Predicting City Safety Perception Based on Visual Image Content

Sergio F. Acosta and Jorge E. Camargo(✉)

UnSecureLab Research Group,
Universidad Nacional de Colombia, Bogotá, Colombia
{sfacostale,jecamargom}@unal.edu.co

Abstract. Safety perception measurement has been a subject of interest in many cities of the world. This is due to its social relevance, and to its effect on some local economic activities. Even though people safety perception is a subjective topic, sometimes it is possible to find out common patterns given a restricted geographical and sociocultural context. This paper presents an approach that makes use of image processing and machine learning techniques to detect with high accuracy urban environment patterns that could affect citizen's safety perception.

1 Introduction

Cities are spatial structures of a significant size. In the case of Bogotá-Colombia, the city has an urban area of 307,36 km^2, and it is divided into 20 well defined zones that are named localities. This attribute makes it difficult to appreciate and experience them completely in just one round. This is the reason why the individual perception of a city is the result of a mixture of own experiences and experiences from others. Neighborhoods often differ in their demographics, such as income, and ethnicity of people that inhabits them, but also on how safe they are perceived [6]. Some of the recent works about automatic urban perception prediction have been based on Convolutional Neural Networks (CNN) [7], sets of between 100,000 and 1,000,000 images and with a wider territory scale approach. This work presents an approach based on a restricted geographical and sociocultural context, a modest image set, and on a technique called transfer learning. According to [3], perception of security can be defined as:

> Perception of security (PoS) refers to the subjective assessment of the risk and the magnitude of its consequences. The risk can be defined as the likelihood that an individual will experience the effect of danger, threats, or any adverse events.

Projects such as [2,8,9] and [10] have been focused on how to structure computational models that make it possible the automatic characterization of cities. These works have based their research on the visual appearance of city streets, and its association with the people perception. Misclassification is mainly caused by the huge variability in the set of images associated to the same group or tag.

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 177–185, 2019.
https://doi.org/10.1007/978-3-030-13469-3_21

In [1] it is proposed that the activation output of inner layer of a CNN trained for some other classification task can be used as a visual feature of an image in a different classification task. Based on this idea, the first urban perception model based CNN was proposed in [10]. CNN approach is also implemented in [2], where a CNN architecture is implemented in order to build a worldwide urban perception model.

The methodology described in [11] involves an image score estimation using the fraction of times this is selected over another image, then this is corrected by the "win" and "loss" ratios of all images with which it was compared during a visual survey. In [8], people perception obtained through visual surveys is converted to a ranked score for each image using the Microsoft Trueskill algorithm [4]. In [2] and [10] it is proposed to predict pairwise comparisons by training a CNN model directly from image pairs and their crowd-sourced comparisons, which is used to generate "synthetic" comparisons by taking random image pairs as input.

Our paper presents an alternative approach in the construction of a computational model whose purpose is to predict how safe may be a given Bogotá city zone. This has been carried out in a restricted sociocultural and geographical context. In order to restrict sociocultural context, just people living in Bogotá was invited to take the visual survey. Safety perception like humor can depend on a particular sociocultural context. This is why sometimes a joke that is fun in the USA may not be fun in Germany. Geographical context restriction means that just local street images were used. One motivation for not using foreign street images is that many of the urban environment found in other countries, e.g. Washington or Germany, does not exist in Bogotá. It is expected that these restrictions reduce the variability of the underling distribution as well as the noisy of data. The presented approach makes use of a technique called transfer learning. This takes a piece of a model that has already been trained on a related task and reusing it in a new model. In particular VGG19 [12] model, loaded with weights trained on ImageNet, has been used for generic image feature extraction. A particular city zone of 40 km^2 was chosen for this work, and local people were asked to participate in the visual survey. Since Bogotá has an extremely heterogeneous urban environment, this restriction still guarantees a significant image variability. As a result, a model with an accuracy of 81% was obtained, and it was used to predict a safety perception score for two other neighboring localities. In order to detect different patterns, prediction on neighboring localities was performed using their particular image sets.

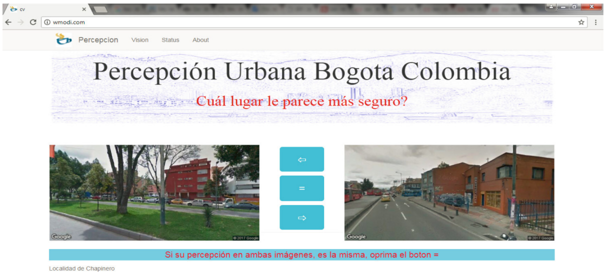
2 Materials and Methods

2.1 Dataset Construction

Street Image Crawler. A city street image crawler was built using the Google Street View API V3.0 along with a file that contained geographical limits of the target zones.

Image Filtering. Since some collected images have no a wide view of a street, but a close-up of a building facade, or a totally black image, it was required to remove these images from the collected set. This was done by extracting SIFT local descriptors from each image, and excluding images with a descriptor count less than 420.

Visual Survey. Visual surveys was carried out through a public web site. On this, users were shown two geotagged street city images, and asked to click on one in response to the questions "which place looks safer?". In this way a bit of citizenship perception is obtained. An image pair and its associated comparison are the unit of information that will be used during the training task. The visual survey tool is online in http://wmodi.com/. Figure 1 shows this site.



Percepción Urbana Bogota Colombia

Cuál lugar le parece más seguro?

Si su percepción en ambas imágenes, es la misma, oprima el boton =

Localidad de Chapinero

Fig. 1. Visual survey site

Image Serving Policies. Image pair serving policy and database housekeeping try to fulfill the following requirements.

- Each image should have the same vote share.
- Repeated image pair comparisons should be reduced to a single or no vote.
- Same image comparison is not allowed and should be removed.
- Comparisons of near images should be removed.

Vote Coding and Database Structures. In each visual survey session, two street images were presented to the user. The user was asked to click on one image or on the equal button in order to answer the question: Which place looks safer? Once the user click on an image or a button, the vote is coded as follows: If click on image 0 (left side image), vote is coded as 1; if click on button *equal* vote is coded as 0. Finally, if click on image 1 (right side image), vote is coded as 2. For each visual survey session, the resulting vote code along with the involved left and right image identifiers were stored. There is also a database structure associated to each published image. This holds the image identifier, a positive perception counter, a negative perception counter and a neutral perception counter. If user click on image 0 (left side image), this image positive perception counter is increased by 1 and image 1 (right side image) negative perception counter is increased by 1. Same logic applies if user click on image 1 (right side image). Finally, if user click on button *equal* both images neutral perception counter is increased by 1. This counters are used for each image basic perception percentage estimation.

Collected Data. 5,505 images of the target zone were published. In one year, 17,703 image pair votes were collected. Each image participated in 6 ± 1 image pair comparison session. The collected vote distribution based on its code is as follow: 5,657 code 0 votes, 5,946 code 1 votes, and 6,100 code 2 votes.

2.2 Image Feature Extraction

IN THE VGG19 Keras[1] model it was removed its fully-connected layer at the top of the network and loaded with the weights trained on ImageNet. Then, this model was used to obtain a row vector representation of each published image. This is a 512 values vector, i.e, the model is used like an image feature extractor and no training is required for this task. It is important to note that the goal of this research is to train a model that will be able to predict a visual survey vote based on a vector representation of a pair of city street images. That is, each training item is composed of two image feature vectors and their associated vote code.

2.3 Training Set Construction

The construction of the training set involved the transformation of the collected visual survey votes into number vectors and its associated label (vote code). This task starts with the elaboration of a list of all the votes found in the database. From this list, image identifiers are replaced by their associated vector descriptors. Finally, images descriptors associated to each single vote are concatenated and annotated with the respective $\{1, 2\}$. Ties (code 0 votes) were not used, but just charged votes were used. If an actual vote has been annotated with vote code 1, it means that left image was better perceived than the right image. This means also that if image positions are exchanged, the resulting vote should be annotated with vote code 2. This fact has been used to double the initial data set size. The resulting set of 24,092 $(5,946 * 2 + 6,100 * 2)$ votes was split into training, validation and testing sets. A distribution of 65-7-28 was used. In this way 15,636 votes were used for training, 1,754 for validation and 6,702 votes for testing. Each descriptor vector corresponding to each image of the VGG19 was normalized applying the following expression,

$$f_{i,j} = (f_{i,j} - \mu_i)/\sigma_i,$$

[1] https://keras.io/.

where $f_{i,j}$ is the j-th component of the i-th vector, (μ) the mean, and (σ) the standard deviation. Each feature i of each row vector j is normalized by subtracting the associated mean and scaled by the associated standard deviation. Same normalization scheme was used on the image set of the neighboring localities, the model was used to predict on.

Training Phase. Training sessions were performed using the TensorFlow[2] machine learning framework. The implemented neural network configuration is shown in Fig. 2.

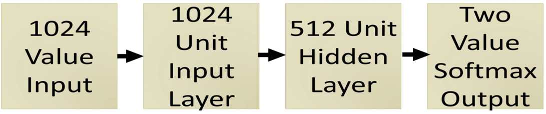


Fig. 2. Full connected neural network setup

The input data is the concatenation of image 0 and 1 VGG19 based on a 512 vector descriptor. This is a 1024 bin vector. A dropout technique was used as regularization method. Dropout rates of 0.5, 0.45 and 0.3 were applied to the input, hidden and output layers, respectively. The learning rate was set to 0.00001, and a mini-batch size was defined as 64. The AdamOptimizer [5] method was used along with Cross Entropy Loss. Figure 3 shows training session cost curves.

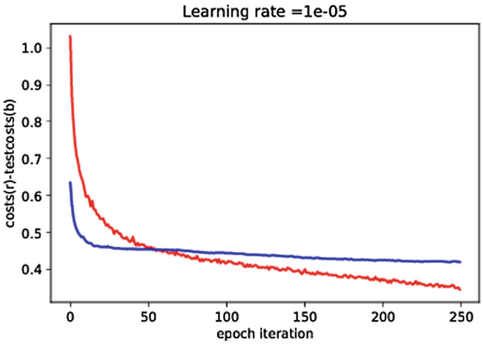


Fig. 3. Training session cost curves

Accuracy Report. Table 1 presents the confusion matrix for the testing set, for which an accuracy of 81.5% was achieved.

[2] https://www.tensorflow.org/.

Table 1. Testing confusion matrix

	Predicted Vote Code		Total
	1	2	
1	2,777	574	3,351
2	663	2,688	3,351
Total	3,440	3,262	6,702

Synthetic Vote Generation. Synthetic votes were generated dividing each locality collected image set into two same size groups. Each group holds images homogeneously distributed over the target locality area. Each group was divided into 10 subgroups. Then, each image from one group was paired with a randomly (uniform) selected image from each subgroup in the other group.

Synthetic Vote Prediction. In order to be able to make a map based on synthetic votes, the initial dummy vote label must be substituted by one based on the model prediction. The output of the softmax layer was used for this purpose. At first, the option with the higher probability was used to annotate the associated vote as 1 or 2. However, the absolute difference between the two probabilities must be higher than 0.25, otherwise the vote was annotated with the code 0. At the same time each image positive, negative and neutral perception counters were updated.

Image Score Based on Perception Counters. In section "Vote Coding and Database Structures" it was mentioned that published images had an structure associated to them and how its fields are updated during a visual survey session. This holds the image identifier, a positive perception counter, a negative perception counter and a neutral perception counter. Every image used in the synthetic vote prediction task has the same structure, and its perception counters are updated in the same way as during a visual survey session. If an image neutral perception counter is greater than 0, this value is redistributed between this image positive and negative counters. This redistribution is done by factors that are worked out form the perception counter summation of all images, which this image tied with. Finally, each image counter summation is normalized [0, 1] and each counter turned into positive, negative and neutral safety perception percentages.

3 Results

3.1 Actual and Synthetic Vote Maps of Same Zone

For the initial target zone training images were obtained, and a perception maps was built based on both, actual and synthetic votes. This was done in order to verify that colors patterns found in actual vote perception map are present in the synthetic vote perception map. For further exploration, actual[3] and synthetic[4] vote perception map are available in the project web site.

Figure 4 shows the color gradient used on the map set. Here left green end indicates a 100% percent of positive safety perception, and right red end 0%.

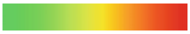


Fig. 4. Color gradient reference for image safety perception score (Color figure online)

3.2 Safety Perception Score Prediction for Other Localities

Model was used to generate synthetic votes on image pairs of different city zones. These image maps are available on line at left[5] and right[6] image map links (Fig. 5).

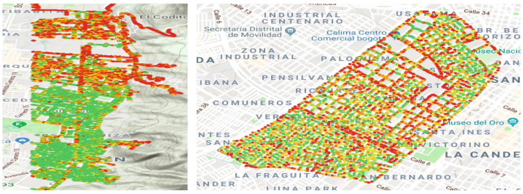
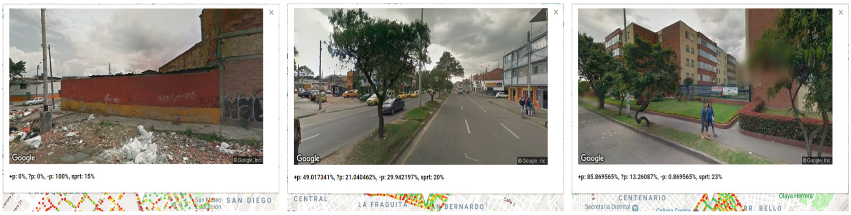


Fig. 5. Left: *Usaquen* zone 94,780 synthetic votes safety perception score map. Right: *Martires* zone 37,880 synthetic votes safety perception score map

3.3 Predicted Image Set

Figures 6 and 7 are samples of other zone images whose score has been predicted by the system based on synthetic votes. It is worth noting that the trained model predicts with high precision the perception of test images. For instance, the predicted vote of left image in Fig. 6 captures negative perception safety characteristics such as dirty houses, lonely streets, trash, etc., whilst the predicted vote of image in right image in Fig. 7 captures positive perception safety characteristics such illumination, green zones, clean streets, etc.

[3] http://wmodi.com/chapinero_17703actualvote_jun04_2018_imgscore.
[4] http://wmodi.com/chapinero_55040NNsyntheticvote_jun04_2018_imgscore.
[5] http://wmodi.com/usaquen_94780NNsyntheticvote_jun04_2018_imgscore.
[6] http://wmodi.com/martires_37880NNsyntheticvote_jun04_2018_imgscore.

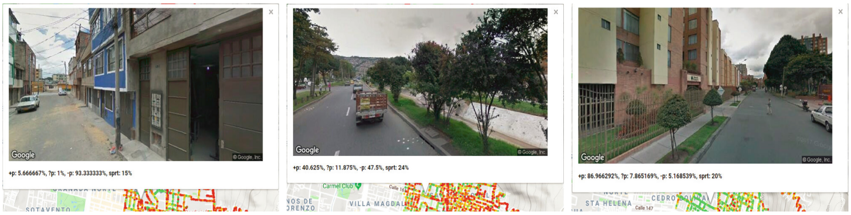


+p: 0%, ?p: 0%, -p: 100%, sprt: 15%

+p: 49.017341%, ?p: 21.040462%, -p: 29.942197%, sprt: 20%

+p: 85.869565%, ?p: 13.26087%, -p: 0.869565%, sprt: 23%

Fig. 6. *Martires* zone: 0%, 49%, 86% positive safety perception images



+p: 5.666667%, ?p: 1%, -p: 93.333333%, sprt: 15%

+p: 40.625%, ?p: 11.875%, -p: 47.5%, sprt: 24%

+p: 86.966292%, ?p: 7.865169%, -p: 5.168539%, sprt: 20%

Fig. 7. *Usaquen* zone: 6%, 41%, 87% positive safety perception images (Color figure online)

4 Conclusion and Future Work

This paper presented a model that allows to predict citizen's safety perception using visual information of street images. The obtained results show a prediction accuracy of 81%, which is higher than results obtained in recent state of the art methods. Up to our knowledge, this is the first time that this analysis is performed to Bogotá city. The presented method does not require a high computing capacity, that is, 1 model iterations per hour can be performed. This feature makes this approach appropriate for the development of an online tool. It is expected to carry out more exhaustive evaluations in order determine the robustness of the predictions as well as statistical stability. All these results and the results of an earlier model evaluated by us based on SVM with a smaller amount of votes are available at http://wmodi.com/.

References

1. Donahue, J., et al.: DeCAF: a deep convolutional activation feature for generic visual recognition (2013)
2. Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A.: Deep learning the city: quantifying urban perception at a global scale. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 196–212. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_12
3. Gómez, F., Torres, A., Galvis, J., Camargo, J., Martínez, O.: Hotspot mapping for perception of security. In: IEEE 2nd International Smart Cities Conference: Improving the Citizens Quality of Life, ISC2 2016 - Proceedings, pp. 0–5 (2016)

4. Herbrich, R., Minka, T., Graepel, T.: TrueSkill: a Bayesian skill rating system. In: Advances in Neural Information Processing Systems, vol. 20, pp. 569–576 (2006)
5. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR, abs/1412.6980 (2014)
6. Kominers, S.D., et al.: Do People Shape Cities, or Do Cities Shape People? The Co-evolution of Physical, Social, and Economic Change in Five Major U.S. Cities (2015)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1–9 (2012)
8. Naik, N., Philipoom, J., Raskar, R.: Streetscore - Predicting the Perceived Safety of One Million Streetscapes (2014)
9. Ordonez, V., Berg, T.L.: Learning high-level judgments of urban perception. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 494–510. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_32
10. Porzi, L., Buló, S.R., Lepri, B., Ricci, E.: Predicting and Understanding Urban Perception with Convolutional Neural Networks, pp. 139–148 (2015)
11. Salesses, P., Schechtner, K., Hidalgo, C.A.: The collaborative image of the city: mapping the inequality of urban perception. PLoS ONE **8**, e68400 (2013)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)