Title

Bikers are Like Tobacco Shops, Formal Dressers are like Suits: Recognizing Urban Tribes with Caffe

ISBN
9781479966820

Authors
Wang, Yufei
Cottrell, Garrison W

Publication Date
2015

DOI
10.1109/wacv.2015.121

Peer reviewed

Bikers are like tobacco shops, formal dressers are like suits: Recognizing Urban Tribes with Caffe

Yufei Wang Garrison W. Cottrell
University of California, San Diego
{yuw176, gary}@ucsd.edu

Abstract

Recognition of social styles of people is an interesting but relatively unexplored task. Recognizing "style" appears to be a quite different problem than categorization; it is like recognizing a letter's font as opposed to recognizing the letter itself. Similar-looking things must be mapped to different categories. Hence a priori it would appear that features that are good for categorization should not be good for style recognition. Here we show this is not the case by starting with a convolutional deep network pre-trained on ImageNet (Caffe), a categorization problem, and using the features as input to a classifier for urban tribes. Combining the results from individuals in group pictures and the group itself, with some fine-tuning of the network, we reduce the previous state of the art error by almost half, going from 46% recognition rate to 71%. To explore how the networks perform this task, we compute the mutual information between the ImageNet output category activations and the urban tribe categories, and find, for example, that bikers are well-categorized as tobacco shops by Caffe, and that better-recognized social groups have more highly-correlated ImageNet categories. This gives us insight into the features useful for categorizing urban tribes.

1. Introduction

In the past few years, there has been impressive progress in automatically understanding the content of images in tasks such as object recognition, scene recognition, and object detection. However, the analysis of the social features of images of groups of people has not attracted a great deal of attention. This is an important unsolved problem, because current image search algorithms, given a picture of surfers, for example, fail to capture information about personal styles or social characteristics of groups of people, but instead retrieve images with similar global appearance [1]. Recognition of groups of people from a social perspective provides many other potential applications. With more accurate group recognition results, more accurate recommendations can be made in social networks, and more relevant advertisements can be used to target particular groups of people. However, the analysis of groups of people is difficult in that the group categories are semantically ambiguous, and have high intra-class variance.

Kwak *et al*. studied this problem of group recognition ([1], [2]). They created the urban tribe dataset that we are using in this study. The classes are defined from social group labels provided by Wikipedia. They selected the eight most popular categories from Wikipedia's list of subcultures, and added three other classes corresponding to typical social venues (formal events, dance clubs, and pubs). For each class, images of groups of people were discovered with different search engines, and about 100 images for each class were collected. They proposed a group recognition pipeline that extracted hand-designed features of individuals and groups, such as the relative amount of skin in an image, the arrangement of faces in the photo, HOG features, etc., and used a bag of words approach or an SVM to classify the images. They achieved 46% correct using this approach, setting a benchmark for future research. Related research has shown that the visual structure of a group is useful [3] that modeling the social relationships aids event recognition [4], and that both local and global factors are useful for group level expression analysis [5].

In this paper we use Convolutional Neural Networks (CNNs) on this problem. A convolutional neural network is one where hidden units share weights and training signals across an image, so that the hidden unit features are "convolved" with the image. A simple version of a CNN with back-propogation was introduced in 1986 by Rumelhart et al. [6]. However, large, multi-layer convolutional neural networks for real-world problems were not developed until LeCun et al. applied them to hand-written digit classification [7]. Since then, CNNs have shown state of the art results on various computer vision tasks, such as document classification ([8],[9]), object categorization ([10], [11]), object detection ([11], [12], [13]), object localization ([12]). Many variations of CNN architectures have been investigated ([9], [14], [15]).

Recently, many researchers have shown the utility of applying CNN features learned on the ImageNet task to novel tasks. Zeiler and Fergus showed state-of-the-art results on Caltech-101 and Caltech-256 using pre-trained CNN features combined with a softmax classifier [16]. Donahue *et al.* used different layers of a pre-trained CNN as input to simple classifiers such as SVMs and Logistic Regression, and outperformed the state-of the-art on several vision challenges such as scene recognition and domain adaptation[17]. Most relevant to the present work, Karayev *et al.* compared different approaches for photographic and painting style recognition, and pre-trained CNN features generally give the best results [18].

In this paper, we investigate the generalization ability of pre-trained (and fine-tuned) CNN features to social group recognition. We propose a CNN feature-based architecture that combines individual features and global scene features. Even without fine tuning of the network, we achieve 69% correct using our methods; fine-tuning adds an improvement of only 2%. This is a large boost of performance over the previous classification method provided by [1] of 46%. We show that both individual information and global scene information contribute to a social group's characteristics, and that different feature extraction schemes for individual and global information are necessary.

We further investigate why features extracted from a pre-trained CNN are useful for the urban tribe recognition task. For an input image, there is a correlation between the probability of it being mapped to ImageNet classes and being in different urban tribe classes. Moreover, the degree of correlation is related to the recognition rate of different urban tribes classes - better-recognized categories have more highly-correlated ImageNet categories. This suggests that the performance on the novel task is related to how sharply (and obviously incorrectly) the new image classes are “mapped” to ImageNet classes. However, the actual relationship between the two types of categories is still mysterious in most cases, and will require future work to sort out.

2. Methods

This section describes the urban tribe dataset and elaborates on the model architecture.

2.1. Urban tribes dataset

Urban tribes are groups of people who have similar visual appearance, personal style and ideals [19]. The urban tribes dataset consists of 11 different categories: *biker, country, goth, heavy-metal, hip-hop, hipster, raver, surfer, club, formal, casual/pub*, with an average of 105 images from each category.

Unlike conventional visual classification problems, urban tribe categories are more ambiguous and subjective. Also, each class contains a broad range of scenarios. The high intra-class variation of the urban tribe dataset makes the classification task challenging. The urban tribe dataset also has some interesting properties. The number of people in each urban tribe image varies. Members of one tribe often have similar visual styles, including their clothes, accessories, and even demeanor. For example, surfers possibly carry surfboards, and the goth often have dark attire, makeup and hair. The environment they are in also contributes to each tribe characteristics: pictures of country tribes are more likely to be taken outdoors with grassland, while pictures of clubbers are often photographed in clubs with dim lighting.

2.2. Classification hierarchy

To utilize the properties of urban tribes fully, our final feature vector consists of the features extracted from individuals and features extracted from group pictures, in order to include contextual features. For each feature type, we use a similar extraction strategy. Individual features and environmental features are hierarchically combined to form the final decision function. The network hierarchy is shown in Figure 1.

For each group image, we represent the group G as the combination of a set of individual people and the group scene. To give the prediction of class C, the individual feature vectors and scene feature vectors are extracted separately. These are then passed through two CNNs using the CAFFE pre-trained network (described in the next section).

For the individual feature vectors, first, candidate person images are detected with a poselet based person detection algorithm. This is far from perfect, as can be seen in Figure 1. The candidate person images $H = \{H_1, H_2, ..., H_p\}$ are used as a whole. Each candidate person is resized to 256×256, and ten 227×227 patches $\{h_{ij}\}, i \in \{1, 2, .., p\}, j = 1, 2, ..., 10$ are extracted (patches from the four corners of the image patch and the center, and their horizontal reflections).

Each individual image patch h_{ij} then passes through the Convolutional Neural Network for person images CNN$_{Person}$, generating activations from the 6th and 7th hidden layer of the CAFFE network. The activations from 6th and 7th layer are both 4096 dimensional (we also tested using just the 6th and 7th layers, see Results). They are concatenated to form an 8192-dimensional vector f_{ij}, where $i \in \{1, 2, .., p\}, j \in \{1, 2, ..., 10\}$.

The feature vectors are then fed into a multi-class SVM$_{Person}$. We use LIBLINEAR[20] to train the SVM on individual patches, and to estimate probabilities for each category given individual patch h_{ij}: Pr$_{ij}(C|h_{ij}), C \in \{1, 2, .., 11\}$. The individual patches h_{ij} in one group image

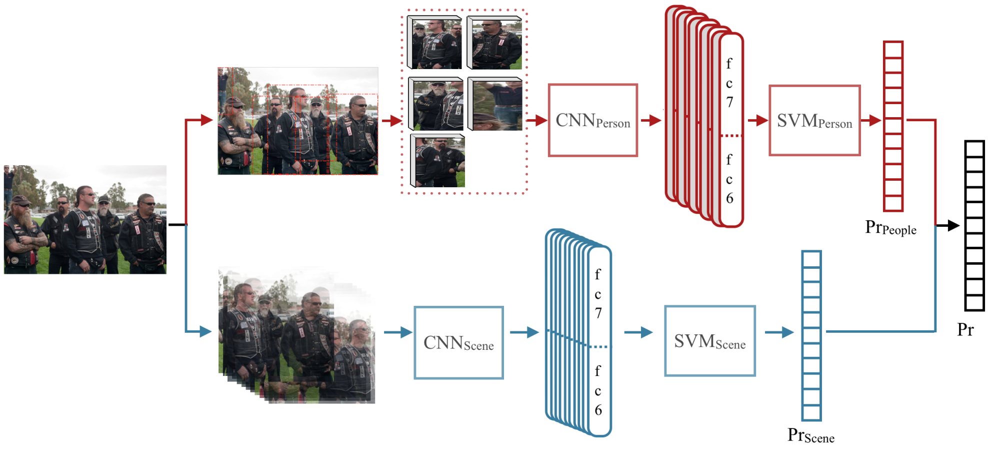


Figure 1: Architecture of classification algorithm using Nets$_{SDense}$ which is introduced in Section 2.3.3. The upper half estimates the probability given people candidate images, and the lower half estimates the probability given the entire scene. Dense crop CNN$_{Scene}$ and distorted crop CNN$_{Person}$ are used, as described in Section 2.3.1.

are usually highly correlated. Therefore, in order to obtain a reliable probability estimate from the noisy yet correlated set of probabilities Pr$_{ij}$, a simple but effective average pooling is performed to Pr$_{ij}$:

$$\text{Pr}_{People}(C|H_1, ..., H_p) = \frac{1}{10p}\sum_{i,j}\text{Pr}_{ij}(C|h_{ij}) \qquad (1)$$

Pr$_{People}(C|H_1, ..., H_p)$ is the probability estimate of class C given the set of people candidate images H.

In addition, the entire scene image, denoted by S, is used for probability estimation. The procedure to generate probability estimate of class C given the scene Pr$_{Scene}(C|S)$ is similar to that for Pr$_{People}(C|H)$. The difference is that the input of the network is 227×227 patches extracted from the entire scene image, and the fine-tuned network: CNN$_{Scene}$ and SVM: SVM$_{Scene}$ are trained with the entire set of scene images. Several different strategies to extract patches from scene images and corresponding Convolutional Neural Network architectures are explained in Section 2.3.

Therefore, the probability estimate of a class C given observation of scene S is:

$$\text{Pr}_{Scene}(C|S) = \frac{1}{K}\sum_{k=1}^{K}\text{Pr}_k(C|s_k) \qquad (2)$$

where s_k is the kth scene patch, Pr$_k(C|s_k)$ is the probability for class C given kth scene patch, and K is the number of scene patches extracted from a group image. K varies with different patch extraction strategies. We again use average pooling, because the assumption of high correlation between patches holds.

Now we have the estimates of two kinds of conditional probability Pr$_{People}(C|H)$ and Pr$_{Scene}(C|S)$. We make a strong assumption that the final category is conditionally independent given H and S, and that the prior probability distribution of the urban tribes Pr(C) is a uniform distribution. The final classification can be expressed as:

$$C^* = \text{arg max}_{i=1,\dots,c}\text{Pr}(C = i|G) \qquad (3)$$

where

$$\begin{aligned}\text{Pr}(C = i|G) &= \text{Pr}(C = i|H, S)\\ &= \frac{\text{Pr}_{People}(C = i|H_1, ..., H_p) \cdot \text{Pr}_{Scene}(C = i|S)}{\text{Pr}(C = i)}\\ &\propto \text{Pr}_{People}(C = i|H_1, ..., H_p) \cdot \text{Pr}_{Scene}(C = i|S)\end{aligned} \qquad (4)$$

and C^* is the predicted label for the group image.

2.3. Convolutional network feature extraction

It has been shown in many experiments that the set of weights of a convolutional network trained from ImageNet can generate a set of generic visual features. Following [21]'s work, we use the network framework called Caffe. The network architecture is described in [10], and won the ImageNet Large Scale Visual Recognition Challenge 2012. We take the activations from the 6th and 7th hidden layer of the convolutional neural network, which are two fully connected layers before the class prediction layer. We also take the activations from 6th or 7th layer alone as comparison. We choose these two layers because as the layers ascend, the features extracted show increasing invariance and semantic meaning.

2.3.1 Pre-processing of the dataset

The urban tribe dataset is a relatively small dataset, and both people candidate crops and scene images are of various resolution. Our convolutional neural networks requires constant input image size of 227×227, so pre-processing of the dataset is necessary.

There are several strategies to make one image compatible with the CNN:

1. *Distort cropping:* As in [10], resize the image to a fixed resolution of 256×256, and crop five 227×227 patches (from the four corners and the center) and their horizontal reflections to generate ten patches from a single image. This way, the aspect ratio of the original images is lost, but for each crop, the portion it takes from the original image is fixed, so that the amount of information all the crops have is relatively stable.

2. *Sparse cropping:* Keep the aspect ratio of the original image, resize the shorter side to 256, and then crop four corner 227×227 patches a middle patch, and their horizontal reflections as before. This method avoids distortion of the image and objects in it, but the crops will possibly lose information when the aspect ratio of the original image is far from 1.

3. *Dense cropping:* Keep the aspect ratio of the original image, resize the shorter side to 256, and then densely crop multiple 227×227 patches and their horizontal reflections. This way, the information of original image is kept by dense cropping process, and the distortion is avoided. The number of crops attained with this method is larger than the previous two methods.

2.3.2 Network Fine-tuning

Although the pre-trained network from [21] can already generalize well to many datasets, the urban tribe categorization problem has unique properties. In particular, it emphasizes the style of the clothing, surrounding objects (e.g., surfboards, motorcycles, bright lights, etc.) rather than distinct categories. To rearrange the importance of the features and adapt the features to the urban tribe dataset, the network can be fine-tuned.

The dataset used for fine-tuning is the same set used for SVM training. In our fine-tuning process, the last layer is replaced by 11 softmax outputs, and the initial weights of the last layer connections are drawn from a zero mean gaussian distribution. Back propagation is used, and the learning rate is set to be small so that the fine-tuning process adapts the extracted features to the urban tribe dataset while preserving the initial properties of the network. The initial learning rate used for the output layer was 0.01, and 0.001 for the rest of the network. We trained for 6,000 epochs, and at the end of every 1000 epochs, we divided the learning rates by 10.

2.3.3 Choices of network combination

Scene images and individual images have different properties, and need different strategies for pre-processing and separate fine-tuning. For the scene network, due to the small size of the dataset, we use the *dense cropping* technique, the third technique in Section 2.3.1, to increase the dataset size. For the person network, we use the *distort cropping* technique, because the subimages are of greater height than width, and the second and third strategies using squared crops of a person image may lose too much information, no matter which location we choose to crop them; whereas the first method ensures each crop keeps the essential features for classification.

The combination of dense crop CNN_{Scene} and distorted crop CNN_{Person} are denoted as Nets_{SDense}.

We also construct other combination of networks for comparison:

1. Nets_{NoTune}: Directly use the pre-trained network by [21] for both scenes and persons, and use the *distort cropping* technique (distorted crops) as input patches for both networks. This choice of cropping strategy is in consistent with the way the network is pre-trained.

2. $\text{Nets}_{SSparse}$: Use the *sparse cropping* strategy for scene features, and the *distort cropping* strategy for person features.

3. $\text{Nets}_{SDistort}$: Use the *distort cropping* strategy for both scene features and person features.

3. Experiments and Results

In this section, the performance of the proposed classification scheme is evaluated and analyzed. In the experiments, six rounds of 5-fold cross validation are performed, therefore we have 30 training experiments in total. The dataset is partitioned into 5 equal sized subsets, containing one-fifth of the data points from each category. One of the subsets is used as test set, and the remained 4 subsets are used as training data.

3.1. Urban tribe classification performance

Table 1 shows the comparison of performance using different approaches. The 30 segmentations of datasets are used for all the approaches tested in this section, and 30 test results are averaged for each approach. The standard error is shown with the accuracy in Table 1. We also compare our result with the result achieved by [1] using their best model. The advantage of CNN pre-trained features is obvious.

The confusion matrix is shown in Figure 2 for Nets_{SDense}, and all the 30 training experiments are averaged. We can observe there is a obvious difference of difficulty of different categories. Class *formal* has accuracy as

Table 1: Performance of different approaches. Note that the bold numbers in columns 2 and 3 are identical because they are from the same networks.

Accuracy (%)	Individual candidate	People	Entire scene	People+Scene
Nets$_{NoTune}$ with concatenated features	40.03 ± 0.31	64.18 ± 0.63	62.61 ± 0.58	69.19 ± 0.51
Nets$_{SDense}$ with fc7 features	46.95 ± 0.33	67.42 ± 0.50	64.87 ± 0.39	70.68 ± 0.47
Nets$_{SDense}$ with fc6 features	45.80 ± 0.37	66.31 ± 0.47	66.72 ± 0.56	70.68 ± 0.44
Nets$_{SDense}$ with concatenated features	**47.06 ± 0.37**	**67.46 ± 0.51**	**67.01 ± 0.52**	**71.45 ± 0.48**
Nets$_{SSparse}$ with concatenated features	**47.06 ± 0.37**	**67.46 ± 0.51**	66.48 ± 0.41	71.26 ± 0.49
Nets$_{SDistort}$ with concatenated features	**47.06 ± 0.37**	**67.46 ± 0.51**	65.06 ± 0.48	71.15 ± 0.53
SVM_8[1]	-	-	-	46(std: 2)

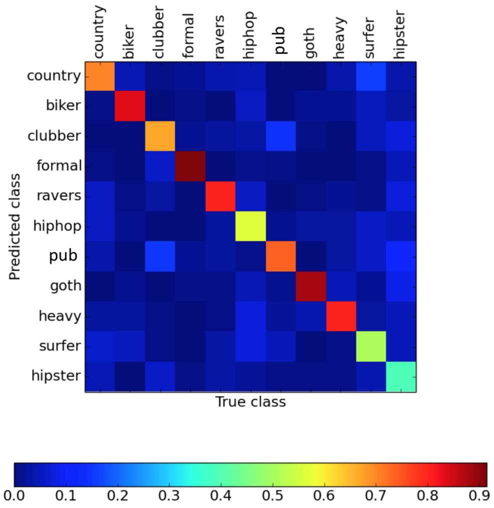


Figure 2: Confusion matrix for classification results with Nets$_{SDense}$, using people and scene features.

high as about 90%, while the category hipster is the most difficult class, having less than 60% accuracy (which is still far above the chance performance of 9%. Hipsters are most confused with casual/pub, clubbers, ravers and goth, which are reasonable confusions, especially since hipsters are a famously vaguely-defined class [22].

Comparing the result of using different features in the same approach shows the necessity of every step of our architecture. In results using Nets$_{SDense}$ with concatenated features, average accuracy for each candidate person is 47.06%. Average pooling of candidate person probability estimates produces a large accuracy increase of about 20%. Accuracy using the entire scene only results in 67.01% accuracy. Combining probabilities $\Pr_{People}(C|H_1, ..., H_p)$ and $\Pr_{Scene}(C|S)$ achieves accuracy as high as 71.45%, which verifies the complementary role of people candidate features and the environment features in a group image.

We also compare the accuracy using only 6th or 7th layer activation from the networks. Nets$_{SDense}$ produces nearly identical results, showing that both layers' activations can generate excellent features, but that using layer 7 instead of 6 does not increase the performance. Indeed, concatenating both layers' activations increases the accuracy by only 0.77%.

To see the role of fine-tuning, we can compare the result of Nets$_{NoTune}$ and Nets$_{SDistort}$. These two approaches both use resizing that causes distortion, and they only vary in whether the networks were fine-tuned or not. The individual candidate performance is increased by 7%, and the combined candidate accuracy ("People") and the entire scene are also improved considerably. These accuracy boosts diminish when the two are combined, giving less than 2.5% performance improvement overall. This suggests exploring other, perhaps more aggressive, adaptation approaches.

Nets$_{SDistort}$, Nets$_{SSparse}$, and Nets$_{SDense}$ use different patch extraction strategies. Note that we use the same distorted patch extraction method for person images, as mentioned in Section 2.3.3, while we use three different methods for scene images. The results for scene images show the advantage of keeping the aspect ratio of scene images, and the slight advantage of using dense crops. However, the final results with People+Scene for the three methods aren't significantly different, due to the combination with people information.

The last three rows of Table 1 indicate some interesting features of different patch extraction methods. The entire scene accuracy of Nets$_{SSparse}$, and Nets$_{SDense}$ is more than 1.4% higher than Nets$_{SDistort}$, which verifies that preserving the aspect ratio of scene images is better than using distorted patches. This suggests exploring using dense sampling for both scene and individual networks, which we did not try due to the excessive training time required for the person networks.

In work not reported in detail here, we used our model with the recently released Very Deep Convolutional Network model with 19 layers ([23]). The same set of cross validation is used, and the preliminary result of our Nets$_{NoTune}$ model is 73.36%. This result already outperforms our best model using Krizhevsky's architecture. This result shows the Very Deep Convolutional Network model with higher recognition rates on the ImageNet dataset also has more generalization ability. We believe there will be promising results applying our Nets$_{SDense}$ model with the Very Deep Convolutional Network architecture.

3.2. Urban tribe classes vs. ImageNet classes

As described in the introduction, it has recently been found that CNN pre-trained features are generic and can be used for many new tasks. In this section we investigate the relationship between the new tasks and original ImageNet task, using the urban tribe dataset,which gives us some insight about the features extracted from the pre-trained network.

The urban tribe dataset contains groups of people, and the important features for categorization are mainly human related features, such as attire, make up, posture and expressions. However, the ILSVRC dataset used for pre-training contains few images of humans. Instead of examining the output of the layers directly, we try to find the relationship between the 1000 classes in ILSVRC dataset and the classes in two image sets: urban tribe dataset, and candidate person images extracted from urban tribe dataset.

We use the parameters of pre-trained CNN network as our feature extraction model, and train a softmax layer on top of its 7th layer to predict the probabilities of one input image(either scene image or candidate person image) being in certain urban-tribe class Pr(l_{urban}), where l_{urban}pre-trained is the 11 urban tribe categories. The output layer is trained for 3000 training iterations. We can also use the output of the pre-trained CNN network to predict probabilities of one input image being in certain ImageNet category, denoted as Pr$(l_{ImageNet})$, where $l_{ImageNet}$ is the 1000 ImageNet categories. We use one round of 5-fold cross validation, and use all the images in urban tribe dataset for analysis.

We first check the relationship of Pr(l_{urban}) and Pr$(l_{ImageNet})$ of candidate person images and scene images. We calculate the 1000 mutual information scores of the predicted urban tribe and ImageNet (where predicted score is 1 if the predicted label is the category being tested, 0 otherwise), denoted as $I(l_{urban}; l_{ImageNet})$.

In Figure 3, we choose two urban classes: *biker* and *hipster*, and plot the mutual information I.The first row shows the result of *biker*, and second row *hipster*. The left column is the results of scene images, and right column is of candidate person images. For *biker*, there are several spikes in scene images, and one significant spike for person images. *bulletproof vest* and *tobacco shop* have high mutual information with *biker* from person and scene images respectively. Meanwhile, for *hipster*, which is the most difficult class, the mutual information is both low for all ImageNet classes.

To confirm the correlation,we use the output vector of the original network and train a classifier directly based upon that. The accuracy is 52.68%. This is a decent result, and indicates the relationship between l_{urban} and $l_{ImageNet}$. Then, we check $l_{ImageNet}$ with highest mutual information with l_{urban}. In Figure 5, we choose four l_{urban}: *formal*, *raver*, *biker*, *hipster*, and choose some examples of candidate person images that have both high Pr$(l_{ImageNet})$ and high Pr(l_{urban}). We also show examples of the images in $l_{ImageNet}$. We can see the shared features between corresponding person images and ImageNet images, for example, similar visual features for bulletproof vests and bikers(Figure 5c).

We also use the visualization method proposed by Zeiler and Fergus ([16]) to examine the features more carefully. From 5th layer features maps, we select the feature maps that are highly correlated with the Pr(l_{urban}) and corresponding Pr$(l_{ImageNet})$. In Figure 4, we visualize the feature maps for *biker* - *bulletproof vest* from candidate person images, and *biker* - *tobacco shop* from scene images. As shown, the useful feature for *biker* - *bulletproof vest* is apparently about the stripe pattern, and the useful feature for *biker* - *tobacco shop* is about the bar-like building structure. The results show that the features that the network catches are semantically important features to human.

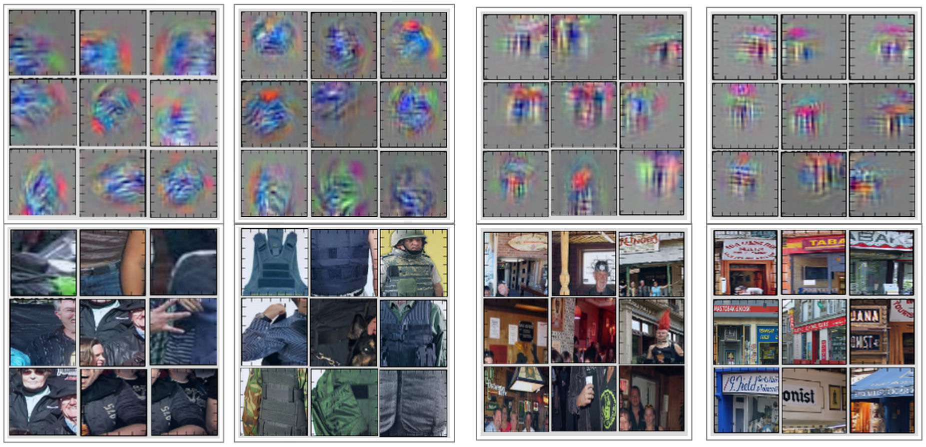
There is a correlation between class-wise accuracy of predicted l_{urban} and the degree of relationship between predicted Pr(l_{urban}) and Pr$(l_{ImageNet})$, as shown in Figure 6. Class-wise accuracy is calculated for candidate person images(Figure 6a) or scene images(Figure 6b). For each l_{urban}, the maximum mutual information over 1000 ImageNet classes are used to indicate the degree of its relationship with $l_{ImageNet}$.

The correlation between ImageNet class and urban tribe class and its relationship with class-wise recognition rate may indicate that the "generic" features extracted by pre-trained CNN networks are not so generic. The network is trained to separate the ImageNet classes most, and if we use the features for a new classification task, the performance of the task is related to how well the new classes can be "mapped" to the ImageNet classes.

4. Conclusion

In this work, we proposed a framework for social group recognition. The framework uses both individual and global features. The features are extracted from CNN networks that have been pre-trained on the ImageNet dataset, and then combined. Our results show the success of our framework by achieving much higher accuracy than the previous state of the art.

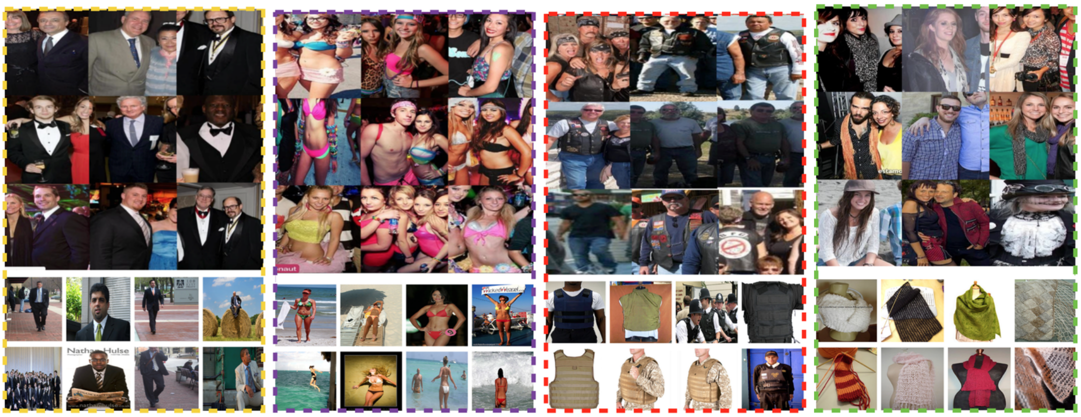
We also investigated the pre-trained CNN features. Both visualization and numeric results showed the generaliza-



(a) Biker & Bulletproof vest

(b) Biker & Tobacco shop

Figure 4: Two selected feature maps from *conv5* layer which are highly correlated with the indicated Pr(l_{urban}) and Pr($l_{ImageNet}$). Top images: Visualization of top 9 activations in the selected feature map across the validation person candidate images(a) and scene images(b), and across the Imagenet dataset within class vest(a) and tobacco shop(b). Bottom images: Corresponding image patches.



(a) Formal - Suit

(b) Raver - Bikini

(c) Biker - Vest

(d) Hipster - Stole

Figure 5: Selected urban tribe classes and the corresponding highest correlated $l_{ImageNet}$. Upper nine images: candidate person images with high Pr($l_{ImageNet}$) and high Pr(l_{urban}). Lower eight images: example of images in $l_{ImageNet}$.

tion ability of pre-trained CNN features to features of people's social styles. Meanwhile, we found correlations between the probability of an image being categorized as an ImageNet class and a social group class, and that better-recognized urban tribe categories are more correlated with ImageNet categories.

In future work, we intend to improve the classification performance by adapting convolutional networks more to social group datasets. The relationship between ImageNet categories and urban tribe classes also suggests an interesting future topic for research.

References

[1] I. S. Kwak, A. C. Murillo, P. Belhumeur, S. Belongie, and D. Kriegman, "From bikers to surfers: Visual recognition of urban tribes," in *British Machine Vision Conference (BMVC)*, (Bristol), September 2013.

[2] 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)*, (Providence, RI), June 2012.

[3] A. C. Gallagher and T. Chen, "Understanding images of groups of people.," in *CVPR*, pp. 256–263, IEEE, 2009.

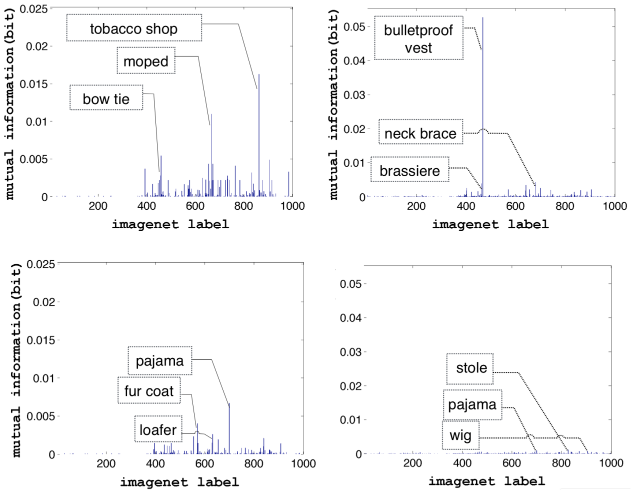
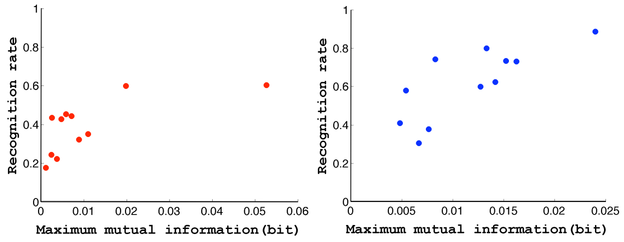


Figure 3: Mutual information of l_{urban} and 1000 $l_{ImageNet}$. The first row is for l_{urban}=Biker. The second row is for I_{urban}=Hipster. Left column is for scene images, right column is for candidate person images.Top three ImageNet classes are marked.



(a) Individual accuracy - maximum of I

(b) Scene accuracy - maximum of I

Figure 6: The relationship between class-wise recognition rate and maximum of mutual information $I(l_{urban}, l_{ImageNet})$

[4] G. Wang, A. C. Gallagher, J. Luo, and D. A. Forsyth, “Seeing people in social context: Recognizing people and social relationships.,” in *ECCV* (K. Daniilidis, P. Maragos, and N. Paragios, eds.), vol. 6215, pp. 169–182, Springer, 2010.

[5] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, “Finding happiest moments in a social context.,” in *ACCV 2* (K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, eds.), vol. 7725, pp. 613–626, Springer, 2012.

[6] D. Rumelhart, J. McClelland, and the PDP Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1. MIT Press, 1986.

[7] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural Computation*, vol. 1, pp. 541–551, 1989.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” in *Proceedings of the IEEE*, vol. 86(11), pp. 2278–2324, 1998.

[9] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “Maxout networks,” 2013, arXiv:1302.4389.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks.,” in *NIPS* (P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1106–1114, 2012.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. R. adn Dragomir Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions,” 2014, arXiv:1409.4842.

[12] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” in *International Conference on Learning Representations (ICLR2014)*, 2014.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[14] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR*, vol. abs/1207.0580, 2012.

[15] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “Network in network,” 2013, arXiv:1312.4400.

[16] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks.,” *CoRR*, vol. abs/1311.2901, 2013.

[17] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition.,” *CoRR*, vol. abs/1310.1531, 2013.

[18] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, “Recognizing Image Style,” 2013, arXiv:1311.3715.

[19] M. Maffesoli, *The Time of the Tribes: The Decline of Individualism in Mass Society*. SAGE Publications Ltd, 0 ed., 1996.

[20] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “Liblinear: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[21] Y. Jia, “Caffe: An open source convolutional architecture for fast feature embedding.” http://caffe.berkeleyvision.org/, 2013.

[22] J. Plevin, “Who’s a hipster?.” http://www.huffingtonpost.com/julia-plevin/whos-a-hipster_b_117383.html.

[23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.