Recognizing places impressions from images

Luis Emmanuel Medina Ríos
EPFL
luis.medinarios@epfl.ch

ABSTRACT

Getting the impression of an scene is quite relative since several factors affect it. An scene could be interpreted in different ways depending of its context and the background of the interpreter. Since the problem of the latter could be very wide, we try to focus on the physical objects. In this report, we attempt (1) to answer to the question of what outdoor physical elements could explain different urban awareness labels by inferring them with different datasets with fixed features (objects) and (2) to compare impressions between two different groups of people (locals and non-locals) based on an image corpus grounded in 3 different cities in central Mexico. We perform both regression and classification tasks via Random Forest to analyze how well the objects from the datasets relate to each of the labels. Other tasks such as correlation between features and labels, and visualization of the features are done in order to understand more our data.

KEYWORDS

Inferring Places Impressions, Outdoor Objects, Manual Annotation Task, Correlation Between Visual Objects and Urban Awareness Labels, fine-tuning task, Urban Impressions, Crowdsourcing, Understanding Mexican Scene

1 INTRODUCTION

Trying to understand how people perceive an scene is not an easy task, since a lot of factors have to be taken into account. Such factors could include the gender, level of studies, country, socioeconomic status, neighborhood, time of day, experiences, etc. Therefore, there is no way to generalize this: Two people can have a totally different perception of the same scene. A great example of this could be if we tell two people from different countries and backgrounds (let us say a Mexican and a Swiss) to compare the same scene: The outside of the Shrine of "Holy Death" within the Tepito Barrio (neighborhood) in the heart of Mexico City [11]. On the one hand, the Mexican would react fast and say "this is dangerous", since most of the Mexican people link "Holy Death" and Tepito Barrio to crime [20], on the other hand, a Swiss would say, "it is not dangerous and it could be safe since it is a 'shrine' and people are praying". This was an extreme example, but it is just to illustrate how the context of the people affects the perception.

We attempt to do a follow-up of the work done by Dr. Darshan Santani, Dr. Daniel Gatica-Pérez and Dr. Salvador Ruiz-Correa during the SenseCityVity project [1] by developing the following objectives:

- Understand the relation between different objects and specific urban awareness labels.
- Compare impressions of two groups with completely different contexts and backgrounds.
- Train models to infer impressions of scenes within pictures.
- Compare two different type of datasets: one based on general objects and the other based on a visual vocabulary grounded on the image corpus under consideration.

This report is organized as follows: Firstly, we explain the SenseCityVity project, which is part of the previous work, followed by a description of the datasets that we use for the analysis, then we explain all the methodology we are based on to get the results. Afterwards, we perform a feature analysis on the different datasets, in which we include a t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis for a better visualization of the features and the labels and a correlation task between datasets with the visual objects and the urban awareness labels to understand the features that contribute more to the latter. Then, we do a manual annotation process on different sets of random pictures of the image corpus for two urban awareness labels: Dirty and Dangerous, afterwards we compare the impressions based in an image corpus between two groups with different backgrounds. We run classification and regression tasks to see how well we can infer the different urban awareness labels with the different objects we find in the image corpus. Finally, we perform a fine-tuning task to try to get models based on different CNN architectures.

2 PREVIOUS WORK

2.1 SenseCityVity project

SenseCityVity is a collaborative project, with both scientific and human sides, between the Idiap Research Institute and the École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland and the National Center for Supercomputing of the Instituto Potosino de Investigación Científica y Tecnológica (CNS-IPICYT) in Mexico, supported by EPFL's Cooperation and Development Center (CODEV) [1] [16] [15]. The idea of this work was to understand how people perceive places in their daily life, trying to make people reflex about different problems that are happening around them, this in order to raise awareness and propose solutions. In such project, an image corpus was generated during an Urban Data Challenge (UDC) in 3 different cities in the state of Guanajuato in Mexico. In this UDC, high school students from Guanajuato city participated documenting the Guanajuato scene in their perception through approximately 7,000 pictures. After getting the image corpus, which was based on the selection of 1,200 images, an online crowdsourcing study was performed via Mechanical Turk to get the annotations of the impressions of these images so that a consolidate file was created with the aggregated values in a seven-point Likert scale for each of the 12 urban awareness labels (or simply, labels) that are being analyzed: Accessible, Dangerous, Dirty, Happy, Interesting, Pleasant, Picturesque, Polluted, Preserved, Pretty, Quiet and Wealthy.

By performing a correlation analysis in a pair-wise way between the aggregated values of all the 12 labels, three groups of labels

were found [16]: negative labels (those that generate negative impressions: Dangerous, Dirty and Polluted), a neutral label (a label that generates either a positive or a negative impression, depending on the context: Quiet) and positive labels (those labels that generate a positive impression: Pretty, Preserved, Accessible, Interesting, Picturesque, Wealthy, Quiet, Polluted, Pleasant and Happy). From the correlation matrix shown in figure 1, one can appreciate that the negative labels are inversely correlated with the positive labels, and the neutral label presents a slight bias towards the positive labels.

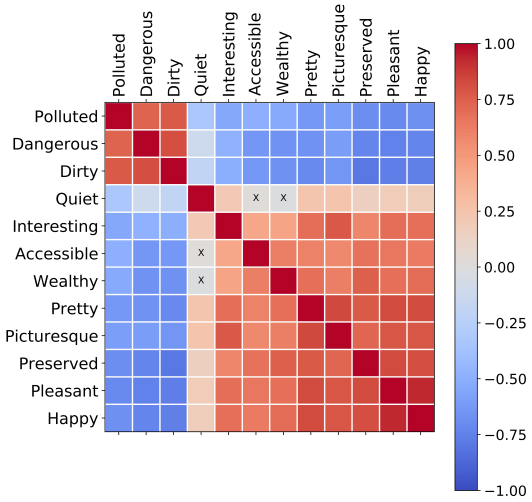


Figure 1: Correlation matrix of the 12 labels. Please note that the "X" indicates that the cell is not statistically significant at p<0.05.

A PCA analysis was also done in order to complement the correlation analysis: It was demonstrated that the two first principal components of the 12 labels explain approximately 77% of the variance [16] and by looking at the figure 2, one can notice that both the negative and the positive labels mainly rely on the first principal component, while the neutral label relies mostly on the second principal component.

In order to analyze the differences between two different groups of observers (one composed of local people and the other of non-local people), another study was done [17]. They got a new image corpus based on 99 images taken in Guanajuato city and conducted 2 crowdsourcing studies: one with the help of volunteer students from a local high school in Guanajuato city (CECYTE campus Guanajuato) and the other was done via Amazon's Mechanical Turk (AMT) with US-based workers. At the end, they analyzed the two groups based on the annotations made by both groups and 6 urban awareness labels: Accessible, Dangerous, Dirty, Interesting, Preserved and Pretty. These results will be shown along this report when comparing to our results.

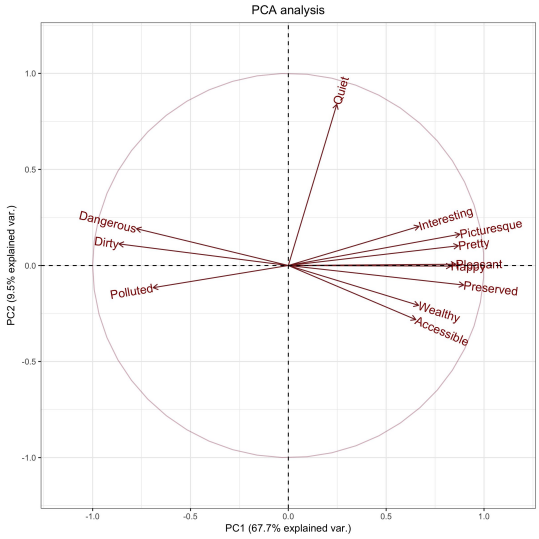


Figure 2: Plot of the loadings for the two first principal components of the 12 labels.

3 IMAGE CORPUS

We base our study on the image corpus generated in the Urban Data Challenge (UDC) during the SenseCityVity project [1] [16] [15] [14] where 7,000 images were approximately taken in the state of Guanajuato, which is located in central Mexico, and whose population reaches almost 6 million inhabitants, most of them urban (around 70%). From these 7,000 images, 1,200 were selected in groups of 400 images for each of the three cities that were chosen to collect the images: (1) Guanajuato (approx. 170,000 inhabitants), which is a UNESCO world heritage site and whose economic activity relies mostly on tourism, fact that makes people from there to care about the image of their city, trying to keep it clean and safe from crime; (2) Leon (approx. 1.6 million inhabitants, considered as one of the most populated cities in Mexico), is an industrial and business city with several factories specialized mainly on leather and footwear products; and (3) Silao (approx. 147,000 inhabitants), which is a very industrial city, with industrial parks and several automotive spare parts companies because of the presence of one of the major car assembly plants, General Motors.

The high school students tried to captured in pictures the characteristics of each of the three cities, documenting different neighborhoods and iconic places [15]. Examples of the the pictures taken can be seen in figure 3.

4 DATASETS

In order to make a computational analysis of the factors that are involved when trying to explain an scene impression and to compare impressions between local and non-local people based on our image corpus, we use six different datasets: four contain features (objects) and the other two datasets contain impressions of the urban awareness labels based on different groups of raters.

Figure 3: Samples from the image corpus. Please note that the images were selected randomly and each row represents each city: Guanajuato at the top, Silao in the middle and Leon at the bottom.

4.1 DilatedNet Semantic Segmentation: 150 features

This dataset consists of 150 visual features resulted from an extraction at a level object by applying Deep Learning techniques and using a pre-trained segmantic segmentation network called DilatedNet [24] on the image corpus. The features in this dataset include both indoor and outdoor generic elements such as wall, building, sky, floor, tree, etc. and the images are described with the actual proportion of each of the features.

4.2 GoogLeNet places205: 205 features

This dataset is based on an extraction of 205 visual features of each of the pictures from the image corpus by applying Deep Learning techniques and using the final layer with class probabilities of a pre-trained CNN based on the GoogLeNet architecture trained on the places205 database [27] [15]. The images were re-sized to 256x256 pixels and subjected to mean image subtraction [15]. The features included in this dataset are both indoor and outdoor objects like in the semantic segmentation dataset. However, we have more objects that can help us to describe an scene impression and that are more related to our image corpus, for example, alley, basilica, corridor, church, residential neighborhood, etc.

4.3 GoogLeNet places365: 365 features

places365 is an update of the places205 database and consists of around 1.8 million of training pictures. In this case, it uses the same CNN architecture (GoogLeNet), but trained on the new places365 database [26]. The images were re-sized to 256x256 pixels and subjected to mean image subtraction [15]. As we can tell from the name of the database, this dataset contains 365 visual features that were extracted from the image corpus by using the final layer with class probabilities. The features include almost the 205 from the places205 database and approximately 160 more like bazaar, downtown, flea market, house, industrial area, junkyard, park, promenade, etc. We use places365 to analyze the behavior of the models by using a dataset with more objects.

4.4 MTurk: Urban awareness labels

The annotations of the 12 labels, that represent urban awareness, were obtained via an online crowdsourcing study on Mechanical Turk as explained in section 2 and [16]. For each of the pictures from the image corpus, each label has a value according to a seven-point Likert scale, where the values go from 1 (strongly disagree) to 7 (strongly agree). This dataset will be used as the dependent variable when making the predictions and contains the aggregated values for each label.

4.5 CECYTE: Urban Awareness labels

This dataset consists of the aggregated values of the annotations of 6 urban awareness labels (in alphabetical order: Dangerous, Dirty, Interesting, Pleasant, Polluted and Pretty) that we have been using and was obtained via a crowdsourcing study with the help of students from the CECYTE campus Guanajuato in Mexico, which is a technical high school located in the urban scene to analyze. This dataset contains values on the same seven-point Likert scale and unlike the one obtained by Mechanical Turk in 4.4, this is a new dataset that will help us to compare impressions between locals and non-locals. We will discuss this dataset in details in the following sections.

4.6 CECYTE: Semantic Descriptors

The annotations of defined semantic descriptors were obtained in the same crowdsourcing study as the urban awareness labels with the students from the CECYTE campus Guanajuato. This dataset, which will be discussed in the following sections, contains the aggregated values of 10 semantic descriptors based on clusters defined with the help of the Block Environmental Inventory [13].

5 METHODOLOGY

5.1 Manual Annotation

We identified in section 4, that the datasets we use contain generic objects. However, trying to characterized an urban awareness label with such datasets could be difficult, even though we will see in section 7 some objects, with a relative good correlation, that can contribute either in a positive or in a negative way to the label. As an ilustration, a good explanation of how people could perceive "danger" is found in [2], in which the authors explain the concept of "Perceived Personal Danger" as the general fear of people to become a victim. For example, if a person walks alone during the night, he would look at all the elements that can avoid himself from escaping in case of danger as well as the darkest places where someone else could be hiding. To have a better understanding of what some of the objects that can characterized an urban awareness label are, we make a manual annotation analysis, image by image

on different sets, for two of the labels: Dirty and Dangerous. With that, we also aim to create a visual vocabulary to describes those labels. We consider all the physical elements that can contribute to the label and also all the elements that produce the opposite effect (i.e. elements that describe non-Dirty and non-Dangerous labels) paying attention to be more specific with the description of the features. Please note that this analysis was made by the author of this document who is Mexican-born and was raised in Mexico and therefore, understands more the context of the pictures. Finally, in order to do this analysis, we get 50 random images from each of the following values ranges of the Likert scale: 1-3, 3-5 and 5-7.

5.2 Crowdsourcing impressions

5.2.1 Mechanical Turk (MTurk). This was a prior work made by Darshan et al. in [16] and [15]. An online crowdsourcing study was conducted via Amazon's Mechanical Turk (AMT). In such study, US-based workers with a very high approval rate were chosen to complete the correspondent HITs (Human Intelligence Task). Every HIT consisted in observing every picture and rate the personal impression based on each of the 12 urban awareness labels. All the annotations were in a seven-point Likert scale (where the values went from 1: Strongly disagree to 7: Strongly agree). The workers were not told about the source, location of the images or any other information about the cities in the study and this was done in order to avoid any potential bias towards the answers. At the end, 10 annotations, were gathered per image and per label, making a total of 144,000 individual judgments. Every worker was reimbursed 0.10 USD per HIT.

5.2.2 CECYTE. Based on the same image corpus, another crowdsourcing study was held with the help of Dr. Salvador. Images were annotated by volunteer students from CECYTE campus Guanajuato, aged 16-18 years old and living in Guanajuato city or suburban areas and therefore, they knew that the taken pictures were from Guanajuato. A website was designed to collect the impressions from the students like in [17]. Unlike the online crowdsourcing study explained above and based on previous work [14], we only took into account 6 (out of the 12) urban awareness labels that we have been studying to characterize an urban scene: Dangerous, Dirty, Interesting, Pleasant, Polluted and Pretty. Besides the labels, we also included 10 semantic descriptors based on the labels Dirty and Dangerous that resulted from the manual annotation task described in sections 5.1 and 8.1:

- S1: Vandalism
- S2: Lack of maintenance
- S3: Unkempt houses/buildings
- S4: Littering
- S5: Bad urban planning
- S6: Lack of security elements
- S7: Lack of outdoor lighting
- S8: Neglected vegetation
- S9: Deteriorated roads
- S10: Vacant lots

At the end, 5 annotations were gathered per image, per label and per semantic descriptor, making a total of 96,000 individual judgments. The annotations of the semantic descriptors were simply based on a binary output: *1* (resp. *-1*) for "the semantic descriptor is (resp. *not*) perceived within the picture" and for the urban awareness labels, the annotations are based on the same seven-point Likert scale as with MTurk. Finally, the students didn't receive any financial incentive for the annotations, but they were given a small present for their help and the satisfaction that they were contributing to the study.

5.3 Regression task

We follow a same approach as developed in [15]. We run a Random Forest algorithm to perform a regression task in which the dependent variables that we want to predict are the urban awareness labels and the independent variables are the datasets with features and semantic descriptors that were explained in the previous section. We use the cross-validation technique with k=10 folds with one repetition to avoid overfitting and get the mean of these 10 runs. In order to analyze the performance of the Random Forest, we get two metrics: the Coefficient of Determination (R^2) and the Root-Mean-Square Error (RSME). Finally, for the baseline model we take the average of the annotated values of each label as the predicted value.

5.4 Classification task

We follow a similar approach to what we do with the regression task, we run a Random Forest algorithm to perform a classification task. In this case, we only consider the labels dataset obtained by Mechanical Turk (explained in 4.4), where all the values of the labels are binarized. However, if we choose to have a threshold of 4.0 (since it is the middle point of the scale we are using), where the class 1 will say that the label corresponds to the picture and the class 0 will say that the label does not, we will have some unbalanced classes as we can see in table 1.

To attack this and because the distribution of the classes is very disproportional among some of the labels, we have to choose different thresholds for the latter as shown in table 2. Please note that due to the new thresholds, we can have "more" balanced classes (between 40% and 60% for both 1's and 0's for each label). However, there is no threshold that make the classes of the label Pretty balanced.

Label	1's	0's	Threshold
Dangerous	24%	76%	4.0
Dirty	31%	69%	4.0
Polluted	21%	79%	4.0
Interesting	54%	46%	4.0
Accessible	87%	13%	4.0
Happy	51%	49%	4.0
Quiet	44%	56%	4.0
Pretty	26%	74%	4.0
Picturesque	24%	76%	4.0
Pleasant	57%	43%	4.0
Preserved	57%	43%	4.0
Wealthy	13%	87%	4.0

Table 1: Unbalanced classes.

Label	1's	0's	Threshold
Dangerous	59%	41%	3.0
Dirty	42%	58%	3.5
Polluted	55%	45%	3.0
Interesting	54%	46%	4.0
Accessible	51%	49%	5.0
Happy	51%	49%	4.0
Quiet	56%	44%	3.5
Pretty	62%	38%	3.0
Picturesque	60%	40%	3.0
Pleasant	57%	43%	4.0
Preserved	57%	43%	4.0
Wealthy	44%	56%	3.0

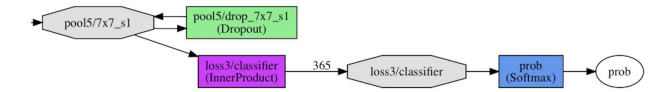
Table 2: New thresholds to get balanced classes for each label.

We consider the binarization of the values of the labels (taking into account the thresholds in table 2) as the dependent variables and the three datasets of the extracted features as the independent variables. Even though we can get more balanced classes with the new thresholds, this is not enough to have a proportion of 50% for each class, for this we apply undersampling techniques to try to consider all the samples for both classes. In order to avoid overfitting, we apply cross validation with k=10 folds and do 4 repetitions so at the end we run the algorithm 4*10 times, and we get the mean of these runs. For the evaluation of the classification Random Forest we get two metrics: the accuracy of the model and the Cohen's Kappa value. Finally, for the baseline model, we do exactly the same as with the regression task: we take the average of the annotated values of each label as the predicted value, and since we classify between two classes and we try to be as fair as possible with both classes (balanced classes), we expect a value of 0.5 for this metric.

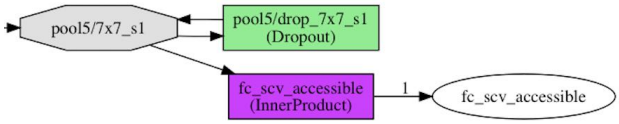
5.5 Fine-tuning task

We attempt to do a fine-tuning task on three of the most popular CNN architectures pre-trained on the places365 [26] database: AlexNet [10], GoogLeNet [22] and VGG16 [19]. We choose to use places365 over places205, since it offers more outdoor objects that can be extracted from the images and that can explain better our image corpus.

In order to do this task, we follow a general approach for fine-tuning based on the examples provided by the deep learning framework Caffe [8]: We remove the last layer of every CNN (figure 4a) and we add a fully connected layer with a single output (figure 4b), we reduce the learning rate (lr) and boost the learning rate multiplier of the layer that we added as well as decrease the step size. We make this because we want the new layer to learn more and much faster than the other layers that have been already trained with the places365 database. For the training deploy, we add an euclidean loss layer so that we can do a regression task and predict continuous values based on the seven-point Likert scale. Finally we create a data layer at the beginning of the CNN, so that we can feed the network with our data. All images have been re-sized to 256x256 pixels, stored in lmdb files (one for training data and another for validation data) and subjected to mean image subtraction [15].



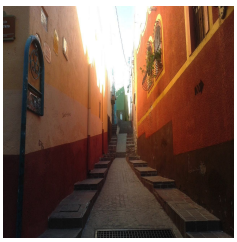
(a) Original deploy



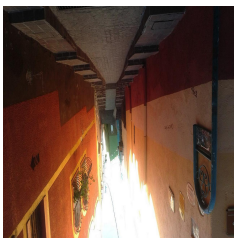
(b) Modified deploy

Figure 4: An example of modification of the CNN architecture deploy for the fine-tuning task: GoogLeNet places365 with label Accessible.

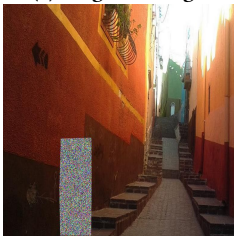
Since the image corpus is relatively small (1,200 images) for the fine-tuning task, and because we need to split our data into a train dataset (1,020 images, 85% of the data) and a validation dataset (180 images, 15% of the data), we perform data augmentation [12] [3] in our train dataset (figure 5) to artificially create more images from our original images and with this to have a bigger image corpus that will help us to have better results and avoid overfitting. At the end, we augmented the train dataset to 5,020 (1,020 real-world + 4,000 augmented images). Some of the modifications in the augmented train dataset include: rotations, flips, random distorsions, random erasing, cropping, etc.



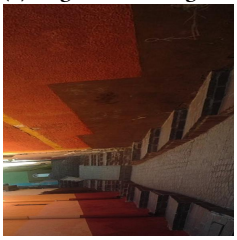
(a) Original image



(b) Augmented image #1



(c) Augmented image #2



(d) Augmented image #3

Figure 5: Example of data augmentation on one of the images of the image corpus.

Finally, the fine-tuning task is performed 12 times multiplied by 3, one time per each urban awareness label by each of the three CNN architectures, and the maximum number of iterations per fine-tuning task was 76 epochs since we did not see any improvement of the results. The deep learning framework that we decided to use is Caffe [8] due to the fact that the pre-trained models that we need are available only for that framework.

6 FEATURE ANALYSIS

6.1 DilatedNet Semantic Segmentation dataset

As explained in section 4, the DilatedNet Semantic Segmentation dataset with the 150 features contains the annotations corresponding to the proportion of pixels of each object in each image. These 150 features include both indoor and outdoor objects. However, when analyzing these objects, one can see that the outdoor objects are too general, e.g. building, sidewalk, road, sky, tree, wall, etc. and not specific like graffiti on walls, trash on streets, sidewalks or buildings in poor condition, etc. Then, it is more difficult to say whether an urban awareness label is seen within an scene of an image.

In figure 6, we can see that the 10 objects with more presence are still too general and they represent around the 92% of the pixels for all the images. Despite this, we note that even with the general outdoor objects, we can explain some labels within the images. For this, we analyze 2 different images (one with a high value in the Dangerous label and one with a low value in the same label) and their corresponding segmentation.

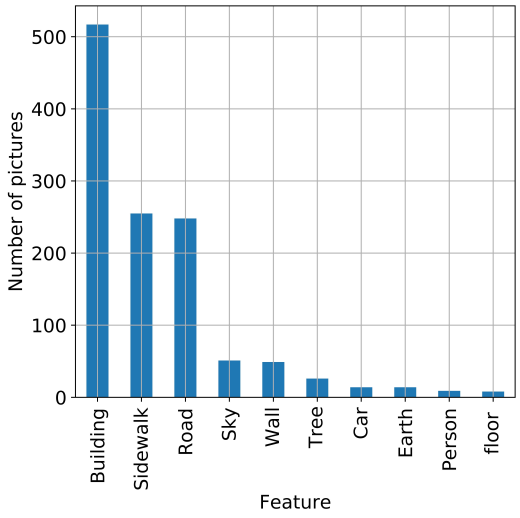
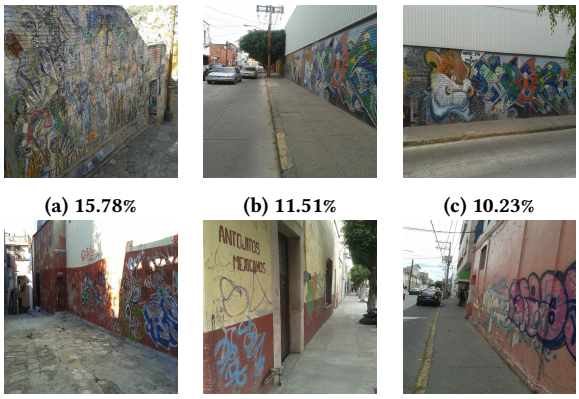


Figure 6: Top-10 objects from DilatedNet Semantic Segmentation dataset with more presence in the image corpus.

One interesting thing with this dataset is that, even though we mentioned above that "there are no specific objects like graffiti on walls", we find that the object "painting" might classify the graffiti as they share similar characteristics (use of colors, lay on a surface, etc.). We can appreciate this by looking at some of the pictures from the image corpus (figure 7) that contain "painting" with a high proportion of the total pixels. This is important, since the object "graffiti" is seen in a lot of pictures and in general, it is part of the scene in Mexico. This might be interesting when trying to infer Dangerous label, since people may feel unsafe because of the presence of graffiti [25].



(a) 15.78% **(b) 11.51%** **(c) 10.23%**

(d) 7.95% **(e) 6.34%** **(f) 6.08%**

Figure 7: DilatedNet Semantic Segmentation: Pictures from the image corpus with the highest proportion of pixels in the feature "painting" and their corresponding values.

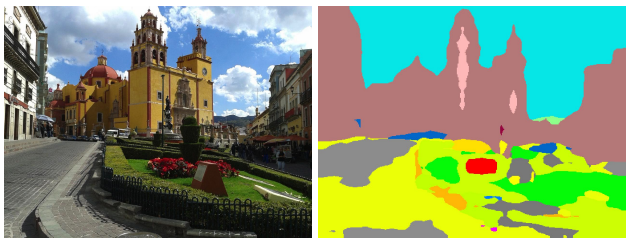
(a) Original image **(b) Segmented image**

Figure 8: Labels: Dangerous: 5.5, Dirty: 5.0, Pretty: 2.0, Preserved: 3.0, Accessible: 3.5, Interesting: 4.0, Picturesque: 2.0, Wealthy: 1.0, Quiet: 4.0, Polluted: 3.5, Pleasant: 3.0 and Happy: 3.0.

By looking at the values of the labels of figure 8, we notice that the image is both Dangerous (5.5) and Dirty (5.0). The original image (figure 8a) shows different objects like graffiti, undergrowth, worn paint or small alley that could characterized the scene as dangerous or dirty. However, as explained above, these elements are not part of the 150 features of the segmentation (even though, we saw that "painting" can be interpreted as "graffiti"). By analyzing the objects with respect to the proportion of pixels (given by the dataset) in the segmented image (figure 8b), we find objects such as wall (27%), building (38%), stairs (8%), sky (4%), floor (8%), sidewalk

(4%), mountain (2%), path (1%), etc. These objects make little sense for the Dirty label, but can make sense for the Dangerous label, for example with the object "sky", since the less "sky" one sees in an scene, the more closed a place is and, therefore the place could be more dangerous, because in case of danger a person would not know where to go (this is explained by Blobaum and Hunecke in [2]). Another example could be the object "stairs", in which one can think is "dangerous" in some sense, since one can fall from it.

We now analyze the counterpart of the image in figure 8a, i.e. a picture with the opposite value in the Dangerous label. For this purpose, we choose the image in figure 9a. By looking at the values and the original image (figure 9a), one can see that it is a nice scene, clean, not dangerous, very pleasant, which more representative elements are flowers, plants, a church, people, clean and well-preserved streets and a lot of sky which means that it is an open space. In the semantic segmentation (figure 9b), on the other hand, we find more general objects such as wall (2%), building (34%), sky (23%), road (6%), grass (6%), sidewalk (12%), plant (9%), flower (0.6%), bench (2%), tower (1%), etc. Since most of the pictures from the image corpus, regardless their label values, contain very general objects like building, wall, sidewalk,road, etc., some objects like grass, road, sky, bench, or flower, even with low percentage of proportion of pixels, are enough to describe that, for example, the image is not dangerous or dirty but picturesque, clean, pleasant, etc.



(a) Original image **(b) Segmented image**

Figure 9: Labels: Dangerous: 1.0, Dirty: 1.0, Pretty: 6.5, Preserved: 6.5, Accessible: 7.0, Interesting: 6.0, Picturesque: 7.0, Wealthy: 5.0, Quiet: 4.0, Polluted: 1.0, Pleasant: 6.5 and Happy: 6.0.

6.2 GoogLeNet places205 dataset

As explained before, we extract the 205 features from the image corpus by using the final layer of the GoogLeNet CNN trained on the places205 database with the class probabilities. The features (objects) we find in this dataset are a bit more specific to our image corpus than the semantic segmentation dataset, then we expect better results during the regression and classification tasks when comparing to the semantic segmentation ones.

Figure 10 shows the top-10 recognized objects in our image corpus [15]. The fact that the feature "alley" is the most dominant one makes totally sense, since the cities from the state of Guanajuato are well-known by their alleys. "medina", "plaza" and "market" were also expected because of the great amount of these elements in the Mexican cities. The rest of the top-10 features are good to explain

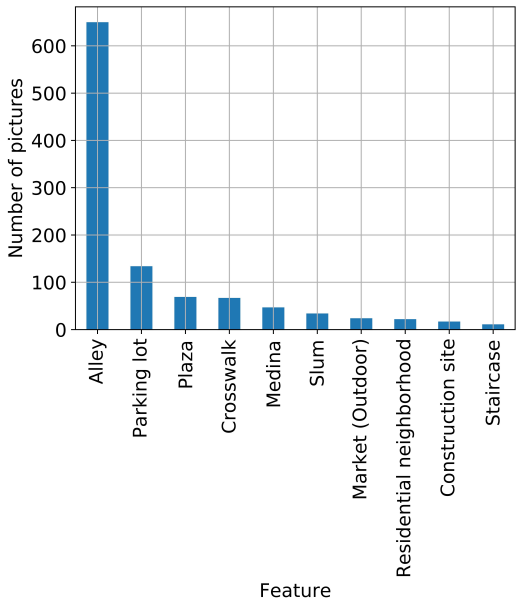


Figure 10: Top-10 objects from GoogLeNet places205 dataset with more presence in the image corpus.

several urban awareness labels, like Dirty with "slum", or Dangerous with "alley" or "construction site".

By analyzing one of the pictures that we explained with the semantic segmentation dataset (figure 9a), we find the following distribution of objects according to places205: basilica (19%), cathedral (16%), plaza (13%), church (10%), castle (10%), palace (7%), tower (5%), monastery (5%), abbey (4%), fountain (1%), etc. We see that more than 76% of the objects are church-oriented (including "palace" and "castle") and it makes totally sense respect to the actual picture, and like the semantic segmentation, we have, in less percentage, some features that can help us describing more the picture, courtyard (1%), apartment building (0.36%), formal garden (0.27%), sky (0.1%), etc. At least, from the features we just mentioned, we can imagine a pretty, interesting, accessible, pleasant and picturesque scene.

6.3 GoogLeNet places365 dataset

This dataset was obtained by using the same techniques as with the GoogLeNet places205 dataset, but with the places365 database instead. We find around 160 more features than its predecessor database (places205). Figure 11 shows the top-10 recognized objects in our image corpus. We see that most of the features are still the same as in the top-10 features of the places205 dataset ("Bazaar" replacing "Market", instead of "Crosswalk" we have "Street" and instead of "Construction site" we have "Industrial area"), and that two new features are included: (1) "gas station", which was unexpected since in the image corpus we do not see pictures with gas stations and (2) "Loading dock", which makes totally sense since a loading dock seems to be a metallic curtain door that many stores in Mexico have. From figure 12, we can appreciate that we have a misclassification

for the feature "gas station" since we only see avenues or streets and cars, object that maybe led to the misclassification. On the other hand, for the case of the "Loading dock" we can see that the mistake in the classification was due to the shape of some metallic doors that most of the houses in Mexico have as protection because of the crime as well as the already mentioned metallic curtain that some shops have. One difference with the most dominant feature, is that the number of pictures with "street" is half the number of pictures with "alley" (places205), and that now "alley" is the second most dominant feature (places365).

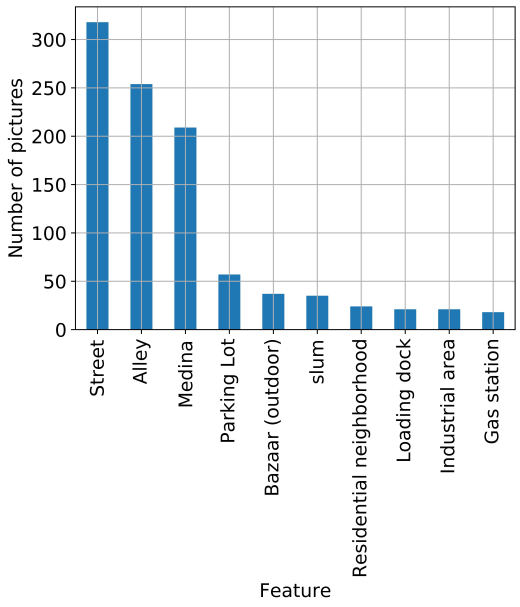


Figure 11: Top-10 objects from GoogLeNet places365 dataset with more presence in the image corpus.

By doing the same analysis on the same pictures explained with the places205 dataset (figure 9a), we find the following distribution of objects: palace (27%), castle (19%), church (13%), plaza (8%), tower (4%), courtyard (3%), formal garden (3%), mansion (3%), campus (3%), building facade (2%), topiary garden (2%), fountain (1%), etc. We can tell that, unlike the previous analysis, we have 59% of objects regarding religion (including "palace" and "castle" that could classify these objects), and the rest of the picture presents other features that could explain more urban awareness labels: besides the ones we mentioned with the places205 database, now we could also explain non-Dangerous non-Dirty and Non-Polluted, since we know that the area has vegetation (8% compared to the 1.3% found with places205), a plaza and a fountain. The problem with this database is that, perhaps we are facing more misclassified elements like "campus" or "mansion" (that could be classified due to the shape of the buildings) or the already mentioned "palace" and "castle". In the end, we expect to get the best results in the classification and regression task when comparing to the ones obtained with the other two datasets.

(a) 67.58% **(b) 40.11%** **(c) 38.20%**

(d) 84.33% **(e) 73.74%** **(f) 65.27%**

Figure 12: GoogLeNet places365: Pictures from the image corpus with the highest class probability in the features "Gas station" (top row) and "Loading dock" (bottom row) and their corresponding values.

6.4 t-SNE Analysis

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique designed for dimensionality reduction that can be used to visualize high-dimensional datasets [23]. We apply t-SNE on the three features datasets in order to have a visualization in two dimensions. The results show that the 1,200 images are distributed along the plan, and that they are grouped in clusters with similar features. Please note that due to the limitations of the length of this document we do not include the complete plot.

DilatedNet Semantic Segmentation: We already found that this dataset is the one with the most general outdoor objects. The t-SNE plot shows that the images are very dispersed and it is a bit difficult to find clusters based on the separation between each image. However, we can still see groups that are organized by similar values in their labels, like we observe in the figure 23 in the appendix A, which is a section that we crop from the whole plot. As expected, we have similar results as in the correlation analysis of the labels made in [16] and section 4.4, low values for negative labels, dispersed values for the neutral label and high values for positive labels. If we take a look at the images (figure 13) that correspond to the exact section of the t-SNE plot showed in 23 in the appendix A, we note that these share several objects in common such as park elements (benches, trees, etc.), open and clean spaces (plazas, avenues, etc.) and other elements that contribute to positive impressions of scenes (churches, a clear sky, etc.), some of these objects are equally shown as generators of aesthetic impressions in [25].

GoogLeNet places205: We find similar results to the t-SNE analysis on the semantic segmentation dataset. However, we can note more defined clusters in the plot based on the distance between elements. We crop a section of the t-SNE plot (figure 24 in the appendix A) and we observe again that it is a group of images with positive impressions (high values in positive labels and low values in negative labels), we corroborate this by looking at the figure 14, in which we can see park-oriented objects like trees, benches,

Figure 13: DilatedNet Semantic Segmentation: A section of the visualization of the pictures by applying t-SNE.

etc. and religion-oriented objects like churches. The label Quiet, in which in section 4.4 we spotted that it has a little bias towards the positive labels, this time has a bias towards the negative labels. The latter can be explained considering that we find active elements [6] like the restaurants or people that affect the quietness.

Figure 14: GoogLeNet places205: A section of the visualization of the pictures by applying t-SNE.

GoogLeNet places365: Like the other two cases above, we appreciate some clusters based on objects that the pictures share. Following the same procedure as before, we crop a section of the complete plot (figure 15) and we observe different pictures regarding parks and plazas and actually we can appreciate some of the same pictures that we got with the t-SNE analysis of the GoogLeNet places205 dataset (figure 24). By looking at the values of the labels in figure 25 in the appendix A, one can note the same pattern as with the t-SNE analysis done before: the negative labels have low values and positive labels (including label Quiet) have high values. This is not surprising since we can explain this with an example: The pictures show a prominence of trees and open spaces , fact that increases the pleasant factor [25] and decreases the dangerous factor [7] or the dirtiness, or the religion-oriented objects that make the scene interesting.

Figure 15: GoogLeNet places365: A section of the visualization of the pictures by applying t-SNE.

Finally, with the results we get with t-SNE visualization analysis on the features datasets, we can say that it is a very good approach to explain similar pictures of the image corpus based on the visual objects that they share.

7 CONNECTION BETWEEN OBJECTS AND URBAN AWARENESS LABELS

In order to see the connection between the different objects presented in the datasets and the 12 urban awareness labels (based on the MTurk annotations explained in sections 4.4 and 5.2.1), we make a pair-wise Spearman correlation analysis between the objects and the labels. Results are shown in tables 9 and 10 in the appendix B. Please note that the thresholds for considering that an object is correlated to a label are different between datasets (an absolute value of 0.12 for semantic segmentation dataset and an absolute value of 0.3 for both places205 and places365 datasets) and due to limitation of space, we only report the 5 objects with highest values.

Most of the positevely correlated objects with the positive labels are religion-oriented objects such as cathedral, basilica, church, etc. and open space objects like plaza which was kind of expected, since

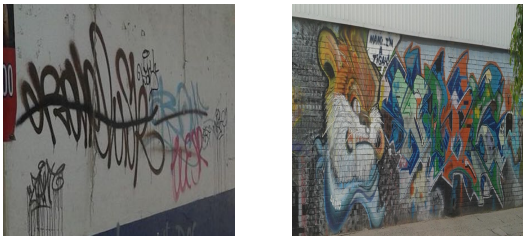
Mexico as one of the most catholic countries, has a lot of churches in their cities, these are normally located in plazas and most of them are considered as a tourist attraction that arouses positive feelings and emotions in people, not only because of the religion but also because of the different elements that they keep such as architecture (building texture) and location (normally in open spaces like plazas). With the naked eye, we can note that some urban awareness labels are correlated with their characteristic objects, such as Dirty and Polluted labels which share the positively correlated objects dirt track and slum. Dirty, for its part, correlates positively with painting (which can be considered as graffiti as we described in section 6.1) as some people consider that they increase the visual pollution. On the other hand, Polluted shows a positive correlation with junkyard, garbage dump, slum and landfill; Interesting label with positively correlated objects like the religion-related ones, or negatively correlated objects like garbage dump or loading dock; or Preserved label in which we find objects such as religion-oriented, street elements (sidewalk), or park-oriented elements like plaza or tree. Label Dangerous shows a negative correlation with nature-oriented objects (such as tree or botanical garden) as well as open spaces (such as plaza) as explained in [7], and it is quite surprisingly that an object like street light does not appear. Painting again, is a positively correlated object of Dangerous, as explained in section 6.1 and in [25]. Another finding is with the label Pleasant, which positively correlated objects are shown in [25] and are related to open spaces (plazas), prominence of trees (trees) as well as building texture (religion-oriented objects). From the label correlation analysis made in section 2 and [16], we know that Pleasant, Pretty and Happy labels are very correlated and we note this, because they also show same positively correlated objects. We observe that both Picturesque and Pretty labels share positively correlated objects such as the religion-oriented or park-oriented like courtyard and negatively correlated objects such as dirt track or slum. For an accessible impression of a place we need open spaces like we find in the results such as plaza or sidewalk and people, and the latter shows that a place could be accessible since there are people walking. On the other hand, it is obvious to think that the smaller and the more closed the space is, the less accessible it is, and we find this with the negatively correlated objects such as alley, wall or door (which can mean that the access is blocked by these elements). One expected result is the negatively correlated objects regarding the label Quiet, since most of them allude to objects that are in relation with activities and active elements [6] (vehicles or supermarkets) which can be translated as movement which opposes to a quiet scene. Although their positively correlated objects do not offer information at all since they could be a misclassification error (it shows objects like jail cell, which can be interpreted as the window protection in houses; corridor, catacomb, etc.). From the Wealthy label, we observe some objects that in their general description offer no meaning for the label, one may think that a cathedral, a hotel, could contribute a positive impression for the label. Perhaps beer garden or plaza can contribute to the label in a positive way as we spotted in the correlation results since they represent financial activity (normally in plazas one can find restaurants and different kind of shops). However, nowadays these are elements that are presented in any kind of neighborhood, middle-class, wealthy or poor.

It is important to mention that even if we find similar objects in both places205 and places365 datasets, with the latter we can add some of objects to the labels, e.g. Polluted, which we have garbage dump and slum (places205) as positively correlated objects and we can complement that with junkyard and landfill (places365). However, even with more objects (places365) it is still difficult to have a good and consolidated database of objects to explain every urban awareness label. In the end, if we wanted to generalize, it would be difficult to find or create a good database with the different objects that can explain every urban awareness labels, since every city is different and so is the mentality of the people.

8 RESULTS AND DISCUSSION

8.1 Manual annotation

The results show that there is a tendency towards the objects that characterized the labels (tables 11 and 12 in the appendix C), same objects are repeated with a frequency depending on the range value (labels: the higher value, the higher frequency; Non labels: the less value, the more frequently) and as expected, some of these objects are contradicted by those that correspond to the *Non-labels* (i.e. Non-Dirty or Non-Dangerous labels, tables 13 and 14 in the appendix C) such as artificial lighting and no artificial lighting in the Dangerous and non-Dangerous labels. We also observe that some of the objects like graffiti or public road in bad condition are repeated in both Dangerous and Dirty labels. This can be explained by the correlation analysis of the labels explained in section 4.4. Tag graffiti (figure 16a is an element that can describe the most dangerous scene (highest values in the scale) since it could refer to different gangs in the zone; and surprisingly, litter (trash on street) is not the dirtiest element but graffiti (figure 16b).



(a) Tag graffiti **(b) Graffiti**

Figure 16: Examples of a tag graffiti and a graffiti.

To go deeper in our analysis of manual annotation, we create different clusters that group the different objects for the labels. We based this clustering task on the Block Environmental Inventory (BEI) explained by Perkins et al. in [13], in which they work with three different types of crime and fear related physical cues: (1) Incivilities, vandalism and dilapitated houses; (2) Signs of territorial functioning and (3) Defensible space features. However, in our manual annotation, we only observe elements that correspond to the type (1) and (3). We try to expand the latter by adding more elements that can affect an urban espace in more aspects, since the analysis made in [13] covers only neighborhoods and not other

places of a city/town. Finally, the analysis shows 12 clusters for the dangerous label and 9 clusters for the dirty label. We see that even though the latter analysis refers more to a dangerous scene, some of the clusters from the dirty label fit in the 2 types of crime and fear that we are working with. Therefore, we try to group the clusters in the latter:

Dangerous label:

- Incivilities, vandalism and dilapidated houses:
  - Vandalism
  - Lack of exterior maintenance
  - Unkempt house buildings
  - Vacant lots
  - littering
  - Negative elements regarding people
  - Suspicious elements
  - Bad planning
- Defensible space features:
  - No security house/building elements
  - No outdoor lighting
  - Place for concealment
  - Blocking elements

Dirty label:

- Incivilities, vandalism and dilapidated houses:
  - Vandalism
  - Littering
  - Unkempt houses/buildings
  - Neglected vegetation
  - Negative elements regarding the people
  - Deteriorated road signs
  - Lack of exterior maintenance
  - Remains
- Defensible space features:
  - Blocking elements

Virtually, the labels could be explained by the first type of crime and fear for both labels as we can see in the distribution of the objects corresponding to each clusters in figures 26a and 26c in the appendix C. The three elements for the Dangerous label with more frequency are: (1) Lack of exterior maintenance, (2) vandalism and (3) no outdoor lighting; and for Dirty label: (1) Unkempt houses buildings, (2) Littering and (3) Vandalism. The distribution for the Non-Dangerous and Non-Dirty clusters can be seen in figures 27a and 27c in the appendix C, and they show that an scene can be clean (non-dirty) and safe (non-dangerous) if they have different elements that relate people and vegetation in different aspects: commercial zones, people walking, non-neglected plants, etc.

Finally, while it is true that some of the features that we found when annotating manually are the same from the features included in the datasets (such as street light, trash can, etc.), we found more specific features (such as damaged wall, worn paint, etc. instead of just wall), for the case of graffiti, we saw in section 6.1 that it can be seen as "painting", although with a manual labeling of the image corpus we could corroborate the latter.

8.2 Annotations Quality

8.2.1 Urban Awareness Labels. In order to measure the annotations quality of the urban awareness labels, we proceed to analyze the interrater reliability by computing the intraclass correlation (ICC) among the ratings [9] on each of the urban awareness labels across all the image corpus. Since each of the images is rated by a set of k annotators randomly selected from a larger population of K annotators [17], we chose to use the *One-Way Random-Effects Model* [9] taking an average of the k rater's ratings, i.e. *ICC(1,k)*.

The annotations quality of the MTurk study was detailed by Darshan et al. in [16]. However, since we are only considering 5 annotations per image for the CECYTE crowdsourcing study and to be fair in the comparison, we decided to compute *ICC(1,k)* by randomly sampling 5 (out of 10) annotations, doing this 10 times and get the mean of these 10 values as well as the standard deviation. Results of the computed *ICC(1,k)* with both *k=5* and *k=10* annotations for both crowdsourcing studies are reported in table 3. Please note that we only reported values for 6 urban awareness labels, since these correspond to the CECYTE study.

Label	MTurk			CECYTE	
	k=5 Mean±SD	*k=10*	*k=10* [17]	*k=5*	*k=10* [17]
Dangerous	0.62±0.01	0.76	0.83	0.34	0.63
Dirty	0.65±0.01	0.78	0.85	0.36	0.68
Interesting	0.54±0.01	0.70	0.63	0.52	0.70
Pleasant	0.67±0.01	0.79	-	0.56	-
Polluted	0.46±0.02	0.64	-	0.28	-
Pretty	0.61±0.01	0.73	0.83	0.58	0.80

Table 3: *ICC(1,k)* scores, including standard deviation and mean values for the Mturk (*k=5*) case. All values are statistically significant at p<0.05. Please note that the dash (-) indicates that the label was not included in the study.

According to the results in table 3, we can appreciate that for *k=5* there are more agreement between the ratings given by the MTurkers. Computed values for the latter indicate moderate reliability (values between 0.5 and 0.75) for all labels except for Polluted which indicates poor reliability (value below 0.5) [9]. Analyzing the CECYTE study *ICC(1,k)* values, we have that 3 labels indicate moderate reliability (Interesting, Polluted and Pretty), being Polluted the label with the less agreement between raters, like happened with the MTurkers. Label Pleasant (resp. Pretty) achieved the highest agreement between the students from CECYTE (i.e. locals) (resp. MTurk, i.e. non-locals). Comparing the two groups (MTurk and CECYTE for *k=10*), we find that 2 labels share almost same values (±0.07): Interesting and Pretty, meaning that the agreement is almost the same between both groups. Finally, we see the same tendency as shown in [17]: Non-local raters (MTurk) tend to agree more for most of the urban awareness label than the local raters (CECYTE).

One surprising thing was that, even though we don't have the same values for the *ICC(1,k)* when comparing our study and the results shown in [17], in part due to the difference in the number of annotators, we find that when ordering the labels with respect to their values (from the lowest one to the highest one), we note

exactly the same order each of the groups. Mturk: Interesting, Dangerous (tie), Pretty (tie) and Dirty; and CECYTE: Dangerous, Dirty, Interesting and Pretty. Showing that local people agree less in their perception of danger, and this could be explained by the fact that the local people have experimented every kind of perceptions in those places and maybe during the HIT of the crowdsourcing study, they have associated or recognized some places within the pictures with their personal risk [2] and experience. For example, a student could be mugged near to a big plaza and therefore, associated that with danger despite the fact that other people can consider the place as safe since nothing has happened to them. However, the latter cannot be experimented by a non-local rater for obvious reasons.

8.2.2 Semantic descriptors. We also attempt to measure the annotations quality of the 10 semantic descriptors that were part of the CECYTE crowdsourcing study explained in sections 5.1, 5.2.2 and 8.1. We proceed to analyze the reliability by computing the *Fleiss' Kappa* value, which is a extended version of the *Cohen's Kappa* measurement for more than 2 raters [4] and it is suitable for nominal or binary data. Additional to this measure, we also compute the *Krippendorff's Alpha* value like other studies [18] have done to complement the results.

Semantic Descriptor	***Fleiss' Kappa***	***Krippendorff's Alpha***
S1: Vandalism	0.16	0.16
S2: Lack of maintenance	0.15	0.15
S3: Unkempt houses/buildings	0.16	0.16
S4: Littering	0.12	0.12
S5: Bad urban planning	0.09	0.09
S6: Lack of security elements	0.05	0.05
S7: Lack of outdoor lighting	0.06	0.06
S8: Neglected vegetation	0.04	0.04
S9: Deteriorated roads	0.16	0.16
S10: Vacant lots	0.05	0.05

Table 4: *Fleiss' Kappa* and *Krippendorff's Alpha* values for the 10 semantic descriptors. All values are statistically significant at p<0.05.

The results are shown in table 4 and are very low (showing a slight agreement with values less than 0.20). However, we can explain this by the fact that these semantic descriptors are still very general and could lead to a different interpretation depending of the rater and his/her context. How does one interpret vandalism? by seeing someone mugging someone? or also by seeing graffiti on walls, like we have been mentioned on this report? Or what are security elements? surveillance camera systems? police men at every corner? or simply protections on the windows and doors. Like the previous examples we can say that trying to explain every semantic descriptor without any other information is quite subjective and could lead to our low results and therefore a high disagreement between raters.

It was surprising that both the results of the *Fleiss' kappa* and *Krippendorff's Alpha* values are practically the same (at the end, we rounded to 2 decimals). However, we found in other studies like [5] that they got similar values in both measurements. Finally, we see that *vandalism*, *unkempt houses/buildings* and *deteriorated roads* are the descriptors with highest values (0.16) and *neglected vegetation* is the descriptor that most of the raters disagree.

8.3 Descriptive Statistics

Given the fact that we have many annotations per image, per urban awareness label and per semantic descriptor, it is necessary to create an aggregated score per urban awareness label and semantic descriptor.

8.3.1 Urban Awareness Labels. We follow the procedure described in [17] to aggregate the values of the CECYTE annotations: The annotations rely on an ordinal scale (that also describes a ranking), and knowing that one of the permissive statistics to get the central tendency of an ordinal scale is the median [21], we proceed to extract the latter on each of the 5 ratings of the urban awareness labels per image. Given this median scores, we compute the mean and the standard deviation per label using all the image corpus. Please refer to [16] and [15] to read about the descriptive statistics of the MTurk annotations.

Label	MTurk		CECYTE	
	Mean±SD [15]	Mean±SD [17]	Mean±SD	Mean±SD [17]
Dangerous	2.98±1.00	3.19±1.20	3.78±1.45	4.43±0.91
Dirty	3.16±1.10	3.25±1.26	3.53±1.53	4.33±1.24
Interesting	3.84±0.90	4.14±1.10	3±1.54	3.55±1.23
Pleasant	3.82±1.00	-	3.08±1.58	-
Polluted	2.89±0.90	-	3.39±1.38	-
Pretty	3.11±1.00	3.25±1.36	3.04±1.65	3.47±1.38

Table 5: Means and standard deviations of the annotation scores for each label and group. Please note that the dash (-) indicates that the label was not included in the study. [17] refers to the study with an (99) image corpus.

The table 5 shows the descriptive statistics per urban awareness labels. All the mean scores for both MTurk [15] and CECYTE studies based on the (1,200) image corpus, show a trend towards a disagreement on all the urban awareness labels (values are below 4 which in the seven-point Likert scale are equivalent to a *Disagree somewhat*). When comparing the two groups based on the (99) image corpus [17], we see that CECYTE students (locals) tend to perceive the images more *dangerous*, *dirty* and *pretty* and the MTurkers (non-locals) tend to perceive the images more *interesting*. We find the same pattern between locals and non-locals, except for the label Pretty in the (1,200) image corpus. Regarding the labels Polluted and Pleasant, we see that the locals perceive the images more *polluted* and the non-locals more *pleasant*.

We made a correlation analysis of the CECYTE median scores (figure 17), and we got similar results as with the Mturk study in [15] (see figure 1 in section 2): There are two groups of label which are highly positively correlated between them and negatively correlated between the other group: Positive labels: Interesting, Pretty and Pleasant; negative labels: Dangerous, Dirty and Polluted. We see that unlike the MTurkers, the positive group is more correlated (values around 0.8) than the negative group (values around 0.55,

except for polluted and dirty which correlation value is 0.66 as they have more to do between themselves than with Dangerous). Regarding the negative correlations among the two groups, we can note that these are not so negative (correlation values between -0.20 and -0.30) when comparing to the other studies [17] and [15].

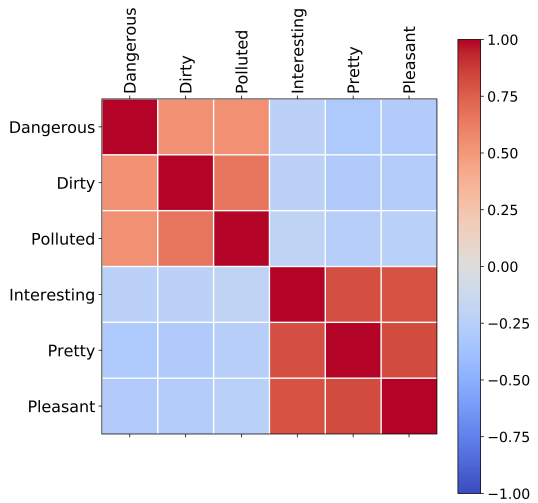


Figure 17: Correlation matrix of the 6 labels of the CECYTE study. All cells are statistically significant at p<0.05.

8.3.2 Semantic descriptors. The annotations are based on a binary output without inherent order and therefore, can rely on a nominal scale. Thus, knowing that one of the permissive statistics to get the central tendency of a nominal scale is the mode [21], we proceed to extract the latter on each of the 5 ratings of the 10 semantic descriptors. Additionally, we also extract the mean of the 5 the ratings per semantic descriptor per image, considering this mean as the probability of the descriptor to appear in the picture. Given the mean and the mode of the semantic descriptors and the median of the urban awareness labels, we compute two co-occurrence matrices: For this, we only consider only those images in which the value of the median of the seven-point Likert scale for each label is 5 (Agree somewhat) or above showing that the label is presented in the perception of the image and two cases: (1) we only consider the mode (only *1*s) to build the co-occurrence matrix and (2) we consider the logical conjunction *OR*, i.e. if at least one rater gave a *1* on the semantic descriptor, then this is counted. This is equivalent of taking those values of 0.2 or greater in the mean.

Figures 18 and 19 show the co-occurrence matrix for both cases of the aggregated values: mode and mean, as described above. We can note that despite the fact that we have more pictures when using the logical conjuction *OR*, we appreciate the same pattern: The semantic descriptors are much more present in the negative group of labels (Dangerous, Dirty and Polluted), fact that makes totally sense, since we based these descriptors on the clusters found in the Block Environmental Inventory (BEI) [13] when manually annotating the Dirty and Dangerous images (as explained in section 8.1). However, we can also appreciate (better in figure 18) that 5

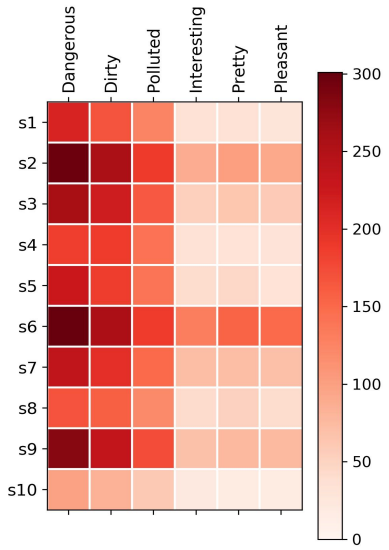


Figure 18: Co-occurrence matrix of the 10 semantic descriptors of the CECYTE study considering the mode of the ratings.

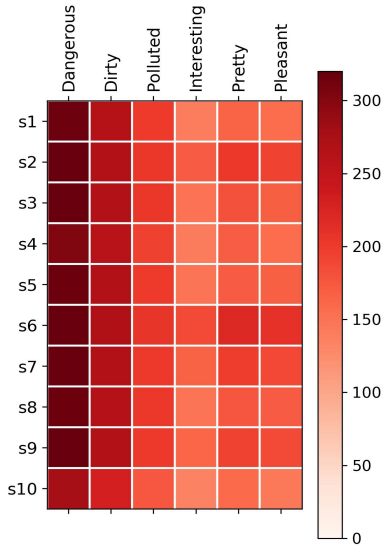


Figure 19: Co-occurrence matrix of the 10 semantic descriptors of the CECYTE study considering the logical conjunction *OR* in the ratings.

semantic descriptors highlight in the negative group of labels and to a lesser extent in the positive group of labels (Interesting, Pleasant and Pretty): *S2: Lack of maintenance*, *S3: Unkempt houses/buildings*, *S6: Lack of security elements*, *S7: Lack of outdoor lighting* and *S9: Deteriorated roads*, being the descriptor *S6* the one with the highest number of occurrences. The latter is surprising due to the fact that

when we were doing the manual annotation task, that was one of the clusters with the less number of pictures (figure 26a in the appendix C). However, subjectivity on how one understands every semantic descriptor could explain why we have that difference of perception, even if the raters come from a same context as the author. We can also say that those (5) descriptors, in some cases, can also affect the perception of the raters in a negative way when evaluating the positive group of labels, and in other cases can mean nothing for the labels. One example is the label Pleasant: in which it is suggested [25] that one of the factors that arouses pleasure is the pavement pattern or the building texture, and being these deteriorated (semantic descriptor *S9: Deteriorated roads*, *S2: Lack of maintenance* and *S3: Unkempt houses/buildings*) as shown in some of the pictures, will make the rater to feel displeasure.

Additional to the co-occurrence matrices, a correlation analysis between the labels and the semantic descriptors was done in order to support our results. The results of the correlation analysis can be seen at figure 20, and show what we mentioned above and what we aimed for when doing the manual annotation for the labels Dangerous and Dirty: We can see two groups of labels again, the 10 semantic descriptors correlate positively with the negative group of labels and negatively with the positive group of labels, being *S4: Littering* the descriptor that correlates, with the highest values, with the labels Dirty (correlation value of 0.43) and Polluted (correlation value of 0.39), and *S3: Unkempt houses/buildings* the one for the label Dangerous (correlation value of 0.38). With respect to the positive group of labels, *S2: Lack of maintenance* is the descriptor with the highest negative correlation values for the 3 labels: Interesting (-0.45), Pretty (-0.49) and Pleasant (-0.47), which makes sense, since if someone sees, for example a building without exterior maintenance, he/she will think that it is not pretty, nor interesting nor pleasant, due to a lack of a nice building texture [25].

8.4 Comparing Impressions between Groups

Now we try to compare the impressions between the two groups of study: Mturkers (non-local people) and students from CECYTE (local people). We saw in the previous section that both groups show a trend towards a disagreement on all the urban awareness labels. Now we want to understand if the mean difference of the labels between the two groups is statistically significant or not. In order to do this, we follow a same approach as in [17]: We perform the Tukey's Honest Significant Difference (HSD) test (table 6) and to complement this study we also show the plot comparing the distribution of the perception ratings between the MTurkers and CECYTE groups. An example of these plots for the label Dangerous can be seen in figure 21, the plots for the rest of the statistically significant labels are in the figure 28 in the appendix D. In order to compare our results, we also include the results of [17] in table 7.

Based on the statistics, we observe that:

- Images were perceived to be more *dangerous*, *dirty* and *polluted* by the CECYTE students (locals) when compared to MTurkers (non-locals). This tendency was seen in [17] for the labels Dangerous and Dirty. When looking at the individual median scores per image, we found that for locals, 73% of the images were rated to be more *dangerous*, 61% were rated to be *dirtier* and 75% were rated to be more *polluted*. In

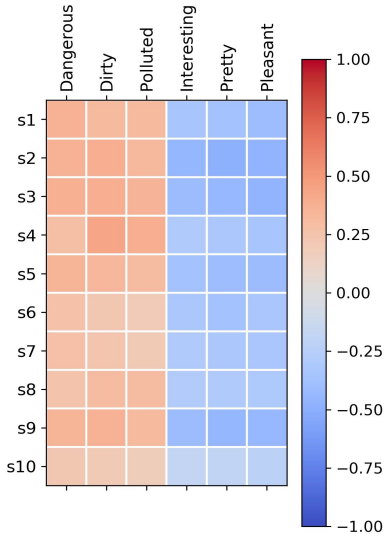


Figure 20: Correlation matrix between 10 the semantic descriptors and the 6 labels of the CECYTE study. All cells are statistically significant at p<0.05.

Label	Group pair	Image corpus: 1,200 images	
		Mean difference	p-value
Dangerous	CE-MT	+0.79	**0.00**
Dirty	CE-MT	+0.37	**0.00**
Interesting	CE-MT	-0.85	**0.00**
Pleasant	CE-MT	-0.74	**0.00**
Polluted	CE-MT	+0.50	**0.00**
Pretty	CE-MT	-0.07	0.21

Table 6: Tukey's HSD statistics for the (1,200) image corpus. CE and MT respectively stands for CECYTE students and MTurkers. Values in bold are statistically significant at p<0.05.

Label	Group pair	Image corpus: 99 images [17]	
		Mean difference	p-value
Dangerous	CE-MT	+1.24	**0.00**
Dirty	CE-MT	+1.08	**0.00**
Interesting	CE-MT	-0.59	**0.005**
Pretty	CE-MT	+0.22	0.24

Table 7: Tukey's HSD statistics for the (99) image corpus. CE and MT respectively stands for CECYTE students and MTurkers. Values in bold are statistically significant at p<0.05. Please note that we only included the labels that are also covered in our study.

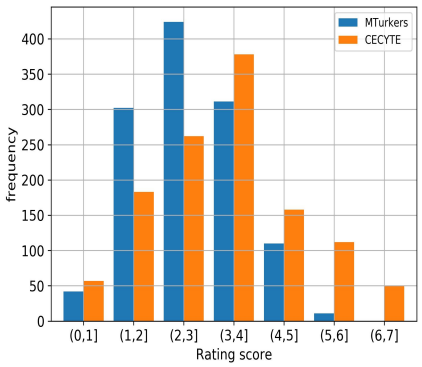


Figure 21: Plot comparing the distributions of perception ratings for the label Dangerous among the two groups.

fact, for these 3 labels, one can see (figure 28 in the appendix D) that they have a very similar distribution for the locals along the Likert scale.

- Images were perceived to be more *interesting* and *pleasant* by the MTurkers as seen in [17] for the particular case of the label Interesting. When looking at the individual median scores per image (figure 28 in the appendix D), we found that for non-locals, 64% of the pictures were rated as more *interesting* and that 58% were rated as more *pleasant*
- Finally, like in [17], we found that the range of perceptions for the label Pretty are not statistically different between CECYTE students and MTurkers.

The main difference between the perception of the images is the fact that the CECYTE students are immersed in the context of the pictures and actually, they knew that image corpus was based on Guanajuato and therefore could rate the images based on their background and personal experiences, that is why the labels Dangerous, Dirty and Polluted are rated higher: Most of the times, local people will better know about their cities and actual situation. On the other hand, MTurkers might not have had this context because probably they haven't visited these kind of urban scenes, and imaging that they have seen these scenes before, we can recall that, even though tourism is one of the main economic activities in the state of Guanajuato, the image corpus was composed of pictures from both touristic and non-touristic places. At the end, non-local people will judge an scene based on the visual cues within it [17]. As an example of analysis, Guanajuato has several colonial era mansions in their historic centers as well as several catholic temples, this makes us think that for a local to see this in a daily basis is not interesting, since he/she is used to it, but the same doesn't happen to the non-local people, which the new different things may arouse an interesting or even pleasant perception. Just as a comment, at the end, it seems that local people will see more negative aspects of their cities than non-local people, like we saw in this analysis.

Finally, like in [17], we analyze some of the comments given by the MTurkers. We filtered those pictures with high values (5 to 7) in the labels Pleasant and Interesting for MTurkers and low values for CECYTE students, we found some comments like *"This would make me feel like I was seeing something out of the ordinary"*, *"Makes me feel like i am exploring an area with a rich cultural heritage"*, *"These buildings are in much better shape than any others I've seen so far; effort has been put into preserving them"* or *"The street texture makes me happy. I love how it looks"*. Those are comments that are concerned with what we mentioned above: new things may be more interesting for people. Additionally, some of these comments support some studies like [25] for the label pleasant (for example, building texture or pavement pattern). We did the same analysis for the other group of labels (Dangerous, Polluted and Dirty) that MTurkers rated with high scores compared to CECYTE students, one comment is *"Those wires look extremely dangerous"*, which we can say that even though the picture received a high score on the label Dangerous for both groups, for the other two labels, MTukers rated it with high scores but CECYTE students with low scores, making us thinking that this is a normal situation (seen in many pictures, refer to the appendix C), that the students may be used to it but for non-locals can generate some visual contamination due to the tangled electrical wiring. Another comment is *"Makes me feel as if i have turned down the wrong path into a back alley, or stairwell, that i shouldnt have"*, which locals rated with a low (3) score and non-locals with the highest value in the scale (7) in the label Dangerous. We can see that in this case, the background of the people will affect the perception of the scenes they see, since for a student from Guanajuato this a very common thing: walking inside an alley (unless something has happened to him/her) for a non-local this could be dangerous since they don't know who or what can be at the other side of the alley. We can then say and reaffirm to what have been studied before [17] and what we have been mentioning along this section: The background and experience of the people will play a crucial role when forming urban perceptions.

8.5 Pair-wise Analysis

In order to understand the variability of the ratings, we performed a pair-wise analysis of these considering the whole image corpus for both groups: Mturkers and CECYTE students. The scatter plots of all the statistically significant labels (based on the Tukey's HSD test as in previous section) of this analysis can be consulted in the appendix E. Based on them, we observe:

- Label **Dangerous**: Most of the points are below the 45° line, which means that CECYTE students perceive images more *dangerous*.
- Label **Dirty**: We see that the points are more distributed along the plane and that the most dense points (more images) are near the 45° but showing a little tendency towards the CECYTE ratings, which can be explained by looking at the mean difference 6: +0.37 and making CECYTE students to perceive images *dirtier*.
- Label **Polluted**: Almost same situation as with label Dirty: most dense points are near the 45° line but showing a more notorious tendency towards the CECYTE ratings. Mean difference of +0.50 and this makes CECYTE students to perceive images more *polluted*.

- Label **Interesting**: We clearly see that most of the points and dense points are above the 45° line. Making MTurkers perceive images more *interesting*.
- Label **Pleasant**: Almost same situation as with label Interesting, most of the points are above the 45° line and on the Mturkers side, making them to perceive images more *pleasant*.

At the end, we could validate our findings in section 8.4 as well as the findings based on the smaller image corpus made in [17]. Finally, we analyze two pictures with opposite scores for the two groups (i.e. while one group rates high the other rates low and viceversa) in two labels like in [17]: Dangerous and Interesting. These pictures are identified with the tags *I-1* and *I-2* within the scatter plots shown in the appendix E.

(a) Dangerous: I1 **(b) Interesting: I-2**

Figure 22: Images where the perceptual ratings differ significantly between local and MTurk population.

Figure 22 shows the 2 selected images with significant difference in the ratings given by the groups (as seen in figure 29), on an specific label. Figure 22a is labeled as *I-1* and ratings are for the label Dangerous: 7 and 1.5 for CECYTE students and MTurkers, respectively. This is very a particular case, as the local people rated it as very Dangerous. However, with a naked eye, we don't agree with such perception and actually, the other negative labels are rated very low (2). Visual cues show a clean and well-preserved street, and seems to be near to cathedral or a church with people walking and parked cars, all these elements could lead that the picture might be taken in the historic center. However, we have to take into account what we have been saying: in a local context, background and experiences play a crucial role, maybe the raters have suffered something negative in that street. Another thing to take into account is the fact that we see this image in the day light. However, we don't see any street light or security element and even worse, in case of danger, there is only one direction to run. On the other hand, MTurkers have given a really low score (1.5) for the label Dangerous, which makes totally sense to us.

The other picture that we analyze is shown in figure 22b, labeled as I-2 and which ratings for the label Interesting are: 1 and 6 for CECYTE students and MTurkers, respectively. In this case, the local score makes more sense to us, since that is just a common neighborhood, nothing interesting for the locals and even worse due to the worn paint and lack of exterior maintenance. However, for non-locals this is interesting, perhaps due to the fact that it is an scene that they have never seen before and that they would like to explore, or maybe the combination of colors make the scene interesting. We checked the comments just to see what the opinion of the MTurkers was and it seems it's an interesting scene for them: *"It's visually interesting but it makes me feel claustrophobic"* and *"This would make me feel like I was seeing something out of the ordinary"*.

8.6 Regression task

For the regression task, we first train Random Forest models using the annotation of each urban awareness label as the dependent variable and the three datasets with features and the one with semantic descriptors as the independent variable as explained in section 5. With these Random Forests, we attempt to infer the impressions of the scenes that are included in the image corpus. We evaluate six cases: (1) Using the semantic segmentation dataset; (2) using the places205 datasets [15]; (3) using the places365 dataset; (4) using the first 2 principal components of the labels; (5) using the 10 semantic descriptors with the 6 labels (Locals: CECYTE annotations) and (6) with the 12 labels (Non-locals: MTurk annotations). Please note that for the first 4 cases, we only use as dependent variables the 12 labels obtained via MTurk. Finally, we compare the obtained results with the best results obtained in [15] by using a fully connected layer of a GoogLeNet CNN pre-trained (CNN-FC) on places205 database.

First case: We train 12 Random Forest with the 150 features of the semantic segmentation dataset. Even though the results (see table 15 in the appendix F.1) are not that good (except for the label Quiet with an R^2 of 0.44, which negatively correlated objects explain well the label), they can be comparable with the results obtained in [15] taking into account that the number of dimensions and more important, the number of objects of the semantic segmentation is far less when comparing to the CNN-FC (since it was pre-trained on the places205 database). The R^2 values are around 0.10 to 0.20 below the results of the CNN-FC. When looking at the most important features, we observe almost the same features with the highest values in the correlation between each label and the features as was explained in section 7.

Second case: This case corresponds to the use of the places205 dataset and it is the same analysis as in [15] and the results can be consulted in table 15 in the appendix F.1. The only thing that we can add in this case is in relation with the most important features of the model, they were objects like plaza, cathedral, botanical garden, garbage dump, slum, etc. which technically were the same as the ones explained in section 7.

Third case: Following the procedure, we train Random Forest on the places365 dataset. In general, the model obtains good results in terms of R^2 when comparing to CNN-FC, especially in two of the labels: Dangerous, which got a higher value than the CNN-FC and whose positively correlated features (table 10 in appendix F.1) do not exceed the threshold we set up (absolute value of 0.3) and negatively correlated features are similar to the ones with places205; and the label Polluted, which has the highest value and whose positively correlated features are a bit more explained than places205 and there were not negatively correlated features that exceeded the threshold. Regarding the other 10 labels, when comparing to the results with places205 dataset, we do not find much difference (approx. ± 0.04), and there are even labels in which the R^2 is higher for the latter.

Fourth case: From section 2 we know that 77% of the variance of the 12 labels can be explained by the first two principal components. Then, we apply PCA on the labels (these were scaled to unit variance previously) and train Random Forests to predict these two components. The results can be seen in table 16 in the appendix F.1, in which we can note an improvement in the R^2 values for all the datasets, specially for the principal component 2 which explains the Quiet label. However, as we know that the rest of the eleven labels are inversely correlated (negatives and positives), we can at least have a moderately good model to explain that an scene within an image has either a negative or a positive impression. Again, even with more objects (places365 dataset), the results were almost the same or even better for the case of the places205 dataset.

Fifth case: We train the Random Forests by using the 10 semantic descriptors annotations (aggregated by mode and by mean) and the two urban awareness labels. Results shown in tables 17 and 18 (in the appendix F.1) indicate that we got better results when aggregating the descriptors by the mean and using the annotations made by the MTurkers, which make some sense, since the non-local people have no context or clue about the cities, they will have no bias towards the pictures and they will rate only based on visual cues they perceive in the pictures as seen in section 8.4 and 8.5. Finally, even though these results are lower than the obtained with the CNN-FC or the places365, we can say that they are good results because we only consider 10 semantic descriptors against 205 or 365 objects that can explain better an scene, and even better because even though we thought of these semantic descriptors to explain a *dangerous* or *dirty* scene, they ended up explaining other urban awareness labels. For the particular case of the label Quiet, we were expecting such results since none of the descriptors has to do with the label.

Even though, the results in tables 15, 16, 17 and 18 are obtained with the datasets without any kind of pre-processing, we also try several pre-processing task such as binarizing the features, applying PCA (getting 26 principal components to explain the 99% of the variance, applying ZCA, or even combining the datasets. However, the results were not satisfactory.

8.7 Classification task

Before doing the classification task, we have to binarize the value of the labels, making sure that we have balanced classes as explained in section 5. Please recall that for classification, we will only use the MTurk scores. After doing so, we perform a classification task via a Random Forest model to infer whether an scene within a picture is either represented by an urban awareness label or not. For this task, we evaluate three cases: (1) Using the semantic segmentation dataset; (2) using the GoogLeNet places205 and (3) using the places365 dataset.

First case: We train 12 Random Forest to make classification and we get the results showed in table 19 in the appendix F.2. We a see a good accuracy and a good Kappa value compared to the R^2 obtained with the regression task (above 0.1). Even though, they are different values with different meanings, they represent the regression and the classification models, respectively. The surprisingly thing was the Kappa value obtained for the Quiet label, since it is the only Kappa value that exceeded the ones obtained in the other two cases.

Second case: We apply the same classification method but this time by using the places205 dataset, as expected, we get the higher values (see table 19 in the appendix F.2) for both the accuracy and Kappa value in all labels except for one as we mentioned in the first case. We appreciate that for 6 labels, we get the highest Kappa values, and there is no much difference to what we get with the places365, in general the difference is ±0.3 for the Kappa values.

Third case: We use the places365 to do classification, even though we have more objects, we only get 5 out of 12 labels with the highest Kappa value, and like we mentioned in the case above, the difference of the results of the other labels between places205 is minimal.

We can say that, even though we get really good results in the classification task, we cannot make sure that we can rely on these results, since the binarization process and the thresholds we set up to delimit the classes are based only on the 1,200 pictures from the image corpus and for some cases like for the Interesting label the threshold is high (a value of 5) for describing that an scene is interesting. We would have to have more samples and from different cities and context to make a better generalization of these thresholds.

8.8 Fine-tuning task

In section 5.5, we explained the methodology we use to do fine-tuning of the last layer of three CNN popular architectures based on the places365 database [26]. Therefore, we analyze 3 cases: (1) GoogLeNet-places365; (2) AlexNet-places365 and (3) VGG16-places365. Results are shown in table 8.

Label	CNN architecture		
	GoogLeNet places365	**AlexNet places365**	**VGG places365**
	RMSE	RMSE	RMSE
Accessible	**0.72**	0.94	0.80
Dangerous	**0.79**	0.92	0.87
Dirty	**0.89**	1.08	0.94
Happy	**0.74**	0.81	0.84
Interesting	**0.72**	0.92	0.81
Picturesque	**0.79**	1.00	0.91
Pleasant	**0.72**	0.91	0.85
Polluted	**0.75**	0.82	0.79
Preserved	**0.86**	1.01	0.87
Pretty	**0.82**	0.95	0.86
Quiet	**0.72**	0.78	0.78
Wealthy	**0.65**	0.74	0.72

Table 8: Fine-tuning task results for the three CNN architectures. Values in bold represent the lowest RMSE per label.

First case: GoogLeNet [22] is the architecture in which we extracted the visual features based on the places205 and places365 databases. The results are very close to all the RMSE obtained with the Random forest and each of the cases that involved the GoogLeNet architecture (table 15). Respecting the computational times, it took 2 hours to train every network for each label.

Second case: AlexNet [10]. We got the worst RMSE values for each of the labels, except for two cases: Labels Quiet and Happy. However, this was the network that took less time to be trained, around 40 minutes per label considering the similar configuration as with the other CNN architectures.

Third case: VGG16 [19]. Even though, the results are not as good as with GoogLeNet, we found pretty close RMSE for some labels, such as Pretty or Preserved. Normally, differences between the values are approximately ±0.1 with respect to GoogLeNet. Respecting the computational time, we noticed that this was the CNN architecture that most time consumed (around 9 hours per label).

We saw that with the Fine-tuning, we can get similar RMSE values as if we first extract the features and then run Random Forest models to get regression. However, we found that this is not very practical in our specific case for several reasons: our image corpus (and even if we use image augmentation) is not big enough for doing deep-learning, and even more important, because of the computational resources that we need, we cannot compare a simple and ordinary computer (that we used to extract the visual features from the CNN architectures) with a GPU cluster (that we used to do the fine-tuning task). Finally, due to this observations, we conclude that the fine-tuning task is not suitable for our particular case.

9 CONCLUSIONS

In conclusion, we have seen that trying to infer an impression of an scene with one or more of the 12 urban awareness labels with general objects is difficult, although not impossible, since we showed that with datasets with general objects (like building, wall, tree, church, sidewalk, etc.) we could well describe the labels as we demonstrated (with the 150 features of the DilatedNet Semantic Segmentation dataset) in the regression task that 5 out 12 labels had a value of R^2 more or equal than 0.30 and during the classification task, 10 out of 12 labels had a Kappa value more or equal than 0.30. However, as we explained in the results section, we cannot rely on the results of the classification task as they have different thresholds for getting the classes merely based on the aggregated labels values of our image corpus. As mentioned in section 8.6, for the label Polluted, with the places365 we could get the highest R^2, however, the difference was not considerable (0.02 above CNN-FC); and for the label Dangerous we could get a higher value as well (R^2 of 0.39). The results showed by the prediction of the 2 principal components are quite interesting because instead of trying to infer labels in a separately way, we merge them in two components, where the first one tells whether an scene has a negative or positive impression and the second component will tell whether the scene is quiet or not. With this, we saw an increase of the R^2 on all the datasets for the component that explains 11 labels (i.e. first component): around 0.50 for both the GoogLeNet places205 dataset and GoogLeNet places365, and 0.36 for the DilatedNet Semantic Segmentation dataset. For the second component (i.e. quiet) we got values between 0.55 and 0.57 with all the datasets.

Even though we thought that having more objects could help to get better results in general, this was not the case as we saw with the regression in which the places365 could not get significant different results when comparing with places205. Therefore, we saw that it is more than enough to have a consistent and complete set of features to describe well an scene as well as its impression in a very general way; and this is even more important when we try to generalize to any city or country, because having a very specific visual vocabulary can maybe explain quite well a very specific context (e.g. Guanajuato scene) but at the end, it is like if we were "overfitting" the features.

With the help of the manual annotation, we got a visual vocabulary to describe both Dangerous and Dirty scenes. We saw that with only 10 semantic descriptors got from this visual vocabulary, we can get good R^2 when 9 (out of 12) labels (values above 0.25) and that these can be comparable with the models that used the 205 and 365 visual objects because unlike those, our semantic descriptors are adequate and kind of specific to our labels. With these descriptors we also saw that they are even present in some scenes with positive perceptions but that won't affect it, such is the case of the lack of security elements.

The t-SNE analysis helped to group different images with same characteristics (objects) and some of these objects were found in the correlation analysis. Based on our knowledge and on some studies like [7] [25] [2] and [13], we explained the existing correlation of some objects and the labels, which made the analysis easier to understand as some objects can contribute in an opposite way to the labels and as some labels are correlated between each other so they share some objects.

When analyzing the impressions between locals and non-locals we confirmed the results of a previous study based on an smaller image corpus but in the same Guanajuato context: we saw a tendency of the local people to disagree more about their perceptions when compared to non-local people and also another tendency of the locals to rate negative labels higher than the positive ones, and finding the opposite with the non-local people. Which makes us think that even further of having on a set of visual cues, the background and personal experiences will play a crucial role and were considered first when perceiving different scenes: locals, of course, are more aware about the real situation of their cities and will be slightly biased only by visual cues like the non-locals.

The work done in this report is completely based on a Mexican context. As a future work, it would be interesting to make this analysis on cities from developed countries, or apply the created models (regression or classification from section 8) to a different image corpus to see the results as well as doing a crowdsourcing study between people with different backgrounds but that share some characteristics, for example age, education level, etc. so that we can complement the results found in our study.

ACKNOWLEDGMENTS

I would like to thank Dr. Daniel Gatica-Pérez and Dr. Darshan Santani for their supervision, help, knowledge and patience in the development of this project. I would also like to thank Yassir Ybenkhedda and the Social Computing Group at IDIAP for their recommendations, tips and help as well as the IDIAP institute for the use of the computing facilities. Many thanks to Dr. Salvador who helped us conducting the crowdsourcing study in Guanajuato and all the students that participated in the study as well as the school authorities of CECYTE for their support.

REFERENCES

[1] 2014-present. SenseCityVity: Mobile Sensing, Urban Awareness, and Collective Action. (2014-present). http://www.idiap.ch/project/sensecityvity/

[2] Anke Blobaum and Hunecke Marcel. 2005. Perceived Danger in Urban Public Space: The Impacts of Physical Features and Personal Factors. *Environment and Behavior* (July 2005), 465–486. https://doi.org/10.1177/0013916504269643

[3] Marcus D. Bloice, Christof Stocker, and Andreas Holzinger. 2017. Augmentor: An Image Augmentation Library for Machine Learning. *CoRR* abs/1708.04680 (2017). arXiv:1708.04680 http://arxiv.org/abs/1708.04680

[4] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* (1971), 378–382. https://doi.org/10.1037/h0031619

[5] Kilem L. Gwet. 2011. On the Krippendorff's Alpha coefficient. (2011). http://www.agreestat.com/research_papers/onkrippendorffalpha.pdf

[6] Kazunori Hanyu. 2000. Visual properties and affective appraisals in residential areas in daylight. *Journal of Environmental Psychology* 20, 3 (2000), 273 – 284. https://doi.org/10.1006/jevp.1999.0163

[7] Thomas R. Herzog and Kristi K. Chernick. 2000. Tranquility and danger in urban and natural settings. *Journal of Environmental Psychology* 20, 1 (2000), 29 – 39. https://doi.org/10.1006/jevp.1999.0151

[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

[9] Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting Intraclass Correlation Coefficients for reliability research. *Journal of Chiropractic Medicine* 15 (2016), 155–163. Issue 2. https://doi.org/doi:10.1016/j.jcm.2016.02.012

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. (2012), 1097–1105. http://dl.acm.org/citation.cfm?id=2999134.2999257

[11] Judith Matloff. 2013. Take a tour of the barrio most Mexicans won't visit — if you dare. (December 2013). http://america.aljazeera.com/articles/2013/12/8/a-cultural-tour-ofthefiercebarriomostmexicanswontvisit.html

[12] L. Perez and J. Wang. 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *ArXiv e-prints* (Dec. 2017). arXiv:cs.CV/1712.04621

[13] Douglas Perkins, John Meeks, and Taylor Ralph. 1992. The Physical Environment of Street Blocks and Resident Preceptions of Crime and Disorder: Implications for Theory and Measurement. *Journal of Environmental Psychology* 12 (March 1992), 21–34. Issue 1. https://doi.org/10.1016/S0272-4944(05)80294-4

[14] Salvador Ruiz-Correa, Darshan Santani, and Gatica-Perez Daniel. 2014. Young and the City: Crowdsourcing Urban Awareness in a Developing Country. (October 2014).

[15] Darshan Santani. 2016. *Computational Analysis of Urban Places Using Mobile Crowdsensing*. Ph.D. Dissertation. EPFL.

[16] Darshan Santani, Salvador Ruiz-Correa, and Gatica-Perez Daniel. 2015. Looking at Cities in Mexico with Crowds. *DEV '15 Proceedings of the 2015 Annual Symposium on Computing for Development* (December 2015), 127–135. https://doi.org/10.1145/2830629.2830638

[17] Darshan Santani, Salvador Ruiz-Correa, and Gatica-Perez Daniel. 2017. Insiders and Outsiders: Comparing Urban Impressions between Population Groups. (May 2017). https://doi.org/10.1145/3078971.3079022

[18] Philipp Schaer. 2012. *Better than Their Reputation? On the Reliability of Relevance Assessments with Students*. Springer Berlin Heidelberg, 124–135. https://doi.org/10.1007/978-3-642-33247-0_14

[19] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556 http://arxiv.org/abs/1409.1556

[20] Claudia Solera. 2014. Crimen rinde culto a la Santa Muerte ("Crime worships Holy Death"). (2014). http://www.excelsior.com.mx/nacional/2014/04/13/953865

[21] S. S. Stevens. 1946. On the Theory of Scales of Measurement. *Science* 103, 2684 (1946), 677–680. https://doi.org/10.1126/science.103.2684.677 arXiv:http://science.sciencemag.org/content/103/2684/677.full.pdf

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. (2015). http://arxiv.org/abs/1409.4842

[23] L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[24] Fisher Yu and Vladlen Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*.

[25] Heng Zhang and Shih-Hsien Lin. 2011. Affective appraisal of residents and visual elements in the neighborhood: A case study in an established suburban community. *Landscape and Urban Planning* 101, 1 (2011), 11 – 21. https://doi.org/10.1016/j.landurbplan.2010.12.010

[26] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[27] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 487–495. http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf

A T-SNE ANALYSIS

A.1 Figures

Please note that each of the values of the urban awareness labels are represented in a seven-point Likert scale (from (1) strongly disagree to (7) strongly agree) and are shown by different colors: 1.0-1.5, 2.0-2.5, 3.0-3.5, 4.0-4.5, 5.0-5.5, 6.0-6.5 and 7.0.

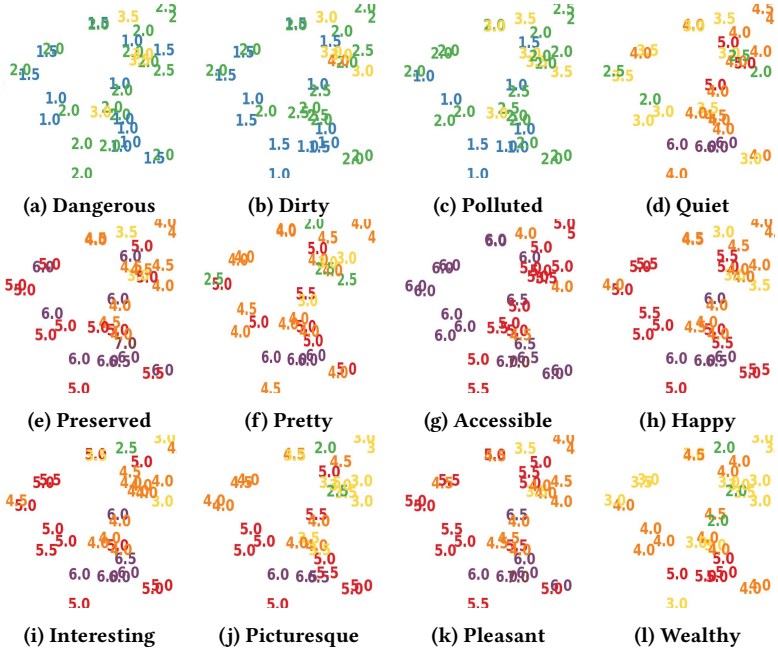


Figure 23: DilatedNet Semantic Segmentation dataset: A section of the visualization of the values of the urban awareness labels by applying t-SNE.

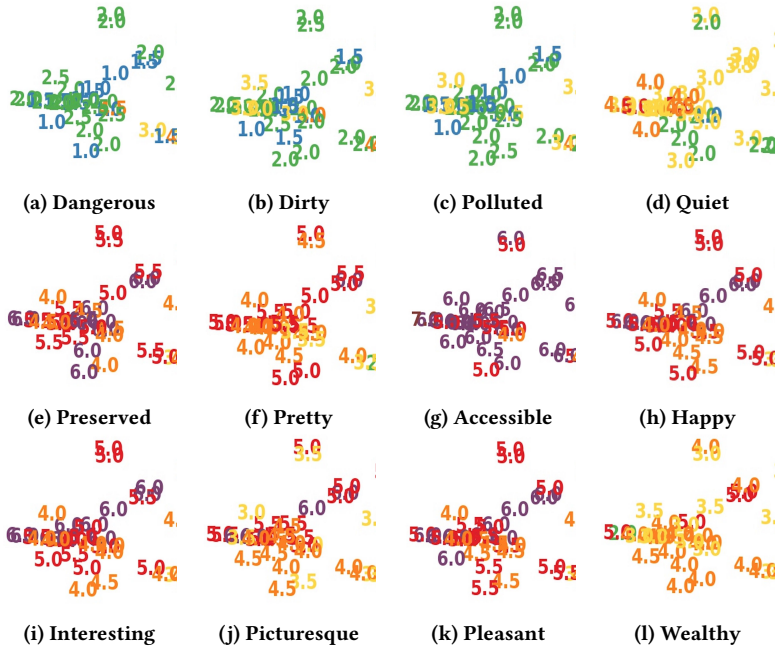
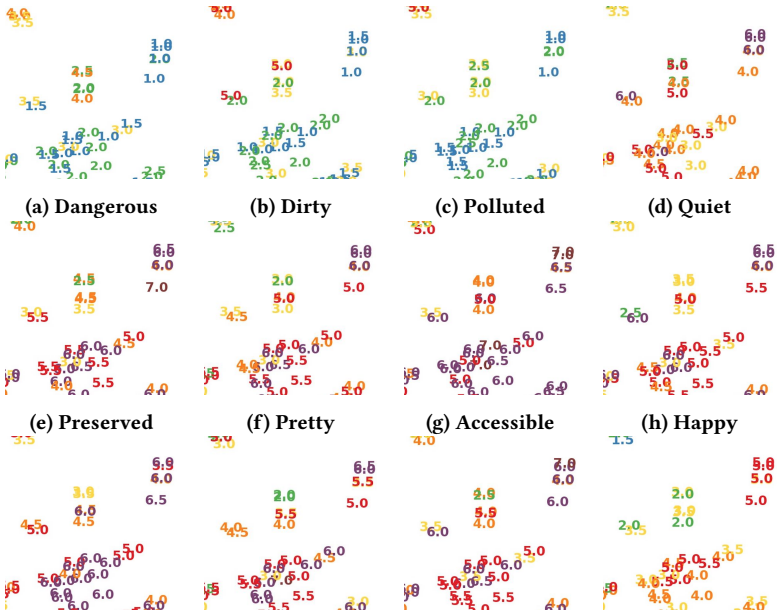


Figure 24: GoogLeNet places205 dataset: A section of the visualization of the values of the urban awareness labels by applying t-SNE.



(a) Dangerous (b) Dirty (c) Polluted (d) Quiet

(e) Preserved (f) Pretty (g) Accessible (h) Happy

(i) Interesting (j) Picturesque (k) Pleasant (l) Wealthy

Figure 25: GoogLeNet places365 dataset: A section of the visualization of the values of the urban awareness labels by applying t-SNE.

B CONNECTION BETWEEN OBJECTS AND URBAN AWARENESS LABELS

B.1 Tables

Label	DilatedNet Semantic Segmentation Correlated objects ($\|\rho\|$>0.12)	
	Positively	**Negatively**
Accessible	sidewalk (0.315), tree(0.297), person(0.263), plant (0.130) and ashcan (0.123)	wall (-0.345), door (-0.211), stairs (-0.193), paintings (-0.179) and dirt track (-0.171)
Dangerous	painting (0.188), wall (0.181), pole (0.138), dirt track (0.134) and building (0.129)	sidewalk (-0.254), tree (-0.253), person (-0.234), bench (-0.184) and plant (-0.155)
Dirty	dirt track (0.190), wall (0.170), painting (0.162) and earth (0.145)	tree (-0.227), sidewalk (-0.216), bench (-0.154), person (-0.151) and plant (-0.147)
Happy	person (0.225), tree (0.222), sidewalk (0.176), plant (0.167) and bench (0.145)	dirt track (-0.175), wall (-0.164), pole (-0.162), painting (-0.157) and road (-0.148)
Interesting	plant (0.211), bench (0.193), person (0.192), floor (0.173) and stairs (0.165)	road (-0.351), car (-0.228), truck (-0.195) and pole (-0.165)
Picturesque	tree (0.214), plant (0.207), person (0.162), bench (0.139) and fountain (0.133)	road (-0.210), pole (-0.169), truck (-0.143), dirt track (-0.138) and car (-0.128)
Pleasant	tree (0.223), person (0.215), sidewalk (0.184), plant (0.162) and bench (0.161)	road (-0.180), dirt track (-0.176), painting (-0.153), pole (-0.153) and wall (-0.150)
Polluted	road (0.244), car (0.218), pole (0.146) and dirt track (0.126)	bench (-0.168), plant (-0.163), floor (-0.141), person (-0.126) and tree (-0.122)
Preserved	sidewalk (0.243), tree (0.227), person (0.198), plant (0.169) and bench (0.134)	wall (-0.207), dirt track (-0.198), painting (-0.180), pole (-0.153) and earth (-0.151)
Pretty	tree (0.242), plant (0.186), sidewalk (0.170), person (0.147) and bench (0.144)	road (-0.160), dirt track (-0.154), wall (-0.141), painting (-0.141) and pole (-0.141)
Quiet	stairs (0.282), stairway (0.271), wall (0.259), building (0.240) and door (0.132)	car (-0.418), person (-0.388), bus (-0.342), sky (-0.246) and signboard (-0.238)
Wealthy	tree (0.291), sidewalk (0.278), person (0.236), plant (0.159) and car (0.140)	wall (-0.284), painting (-0.193), building (-0.183), bed (-0.141) and dirt track (-0.139)

Table 9: Spearman correlation between features of the DilatedNet Semantic Segmentation dataset and the 12 urban awareness labels (MTurk annotations). Due to limitations of space, we only show the top-5 highest values. Negative values are colored in red and positive values in blue. All reported values are statistically significant at p<0.05.

Label	GoogLeNet places205 Correlated objects ($\|\rho\|$>0.3)		GoogLeNet places365 Correlated objects ($\|\rho\|$>0.3)	
	Positively	Negatively	Positively	Negatively
Accessible	plaza (0.585), hotel (0.525, courthouse (0.483), cathedral (0.465) and inn (0.461)	basement (-0.388) and corridor (-0.315)	plaza (0.562), campus (0.508), hotel (0.487), beer garden (0.424) and office building (0.424)	basement (-0.327) alley (-0.326) and trench (-0.308)
Dangerous	basement (0.300)	plaza (-0.437), hotel (-0.364), cathedral (-0.361), fountain (-0.360) and botanical garden (-0.359)	No objects found	plaza (-0.432), campus (-0.410), beer garden (-0.398), banquet hall (-0.388) and gazebo (-0.380)
Dirty	slum (0.332) and basement (0.307)	plaza (-0.395), hotel (-0.352), cathedral (-0.345), palace (-0.334) and botanical garden (-0.321)	slum (0.314) and trench (0.302)	campus (-0.392), plaza (-0.388), hotel (-0.368), gazebo (-0.352) and banquet hall (-0.341)
Happy	plaza (0.446), cathedral (0.424), palace (0.414), hotel (0.399) and basilica (0.399)	basement (-0.358), slum (-0.334) and trench (-0.310)	plaza (0.406), hotel (0.386), campus (0.376), banquet hall (0.345) and beer garden (0.345)	loading dock (-0.328) and trench (-0.308)
Interesting	monastery (0.409), palace (0.389), castle (0.376), basilica (0.368) and cathedral (0.360)	garbage dump (-0.394), highway (-0.334), trench (-0.331), shed (-0.325) and gas station (-0.322)	temple (0.332), church (0.319) and pagoda (0.301)	loading dock (-0.425) and parking garage (-0.307
Picturesque	palace (0.423), cathedral (0.406), basilica (0.404), monastery (0.400) and church (0.388)	basement (-0.330), slum (-0.319), garbage dump (-0.316) and trench (-0.312)	palace (0.340), plaza (0.338), hotel (0.331), campus (0.325) and courtyard (0.308)	loading dock (-0.349)
Pleasant	plaza (0.428), cathedral (0.409), palace (0.399), basilica (0.389) and church (0.384)	basement (-0.350), slum (-0.336) and trench (-0.325)	plaza (0.399), hotel (0.378), campus (0.375), shopping mall (0.351) and beer garden (0.341)	loading dock (-0.329), slum (-0.304) and trench (-0.302)
Polluted	garbage dump (0.368), slum (0.337) and trench (0.300)	courtyard (-0.344)	junkyard (0.338), loading dock (0.338), landfill (0.333) and repair shop (0.320)	No objects found
Preserved	plaza (0.472), hotel (0.433), cathedral (0.430), palace (0.412) and basilica (0.404)	basement (-0.358) and slum (-0.327)	plaza (0.437), campus (0.419), hotel (0.416), apartment building (0.367) and palace (0.351)	slum (-0.333) and trench (-0.301)
Pretty	cathedral (0.415), palace (0.407), plaza (0.402), basilica (0.398) and church (0.393)	basement (-0.337), slum (-0.327) and trench (-0.307)	plaza (0.383), hotel (0.381), campus (0.379), palace (0.348) and banquet hall (0.335)	loading dock (-0.310)
Quiet	No objects found	supermarket (-0.524), market (-0.522), assembly line (-0.495), food court (-0.456) and airport terminal (-0.454)	jail cell (0.397), corridor (0.344), burial chamber (0.329), catacomb (0.326) and alcove (0.307)	market (-0.567), bazaar (-0.562), flea market (-0.548), market (-0.502) and supermarket (-0.497)
Wealthy	plaza (0.516), hotel (0.484), courthouse (0.449), cathedral (0.428) and inn (0.420)	basement (-0.354), alley (-0.312) and corridor (-0.310)	plaza (0.503), campus (0.466), hotel (0.456), banquet hall (0.406) and beer garden (0.405)	medina (-0.333)

Table 10: Spearman correlation between features of the two GoogLeNet datasets and the 12 urban awareness labels (MTurk annotations). Due to limitations of space, we only show the top-5 highest values. Negative values are colored in red and positive values in blue. All reported values are statistically significant at p<0.05.

C RESULTS OF MANUAL ANNOTATION

C.1 Tables

Least Dangerous [1-3)		Moderately Dangerous [3-5)		Most Dangerous [5-7]	
Object	**Frequency**	**Object**	**Frequency**	**Object**	**Frequency**
Cars in motion	26	Tangled electrical wiring	29	Tag graffiti	30
Tangled electrical wiring	23	Cars in motion	22	No artificial lighting	21
People walking on the road	8	Tag graffiti	20	Tangled electrical wiring	21
Tag graffiti	6	No artificial lighting	6	Small alley	17
Small sidewalks	4	Small alley	5	Damaged walls	14
Badly parked motorcycles	2	Badly parked motorcycles	5	Damaged paths	12
Badly parked cars	2	People walking on the road	5	Graffiti	5
Damaged paths	2	Damaged walls	4	Unfinished building	5
No artificial lighting	2	Graffiti	3	Dark alleys	4
Street water	1	Motorcyclist without helmet	3	Rusty house protections	3

Table 11: 10 most frequent objects of Dangerous urban awareness label when annotating in a manual way.

Least Dirty [1-3)		Moderately Dirty [3-5)		Dirtiest [5-7]	
Object	**Frequency**	**Object**	**Frequency**	**Object**	**Frequency**
Worn paint	12	Tag graffiti	28	Graffiti	34
Tangled electrical wiring	12	Worn paint	24	Trash on street	26
Street water	10	Tangled electrical wiring	23	Tangled electrical wiring	25
Tag graffiti	9	Trash on street	14	Worn paint	21
Hawkers	4	Unpainted houses	12	Badly painted houses	20
Damaged paths	4	Street water	11	Unpainted houses	19
Trash on street	3	Undergrowth	8	Covered up graffiti	18
Graffiti	3	Graffiti	8	Moisture on walls	17
Scratched road signs	2	Mud	8	Propaganda	17
Propaganda	2	Damaged path	8	Undergrowth	16

Table 12: 10 most frequent objects of Dirty urban awareness label when annotating in a manual way.

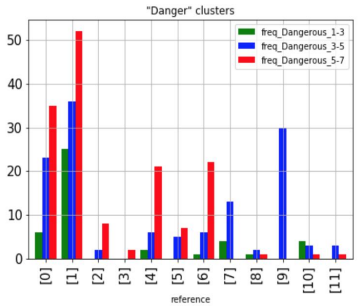
Least Dangerous [1-3) 50 pictures		Moderately Dangerous [3-5) 50 pictures		Most Dangerous [5-7] 44 pictures	
Object	**Frequency**	**Object**	**Frequency**	**Object**	**Frequency**
People walking	41	Artificial lighting	36	Artificial lighting	22
Artificial lighting	35	People walking	35	Normal parked cars	14
Trees	29	Cars in motion	22	People walking	12
Cars in motion	23	Trees	20	Trees	11
Commercial zone	21	Normal parked cars	20	Residential zone	8
Normal parked cars	14	Residential zone	12	Cars in motion	7
Non-neglected plants	14	Stores	10	Non-neglected plants	6
Bus stop	7	Commercial zone	9	Stores	6
Traffic lights	6	Traffic lights	5	People riding bicycles	5
Park benches	5	Public parking	4	Normal parked cars	2

Table 13: 10 most frequent objects of Non-Dangerous urban awareness label when annotating in a manual way.

Least Dirty [1-3) 50 pictures		Moderately Dirty [3-5) 50 pictures		Dirtiest [5-7] 44 pictures	
Object	**Frequency**	**Object**	**Frequency**	**Object**	**Frequency**
No trash on street	43	No trash on street	30	No trash on street	10
Well preserved path	39	Well preserved path	29	Trees	10
Trees	37	Trees	22	Well preserved paths	9
Well painted houses	34	Well painted houses	19	Non-neglected plants	7
Well preserved sidewalks	28	Well preserved sidewalks	16	Well painted houses	6
Lime on trees	18	Non-neglected plants	9	Well preserved sidewalks	5
Non-neglected plants	17	Clean road signs	8	Trash cans	2
Clean road signs	8	Trash cans	3	Well preserved roads	1
Trash cans	7	Lime on trees	2		
Monuments	5	Park fountain	1		

Table 14: 10 most frequent objects of Non-Dirty urban awareness label when annotating in a manual way.

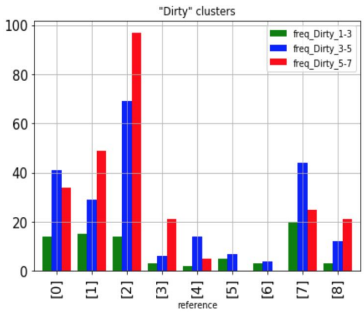
C.2 Figures



(a) Dangerous clusters

reference	cluster
[0]	vandalism
[1]	lack_of_exterior_maintenance
[2]	unkempt_house_building
[3]	no_security_house_building_elements
[4]	no_outdoor_lighting
[5]	vacant_lots
[6]	place_for_concealment
[7]	blocking_elements
[8]	littering
[9]	negative_elements_regarding_people
[10]	bad_planning
[11]	suspicious_elements

(b) Dangerous clusters references

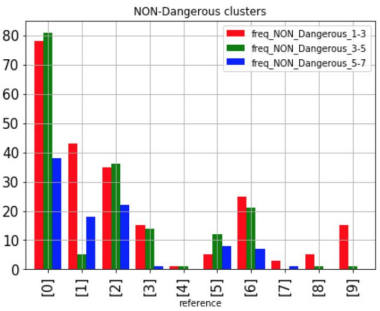


(c) Dirty clusters

reference	cluster
[0]	vandalism
[1]	littering
[2]	unkempt_houses_buildings
[3]	neglected_vegetation
[4]	blocking_elements
[5]	negative_elements_regarding_the_people
[6]	deteriorated_road_signs
[7]	lack_of_exterior_maintenance
[8]	remaining

(d) Dirty clusters references

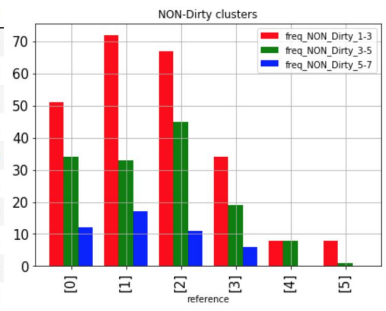
Figure 26: Dangerous and Dirty clusters resulted from manual annotation.



(a) Non-Dangerous clusters

reference	cluster
[0]	positive_elements_regarding_people
[1]	vegetation
[2]	outdoor_lighting
[3]	traffic_elements
[4]	education
[5]	residential_zone
[6]	commercial_zone
[7]	security_elements
[8]	touristic_zone
[9]	park_elements

(b) Non-Dangerous clusters references



(c) Non-Dirty clusters

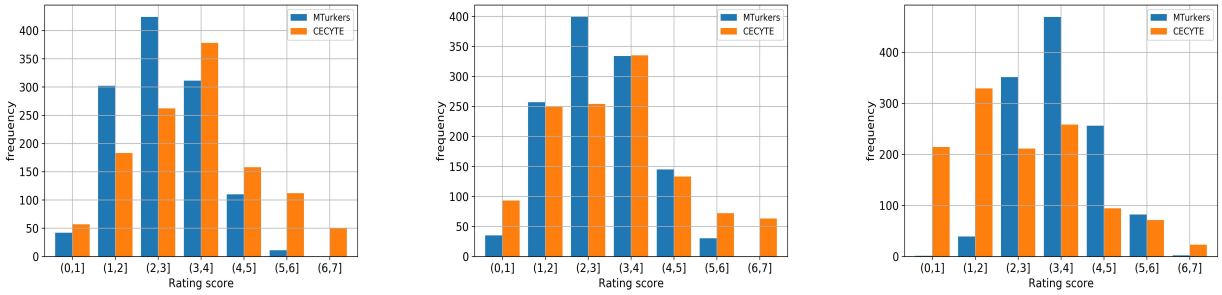
reference	cluster
[0]	no_littering
[1]	vegetation
[2]	exterior_maintenances
[3]	well_maintained_houses_buildings
[4]	well_maintained_traffic_elements
[5]	park_elements

(d) Non-Dirty clusters references

Figure 27: Non-Dangerous and Non-Dirty clusters resulted from manual annotation.

D RESULTS OF THE COMPARISON BETWEEN GROUPS

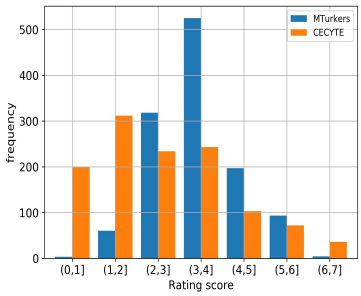
D.1 Figures



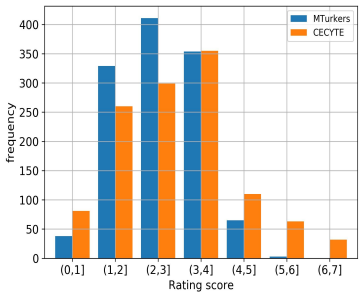
(a) Dangerous

(b) Dirty

(c) Interesting



(d) Pleasant

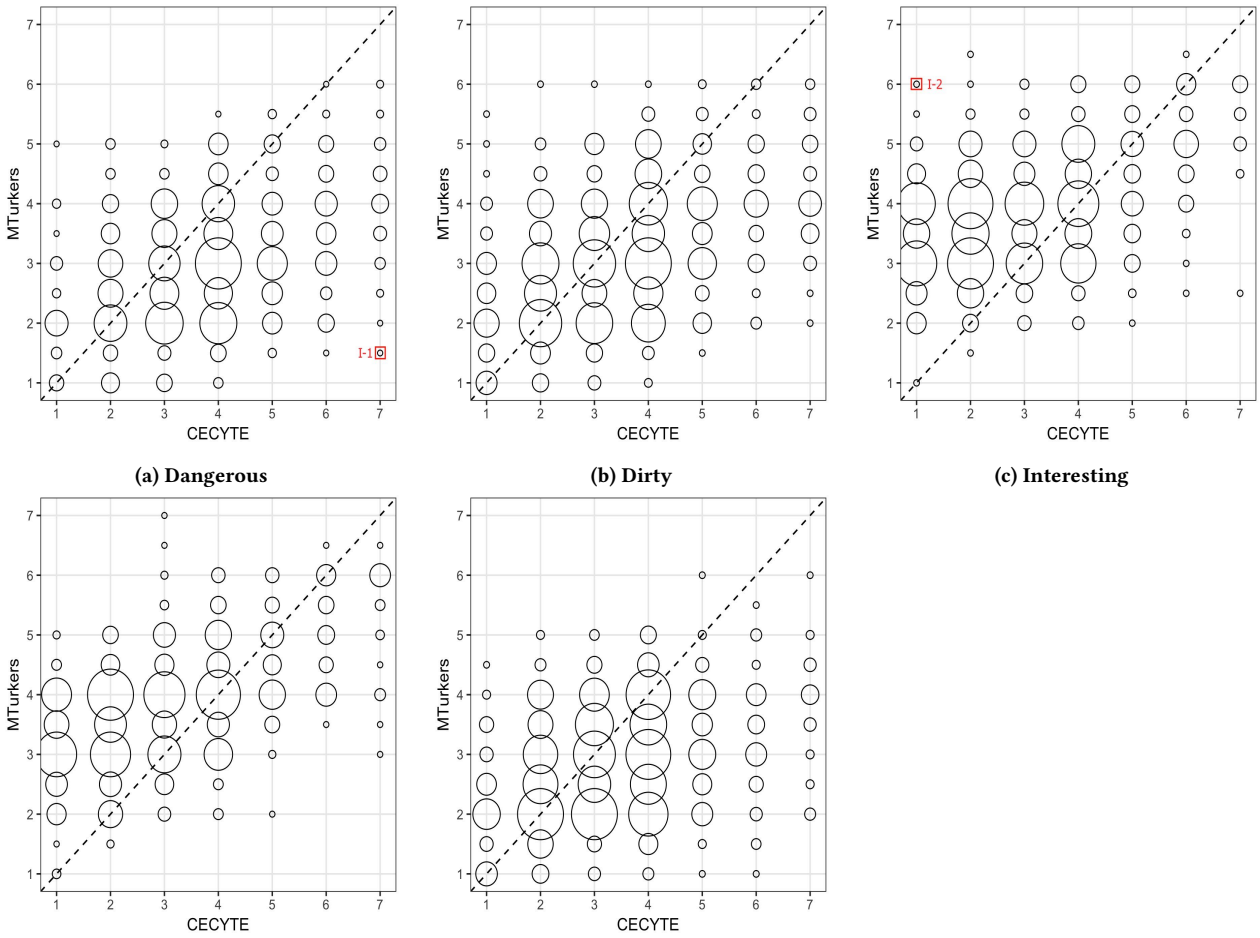


(e) Polluted

Figure 28: Plots comparing the distributions of perception ratings between the Mturkers (non-locals) and CECYTE (locals) study. Please note that we only included the labels (we include the label Dangerous again) which mean difference (in Tukey's HSD test) was statistically significant at p<0.05.

E RESULTS OF THE PAIR-WISE ANALYSIS

E.1 Figures



(a) Dangerous

(b) Dirty

(c) Interesting

(d) Pleasant

(e) Polluted

Figure 29: Scatter plots showing the pair-wise annotator ratings by CECYTE students (locals) and MTurkers (non-locals). Each dot corresponds to an image, with the size of the dots proportional to the number of observations. 45° line is also shown in all the plots. Two dots are highlighted in the plots as I-1 and I-2, these are enlarged in section 8.5. Please note that we only included the labels (we include the label Dangerous again) which mean difference (in Tukey's HSD test) was statistically significant at p<0.05.

F RESULTS OF REGRESSION AND CLASSIFICATION TASKS

F.1 Regression

Label	Baseline RMSE	RF (CNN-FC) [15]		RF (GoogLeNet places365)		RF (GoogLeNet places205) [15]		RF (Semantic Segmentation)	
		RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Accessible	0.94	0.69	**0.48**	0.70	0.46	0.70	0.46	0.76	0.35
Dangerous	0.99	0.79	0.38	0.78	**0.39**	0.80	0.35	0.85	0.28
Dirty	1.06	0.85	**0.37**	0.86	0.35	0.87	0.35	0.93	0.24
Happy	0.95	0.71	**0.46**	0.74	0.41	0.73	0.43	0.80	0.30
Interesting	0.94	0.70	**0.45**	0.73	0.41	0.73	0.41	0.81	0.28
Picturesque	1.06	0.76	**0.49**	0.80	0.43	0.79	0.44	0.91	0.27
Pleasant	0.97	0.73	**0.45**	0.75	0.42	0.75	0.42	0.82	0.31
Polluted	0.92	0.77	0.30	0.76	**0.32**	0.78	0.28	0.82	0.21
Preserved	1.05	0.78	**0.45**	0.81	0.42	0.81	0.40	0.89	0.29
Pretty	1.02	0.75	**0.46**	0.81	0.42	0.77	0.43	0.87	0.29
Quiet	1.00	0.73	**0.48**	0.78	0.44	0.75	0.45	0.75	0.44
Wealthy	0.88	0.65	**0.46**	0.75	0.44	0.66	0.44	0.74	0.31

Table 15: Results of the regression task. The values in bold represent the highest values. Please note that these results were obtained using the MTurk annotations.

Label	Baseline RMSE	RF (GoogLeNet places365)		RF (GoogLeNet places205)		RF (Semantic Segmentation)	
		RMSE	R^2	RMSE	R^2	RMSE	R^2
First principal component	1	0.72	**0.50**	0.72	0.49	0.81	0.36
Second principal component	1	0.67	0.56	0.66	**0.57**	0.68	0.55

Table 16: Results of the regression task taking only the two first principal components after applying PCA on the labels. The values in bold represent the highest values. Please note that these results were obtained using the MTurk annotations.

Label	Baseline RMSE (MTurk)	Baseline RMSE (CECYTE)	RF Semantic descriptors (mode) MTurk (non-local) annotations		RF Semantic descriptors (mode) CECYTE (local) annotations	
			RMSE	R^2	RMSE	R^2
Accessible	0.94	-	0.82	0.24	-	-
Dangerous	0.99	1.44	0.85	**0.27**	1.31	0.18
Dirty	1.06	1.53	0.90	**0.29**	1.34	0.23
Happy	0.95	-	0.79	0.31	-	-
Interesting	0.94	1.53	0.85	0.19	1.34	**0.25**
Picturesque	1.06	-	0.90	0.29	-	-
Pleasant	0.97	1.58	0.80	**0.32**	1.34	0.29
Polluted	0.92	1.38	0.83	**0.20**	1.24	**0.20**
Preserved	1.05	-	0.84	0.35	-	-
Pretty	1.02	1.65	0.84	**0.32**	1.39	0.30
Quiet	1.00	-	0.99	0.05	-	-
Wealthy	0.88	-	0.73	0.31	-	-

Table 17: Results of the regression task using the semantic descriptor (aggregated by the mode of the annotations) dataset as well as the two annotations datasets: MTurk and CECYTE. The values in bold represent the highest values between the annotations datasets. Please note that there are fields with a dash (-), and this is due to the fact that for the CECYTE dataset we only studied 6 urban awareness labels.

Label	Baseline RMSE (MTurk)	Baseline RMSE (CECYTE)	RF Semantic descriptors (mean) MTurk (non-local) annotations		RF Semantic descriptors (mean) CECYTE (local) annotations	
			RMSE	R^2	RMSE	R^2
Accessible	0.94	-	0.81	0.28	-	-
Dangerous	0.99	1.44	0.82	**0.32**	1.32	0.18
Dirty	1.06	1.53	0.87	**0.33**	1.34	0.24
Happy	0.95	-	0.78	0.34	-	-
Interesting	0.94	1.53	0.86	0.19	1.31	**0.28**
Picturesque	1.06	-	0.89	0.30	-	-
Pleasant	0.97	1.58	0.78	**0.37**	1.31	0.32
Polluted	0.92	1.38	0.82	**0.21**	1.23	**0.21**
Preserved	1.05	-	0.82	0.39	-	-
Pretty	1.02	1.65	0.84	**0.33**	1.36	**0.33**
Quiet	1.00	-	1.00	0.04	-	-
Wealthy	0.88	-	0.73	0.33	-	-

Table 18: Results of the regression task using the semantic descriptor (aggregated by the mean of the annotations) dataset as well as the two annotations datasets: MTurk and CECYTE. The values in bold represent the highest values between the annotations datasets. Please note that there are fields with a dash (-), and this is due to the fact that for the CECYTE dataset we only studied 6 urban awareness labels.

F.2 Classification

Label	Baseline RMSE	RF (GoogLeNet places365)		RF (GoogLeNet places205)		RF (Semantic Segmentation)	
		Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Accessible	0.50	0.74	0.47	0.75	**0.49**	0.70	0.39
Dangerous	0.49	0.72	**0.43**	0.70	0.38	0.68	0.34
Dirty	0.49	0.70	**0.39**	0.68	0.37	0.66	0.33
Happy	0.50	0.73	**0.45**	0.72	0.44	0.68	0.36
Interesting	0.50	0.72	0.43	0.74	**0.47**	0.69	0.37
Picturesque	0.49	0.70	**0.39**	0.68	0.35	0.64	0.27
Pleasant	0.50	0.70	**0.40**	0.70	**0.40**	0.66	0.33
Polluted	0.50	0.69	**0.37**	0.68	**0.37**	0.65	0.31
Preserved	0.50	0.72	**0.44**	0.70	0.41	0.67	0.34
Pretty	0.49	0.69	0.37	0.70	**0.39**	0.63	0.26
Quiet	0.50	0.75	0.51	0.76	0.51	0.76	**0.52**
Wealthy	0.50	0.73	0.45	0.74	**0.47**	0.68	0.35

Table 19: Results of the classification task. The values in bold represent the highest values. Please note that these results were obtained using the MTurk annotations.