Urban Perception of Commercial Activeness from Satellite Images and Streetscapes

Wenshan Wang
Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University
Shanghai, China
wswang14@fudan.edu.cn

Su Yang*
Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University
Shanghai, China
suyang@fudan.edu.cn

Zhiyuan He
Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University
Shanghai, China
16210240032@fudan.edu.cn

Minjie Wang
Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University
Shanghai, China
minjiewang12@fudan.edu.cn

Jiulong Zhang
School of Computer Science, Xi'an University of Technology
Xi' an, China
zhangjiulong@xaut.edu.cn

Weishan Zhang
Department of Software Engineering, China University of Petroleum
Qingdao, China
zhangws@upc.edu.cn

ABSTRACT

People can percept social attributes from streetscapes such as safety, richness, and happiness by means of visual perception, which inspires the research in terms of urban perception. To the best of our knowledge, this is the first work focused on revealing the relationship between visual patterns of satellite images as well as streetscapes and commercial activeness. We propose to make use of bag of features (BoF) in the context of computer vision and sparse representation in the sense of machine learning to predict commercial activeness of urban commercial districts. After obtaining the urban commercial districts via clustering, we predict the commercial activeness degrees of them using four image features, namely, Histogram of Oriented Gradients (HOG), Autoencoder, GIST, and multifractal spectra for satellite images and street view images, respectively. The performance evaluation with four large-scale datasets demonstrates that the presented computational framework can not only predict the commercial activeness with satisfactory precision compared with that based on Point of Interest (POI) data but also discover the visual patterns related.

CCS CONCEPTS

• **Applied computing** → *Sociology*;

KEYWORDS

Urban perception; social intelligence; pervasive computing; computer vision; data mining

*Correspondence author. The author is also with School of Computer Science, Xi'an University of Technology, Xi'an, China.

WWW'18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04..
https://doi.org/10.1145/3184558.3186581

ACM Reference Format:
Wenshan Wang, Su Yang, Zhiyuan He, Minjie Wang, Jiulong Zhang, and Weishan Zhang. 2018. Urban Perception of Commercial Activeness from Satellite Images and Streetscapes. In *The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3186581

1 INTRODUCTION

In terms of intelligent information processing over urban big data, an emerging trend is vision-based urban perception. For example, people can judge the richness and safety of a place according to the scene images of this place [20] [22] [24] [26], which shows that there exist correlations between the visual patterns of scene images and the social contexts indicated by such visual patterns. The goal of urban perception is to discover such visual patterns that can act as signals to infer the social attributes of the corresponding scene images, and reveal the association rules to correlate the visual patterns with the social attributes of scene images.

In the literature, Arietta et al. [1] propose to predict location-aware crime rate and population from Google streetscapes based on visual pattern analysis, and learn prediction models with the visual patterns and the corresponding social attributes as input and output, respectively. Khosla et al. [15] propose to predict the distance to McDonald or hospital from the visual patterns of scene images, where they make use of a variety of visual pattern analysis techniques including convolutional neural networks (CNN) and the prediction results turned out from support vector regression (SVR) are comparable to those from human intelligence. Furthermore, they classify city functions into seven categories for urban planning based on about two million scene images collected from 21 cities [33]. Quercia et al. [25] discover the visual patterns to make London beautiful, quiet, and happy based on crowdsourcing and image processing. Recently, Dubey et al. [8] introduce the new dataset focused on liveliness and depression for urban perception. In particular, Nadai et al. [7] investigate the correlation between

Figure 1: Some examples of the street view images indicating different levels of commercial activeness. The levels of images are (highest) B > C > A (lowest).

Figure 2: Some examples of the satellite images in Beijing indicating different levels of commercial activeness. The levels are: (highest) B > C > A (lowest).

the safety and the call frequency of mobile phone users in a region by using convolutional neural network based streetscape analysis. Gebru et al. [10] propose a deep learning-based method to estimate socioeconomic characteristics of regions using 50 million Google street view images. Nikhil et al. [19] use a computer vision method to model the dynamics of physical urban change from time-series streetscapes.

Although a couple of works have been done to reveal the correlation between scene images and certain social attributes, business intelligence in terms of location-aware business planning in regard to surrounding social contexts, a significant issue in terms of socioeconomics and smart cities, remains a missing topic so far. It is known that the visual patterns of streetscapes are informative indicators of population [1], city functions [33], call frequency of mobile phone users reflecting human mobility[7], and psychological perception of beautifulness and happiness [25], which are intuitively key factors to determine commercial activeness of a region, so the correlation between the visual patterns of streetscapes and commercial activeness is analyzed in this study. Aside from streetscapes, satellite images should be a natural means to observe a city from a global point of view and a recent trend is to leverage remote-sensing images to reveal the correlation between the visual patterns and the socioeconomic property of a region. Jean et al. [14] propose to use convolutional neural networks along with transferring learning to learn features for predicting the poverty across a country from nighttime light intensities over satellite images. However, how and what kinds of visual features are correlated to city functions for a broad-spectrum social-economic issues is an open problem yet. In this study, we make use of satellite images to infer commercial activeness due to the intuition that city functions and infrastructures should be exhibited more straightforwardly in the global view of satellite images, which are the key issues among the social contexts to determine the commercial activeness of a region.

In this paper, we investigate the connection between the visual appearances of urban commercial district (UCD) and the corresponding commercial activeness, by applying computer vision and machine learning techniques to fit the visual features of UCD with the corresponding commercial activeness, which is quantized by means of the online reviews on commercial entities as proxy. To the best of our knowledge, this is the first work focused on revealing the relationship between visual patterns of urban regions and commercial activeness based on satellite images and street view images. Fig. 1 and Fig. 2 show some examples of street view images and satellite images in Beijing, respectively, which indicate different levels of commercial activeness. Intuitively, visual features can reflect which one corresponds with high/low commercial activeness. The motivation to fuse satellite images and streetscapes for commercial activeness perception is: It is known that city functions play an important role in attracting people's visiting to a region [30], which enables figuring out regionally geographic profiles in terms of data mining [32], and city infrastructures can be identified in remote-sensing images based on ground objects and structures via deep learning [4]. However, the detailed visual appearances of the ground objects are not visible in satellite images but streetscapes can provide such details complementary to the global view of these objects. Therefore, by fusing visual information from satellite images and streetscapes, higher performance for commercial activeness perception can be expected, which has been experimentally confirmed in this study. The advantage of applying visual perception to prediction of commercial activeness lies in that we can reach a visually straightforward understanding about how commercial activeness is boomed, which reveals the phenomenon from not only the city infrastructure perspective but also the psychology perspective on what kinds of visual patterns affect people's favor.

The problem in regard to urban perception is two-fold: (i) Discovering the physical law that correlates visual patterns with certain social attributes as well as detecting the corresponding visual patterns. (ii) Developing analytical toolkits in the context of computer vision and machine learning to enable sound analysis over visual big data for urban perception.

In order to predict social attributes, most of the existing research works use classifical image features such as HOG, Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBP) [15, 20], and deep image features (convolutional neural network) [24]. On the basis of these works, we present a novel computational framework based on bag of features to predict commercial activeness of urban commercial districts (UCDs), where four image features are incorporated into the bag-of-features model for satellite images and street view images, say, GIST, HOG, Autoencoder, and multifractal spectra, respectively. Specifically, we aim to address the following three questions:

RQ1: Can commercial activeness of UCDs be automatically predicted using classical image features (GIST, HOG, multifractal spectra) as well as deep features (Autoencoder) extracted from satellite images and streetscapes, respectively?

RQ2: Can the combined features from satellite images and street view images predict the commercial activeness more accurately than using those based on either modality of images alone?

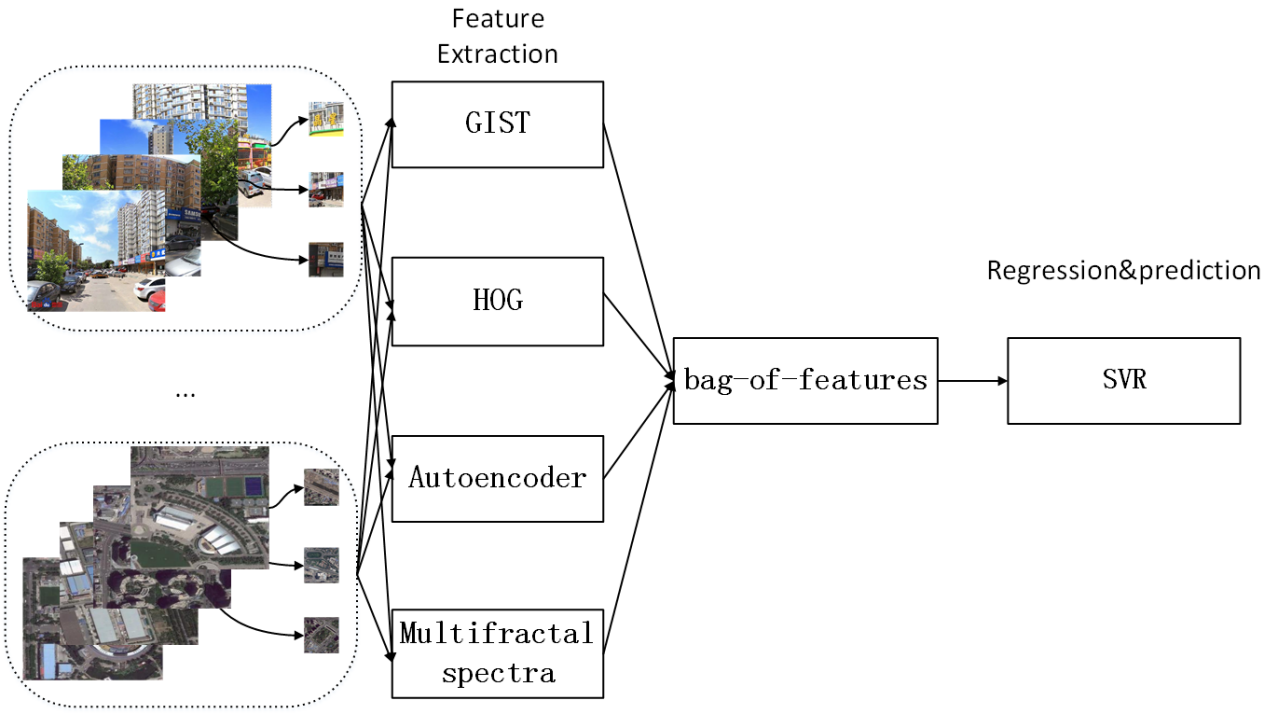


Figure 3: The flowchart of the computational framework based on bag of features.

RQ3: What visual patterns are highly correlated to commercial activeness?

To address these questions, we collect four datasets for Beijing and Shanghai, respectively: a) Point-of-Interest data; b) A dataset of large-scale streetscapes collected from web; c) A dataset of satellite images from Google Earth; d) A dataset of customers' comments on commercial entities. We describe the visual appearances of UCDs using bag of features based on GIST, HOG, Autoencoder, and multifractal spectra for satellite images and street view images, respectively. Furthermore, we reveal the visual patterns related to the commercial activeness based on sparse representation.

The contributions are summarized as follows:

(1) To the best of our knowledge, it is the first work focused on revealing the relationship between visual patterns of scene images and commercial activeness of urban regions.

(2) The experimental results show the feasibility to automatically predict commercial activeness using bag-of-features model over the visual features of satellite images and street view images for each UCD. Based on bag-of-features model, moreover, we are able to find out the visual patterns that relate to commercial activeness.

(3) We aim to discover the correlation between the visual patterns and the commercial activeness in a city so as to predict urban commercial activeness by analyzing large-scale remote-sensing images. As far as we know, it is the first study focused on remote-sensing images and streetscapes rendered business intelligence in the context of urban perception. Interestingly, we identify some semantically meaningful objects only under the supervision of commercial activeness.

2 METHODS

In this work, we consider predicting commercial activeness of UCDs as a regression problem based on BoF [28]. First, we employ the BoF model to obtain a representation of the UCDs' visual appearances of interest. Second, we define and compute the proxy of the commercial activeness using the online review data for each UCD. Third, we apply the regression methods in terms of machine learning to model the association rules between the BoF descriptors and the commercial activeness of the UCDs of interest for the sake of predicting the activeness. Moreover, four generic image features are combined to form a single feature vector prior to computing the BoF model so as to examine how the prediction performance can be improved by fusing different visual features as a whole. The flowchart of the pipelined computational procedure is illustrated in Fig. 3.

2.1 Measures of visual appearances of urban regions

Here, we use bag of features as our model to generate descriptors of urban appearances for each UCD. First, we collect the satellite images and the street view images for every UCD. Second, we partition every image into 50 patches to focus on the local contents. We randomly extract 256×256 patches for each modality of images. Third, we perform four kinds of feature extraction methods on every image patch, namely, GIST, HOG, Autoencoder, and multifractal spectra for street view images and satellite images, respectively. Fourth, we employ the BoF model to obtain a histogram-based overall profile of the UCD of interest, where similar visual features are

grouped into identical histogram bins [5]. The details are described below.

2.1.1 Feature extraction.

GIST: GIST is a biologically inspired feature that globally encodes an image while ignores local details. Previous studies [21] suggest that GIST simulates the process of human visual observation of scenes. Here, we use the popular GIST[1] for scene recognition over satellite images and street view images, respectively, which leads to 512-dimentional descriptor for each patch.

HOG: Some studies [9] suggest that the low-level visual perception mechanism in human visual systems is based on gradient features. Here, we choose to use the powerful feature referred to as HOG [6]. In our experiments, we use 124-dimensional HOG descriptors for 2×2 image cells. We randomly sample 100 HOG descriptors and concatenate them together, which generates 12400-dimentional feature vector for each patch of satellite images as well as streetscapes.

Multifractal spectra: Fractal dimensions provide clues to discriminate natural contexts from man-made objects [23] and can measure the randomness of the object of interest. Multifractal theory extends classical fractal theory to enable more powerful analytical tools in image processing. In particular, multifractal spectra [17] have been applied in a variety of practical applications such as texture analysis [16]. Intuitively, economic activities are correlated to the distribution of infrastructures and natural plants in a city. To characterize the density of man-made infrastructures and natural plants as textures, we employ multifractal spectra to analyze satellite images and street view images. To encode each patch, we make use of multifractal spectra [11] to generate 40-dimentional feature vector.

Autoencoder: Artificial neural networks have been widely used in various challenging computer vision tasks. An autoencoder is a neural network, which is used to train its input to approach its output. It can then be used for feature representation and dimensionality reduction [12]. Sparse autoencoder is one of the variations of autoencoder with an additional sparse penalty term. In addition, autoencoders can be stacked to form a deep network by feeding the output of the previous layer as the input to the current layer.

In our experiments, we train stacked sparse autoencoders [29] for predicting commercial activeness. The first sparse autoencoder learns a representation of the original input. Then, we treat the representation as a new input and use the second sparse autoencoder to learn a new representation for it. The third and fourth sparse autoencoders are learnt from the representations of the previous autoencoders. Here, we train a 4-layer stacked autoencoders. The number of the neurons in the 4-layer neural network is 80, 100, 100, and 100, respectively. We train the model using the stochastic gradient descent with a learning rate of 0.01. Every iteration in the learning procedure treats a collection of 25,000 patches. We extract features from the fourth layer to encode each image patch, resulting in the dimensionality of 100.

2.1.2 Bag-of-features model.

In order to build a dictionary, we use the K-means algorithm to cluster all the descriptors (GIST, HOG, multifractal spectra, and Autoencoder) of the image patches of the UCDs in the training set. Then, we apply vector quantization to assign the descriptors of each UCD to the dictionary to obtain a final BOF vector for each UCD's visual appearance. Here, we achieve optimal hyperparameters of K-means through grid search.

[1] http://people.csail.mit.edu/torralba/code/spatialenvelope/

2.2 Proxy for commercial activeness

We define the proxy that represents the level of commercial activeness or popularity for an UCD as follows: In general, the key attribute of online review data is the number of reviews, which reflects users' interests in commercial entities, namely, popularity [2]. Here, we sum the number of user-generated comments as the indictor of commercial activeness since it reflects how much a UCD attracts the attentions of users, who have truly experienced the interactions with the corresponding entities [13].

2.3 Predicting methods (regression model)

Here, we learn to predict commercial activeness of UCDs in the sense of regression. Given the image feature vector $x_i \in R^n$ and the corresponding target value $y_i \in R$, our goal is to make the predicted value $\hat{y}_i \in R$ approximate the ground truth such that,

$$\hat{y}_i = \tilde{\mathbf{w}}^T \phi(x_i) + b \quad b \in R, i = 1, ..., l \tag{1}$$

where $\hat{y}_i$ is the predicted value, $\phi(x_i)$ maps x_i into a higher-dimensional space, and $\tilde{\mathbf{w}}$ is the optimal weight vector to be learnt.

We use SVR with RBF kernel. In practical situation, the ν-SVR in libsvm [27] tries to minimize the following function:

$$\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{l}\sum_{i=1}^{l}(\max(0, |y_i - \hat{y}_i| - \varepsilon)) \tag{2}$$

where $\mathbf{w}$ is the weight vector, ε the constant, and C the regularization hyperparameter.

3 EXPERIMENTAL RESULTS

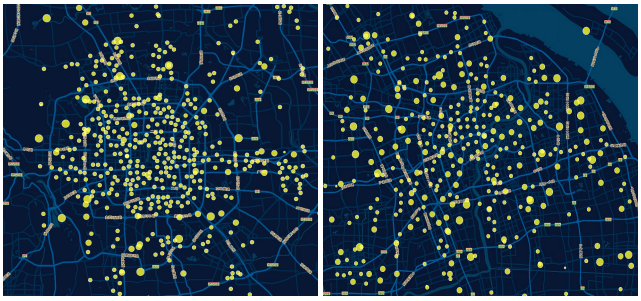
3.1 Datasets

We use four datasets to conduct this research: a) Point-of-Interests data from Baidu Maps; b) A dataset of large-scale streetscapes collected from web using Baidu API; c) A dataset of satellite images from Google earth[2]; d) A dataset with customers' comments on commercial entities collected from http://www.dianping.com.

Point of Interest: A record of point of interest (POI) contains the following items: ID of POI, the longitude and latitude of the location of interest, the address in plain text, and the category to denote the associated city function such as shopping mall, transportation facility, or, hotel. We collect 1,007,373 POI records for Beijing and 1,363,709 POI records for Shanghai via Baidu Maps API[3].

Streetscapes: We hold a large-scale dataset with 272,552 and 204,161 street view images for Beijing and Shanghai, respectively. Here, we use Baidu Maps API[4] to collect images with 400 grid-point intervals within each UCD. We collect four images from per grid-point, which are located 50 meters apart in the four different directions.

[2] https://www.google.com/earth/
[3] http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-placeapi
[4] http://developer.baidu.com/map/viewstatic.htm



(a) UCDs in Beijing (b) UCDs in Shanghai

Figure 4: The areas of the circles represent the relative commercial activeness.

Satellite images: We collect 456 and 429 satellite images for Beijing and Shanghai, respectively, via Google Static Maps API[5]. The resolution of each satellite image is 1024×768 pixels at zoom level 17, covering every UCD with the radius of around one kilometer.

Comments on commercial entities: Some websites such as Yelp archive comments and reviews from customers in terms of their favor on an entity such as a restaurant. In China, the best-known website to record users' comments on commercial entities is Dianping[6], where we collect 100,313 and 110,490 comments for Beijing and Shanghai, respectively. The review data are from December 9, 2014 to February 11, 2015. Each comment includes ID of POI, address of POI, category of POI, total number of users' comments, average price of services or products, and users' rating score. The number of reviews reflects users' interests, say, popularity [2].

3.2 Urban commercial districts

We apply a clustering algorithm to obtain urban commercial districts from POIs data. The POIs with categories to be shopping malls or commercial streets are considered as the seeds for clustering. First, a few seeds are selected to initialize the centroids of the corresponding groups. Then, each POI is assigned to the group the centroid of which is the closest one to the POI among all the competitors. After that, the centroids of the corresponding groups should be updated by taking into account the newly included POIs. A new group will be created if the distance between the POI and the closest centroid is greater than a predefined threshold. Repeat the above procedure until convergence. We obtain the optimal threshold based on grid search.

We find that there are 456 and 429 UCDs in Beijing and Shanghai, respectively, by applying the clustering algorithm. As shown in Fig. 4, the areas of the circles represent the relative commercial activeness.

3.3 Performance in predicting commercial activeness

Here, we evaluate the performance of a regression model using the metric referred to as Mean Absolute Percentage Error (MAPE) by which the precision is defined as follows:

$$Accuracy = 1 - \frac{1}{n}\sum_{i=1}^{n} \mid \frac{\hat{y}_i - y_i}{y_i} \mid \quad (3)$$

where y_i is the true value, $\hat{y}_i$ the predicted value, and n the number of the testing examples.

In this work, we perform four experiments. In general, we randomly partition the dataset into two parts, one for training and the other for testing. We compute the average accuracy over 10 tests with random train/test splits. In order to obtain optimal hyperparameters, we make use of grid search.

3.3.1 Baseline method. It is known that people can percept some socioeconomic attributes from visual indicators of streetscape, such as city function [33], safety [20], and population [1]. These attributes can also be predicted from POI data [31]. Therefore, to demonstrate the effectiveness of the proposed method, we employ POI data to predict commercial activeness as the baseline method. In this work, we utilize bag-of-features model to extract the 13-dimensional POI feature, which reflects the distribution of 13 categories of POIs within each UCD. The 13 different categories of POIs are listed in Table 1. Then, we train the SVR-based predictive model with the 13-dimensional POI feature and the UCDs' commercial activeness as input and output, respectively. The accuracy of the baseline method on Beijing and Shanghai datasets are shown in Table 2.

Table 1: The POI features

No.	Types
1	Catering services
2	Road traffic facilities
3	Address information
4	Famous scenery
5	Company enterprise
6	Shopping services
7	Financial insurance services
8	Science, education, and cultural services
9	Automobile service
10	Living service
11	Sports entertainment services
12	Health care services
13	Government agencies and social organizations

Table 2: The predicted accuracy (%) using SVR with 13-dimensional POI features

City	POI features
Beijing	55.68
Shanghai	54.47

3.3.2 Satellite images. As shown in Table 3, we evaluate the precision of the regression model on four image feature representations for Beijing and Shanghai based on satellite images, respectively.

[5]https://developers.google.com/maps/documentation/static-maps/intro?hl=zh-cn
[6]https://www.dianping.com/

Table 3: Prediction results (%) using regression model on Beijing and Shanghai using satellite images.

City	HOG	Autoencoder	GIST	Multifractal	All
Beijing	56.30	58.80	57.67	57.88	60.17
Shanghai	55.27	57.30	56.78	57.12	59.83

We observe that both Autoencoder and multifractal spectra achieve the superior prediction results on Beijing and Shanghai datasets in the sense of MAPE. This is because Autoencoder is capable of representing high-level concepts and multifractal spectra have the ability to distinguish between natural and artificial image contents. They should be good representations of satellite images. Besides, we find that the combined features achieve the best performance. In addition, the MAPE value is from 55% to 60%, indicating that the regression model can achieve better prediction results compared with the baseline method.

3.3.3 Street view images. In terms of streetscapes, we evaluate the accuracy of the regression model on four image feature representations for Beijing and Shanghai, respectively. As shown in Table 4, we observe that the Autoencoder feature achieves the best prediction result for both Beijing and Shanghai datasets in the sense of MAPE. Autoencoder is a hierarchical feature that can represent abstract concepts well. Besides, we find that the combined features (HOG, Autoencoder, GIST, and multifractal spectra) lead to better performance for both Beijing and Shanghai. Compared with the baseline method, the MAPE is from 56% to 62%, which indicates that the regression model gains advantage over POI based prediction.

Table 4: Prediction of Commercial Activeness (%) in Beijing and Shanghai using street view images.

City	HOG	Autoencoder	GIST	Multifractal	All
Beijing	61.40	62.53	61.10	60.85	62.66
Shanghai	58.18	59.35	57.94	56.97	60.46

3.3.4 Combine satellite images and street view images. Here, we perform the experiment based on feature fusion. We use the aforementioned fusion strategy to concatenate all the four image feature representations including HOG, Autoencoder, GIST, and multifractal spectra. As shown in Table 5, we observe that it improves the prediction accuracy for Beijing and Shanghai, respectively.

Table 5: Concatenate all the image feature representations (%)

City	concatenate all the image features
Beijing	64.07
Shanghai	63.33

4 VISUAL PATTERNS CORRELATED TO COMMERCIAL ACTIVENESS

Here, we aim to answer: What visual patterns are correlated with commercial activeness? What visual cues contribute to predict commercial activeness?

In order to find out the visual patterns that correlated to the commercial activeness, we make use of sparse representation. The goal is to make $\hat{y}_i \in R$ approach the ground truth $y_i \in R$. Then, we add a l^1-norm term to the optimization function as follows:

$$\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}}(\lambda|\mathbf{w}|_1 + \frac{1}{2}|y_i - \hat{y}_i|_2^2) \quad (4)$$

where $\mathbf{w}$ is the weight vectors, $\tilde{\mathbf{w}}$ the optimal weight vectors, and λ the Lagrange multiplier. Specifically, sufficiently large λ makes some of the coefficients in $\mathbf{w}$ to be zeros. Furthermore, the l^1-norm solution has sparse property, which is verified [3].

In the experiment, we train the sparse representation model with the bag-of-features vector and the commercial activeness of each UCD as input and output, respectively, with the help of the open-source software SPAMS [18]. Then, we sort the weights resulting from sparse representation in descending order, which indicates the corresponding contribution of every component in the BOF model. Note that every component in the BOF model corresponds with a cluster of a couple of image patches with similar visual features. We figure out the image patches that are closest to the centroids of the clusters possessing high weights in terms of sparse representation, which are identified as the representatives of the patterns highly relevant to commercial activeness. The visual patterns related to commercial activeness in Beijing and Shanghai are illustrated in Fig. 5 and Fig. 6, respectively.

In Fig. 5, we find that there are several semantically meaningful features such as buildings, houses, urban roads, factories, lakes, farmlands, residential areas, and barren lands. Surprisingly, these semantic meaningful objects are found only under the supervision of commercial activeness. Specifically, the objects that show strong positive impact on commercial activeness are buildings, residential areas, and sports playgrounds. Intuitively, the commercial activeness should be high near these objects, since there are many people living there. Furthermore, factories, urban roads, and houses are the objects that show medium positive impact on commercial activeness while the objects that contribute lowly to commercial activeness are lakes, farmlands, and suburban districts. It is surprising to find that the patch in row 1 and column 2 of Fig. 5 is the airport in Beijing. In general, the airport is built in suburbs, where the commercial activeness should be low.

On the other hand, some semantically meaningful objects related to commercial activeness can be found in Fig. 6. We observe that cars, roads, pedestrians, and brands have strong positive impact on commercial activeness. Intuitively, higher commercial activeness corresponds with more cars and more people. Moreover, the objects that show medium positive impact on commercial activeness are apartments, office buildings, and small business streets while the objects that contribute lowly to commercial activeness are industrial areas and suburban districts.

Figure 5: The visual patterns of satellite images related to commercial activeness. (Beijing: top two rows and Shanghai: bottom two rows)

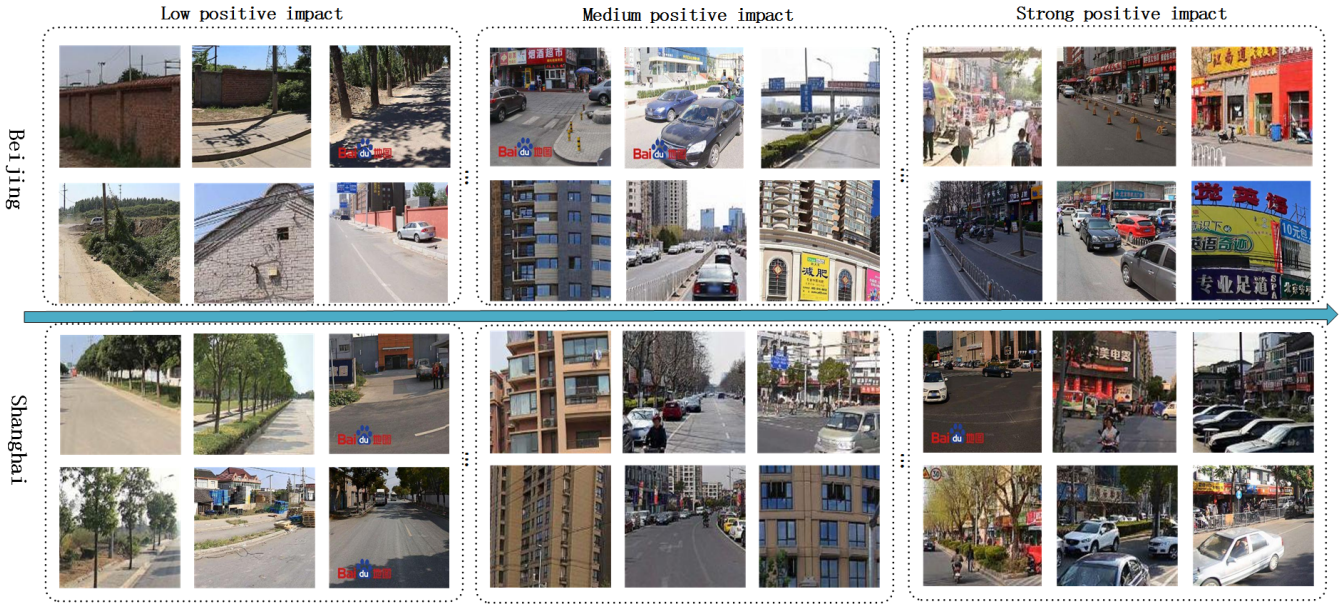


Figure 6: The visual patterns of street view images related to commercial activeness. (Beijing: top two rows and Shanghai: bottom two rows)

5 CONCLUSIONS

It is the first effort to apply heterogeneous image data to infer the correlation between commercial hotness and visual patterns of satellite images as well as streetscapes for the sake of commercial hotness prediction. Besides, we obtain visually straightforward knowledge regarding how city infrastructures and people's psychological favors in terms of visual perception affect commercial hotness. We present a novel computational framework based on bag of features in the context of computer vision and sparse representation in terms of machine learning to predict commercial activeness of urban commercial districts, where we use four image

features, namely, GIST, HOG, Autoencoder, and multifractal spectra for street view images and satellite images, respectively. The performance evaluation demonstrates that the predictor is promising compared with the baseline method and the visual patterns correlated to commercial activeness are revealed.

ACKNOWLEDGMENTS

This work is supported by NSFC (grant NO. 61472087) and Shanghai Science and Technology Commission (grant No. 1751110420).

REFERENCES

[1] S. M. Arietta, A. A. Efros, R. Ramamoorthi, and M. Agrawala. 2014. City Forensics: Using Visual Elements to Predict Non-Visual City Attributes. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec 2014), 2624–2633. https://doi.org/10.1109/TVCG.2014.2346446

[2] S. Bakhshi, P. Kanuparthy, and E. Gilbert. 2014. Demographics, Weather and Online Reviews: A Study of Restaurant Recommendations. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, New York, NY, USA, 443–454. https://doi.org/10.1145/2566486.2568021

[3] Ssb. Chen, S. Ma., and D. Dl. 2001. Atomic decomposition by basis pursuit. *Siam Review* 43, 1 (2001), 33–61.

[4] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu. 2014. Deep Learning-Based Classification of Hyperspectral Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7, 6 (June 2014), 2094–2107. https://doi.org/10.1109/JSTARS.2014.2329330

[5] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. 2004. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision Eccv* (2004), 1–22.

[6] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 886–893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[7] M. De Nadai, R. L. Vieriu, G. Zen, S. Dragicevic, N. Naik, M. Caraviello, C. A. Hidalgo, N. Sebe, and B. Lepri. 2016. Are Safer Looking Neighborhoods More Lively?: A Multimodal Investigation into Urban Life. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. ACM, New York, NY, USA, 1127–1135. https://doi.org/10.1145/2964284.2964312

[8] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo. 2016. Deep Learning the City: Quantifying Urban Perception at a Global Scale. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*. Springer International Publishing, Cham, 196–212. https://doi.org/10.1007/978-3-319-46448-0_12

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (Sept 2010), 1627–1645. https://doi.org/10.1109/TPAMI.2009.167

[10] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei. 2017. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences* 114, 50 (2017), 13108–13113. https://doi.org/10.1073/pnas.1700035114 arXiv:http://www.pnas.org/content/114/50/13108.full.pdf

[11] D. Gimenez, A. Posadas, and M. Cooper. 2003. Multifractal Characterization of Soil Pore Shapes. *Soil Science Society of America Journal* 08901, 12 (2003), 6388–6394.

[12] G. E. Hinton and R. R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (2006), 504–507. https://doi.org/10.1126/science.1127647 arXiv:http://science.sciencemag.org/content/313/5786/504.full.pdf

[13] M. Hu and B. Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*. ACM, New York, NY, USA, 168–177. https://doi.org/10.1145/1014052.1014073

[14] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

[15] A. Khosla, B. An, J. J. Lim, and A. Torralba. 2014. Looking Beyond the Visible Scene. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 3710–3717. https://doi.org/10.1109/CVPR.2014.474

[16] J. Levy Vehel, P. Mignot, and J. Berroir. 1992. Multifractals, texture, and image analysis. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*. 661–664.

[17] R. Lopes and N. Betrouni. 2009. Fractal and multifractal analysis: A review. *Medical Image Analysis* 13, 4 (2009), 634–649.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. 2009. Online dictionary learning for sparse coding. In *International Conference on Machine Learning*. 689–696.

[19] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo. 2017. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences* 114, 29 (2017), 7571–7576. https://doi.org/10.1073/pnas.1619003114 arXiv:http://www.pnas.org/content/114/29/7571.full.pdf

[20] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo. 2014. Streetscore – Predicting the Perceived Safety of One Million Streetscapes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 793–799. https://doi.org/10.1109/CVPRW.2014.121

[21] A. Oliva and A. Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.

[22] V. Ordonez and T. L. Berg. 2014. Learning High-Level Judgments of Urban Perception. In *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*. Springer International Publishing, Cham, 494–510. https://doi.org/10.1007/978-3-319-10599-4_32

[23] A. P. Pentland. 1984. Fractal-based description of natural scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6, 6 (1984), 661–74.

[24] L. Porzi, S. Rota Bulò, B. Lepri, and E. Ricci. 2015. Predicting and Understanding Urban Perception with Convolutional Neural Networks. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. ACM, New York, NY, USA, 139–148. https://doi.org/10.1145/2733373.2806273

[25] D. Quercia, N. K. O'Hare, and H. Cramer. 2014. Aesthetic Capital: What Makes London Look Beautiful, Quiet, and Happy?. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing (CSCW '14)*. ACM, New York, NY, USA, 945–955. https://doi.org/10.1145/2531602.2531613

[26] P. Salesses, K. Schechtner, and C. A. Hidalgo. 2013. The Collaborative Image of The City: Mapping the Inequality of Urban Perception. *PLOS ONE* 8, 7 (07 2013), 1–12. https://doi.org/10.1371/journal.pone.0068400

[27] B. Scholkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. 2000. New support vector algorithms. *Neural Computation* 12, 5 (2000), 1207–1245.

[28] J. Sivic and A. Zisserman. 2003. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*. 1470–1477 vol.2. https://doi.org/10.1109/ICCV.2003.1238663

[29] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol. 2010. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* 11 (Dec. 2010), 3371–3408. http://dl.acm.org/citation.cfm?id=1756006.1953039

[30] M. Wang, S. Yang, Y. Sun, and J. Gao. 2017. Human mobility prediction from region functions with taxi trajectories. *PLOS ONE* 12, 11 (11 2017), 1–23. https://doi.org/10.1371/journal.pone.0188735

[31] J. Yuan, Y. Zheng, and X. Xie. 2012. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, New York, NY, USA, 186–194. https://doi.org/10.1145/2339530.2339561

[32] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. 2015. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. ACM, New York, NY, USA, 295–304. https://doi.org/10.1145/2684822.2685287

[33] B. Zhou, L. Liu, A. Oliva, and A. Torralba. 2014. Recognizing City Identity via Attribute Analysis of Geo-tagged Images. In *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III*. Springer International Publishing, Cham, 519–534. https://doi.org/10.1007/978-3-319-10578-9_34