From Bikers to Surfers: Visual Recognition of Urban Tribes

Iljung S. Kwak[1]
iskwak@cs.ucsd.edu
Ana C. Murillo[2]
acm@unizar.es
Peter N. Belhumeur[3]
belhumeur@cs.columbia.edu
David Kriegman[1]
kriegman@cs.ucsd.edu
Serge Belongie[1]
sjb@cs.ucsd.edu

[1] Dept. of Computer Science and Engineering
University of California, San Diego.
San Diego, CA, USA

[2] Dpt. Informática e Ing. Sistemas - Inst. Investigación en Ingeniería de Aragón.
University of Zaragoza, Spain.

[3] Department of Computer Science
Columbia University, USA.

Abstract

The terms Biker, Punk, Hipster, Goth or Surfer often spark visual depictions of individuals with very distinct fashion styles. These visually salient styles can provide insight into the social identity of an individual. However, despite its potential usefulness, little work has been done to automatically classify images of people into social categories. We tackle this problem by analyzing pictures of groups of individuals and creating models to represent them. We capture the features that distinguish each subculture and show promising results for automatic classification. This work gives vision algorithms access to the social identity of an individual and helps improve the quality of socially motivated image search, relevance of advertisements, and recommendations of social groups.

1 Introduction

In the past few years there has been a massive influx of images provided by social media; Facebook alone receives around 300 million photos a day[1]. The abundance of social media presents a compelling opportunity to analyze the social identity of individuals captured within images. This points to an excellent opportunity for computer vision to interact with other fields, including marketing strategies and psychological sociology [7, 12].

Although there have been major strides in image semantic content analysis (in face, object, scene, and more recently clothing recognition), current algorithms fail to fully capture information from groups of individuals within images. For example, as shown in Fig. 1, visual searches of groups of people often provide uninspiring results. Rather than matching personal style or social identity, the search provided images with similar global image statistics. The mainstream media has noticed this deficiency in some recent discussions [4] and wonders when vision algorithms will catch up to their expectations.

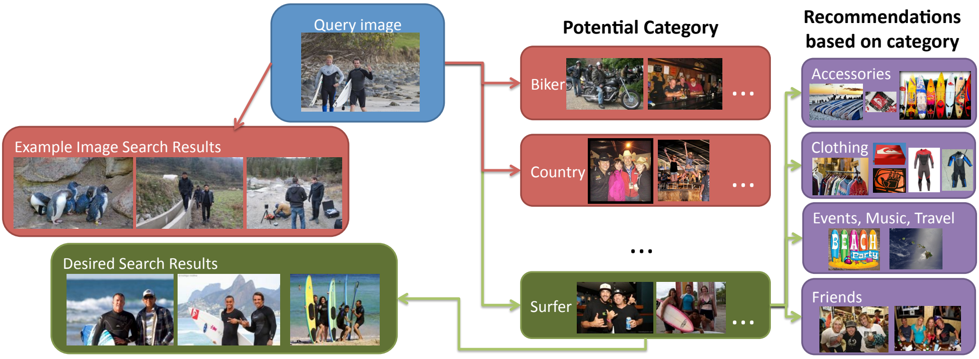


Figure 1: The social groups influence the appearance of their members. This work leverages this intuition to classify images of groups of people into social categories. This can improve recommendation systems and user experience with social media, and image search engines can take advantage of this classification and provide more meaningful search results.

In 1985 Michel Maffesoli described *urban tribes* [16] as a group of people who have similar visual appearances, personal style, and ideals. Among tribe members, similar personal styles often manifest as common accessories such as leather jackets or surfboards. The scene context also provides useful information: surfers are more likely to be photographed outdoors by the sea, whereas bikers may congregate at biker bars or be photographed by their bikes on the road. Though not as discriminative, the overall demeanor between tribes can vary as well, such as the laid back smiling surfers versus the frowning dark subculture members. The visual cues shared by members of these tribes provide the basis for our work; members from the same urban tribe are expected to look more similar than members of different tribes, and they can be easily identified by people just from visual information.

Automatic recognition of these urban tribe categories could provide interesting benefits and applications. More relevant image searches can be conducted; more relevant advertisements can enhance the web experience of both businesses and consumers; social networks can provide better recommendations. Urban tribe classification can also improve surveillance of social demographics. Unfortunately, this categorization problem is difficult because of the ambiguous nature of social categories and the high within class variance. Social categories can evolve and fracture into separate groups; individuals may exhibit features of multiple urban tribes or certain individuals may not present a visually salient style at all.

This work highlights a timely but largely unstudied problem of image group categorization from a social perspective, and contributes towards its solution in several ways. Rather than approaching this problem by classifying isolated individuals (as most recent fashion/style analysis works do), we focus on calculating meaningful group features and models. Following this idea, we present a novel recognition pipeline (see Fig. 2) and we evaluate different modeling approaches, following common frameworks used in other recognition tasks. Finally, we also provide a dataset, Urban Tribes dataset, with around 100 labeled images per class, from 11 different classes. We provide access to this dataset to facilitate further research on social categorization of group pictures[1]. The exhaustive and promising experimental results show that it is possible to extract semantic meaning from social media group photos, opening opportunities for the previously mentioned applications.

[1] `http://vision.ucsd.edu/content/urban-tribes`

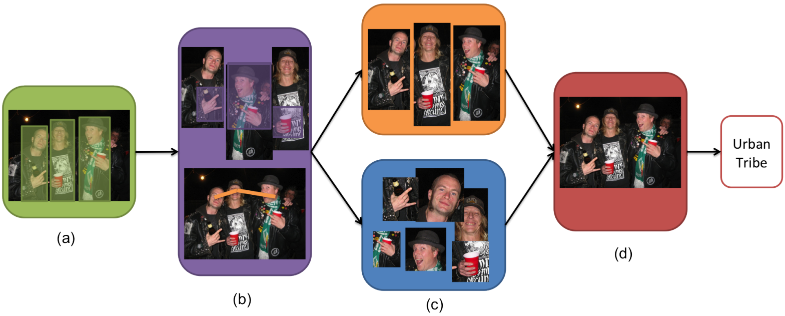


Figure 2: Group image categorization. Our approach consists of: (a) people detection, (b) local and global descriptor computation, (c) group modeling and (d) classification.

2 Related Work

Our work attempts to recognize the content of an image from a social perspective. This is in line with the social signal processing, a growing area of study [22]. Ding and Yilmaz [10] show interesting results for the subjective interpretation of action analysis, proposing how to discriminate positive and negative social relations of individuals in a video sequence. Song *et al.* [21], present a promising approach to predict the occupation of a subject given that individual's clothing and a rough scene context description. Closer to the goals and applications of our work, Yu *et al.* [26] analyze a user's photo album and the associated metadata in order to suggest possible social groups of interest, e.g., flickr or facebook groups about flowers or animals. This is closely related to our goal of analyzing social media, however our approach deals only with visual information and focuses on the analysis of images with groups of people. A common element among all these works and ours is the need for both global image statistics as well as more semantic individual level attributes.

Regarding the analysis of semantic attributes of individuals, we find recent works that recognize the presence or absence of several face attributes to perform higher level tasks such as face verification [13]. Other work, such as [20], model the relative strength or rankings of certain attributes, rather than binary attributes, to provide a more natural and richer modeling, potentially enabling more robust recognition and informative descriptions of novel images.

Finding the urban tribe or social group of an individual is part of a fine grained categorization problem (labeling faces with identity, attributes, or their urban tribe), similar to what we find in other recently studied classification problems, which have focused on classifying various types of living organisms such as plants [14] and animals [23].

An important aspect of our work involves analyzing a group of individuals within an image. The importance of group structure has been researched by Gallagher and Chen [11], who analyzed the layout of the individuals in an image to detect gender, age and even family relationship. Group analysis has also been shown to improve individual identification by Manyam *et al.* [17]. Dhall *et al.* [9] highlights the importance of group analysis as a whole, and uses it to better understand the mood of the group of people in an image. Group analysis methods usually start with individual person detection and description. One of the leading methods is that of Bourdev and Malik [2], which is based on the detection of person parts named *poselets*. Its effectiveness has been demonstrated for human parsing [24] and recognizing semantic attributes such as hair style or clothing type [3]. Clothing recognition itself has become a growing field [6, 15]. One of these works, [25], mentions the interesting link between visual appearance and social identity.

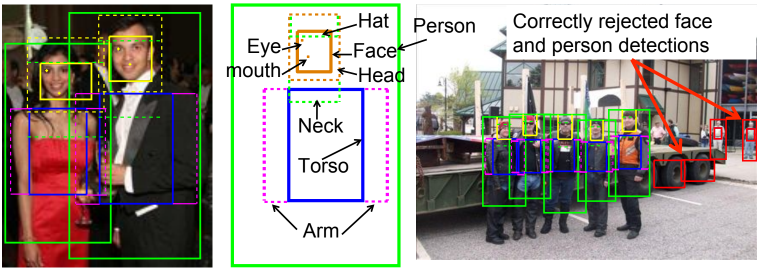


Figure 3: Person part detection. Each person hypothesis can have up to six parts and fiducial points for eyes and mouth. We compute the part descriptors within the bounding boxes of these regions. The examples on the left and middle show a close up of our detections. The example on the right presents two face and two person detections that were correctly rejected thanks to the hypothesis construction process proposed (faces must be aligned with a person hypothesis; person bounding box sizes can't have large deviations).

In our earlier work [18], we obtained preliminary results towards discovering social groups among images from social events with a weakly supervised approach. Image description there includes only face, head and context descriptors to identify and classify images with groups of people into clusters of social groups. The current paper presents an augmented group image description together with more suitable group model representations and additional classification options, which leads to significantly higher accuracy for social group recognition, as shown in our experiments. In addition, we present a strongly labeled dataset, with a larger amount of images and ground truth information which allows a more accurate evaluation of the proposals than the image set presented in previous work.

3 Group description

Our group modeling involves detecting individuals, extracting individual and group features and building the group representation, as key steps for the classification detailed later.

3.1 Person detection and description

We detect individuals in the image and describe each of them as a combination of parts. Similar to [18], individuals are detected by a combination of poselet based person detection [2] and an open source face detector [27]. We obtain six part bounding boxes from the person detector: face, head, upper head (hat), neck, torso and arms. Both face and person detections are merged into a single person hypotheses whenever the overlap of the face region from the two detectors is over a threshold. This simple step filters out many person hypothesis that did not correspond to persons posing for the picture but passing by in the background, as shown in the example in Fig. 3. Thus, an image containing p persons is represented by a set of p hypotheses $\{h_1, h_2, \ldots, h_p\}$, and each person hypothesis is composed of a set of parts (not all parts need to be detected to build a hypothesis, as the torso or the arms may not appear in the image). Therefore each individual h_i is represented by a set with the corresponding parts descriptor vector d_{part_name}: $h_i = \{d_{head},\ d_{face},\ d_{hat},\ d_{neck},\ d_{torso},\ d_{arms}\}$.

We define the descriptor set detailed in Table 1, which is computed for each part. This set is built on the descriptor set we used in [18]. We have initially evaluated an augmented set of descriptors (e.g., larger variety of color spaces, fiducial point information, additional texture descriptors) but we have selected a reduced set, trying to maintain a balance between size and discriminative power, after running standard feature selection techniques [2] and following the intuition behind the features, detailed next.

Ratio of skin pixels vs. total amount of pixels in the patch, obtained with a simple skin-color-based segmentation (normalized to the average face color detected). This descriptor reflects the type of clothing used and how much body is covered with it.
RGB, Luminance and Hue histograms computed for all pixels and computed only for non-skin pixels. This will help modeling the type of clothing used.
Top 3 **dominant values** in Red, Green, Blue, Hue and Luminance color bands. Dominant colors in clothes and accessories are very specific at some social categories.
HoG features [8], which will help capture the different poses.

Table 1: Individual descriptors. They are computed within the bounding box of each part.

3.2 Global group descriptors

In order to account for context and group properties, the image is represented by each of the individuals h_i, as explained above, together with a global group descriptor set, d_{global}: $G = \{h_1, h_2, \ldots h_p, \;\; d_{global}\}$. This set is also an augmented version of previous work proposed descriptor set, again obtained thanks to a combination of feature selection analysis and intuition behind the different descriptors. The global descriptors are split in two sets, low level descriptors (detailed in Table 2) and high level descriptors (detailed in Table 3).

Ratio of pixels within the detected **person bounding boxes** vs. total amount of pixels.
RGB, Luminance and Hue histograms computed on all pixels, on pixels out of the detected person bounding boxes, i.e., background pixels.
Gist [19] and **HOG** [8] descriptors.

Table 2: Low level group descriptors. They are computed over all image pixels and comprise the scene general information, such as lighting, color and gist.

Proximity between individuals in the image. We compute a histogram of distances between faces and a ratio of how much overlap exists between person bounding boxes on average.
Alignment or **pose of the group**. We compute the average angle between a face and its neighboring ones according to a minimum spanning tree computed on the detected faces computed as proposed in [11].
Scene **layout** of individuals. We build a histogram of face locations within the image, using a coarse image grid.

Table 3: High level group descriptors. They represent higher level semantic information and are based on the distribution and pose of the detected persons in the group.

4 Group classification

To classify a group of individuals into a social category, we need to jointly model the features computed for each person hypothesis and the group features. This section describes two different approaches studied in our work. We build on the modeling proposed in [18] to represent the group as a bag of parts, following the typical bag of words representation for object recognition. We also model the group as a set of individuals, combining their responses in a hierarchy of SVM classifiers. Note that as described, hypothesis may have different number of parts detected, which requires a careful classification framework able to deal with heterogeneous descriptor sizes.

Bag of Parts-based classification. Using the Bag of Parts model to represent a group of people, we create a bag of m people parts combined with a global descriptor vector d_{global}.

[2] http://www.cs.waikato.ac.nz/ml/weka/

The combined group model will be referred to as $G = \{p_1, \dots, p_m, d_{global}\}$. This model combines all visible parts and the group description. Let us name this approach BoP_k, to refer to the size k of the vocabulary used. To use this group representation, a vocabulary is built for each part type, by running k-means clustering on all parts of the corresponding type that are visible on the training images. We refer to the vocabulary built for each part type as $V_{part} = \{w_1, ..., w_k\}$. We store the frequencies of each word in each possible class L, building a histogram per word: $w_k \rightarrow hist_{wk} = [count_{L_1}, ..., count_{Lj}]$.

To find the most likely label for a query image, we create the signature of the image as the histogram of word frequencies for each vocabulary V_{part}: $hist_{part} = [count_{w1} ... count_{wk}]$, where $w_1, \dots, w_k$ are the words from the corresponding V_{part}.To be able to deal with missing parts, we first evaluate each part type separately, and later we combine the distances obtained for each part type found in the image to each of the labels into $d_{BoP}(G, L_j)$. We can obtain the distance from each part type p detected in the image to the corresponding model based on the count of the occurrences of each word weighted by its frequency in the training:

$$d_{BoP}(p, L_j) = 1 - \frac{\sum_{i=1}^{k} (count_{wi} \times hist_{wi}(j))}{k}. \tag{1}$$

There are usually several parts of each type in a group image (several faces, arms), and the BoP models somehow how many occurrences of each possible part (i.e. part word) happen in the group. It's not possible to model similarly the group descriptors, since we only have one per image. Therefore, to include the global descriptor information we find simply the nearest neighbor among the reference information for the test image global descriptor, in each of the classes L_j:

$$d_{global}(G, L_j) = min_{j=1}^{t}(\ |g_i, g_j|\ ), \tag{2}$$

where t is the number of training images in class L_j. This distance is normalized between $[0, 1]$ to allow for easy combination with d_{BoP}, which is also normalized. Then, with c possible labels, the label of the group using BoP_k is calculated as:

$$L = \underset{j \in [1...c]}{\arg\min} \, (d_{BoP}(G, L_j) + d_{global}(G, L_j)). \tag{3}$$

SVM-based classification. Alternatively, we represent the group G as a set of persons and model the problem as finding the most likely class C given a particular group image G, i.e., estimating $P(C|G)$ for each possible class. In this case, each person hypothesis gets a probability for each class and the final class estimation is a combination of all of them. In this setting, we used LIBSVM [5] to train a multi-class SVM on the person hypotheses and the global descriptors. We used LIBSVM's built in function to calculate probabilities for each class. More formally, we are considering $P(C|G) = P(C|h_1, h_2, \dots, h_p, d_{global})$, then, the final label L assigned to the query group using this approach is calculated as follows:

$$L = \underset{j \in [1...c]}{\arg\max} \, P(C = j | h_1, h_2, \dots, h_p, d_{global}). \tag{4}$$

We have considered several options to estimate $P(C|G)$, with variations of the following decompositions (details of the considered options can be seen in the experimental section):

$$P(C|G) = P(C|d_{global}) \prod_{i=1}^{p} P(C|h_i) \qquad\qquad P(C|G) = \prod_{i=1}^{p} P(C|h_i\ , d_{global}) \tag{5}$$

5 Experiments and Results

This section evaluates the performance of the proposed algorithms.

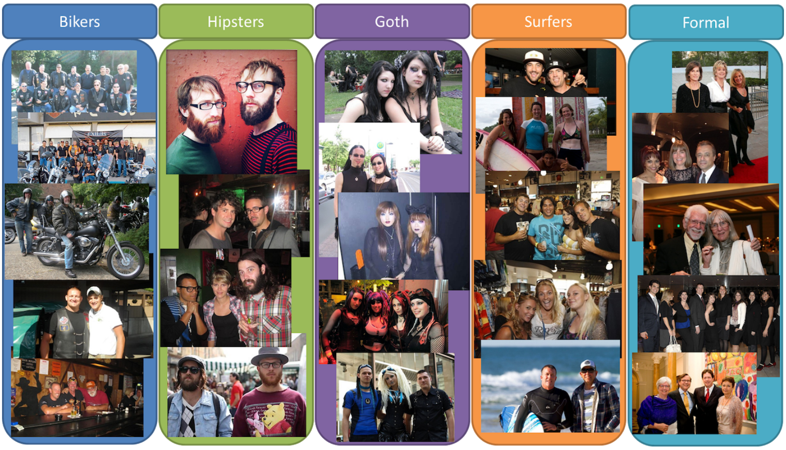


Figure 4: Examples of social groups in our Urban Tribes dataset. The images show a wide range of inter-class and intra-class variability. More details can be seen at the dataset website.

Label	# images	# people	Label	# images	# people	Label	# images	# people
biker	114	443	hip-hop	90	253	club	100	365
country	107	347	hipster	102	288	formal	103	414
goth	99	226	raver	116	305	casual/pub	125	459
heavy-metal	102	266	surfer	100	333			

Table 4: Summary of Urban Tribes dataset.

5.1 Urban Tribes dataset

Creating an urban tribe dataset posed an interesting challenge. Similar to other recently studied problems in computer vision, such as clothing or beauty evaluation, urban tribe categories can be ambiguous and subjective. This is a contrast to other classification problems involving people, where accurate descriptions of each class and its standard appearance can be found, such as age, gender or occupation evaluation. In order to obtain an unbiased dataset, we defined our classes from social group labels provided by Wikipedia. We selected the eight most popular categories from their list of subcultures[3] to facilitate image collection. In addition to these social groups, we added three other classes corresponding to typical social venues (formal events, dance clubs and casual pubs). These classes are intended to include some of the most common social event pictures we can find in social media that may not belong to a clear subculture, but still present common appearances in clothing style and compose a group with similar social interests.

For each of the selected classes, we searched for images of groups of people with different search engines. We used the group labels as search keywords combined with location and event keywords such as *bar*, *venue*, *club* or *concert*. Example search terms include 'bikers' and 'biker bar'. We collected a broad range of scenarios for each class, both indoor and outdoor venues, large group pictures acquired from the distance and close-up images, etc. As shown in Fig. 4, the groups show a variety of realistic conditions and most classes present high intra-class variation. Table 4 shows the class labels as well as the number of images (# images) per class and total amount of detected persons for each class (# people). Although the number of images per class was balanced, the number of detected persons per image was

[3] http://en.wikipedia.org/wiki/List_of_subcultures

different depending on the group.

5.2 Social group recognition experiments

This section evaluates the performance of the proposed algorithms and the most promising paths towards this novel problem framework. We set up classification experiments where training and testing images were randomly selected from each of the categories for 50 different iterations. Note that for the 11 classes in the Urban Tribes Dataset, chance classification is $\frac{1}{11} = 0.09$. For each experiment, a fixed number of the images from each class are used for learning the models, and the rest of images are used for testing. A test is considered correct if the most likely group label is correct according to the ground truth labels.

As explained in section 4, the Bag of Parts modeling builds a visual vocabulary for each part using the training set, with k visual words per vocabulary. After evaluating different values of k, we set $k = 30$ (BoP_{30k}) for the rest of experiments, because the performance increased significantly when increasing k until $k = 30$. The other approach studied, modeling the group as a set of people, uses the training set descriptors to train several SVM classifiers. We evaluated different options: 1) SVM_1, training a single SVM with all the descriptors of each person concatenated, including null values for non-detected parts and replicating the same global descriptor for all hypothesis from a particular image; 2) SVM_2, training one SVM for all part descriptors similarly concatenated and a second SVM for the global image descriptors; 3) SVM_8, training a separated SVM classifier for each part descriptor set and an additional SVM for the global image descriptors. The responses from all the SVMs in each case are simply combined, providing a final probability of each image being of a particular class. The option SVM_1 provided significantly lower performance than the rest during the preliminary tests, therefore we do not use this configuration for the rest of experiments.

Table 5 shows a summary of recognition experiments with different amount of training data and different amount of parts used in the modeling, given the most suitable configurations found for each modeling option considered (BoP_{30k}, SVM_2 and SVM_8). Column $allParts + global$ shows the accuracy when combining all person parts and global descriptors; $allParts$ shows the accuracy when combining only person parts; $global(scene)$ column shows the results if we would only run the equivalent to a standard scene classification approach (using the global descriptors, typically used for scene categorization).

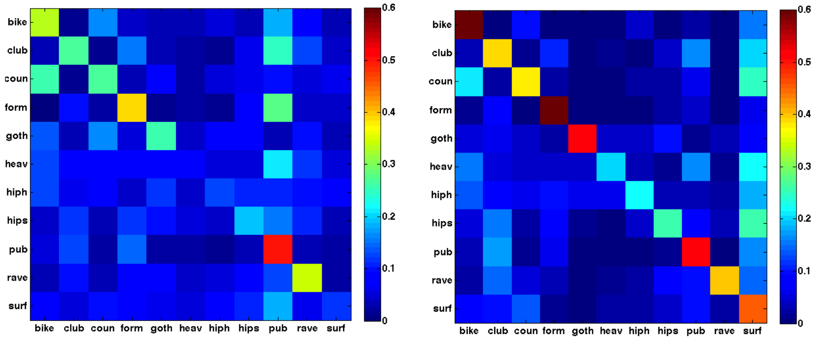
The last columns show additional baseline results: *individual* shows the average accuracy when each person is classified independently from the rest of the group, i.e., there is no consensus from the group nor group global descriptors used at all. This part makes sense when we model the group as a set of individuals, since we can get the response from each person separately. $face + head + global$ [18] shows the results using the Bag of Parts configuration used in the referred previous work. The last rows in the table show the contribution of each type of descriptor separately. This analysis is shown for the BoP approach, since it provides the classification result per part. For all modeling options, even just one set of descriptors was able to classify the images above chance, but the final classification scores are clearly improved when all the parts from all the individuals in the image are combined. Particularly interesting is the increase due to the use of global group information. Additionally, we have noted that in between 10% and 15% of the tests (depending on the modeling option), the global descriptors classifier would guess the correct label while the combination of parts will not, supporting again the idea that group and context provide complementary information to person local parts.

The SVM based classification provided the best results, probably because it is able to

Approach	$allParts + global$ (std)	$allParts$	$global(scene)$	$individual$	$face + head + global$ [☒]
80 random train images per class, 50 iterations, 278 tests per iteration					
SVM_2*	0.43 (0.04)	0.40	0.37	0.34	-
SVM_8	0.46 (0.02)	0.40	0.37	0.38	-
BoP_{30k}	0.37 (0.02)	0.36	0.18	-	0.30
40 random train images per class, 50 iterations, 718 tests per iteration					
SVM_2*	0.38 (0.03)	0.38	0.31	0.33	-
SVM_8	0.41 (0.01)	0.36	0.32	0.35	-
BoP_{30k}	0.33 (0.02)	0.33	0.17	-	0.25
* SVM_2 does not include arm parts because all part descriptors concatenated to train a single SVM was performing significantly worse. This behavior was not observed for the rest of approaches.					

Average accuracy for each type of descriptor considered separately (80 train images per class)								
descriptors used:	d_{face}	d_{head}	d_{torso}	d_{hat}	d_{neck}	d_{armL}	d_{armR}	d_{global}
BoP_{30k}	0.24	0.24	0.22	0.26	0.21	0.21	0.2	0.18

Table 5: Average accuracy for the recognition of all classes using different approaches.



(a) (b)

Figure 5: Confusion matrix for classification results obtained with (a) BoP_{30k} and (b) SVM_8, using 80% of the data for training. The rows show results in alphabetical order of the labels (detailed in Table 4, from top to bottom. To enhance contrast the color scale is set to $[0, 0.6]$.

learn better which components of each descriptor set are more discriminative. We have also experimented with a reduced set of attributes for BoP approaches, using standard attribute selection algorithms, but it did not improve the performance. From the confusion matrices shown in Fig. 5, we can appreciate some interesting hints for future improvements, such as the clear confusion in both matrices between *bikers* and *country* (columns 1 and 3) in both directions, what can point to the necessary additional descriptors or attributes. Analyzing in detail the results, we have confirmed that all images have been included, in average, in the test set of 12 experiments. We have found that some test images are always classified correctly while others are rarely identified as their ground truth category, pointing to the heterogeneous nature of the data. Figure 6 shows some of these sample tests.

6 Conclusions and Future Work

An individual's social identity can be defined by the individual's association with various social groups. Often these social groups influence the individual's visual appearance. In this work we attempt to capture this intuition with a new vision task, social categorization. We provide an exhaustive baseline analysis for the task as well as a dataset to aid future research.

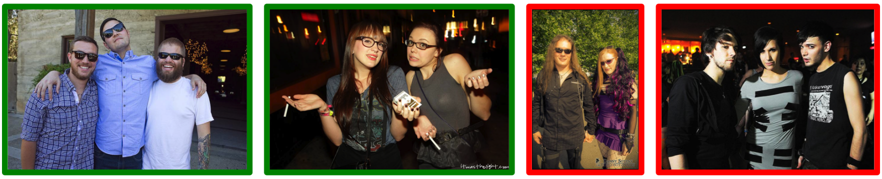


Figure 6: Examples of classification results. These images have been classified as *hipsters* in all their tests. The two left images are correct, but the label for the right two images should be *goth*.

In future work, we intend to incorporate semantic attributes to improve the classification performance and analyze the contribution of different types of attributes. The task introduced in this work opens opportunities for computer vision to improve targeted advertising and social monitoring and provide more tailored experiences with social media.

Acknowledgments. The authors would like to thank the reviewers for their insightful comments. This work was supported by ONR MURI Grant #N00014-08-1-0638 and by projects DPI2012-31781, DPI2012-32100 and DGA T04-FSE.

References

[1] R. Armburst. Capturing growth: Photo apps and open graph. URL http://developers.facebook.com/blog/post/2012/07/17/capturing-growth--photo-apps-and-open-graph/.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. URL http://www.eecs.berkeley.edu/~lbourdev/poselets.

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

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

[5] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012.

[7] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 2010.

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

[9] A. Dhall, J. Joshi, I. Radwan, and R. Goecke. Finding happiest moments in a social context. In *ACCV*, 2012.

[10] L. Ding and A. Yilmaz. Inferring social relations from visual concepts. *ICCV*, 2011.

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

[12] 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.*, 2010.

[13] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*, 2009.

[14] N. Kumar, P. N. Belhumeur, A. Biswas, D. Jacobs, W. J. Kress, I. Lopez, and J. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*. 2012.

[15] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012.

[16] M. Maffesoli. *The Time of the Tribes: The Decline of Individualism in Mass Society*. Sage Publications, 1996.

[17] O. K. Manyam, N. Kumar, P. N. Belhumeur, and D. Kriegman. Two faces are better than one: Face recognition in group photographs. In *Int. Joint Conference on Biometrics (IJCB)*, 2011.

[18] A. C. Murillo, I. S. Kwak, L. Bourdev, D. Kriegman, and S. Belongie. Urban tribes: Analyzing group photos from a social perspective. In *CVPR Workshop on Socially Intelligent Surveillance and Monitoring (SISM)*, 2012.

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJVC*, 2001.

[20] D. Parik and K. Grauman. Relative Attributes. In *ICCV*, 2011.

[21] Z. Song, M. Wang, X. Hua, and S. Yan. Predicting occupation via human clothing and contexts. *ICCV*, 2011.

[22] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009.

[23] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *ICCV*, 2011.

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

[25] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

[26] J. Yu, X. Jin, J. Han, and J. Luo. Collection-based sparse label propagation and its application on social group suggestion from photos. *ACM Trans. Intell. Syst. Technol.*, 2011.

[27] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.