Urban Tribes: Analyzing Group Photos from a Social Perspective

Ana C. Murillo[†], Iljung S. Kwak[‡], Lubomir Bourdev[§,*], David Kriegman[‡], Serge Belongie[‡]
[†]DIIS - Instituto de Ingeniería de Aragón. Universidad de Zaragoza, Spain
[§]Facebook. 1601 Willow Road, Menlo Park, CA 94025, USA
[‡]Computer Science and Engineering Department. University of California, San Diego, USA
acm@unizar.es lubomir@fb.com {iskwak,kriegman,sjb}@cs.ucsd.edu

Abstract

The explosive growth in image sharing via social networks has produced exciting opportunities for the computer vision community in areas including face, text, product and scene recognition. In this work we turn our attention to group photos of people and ask the question: what can we determine about the social subculture or urban tribe *to which these people belong? To this end, we propose a framework employing low- and mid-level features to capture the visual attributes distinctive to a variety of urban tribes. We proceed in a semi-supervised manner, employing a metric that allows us to extrapolate from a small number of pairwise image similarities to induce a set of groups that visually correspond to familiar urban tribes such as* biker, hipster *or* goth. *Automatic recognition of such information in group photos offers the potential to improve recommendation services, context sensitive advertising and other social analysis applications. We present promising preliminary experimental results that demonstrate our ability to categorize group photos in a socially meaningful manner.*

1. Introduction

Punk, Goth, Surfer, Preppy, Hipster, Biker, Hippie – French sociologist Michel Maffesoli coined the term *urban tribe* in 1985 to describe subcultures of people who share common interests and tend to have similar styles of dress, to behave similarly, and to congregate together [9]. The goal of this work is identify urban tribes from a group photograph. Members from the same urban tribe are expected to look more similar than members of different tribes. Sports fanatics are more likely to be wearing T-shirts and caps than opera fans who are likely to be wearing formal dress in group photographs. Even the posture of individuals and the configuration of individuals in a group shot are likely to vary by tribe – consider pictures from a society event vs. a group of bikers or hipsters. Where the tribes frequent is also likely to vary – surfers are more likely to be photographed outdoors which looks different than the inside of biker bar. In order to compare group shots, we introduce a metric and determine similarity in a manner that can be used to classify images as belonging to a particular category (e.g., a particular urban tribe) or could be used to cluster images (e.g. unsupervised learning of tribes). We focus here on the classification problem, and note that the group is more powerful than the individual. The collection of social signals coming from the individual appearances, the configuration of the individuals in the photograph and the photograph's setting are all cues contributing to the comparison metric.

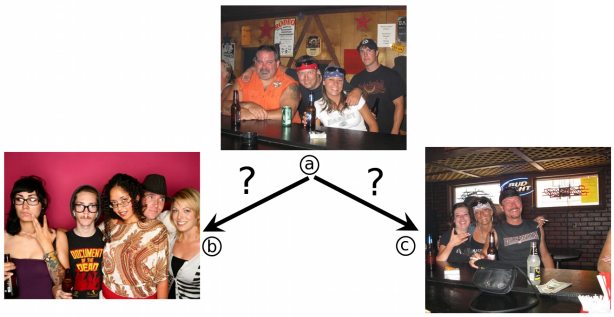


Figure 1. Which groups of people would more likely choose to interact socially: (a) and (b) or (a) and (c)? This is one of the questions we wish to address in this work. In this example, the answer is most likely (a) and (c).

Some individuals may not feel that they associate with any particular urban tribe, while others participate in multiple subcultures. Even so, identifying the urban tribe from a photograph and the imputed membership of individuals can be useful in a number of contexts. In social networking, it can be used to introduce potential friends, people to follow or groups to join. In social media, it can be used to surface content that is likely to be of interest. In advertising, it can help determine an ad's relevance. In surveillance and monitoring, it can provide a social indicator.

*This work was done when the author was at Adobe Systems.

A tremendous amount of social media takes the form of images or videos, and is a largely untapped resource. The analysis of this imagery offers the potential for fruitful interaction between computer science and psychological sociology [4, 8]. The challenge inherent in such interactions is underscored by a recent Forbes article [3], which notes that current search engines fail to capture the contextual information within images. Visual searches often provide unsatisfactory results for queries motivated by a more social perspective (e.g., a trend in fashion or personal style). For example, while search engines such as Google Goggles or TinEye can obtain impressive results for near duplicate retrieval and in certain cases of object and scene retrieval, the results for query images depicting groups of people are largely disappointing. While the dominant colors or distribution of elements may bear a resemblance to the query image, a notion of personal style matching remains elusive.

Problem definition. This work is focused on the analysis of group photos, a very common type of picture in social media; see Fig. 1. Such photos often depict people belonging to the same urban tribe. We attempt to detect these categories of people through their appearance. We focus on the problem of how to represent the images in a way that facilities meaningful comparison, i.e., a metric that captures tribal characteristics. To this end we make use of recent advances in computer vision and machine learning involving person detection and attribute recognition, and we show that it is possible to extract social semantic meaning from group photos.

Social perspective for image analysis. *Social signal processing* is an emerging domain [15] with numerous possible applications. In particular, semantic interpretation of images from social media sources has recently produced promising results for subjective interpretation of action analysis [12]. Other recent work proposes a system that recognizes the occupation of the people in an image [6]. These works are related to our goals of obtaining social information from pictures of people. Some aspects of our processing pipeline are similar, however the categories we want to recognize are completely different and we base our approach on the joint analysis of groups instead of particular individuals.

Our study bears some similarity to recent work on the analysis of group images. This includes an approach for capturing the structure of the group to provide better interpretation of the image [7] or the extension of this work to recognize social relationships between pairs of people, such as family relationships [16]. Group analysis methods usually start with person detection and description, for which one of the leading methods is that of Bourdev and Malik [2]. It is based on the detection of a type of automatically discovered person parts named *poselets*, the effectiveness of which has been demonstrated for person analysis in images, such as human parsing [17] or person description by recognizing semantic attributes such as hair style or clothing type [1].

Contributions. This work presents a novel framework to categorize groups of people from a social perspective, with two main contributions. Firstly we propose a method for group detection and description, based on existing state-of-the-art techniques. One of the key issues is how to model the group jointly, as opposed to dealing with isolated people. Another important part of our proposal is the use of attributes that help describe high level semantic information, e.g., wearing sunglasses, in combination with low level descriptors, e.g., color histograms. The second contribution is a multiple classifier based framework to learn and recognize types of social categories. Additionally, for experimental purposes, we have collected a dataset of group photos depicting people from a variety of urban tribes.

The remainder of this paper is as follows. Section 2 describes the dataset together with our proposed weakly supervised process to build the ground truth categories from a few examples. Section 3 describes the detection and description of groups of people and Section 4 describes the classification methods used to recognize urban tribe categories. Section 5 presents our initial classification results, which serve as a proof of concept of the possible applications of computer vision in the domain of group photos on social networks. We conclude in Section 6.

2. Urban Tribes Image Dataset

This section details the collection and setup of the dataset used in our experiments. This work is focused on analyzing typical group photos, where most of the individuals face the camera.

At the time of experimentation, there was no public dataset to handle our problem. Therefore, we created a dataset by manually collecting images resulting from web searches of two types: public websites of social venues (e.g., club websites) and image searches with keywords such as "group picture" or "party picture." We only collected images where there were two or more people facing the camera. We collected 340 images, 65 of which were annotated by our weakly supervised labeling approach, described next.

2.1. Weakly Supervised Ground Truth Generation

The classification problem studied in this work is difficult due to the subjective nature of social categories. An urban tribe will not always follow its stereotypical depictions and the individuals within a group photo can belong

Figure 2. Images from some of the ground truth categories obtained.

to multiple social categories. This can cause the overall group appearance to be very heterogeneous. We believe that an image of a group of people will have some similar features that can be used to classify the photograph and learn its classes.

It is possible to frame this problem as a form of automatic topic discovery, using techniques such as the simple k-means based approach to the more sophisticated latent Dirichlet allocation. However, this line of research is in its early stages and it is unclear the best way to describe the image and its features in the topic discovery framework and how to evaluate the correctness of the automatic cluster/topic discovery.

Therefore, we decided to focus on classification with human-labeled ground truth. In order to minimize personal biases, we set up a weakly supervised labeling process to determine the classes and representative images of each class. There were no predefined number of clusters, nor names for any cluster. The following process uses human input to produce the reference ground truth categories (tribes):

- Given n training images, we generate a random sampling of pairs (i, j) of images and a human subject answers the question: "Do the groups of people in the two pictures appear likely to interact socially?" For this paper, two co-authors of this paper provided the answers. We only accepted very obvious cases: "don't know" or "maybe" counted as negative, to avoid adding too much noise that may turn into largely heterogeneous groups.

- We create an $n \times n$ connection matrix C where $C(i, j) = 1$ for positive answers to this query and 0 otherwise. The connections refer to the entire image rather than individuals.

- Finally, we find the connected components in C, which provides a set of groups and sample images from each social category. Note that due to conservative labeling the resulting matrix is sparse.

Figure 2 shows some of the social categories and sample images generated by this process. In our experiments, we obtained 14 categories from 340 images. The images provide a perspective on the type of social classes we would like to categorize. In addition, Table 2 includes a short description of each reference category. After inspecting the obtained clusters, it was possible for a human observer to assign a semantic description to each urban tribe in spite of not using predefined labels for the categories.

Because the process is weakly supervised, some of the resulting clusters were less than ideal, such as the two first categories in Fig. 2. L3 is a cluster that contained only 2 images and looks like it should be merged with a larger category, L9. L6 is another class with very few images. In this class all the pictures were underexposed. It would have been possible to manually merge the smaller clusters with

larger clusters, or choose to completely ignore them. Instead, we decided to use the automatically built reference categories to avoid introducing undesired bias and keep the social category formation process as automatic as possible.

3. Person Detection and Description

This section describes our method for building a representation of a group photo. This representation is based on detected people parts and attributes combined with the context of the scene. A *person hypothesis* is a the combination of a detected person and their associated attributes. A set of person hypotheses provides us with a *group hypothesis*. To create a group hypothesis, we combine information from person hypotheses in the same image with global and context information. This group hypothesis is the element used for the social category classification later on.

3.1. Building person hypothesis

Person detection. The first step in group picture analysis is person detection. We run a state-of-the-art person detector followed by an additional face detector to build a robust person hypothesis set.

- To detect an individual in a scene, we use the approach from [2]. In addition to robust person detection, this approach can be used to obtain a set of semantic attributes, described in the next section.

- We use one of the recent top performing face recognition systems [13], through its publicly available API.[1] This face detector provides accurate facial fiducials and high level attributes, described in the next section.

Hypothesis generation. To establish hypotheses, we attempt to match the detected faces with the detected persons. These correspondences allow us to create more robust hypotheses and avoid redundant representations of individuals. In addition, the merged detections form a richer description of the person by combining both face and people descriptors.

However, as shown in Fig. 3, not all face detections will have a corresponding person bounding box and vice-versa. Therefore, to allow richer group descriptions, we consider three types of person hypothesis: (1) person detector + face detector information, (2) only the face detector information and (3) only the person detector information.

3.2. People parts and attributes

We represent each person hypothesis h in a part-based manner, $h_i = \{p_{i1}, \ldots, p_{im}\}$, obtained as follows.

[1] http://developers.face.com

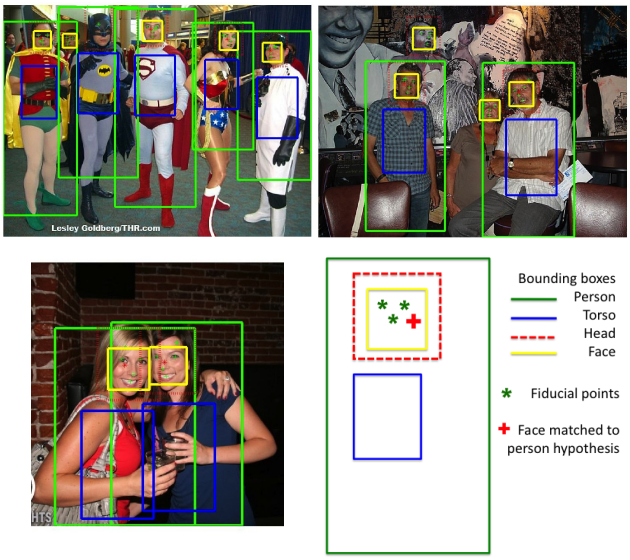


Figure 3. Person detection. Person hypotheses typically consist of matched person (green box) and face (yellow box) detections, but sometimes we fail detecting a face (top left) or a person bounding box (top right). A person hypothesis will always have face and head parts, and optionally full person and torso parts.

Parts detection. The parts we consider in our approach are the face, head and torso. It is possible that some person hypotheses will not have all of parts visible or detected in an image. In the current framework, we use simple bounding boxes to segment these parts, defined as shown in Fig. 3.

For hypotheses of type 1, the face detector provides the face bounding box, and person detector provides the person and torso bounding boxes. Using the detected bounding boxes, we create a head bounding box. The height of the bounding box is defined to be within the top of the person bounding box and the torso bounding box, with a fixed margin from both box boundaries. The width of the head box is set to be wider than the face bounding box with another fixed seperation. If we have a hypothesis without a torso bounding box (type 2), we define the head by augmenting the face bounding box by a fixed amount of pixels in each direction. If we have a hypothesis without face detector information (type 3), we guess the face location to be in the central part of the head region.

Parts description. The description used for each part p is a combination of *low level* commonly used image descriptors, f_{low}, and higher level features, f_{attr}, that we will refer to as *attributes*: $p = \{f_{low}, f_{attr}\}$.

Low Level Features. We compute this set of descriptors f_{low} within each part bounding box. It comprises five

different descriptors,

$$f_{low} = [skin_r,\ rgb_h,\ hue_h,\ HoG,\ textons],$$

for which $skin_r$ is the ratio of detected skin pixels [14] relative to part size, rgb_h is a color histogram of 8 bins per color band, hue_h a histogram of 16 bins for hue values, HoG are the features proposed in [5] and $textons$ are the features computed using the approach and code provided by [10].

High Level Features – Attributes. Attributes are different for each part and we use two sources to obtain them. Note that these attribute computation methods make use of more complex image segmentations than the obtained part bounding boxes. The approach from [2] provides f_{attr1} and the Face.com API provides f_{attr2},

$$f_{attr1} = [gender,\ glasses,\ hair,\ hat,\ tshirt, sleeves],$$
$$f_{attr2} = [gender,\ smiling,\ glasses],$$

where $gender$ is the probability that the individual is male or female; $glasses$ represents the likelihood that the subject is wearing glasses; $hair$ is the likelihood that the individual has long hair; hat is the probability that the person is wearing a hat. $tshirt$ is the likelihood that the subject is wearing a t-shirt; $sleeves$ is the probability that the person has long sleeves on; $smiling$ is the probability that the person is smiling. f_{attr2} attributes are related to the face part, while f_{attr1} comprises $gender$, $glasses$, $hair$ and hat regarding the head part and $tshirt$ and $sleeves$ related to torso part.

3.3. Group and Context

In order to create group level descriptors, we combine the person hypotheses with each other and with global and context information. As mentioned before, our intuition is that the descriptive power of the group may achieve more robust categorization from a social perspective.

Group and context descriptors. Similarly to person part description, we make use of low-level and high-level descriptors:

$$g = [rgb_h, gist, facedist, concomp],$$

where rgb_h is a color histogram of 8 bins per color band of the whole image; $gist$ is the descriptor proposed by [11], computed on the whole image with the code provided by the authors. High level descriptors in this case are $facedist$, which is a histogram of distances between faces, and $concomp$, which is the number of overlapping person bounding boxes. The $facedist$ and $concomp$ features measure the proxmity of individuals within an image.

Group hypothesis. We propose two ways of modeling the whole group with hypotheses and descriptors obtained from previous steps detailed in this section.

Set of People - SoP. In this case, we perform classification on a per person basis, and we use a voting based method to classify the group. More formally, this approach considers a group of people, t, as a set of n person hypotheses, and one context descriptor:

$$t = \{h_1, ..., h_n, g\}. \quad (1)$$

Bag of Parts - BoP. The bag of parts model takes advantage of the attributes visible for each person, as well as group descriptors. We performed classification on a per group/image basis. More formally, this second approach considers a group of people, t, directly a bag of m people parts p, and a context descriptor:

$$t = \{p_1, ..., p_m, g\}. \quad (2)$$

4. Learning Urban Tribes Models

This section describes our proposed framework to learn models for the different social categories. Numerous classification strategies have been considered to train a recognition system from a few samples. The most promising results were obtained with a hierarchy of simple nearest neighbor based classifiers, probably due to the small training set available. We follow a similar strategy for both proposed group representations: run several classifiers, one for each component of the group hypothesis, and then merge their normalized responses to get a consensus on the most similar reference social category.

4.1. Classification using set of people model - SoP

Following the representation in (1), the group hypothesis contains n person hypotheses. We incrementally compute distances from parts, person and group to the different urban tribe categories L_j, and assign to the group hypothesis the label L according to nearest neighbor obtained with d_{SoP}.

Part to part distance. To obtain a distance between two parts of the same type, we compute Euclidean distance between each corresponding descriptor. We normalize the features to be between zero and one. The distance between parts will be the the average part to part distance:

$$d(p_i, p_j) = mean(d_{descr1}, \ldots, d_{descrD}),$$

being $d_{descrD} = ||p_i(D) - p_j(D)||$ for the D^{th} element of the parts descriptor vector defined in the previous section.

Part to category distance. Each part p_i from the hypothesis will find its k nearest neighbors within reference parts of the same kind from category L_j:

$$d_{part}(p_i, L_j) = mean\left(d(p_i, p_{nn1}) \ ... \ d(p_i, p_{nnk})\right),$$

where $part \in \{face, head, torso\}$ and p_{nnk} it the k^{th} nearest neighbor found for p_i. In the experiments, we found that, due to the small size of the dataset available, the best option is $k = 1$.

Person to category distance. We next obtain the distance of individual person hypothesis h_i to each category L_j.

$$d_{cat}(h_i, L_j) = d_{face}(p_i, L_j) + d_{head}(p_i, L_j) + d_{torso}(p_i, L_j)$$

Group to category distance. The distance from a group hypothesis, t, to each category L_j is a combination of the classification results of each person hypothesis in the group, combined as follows:

$$d_{SoP}(t, L_j) = \frac{\sum_{i=1}^{n}(d_{cat}(h_i, L_j))}{n}. \quad (3)$$

4.2. Classification using bag of parts model - BoP

Alternatively, following the representation from (2), we evaluate directly the m people parts obtained. They are used to build a representation similar to the typical bag-of-words and inverted file index commonly used in object categorization.

Multiple vocabularies construction. First, we compute a vocabulary for each part type (head, face, torso) by running k-means clustering on all the detected parts in all the images. This vocabulary will be refered to as $V_{part} = \{w_1, ..., w_k\}$, where w_k is the centroid for cluster k. For each vocabulary, we count how many components in each of the k words belong to each of the j urban tribe categories: $w_k \rightarrow hist_{wk} = [count_{L_1} ... count_{Lj}]$.

In this case, we classify the group hypothesis t with a histogram representing the *bag* of its m parts. We assign each part the closest word in the corresponding part vocabulary and build a normalized histogram that counts the occurrences of each word for each $part$: $sign_{part} = [count_{w1} ... count_{wk}]$.

Group to category distance. We obtain the distance from group hypothesis t to each class L_j as

$$d_{BoP}(t, L_j) = 1 - \frac{\sum_{i=1}^{k}(count_{wi} \times hist_{wi}(j))}{k} \quad (4)$$

As described in the previous subsection, in this representation each group hypothesis gets the label L according to nearest neighbor found using d_{BoP}.

4.3. Combining output of multiple classifiers

The group hypotheses are a combination of person hypotheses and global and context descriptors. We also define a classifier for the global and context descriptors. The output of this classifier is merged with the d_{SoP} or d_{BoP} for a final classification.

Classification using context and global descriptors - Gl We classify the global and context descriptors using nearest neighbor search. We compare the query image global descriptor g to the same descriptor from all images e that belong to each category L_j. As in previous steps, we normalize distance corresponding to each descriptor to $[0, 1]$, to allow fair combination of all of them.

$$d_{global}(t, L_j) = min_{j=1}^{e}(\ |g_t, g_j|\ ) \quad (5)$$

Combination of classifier results. Each of the previously described classifiers provides a distance to each possible urban tribe category c. We obtain the category L_j assigned to a query group image t as a combination of results from classifiers *SoP + Gl* or *BoP + Gl*:

$$L = \arg\min_{j \in [1...c]} (d_{SoP}(t, L_j) + d_{global}(t, L_j)),$$
$$L = \arg\min_{j \in [1...c]} (d_{BoP}(t, L_j) + d_{global}(t, L_j)).$$

5. Experiments

In this section, we describe our experiments to validate the proposed representation and classifiers for urban tribes. We have tested the methods using the labeled images that were used to define the urban tribe categories as well as unlabeled images. For the labeled images, we used cross validation testing to quantitatively evaluate accuracy. Tests on the unlabeled data provide an idea of how well the representation generalizes to harder test data, and these results could only be verified by human inspection.

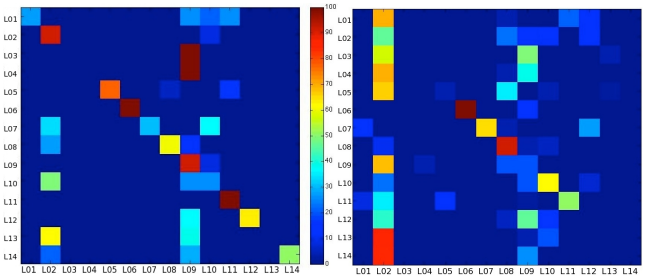
5.1. Experimental Setting

All experiments used the social categories that were determined using the method described in Sec. 2.1, and the classification techniques described in Sec. 4. We used the descriptors from Sec. 3 for both methods. For the BoP approach, we created vocabularies using k-means, with $k = 20$, for each part (head, torso, face).

For the SoP approach we experimented with different variations of nearest neighbor search to assign the closest social category to each person hypothesis in the group. Besides k-nearest neighbors with different values of k, we used a search weighted with the variance of the feature values within a group. However, the best performing method from these variations was basic nearest neighbor search

Table 1. Summary of classification results with different sets of features.

parts used:	face + head + background/context	only face	only head	only context
descriptors used:	$P_{attr} + P_{low} + Gl : P_{attr} + P_{low} : P_{low} + Gl$	$P_{attr} + P_{low} : P_{low}$	$P_{attr} + P_{low} : P_{low}$	Gl
SoP (NN)	0.28 : 0.25 : 0.23	0.37 (0.35) : 0.13	0.26 (0.25) : 0.23	0.16
BoP (k=20)	0.51 : 0.46 : 0.44	0.43 (0.43) : 0.40	0.31 (0.30) : 0.31	0.16



a. BoP b. SoP

Figure 4. Classification confusion matrix for the BoP and SoP methods using low level descriptors, attributes and context descriptors. Each row corresponds to a ground truth social category, and each column corresponds to the classification output. The cell color denotes the percentage of images with the given ground truth label that were classified to the column's label.

(i.e., $k = 1$), probably due to the small size of the training set.

5.2. Performance evaluation

To quantitatively evaluate the classifiers, we ran leave-one-out cross validation tests, i.e., remove one sample from each of the 14 social categories in the reference data and classify those 14 samples with regard to the rest. We generated 100 different test combinations, and the results are summarized in Table 1. We evaluated performance of using all the features or different subsets of them. P_{attr} corresponds to high level parts descriptors (attributes); P_{low} corresponds to low level part descriptors; Gl corresponds to global and context descriptors. Note that the inclusion of high level attributes in the parts description provides significant improvements in the classification.

From the overall results shown in Table 1, BoP is more accurate than SoP and potentially is a good way to model images of urban tribes. Chance classification is 1/14 = 0.07, and all results in Table 1 are well above chance. These results show that this task is not as impossible as it once may have seemed. By investigating the confusion matrix in Fig. 4 and results in Table 2, we can better understand the two methods. The BoP technique was confused between L09 and the clusters L03 and L04. Interestingly, visual inspection of these images revealed that these social categories could have been merged with L09. They depict club shots with more formal, well-dressed people. Although our weakly supervised labeling did not include these images in cluster L09, the classifier determined that these images were similar ones in L09. On the other hand, the SoP nearest neighbor classification presents a clear bias towards group L02, classifying many test images as L02. It showed high confusion between L06 and L08, two categories that could have been merged during training as mentioned previously.

Finally Table 2 shows a comparison of the accuracy for each social category. From this table, there does not seem to be a technique that clearly does better at classifying all groups. This is possibly due to difference of homogeneity of the social categories, and that one technique lends itself to a more homogenous group while the other to a more heterogeneous group. In the future it would be interesting to model this and potentially tailor classification schemes to this information.

Table 2. Social Categories: Fourteen categories were automatically determined based on 65 reference images. Accuracy of the BoP and SoP methods for each social category

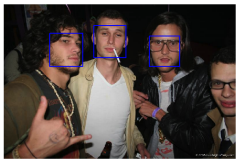
Label	Description	Reference Images	BoP	SoP
L01	Cosplayers	4	-	- -
L02	Informal pub	8	++	-
L03	Pub-Club	2	- -	- -
L04	Pub-Club	2	- -	- -
L05	Beach Party	3	++	- -
L06	Pub - no light	2	++	++
L07	Hippie - outdoors	3	-	+
L08	Hipsters	10	+	++
L09	Club	10	++	- -
L10	Bikers Pub	4	-	+
L11	Formal event (e.g. opera)	5	++	-
L12	Japanese girls	5	+	- -
L13	Country Bar	3	-	- -
L14	Sports Bar/event	4	+	- -

++: very good (>75%), +: good (>50%), -: bad (>25%), - -: very bad

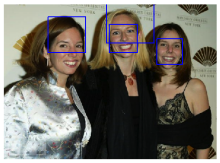
5.3. Additional tests with unlabeled examples

The previous experiments showed that the proposed methods are promising ways to classify group shots into social categories. Because we performed cross validation with a small set of images, we wanted to evaluate the methods on completely unseen data. Recall that while the classifiers are not trained on test samples in the cross validation experiment, the categories themselves were defined using this set of images.

In this experiment, we provide a classifier an unlabeled image as input, and it returns the closest match among the fourteen social categories. This does not provide quantitative results, but can provide insight into the the ability of the methods to deal with unseen images containing a group of people comprising an unseen urban tribe. We show some



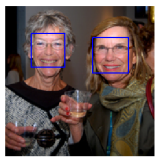
correctly classified as L08 (SoP)



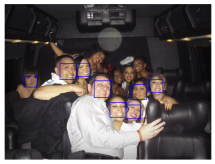
correctly classified as L11 (SoP)

correctly classified as L02 (BoP)

correctly classified as L09 (BoP)



incorrectly classified as L05 (BoP)



incorrectly classified as L10 (BoP)

Figure 5. Additional experiments with unlabeled data. For clarity, only face parts detected are marked with bounding boxes. The first two examples were classified with SoP while the rest were obtained with BoP (best viewed in color).

additional test images in Fig. 5. Here we see four examples that the human observer felt were correctly classified, and two examples that the observer felt were incorrectly classified. The fifth image was probably misclassified because of the detected glasses on all faces whereas the sixth image may have been misclassified due to the dark background.

6. Conclusions

In this paper we examined the question of what can be determined about the social categorization of people in group photos. The framework we propose takes advantage of existing state of the art techniques for parts and attributes as well as global scene descriptors. The people group models we propose are able to capture social dynamics within an image. We show that these techniques were able to classify images in a socially meaningful way. Currently we are working on a larger dataset of labeled images in order to take steps towards more exhaustive experimental validation of people categorization. With larger training data additional classification techniques may be learned and evaluated. In addition to more data, more parts and attributes, such as detecting specific accessories or objects, may help improve recognition performance.

Acknowledgments

This work was supported by ONR MURI Grant #N00014-08-1-0638 and Spanish projects DPI2009-14664-C02-0, DPI2009-08126, FEDER, and DGA mobility grant. The authors wish to thank Manuel Cebrian, Ryan Gomes, Peter Welinder and Pietro Perona for helpful discussions.

References

[1] L. Bourdev, S. Maji, and J. Malik. Describing people: Poselet-based attribute classification. In *ICCV*, 2011. 2

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2, 4, 5

[3] M. Carroll. How Tumblr and Pinterest are fueling the image intelligence problem. *Forbes*, January 17 2012. Web http://onforb.es/yEfDmM. 2

[4] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107(52), Dec. 2010. 2

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2005*, pages 886–893, 2005. 5

[6] L. Ding and A. Yilmaz. Inferring social relations from visual concepts. *ICCV*, pages 699–706, 2011. 2

[7] A. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2009. 2

[8] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The Wisdom of Social Multimedia: Using Flickr for Prediction and Forecast. In *ACM Multimedia Int. Conf.*, October 2010. 2

[9] M. Maffesoli. *The time of the tribes: the decline of individualism in mass society*. Theory, culture & society. Sage, 1996. 1

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–423, July 2001. 5

[11] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. Journal of Computer Vision*, 42(3):145–175, 2001. 5

[12] Z. Song, M. Wang, X. sheng Hua, and S. Yan. Predicting occupation via human clothing and contexts. *Computer Vision, IEEE International Conference on*, 0:1084–1091, 2011. 2

[13] Y. Taigman and L. Wolf. Leveraging billions of faces to overcome performance barriers in unconstrained face recognition, 2011. Face.com. 4

[14] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. In *Proc. Graphicon*, volume 3. Moscow, Russia, 2003. 5

[15] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009. 2

[16] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: recognizing people and social relationships. In *Proc. of the 11th European Conference on Computer vision*, pages 169–182, 2010. 2

[17] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. *CVPR*, 0, 2011. 2