in this paper attribute to various Flickr users under Creative Commons license, and the images for experiments are downloaded from the public dataset [9].

vertising, 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, Cheng et al. [4] 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 [16], 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 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. 8 (a)).

It has been evidenced that the social interactions and relationships can be observed from the social contexts in a photo [21, 9, 25]; 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 [7]. For example, the basic properties of a social group (e.g., a family as Fig. 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 [8]. For example, if we have identified a social subgroup as a "couple" relation (as Fig. 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 [2].

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. 1 (c)) and does not limit the social contexts to one or two individuals (e.g., Fig. 1 (a)(b)). Therefore, we represent the faces in a photo by their gender and age attributes [3], 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.

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.

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 [15], personalized recommendation [4], and face verification [14]. 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.

[2]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.

[3]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%) [14].

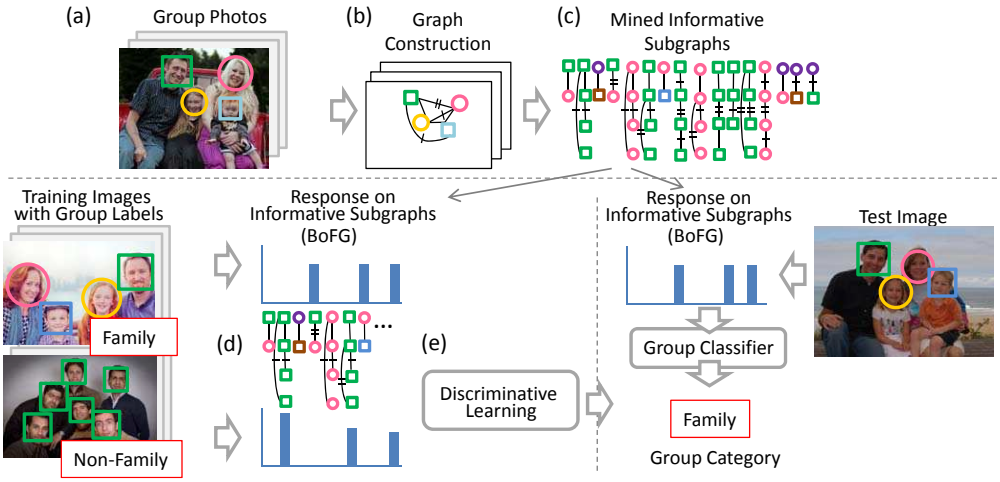

Figure 2: Framework – The inputs (a) of our approach are consumer photos containing faces with automatically estimated gender and age attributes (extendable to other attributes as well). The faces in a photo are modeled as a face graph (b) by the proposed graph construction method. From the face graphs, 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 phenomenon is getting obvious when the group becomes larger and more diverse in attributes.

The early studies tend to predefine several typical pairwise relationships (e.g., mother-child, sibling) between people to compensate the lack. Singla et al. [21] used rule-based approach to identify pairwise relationships in photos by a predefined knowledge base. Afterwards, Gallagher [9] gathered real statistics of facial attributes, positions, face size to correlate the social contexts with certain pairwise relationships in consumer photos. Wang et al.[25] 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 [31, 26] and aesthetic assessment [17] 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. 1 (a) are a couple or just friends. Similarly, we have not enough cues to identify the relationship between the two kids in Fig. 1 (b). Interestingly, the ambiguity extremely drops when we observe the holistic faces in the photo (Fig. 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 [20]. 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 [24].

Resembling to mining the subcomponents in chemical compound [6], we enumerate all the substructures in consumer photos by subgraph mining [27] to preserve pattern regarding both the facial attributes and the topological proximity. Furthermore, subgraph selection is introduced to reduce the representative dimensionality [19], 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. 2 (a)) to a face graph (Fig. 2 (b)) in Sec. 3 and how to discover informative face subgraphs (Fig. 2 (c)) as a vocabulary over a set of face graphs. We will further represent every photo as a bag of face subgraphs (Fig. 2 (d)) for profiling group types or events in Sec. 4 and predict pairwise relationships (Fig. 6) in Sec. 5. Finally, we demonstrate the effectiveness of BoFG for recognizing family-type photos (as Fig. 2 (e)) and the superior performance in predicting pairwise relationships in Sec. 6.

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 [7, 8]. 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 a group photo. We first establish a face graph to model a group of faces as shown in Fig. 3 (c). Then, we enumerate the potential subgraphs (as Fig. 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

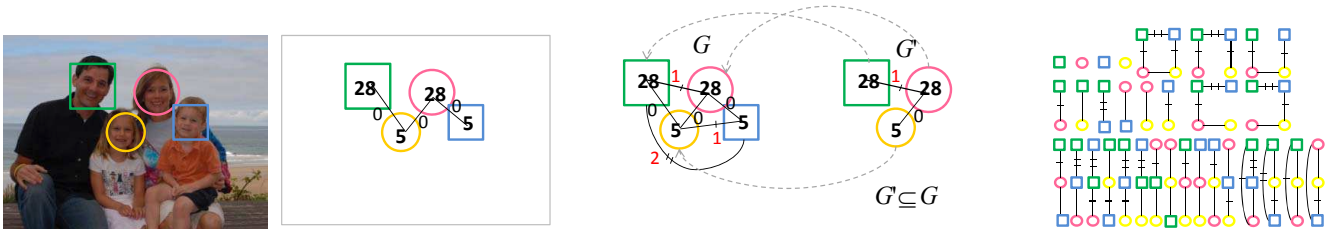

Figure 3: Once the faces in a photo are detected as (a), we depict the basic skeleton of a group as a minimum spanning tree (MST) (b) weighted by pixel distance of any two faces. The face vertices are then fully connected as (c), where an edge of two vertices are labeled by the order distance (numbers 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].)

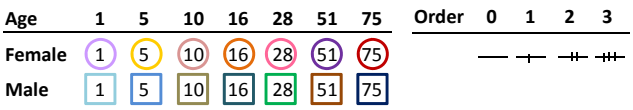

Figure 4: For describing a face vertex, the ages are quantized into seven clusters coupled with gender attribute, thus resulting in fourteen vertex labels as (a). The label of an edge between two face vertices is the order distance between them. (b) denotes the edge labels with order distance equal to 0, 1, 2, 3.

appearing in consumer photos, as a vocabulary for semantic representations.

3.1 Graph Construction

We establish a face graph by all the faces in a photo (as Fig 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 [4] coupled with the gender attribute, thus resulting in fourteen vertex labels (cf. Fig. 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 [10]. Therefore, pixel distance is adopted as an informative cues to measure the interpersonal relation in a photo [25, 31, 26]. Unfortunately, pixel distance is sensitive to environment factors like obstacles, atypical poses and culture differences [2]. 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. 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 [23]. 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 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. 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 3 (c)), all the faces are fully connected using 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 (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 [5]. After graph construction, a group photo would be translated into a face graph (as Fig. 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 $l : V \cup E \rightarrow L$.

3.2 Enumeration of Subgraphs

In real life, a group of people comprises many smaller subgroups, which are important characteristics of the group

[4]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. Note that the framework can be extended to other attributes such as race, etc.

[5]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.

itself [7]. The subgroups resemble the subgraphs in the face graphs constructed from the numerous consumer photos. For example, Fig. 3 (a) is a family, and the faces of the family form a face graph G in Fig. 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. 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. 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 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. 5 (a) and the subgraph G_2 in Fig. 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. 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 [27] 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. 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.

4.1 Subgraph Selection

We conduct feature (subgraph) selection for reducing the substantially large subgraph vocabulary generated in Sec. 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.

4.1.1 Document Frequency (DF)

Document frequency (df), is a manner of feature selection commonly used in text categorization [30] and visual-words based image classification [28]. 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.

4.1.2 Sequential Covering (SC)

In addition to document frequency, we introduce a feature selection approach, sequential covering [19], 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.

$$\begin{aligned} & s^* \leftarrow \max_s \frac{\sum_m^{|M|} C_m(s)}{|M|}, \\ & S \leftarrow S \backslash s^*, \\ & W \leftarrow W \cup s^*, \end{aligned} \tag{1}$$

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) [18]. The prefiltering step greatly reduces the number of checking processes on the training images.

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 [5] shows its superiority to extract subtle edge features for human detection. Pyramid HoG (PHoG) [3] further preserves the traits

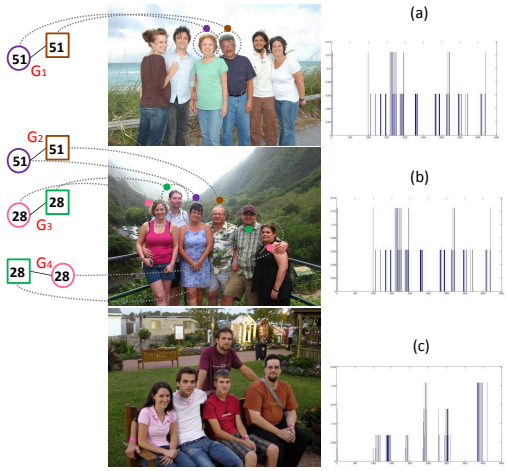

Figure 5: Representativeness of BoFG for different social groups (e.g., family vs. non-family). The first and second photos are with the same group type (e.g., family), thus generating very similar BoFG features ((a) and (b)). The third group photo contains much different social subgroups, therefore, the feature vector (c) generated from the photo is quite different.

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. 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 [22] 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-subgraphs is analogous to that of the bag-of-visual-words [28] and is applicable for group photo classification. The bag-of-face-subgraphs of a group photo are represented by a feature vector f_j,

$$f_j = (t_1, ..., t_i, ...t_{|W|})^T, \quad (2)$$

$$t_i = \frac{n_{ij}}{n_j}, \quad (3)$$

where W is the selected subgraphs in Eq. 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. 5 (a)(b)(c). The first and second photos are of the same group type (i.e., family) and possess similar social 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.

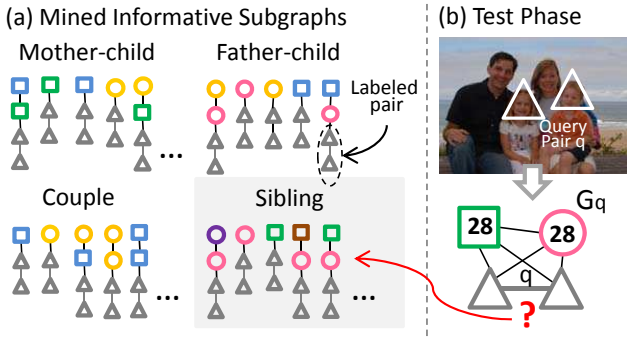

Figure 6: (a) shows the mined informative subgraphs (from supervised learning) containing different pairwise relationships including mother-child, father-child, couple and sibling (denoted by gray triangles and their connected line). For "sibling" relation, the informative subgraphs often contain a woman 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.)

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 [21, 25]. As mentioned in Fig. 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. 3.2) from the labeled face graphs, we discover the informative subgraphs containing the edges labeled with the designated relationship. As shown in Fig. 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 6 (b)), we first construct the face graph G_q as the process in Sec. 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:

$$r^* = argmax_{r_l} \prod_i P(s_i|r_l), \quad (4)$$

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.

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. 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 [14] 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 [9], 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 [25]. 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.

6.1 Classification

The analysis from text categorization [12] 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

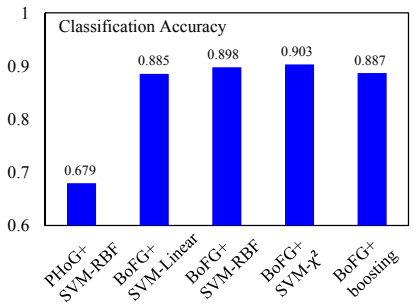

Figure 7: Performance comparisons for social group type classification (family vs. non-family) by different features. Chi-square kernel shows its superiority over both linear and RBF kernels as it has been found excellent in histogram representations (e.g., BoW [28], BoFG). Note that, the accuracy for using low-level feature PHoG is only 67.94 %.

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. [13] 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.

6.2 Effects from Learning Approaches

As shown in Fig. 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 [29]. The classification accuracy of the boosting-based approach also 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.

6.3 Mined Informative Subgraphs for Family

In Fig. 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. Ob-

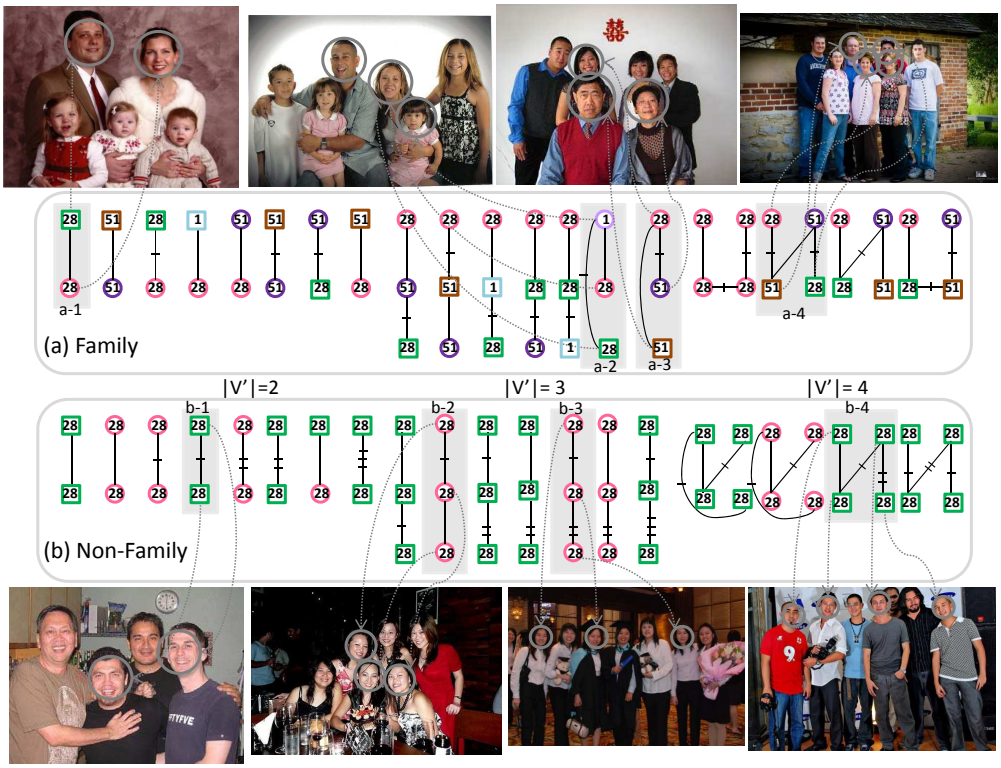

Figure 8: Block (a) is the most informative subgraphs (G') in family photos and block (b) holds the counterparts. Both of them are grouped by the number of vertices ($|V'|$). Obviously, the informative subgraphs in family photos contain faces with larger age gaps (e.g., a-2, a-3, a-4). Besides, the order 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.)

viously, the informative subgraphs in family photos contain faces with larger age gaps (e.g., Fig. 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 photos. They might like to arrange themselves in a row; therefore, the order distance is relatively larger (e.g., b-3, b-4).

The classification results in Fig. 9 confirm the same discoveries in the mined subgraphs. The positions of people in family photos are much concentrated, comparing to the structure of the non-family group (e.g., friends or colleagues). Furthermore, the people in family photos usually form a smaller subgroup implicitly, such as a couple or a parent-child relations, therefore result in many subgraphs with members of opposite sexes. We also show some failure cases not consistent with the learned rules. Fig. 10 (a)(b) are two photos misclassified as families. Those photos also match the subgraphs existing in most family photos, which contain closer subgroups or more subgroups with people of different genders. On the other hand, Fig. 10 (c)(d) show two false negatives misclassified as non-families, where they have relatively smaller age and gender differences in most subgroups. Besides, their positions are less centralized.

We also found that, in some photos, family groups are mixed with non-family groups (e.g., Fig. 10 (b)) thus resulting in uncertainty. Actually, a group of mixed types is commonly seen in certain events such as graduation ceremony or wedding party. Analyzing photos by social subgroups can further clarify the organizations of a group. In such cases, the detailed subgroup information would be useful for softly categorizing a group in a photo.

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. 11 shows the classification accuracy using BoFG constructed by pixel distance and constructed by order distance. All of them are learned by the boosting-based approach. As

(a) Family

(b) Non-Family

Figure 9: Example photos for social group classification. The positions of people in family photos are much concentrated, comparing to the structure of the non-family groups (e.g., friends or colleagues). Furthermore, the people in family photos usually contain smaller subgroups, such as a couple or a parent-child relationship.

(a) (b) (c) (d)

Figure 10: Failure cases for social group classification. (a)(b) are the examples misclassified as families, where the photos contain more closer subgroups with people of opposite sexes. (c)(d) are two cases misclassified as non-families, where they have smaller age and gender differences in most subgroups.

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.

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. 4.1). As Fig. 12 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. 12, DF+SC) is slightly better than document frequency (Fig. 12, 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

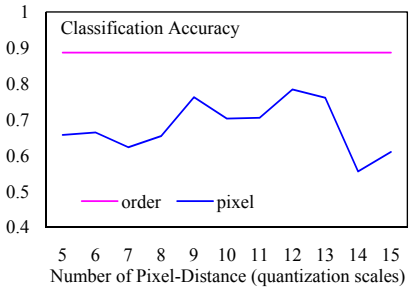

Figure 11: The pixel distance adopted in prior work suffers from the high variations in photo sizes, face scales, number of people, etc. The proposed order distance is more robust to the variances.

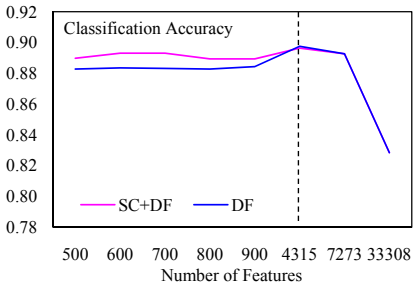

Figure 12: Both the subgraph selection methods, document frequency (DF) and sequential covering (SC), can effectively retain only 10% subgraphs but still ensure the classification accuracy and therefore make the proposed framework more scalable. Notably, besides efficiency, subgraph selection is vital since avoiding the overfitting problem commonly observed in learning from high-dimensional features.

while the number of features is larger than 30,000. The drops should be attributed to the overfitting problem in learning from high dimensional features.

6.6 Performance of Predicting Pairwise Relationships

We use the family photos in [9] 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. 13 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 [25] which are experimented on the same database [9]. For example, the recognition of "sibling" relationship in [25] is less accurate and is probably due to

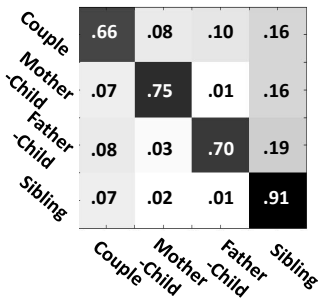

Figure 13: The confusion matrix for predicting pairwise relationships. The results outperform those reported in [24] since the informative subgroups provide supplemental supports for determining the pairwise relationship. For example, the most gain is in "sibling" since the co-occurring parent-like subgroups bring more supports.

the 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.

7. CONCLUSION

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.

8. REFERENCES

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

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

[3] A. Bosch et al. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.

[4] A.-J. Cheng et al. Personalized travel recommendation by mining people attributes from community-contributed photos. In *ACM Multimedia*, 2011.

[5] N. Dalal and B. Trigg. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] M. Deshpande et al. Frequent sub-structure-based approaches for classifying chemical compounds. In *TKDE*, 2005.

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

[8] 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.

[9] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2009.

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

[11] P. Isola et al. What makes an image memorable? In *CVPR*, 2011.

[12] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *ECML*, 1998.

[13] T. Kudo et al. An application of boosting to graph classification. In *NIPS*, 2004.

[14] N. Kumar et al. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[15] Y.-H. Lei et al. Where is who: Large-scale photo retrieval by facial attributes and canvas layout. In *ACM SIGIR*, 2012.

[16] B. Li et al. Towards a theory model for product search. In *WWW*, 2011.

[17] C. Li et al. Aesthetic quality assessment of consumer photos with faces. In *ICIP*, 2010.

[18] B. Liu et al. Integrating classification and association rule mining. In *KDD*, 1998.

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

[20] S. Nowozin and K. Tsuda. Weighted substructure mining for image analysis. In *CVPR*, 2007.

[21] P. Singla et al. Discovery of social relationships in consumer photo collections using markov logic. In *CVPRW*, 2008.

[22] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

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

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[25] G. Wang et al. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, 2010.

[26] P. Wu et al. Close & closer: Discover social relationship from photo collections. In *ICME*, 2009.

[27] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *ICDM*, 2002.

[28] J. Yang et al. Evaluating bag-of-visual-words representations in scene classification. In *MIR*, 2007.

[29] J. Yang et al. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[30] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, 1997.

[31] T. Zhang et al. Consumer image retrieval by estimating relation tree from family photo collection. In *CIVR*, 2010.