Human Computation (2018) 5:1:13-30

ISSN: 2330-8001, DOI: 10.15346/hc.v5i1.2

Human Computation vs. Machine Learning: an Experimental Comparison for Image Classification

GLORIA RE CALEGARI, CEFRIEL

GIOELE NASI, CEFRIEL

IRENE CELINO, CEFRIEL

ABSTRACT

Image classification is a classical task heavily studied in computer vision and widely required in many concrete scientific and industrial scenarios. Is it better to rely on human eyes, thus asking people to classify pictures, or to train a machine learning system to automatically solve the task? The answer largely depends on the specific case and the required accuracy: humans may be more reliable – especially if they are domain experts – but automatic processing can be cheaper, even if less capable to demonstrate an "intelligent" behaviour.

In this paper, we present an experimental comparison of different Human Computation and Machine Learning approaches to solve the same image classification task on a set of pictures used in light pollution research. We illustrate the adopted methods and the obtained results and we compare and contrast them in order to come up with a long term combined strategy to address the specific issue at scale: while it is hard to ensure a long-term engagement of users to exclusively rely on the Human Computation approach, the human classification is indispensable to overcome the "cold start" problem of automated data modelling.

1. INTRODUCTION AND MOTIVATION

Light pollution is the excessive, poorly directed or unnecessary artificial light at night (Hollan, 2009); it leads to negative effects on human well-being, biodiversity, visibility of stars, safety and energy consumption. We are witnessing an increasing awareness about the light pollution issue and the strategies to fight it: overhead lamps should never emit light directly above the horizontal, blue light should be avoided in favour of a colour temperature of 3000 K or lower, light should be used only as much as needed for the specific purpose. On the other hand, the study and measurement of the light pollution phenomenon is still partially an open research question (Sánchez de Miguel, 2015).

One indirect method to quantify the phenomenon is by analyzing pictures of the Earth at night, and especially of cities at night. The astronauts of the International Space Station (ISS) take tens of thousands of pictures of the Earth from above during their space missions; NASA makes this wealth of images available for free on the Web. Those pictures have the right level of resolution to detect the brightness and to distinguish the light sources; moreover, they are continuously taken across different missions, therefore they can be used also to study the light pollution evolution over time.

However, those images come unlabeled and their subject can vary greatly: photos can include not only cities at night and stars, but also images of astronauts floating in space, or of the ISS itself; cities and portions of Earth during the day; astral phenomenons like the Aurora Borealis; and mixed-subjects images not of interest for light pollution research. Therefore, we address the task of classifying those images in order to make them suitable for subsequent analysis of light pollution levels.

In this paper, we illustrate and compare different approaches to solve the classification problem, one based on Human Computation methods and two based on the employment of Machine Learning techniques. We illustrate the experiments and analyze their results to come up with suggestions on the possible optimal setting to address this issue in the long term.

2. IMAGE CLASSIFICATION APPROACHES

To solve the image classification task, two alternative approaches have traditionally been used in literature: people-based human computation and computer-based machine learning. A more recent trend addresses the possibility to take the best of both worlds.

2.1. Human Computation

Human Computation (Law and Ahn, 2011) is a computer science technique in which a computational process is performed by outsourcing certain steps to humans. Unlike traditional computation, in which a human delegates a task to a computer, in Human Computation the computer asks a person or a large group of people to solve a problem; then it collects, interprets and integrates their solutions. The original concept of Human Computation by its inventor Luis von Ahn derived from the common sense observation that people are intrinsically very good at solving some kinds of tasks which are, on the other hand, very hard to address for a computer; this is the case of a number of targets of Artificial Intelligence (like image recognition or natural language understanding) for which research is still open.

A specific kind of Human Computation is crowdsourcing (Howe, 2008), i.e. the process to outsource tasks to a “crowd” of distributed people, usually through online micro-work platforms (e.g. Amazon Mechanical Turk), which often imply a monetary (micro-)reward. When the accent is more strongly related to the involvement of people in scientific campaigns or experiment, the term Citizen Science (Irwin, 1995) is used to stress the contribution to public awareness.

With specific reference to non-monetary incentives to people, entertainment and fun are often used to motivate participation. A widely used approach (Von Ahn, 2006) is based on Games With a Purpose (GWAP). A GWAP is a gaming application that outsources steps of a computational process to humans (Human Computation task) in an entertaining environment. To be effective, a GWAP should be carefully designed (1) to provide an effective mechanism to solve the task and (2) to assure a continuous involvement and contribution of users/players.

2.2. Supervised Machine Learning

Many different supervised classification methodologies exist and can be applied to solve the image classification task. Those approaches can be roughly distinguished between manual selection of the image features to train the model and a so-called deep learning approach which also include a feature learning process.

With respect to manual selection of features, in literature different techniques are used for image classification: detection of edges and curvatures (Nixon and Aguado, 2008), extraction of texture to identify objects or regions of interest in an image (Lin et al., 2011), image indexing by the color contents (Smith and Chang, 1995), color quantization to reduce the color levels as in (Debevec, 2008). When features are identified, any traditional supervised learning algorithm can be used (Kotsiantis et al., 2007).

With respect to the deep learning approach, Convolutional Neural Networks (CNN) are widely used in the context of computer vision and image recognition, because they are built to resemble visual perception in humans (Krizhevsky et al., 2012). They are made of several layers of artificial neurons that process portions of the original image: they manage to maintain information related to the spatial structure of the data, keeping track of pixels that are nearby to each other in order to recognize visual patterns. Deep Convolutional Neural Networks have been widely used in the last years to solve image recognition tasks with performance close to that of humans (Russakovsky et al., 2015).

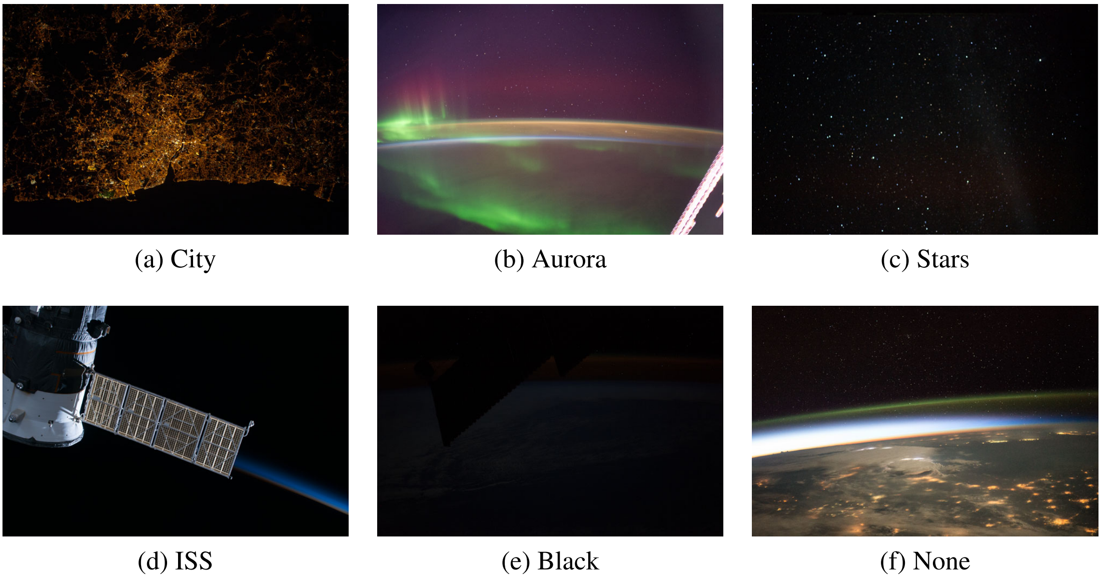
2.3. Active Learning

Supervised learning methods always need a training set to "learn" the classification model. The key idea behind active learning (Settles, 2012), instead, is that a machine learning algorithm can perform better with less training if it is allowed to dynamically choose the data from which it learns. An active learner may pose "queries" (in the form of unlabelled data instances) to an "oracle" (e.g., a human annotator) which understands the nature of the problem and can output labelled data instances.

While not a full mixed human-machine approach, active learning is a successful example of a trade-off between purely-manual and purely-automatic classification. This approach is well-motivated in many applications in which unlabelled data is abundant or easy to collect, but training data is difficult, time-consuming, or expensive to obtain.

3. DATA AND OBJECTIVE

Looking at the pictures made available by NASA, we define six categories: *CITY* (images of cities at night), *AURORA* (images of the aurora borealis or northern lights), *STARS* (images of starry space), ISS (images of the spacecraft or astronauts), *BLACK* (completely dark images) and *NONE* (for any other cases); an example of each category is shown in Figure 1. The goal of the image classification task is to label each picture with the corresponding category.



(a) City (b) Aurora (c) Stars

(d) ISS (e) Black (f) None

***Figure 1.* Examples of images belonging to the six classes**

Not all those classes are equally important for light pollution research. We are particularly interested in identifying images of cities at night (to quantify the light pollution phenomenon) and pictures of stars and black images, since they are useful for the calibration of the measurement system (Sánchez de Miguel, 2015).

Therefore the objective is to obtain very good results in predicting at least those three classes. We would like to retrieve as many pictures of cities at night as possible (high values of both precision and recall) to potentially study all the cities on Earth. On the other hand, for calibration purposes, we need to select pictures that are actually black and images with actual stars (high values of precision, no requirement on recall). As regards the other classes (ISS, AURORA and NONE), they are not directly related to the light pollution goal, but they are useful to be distinguished both by humans and machines.

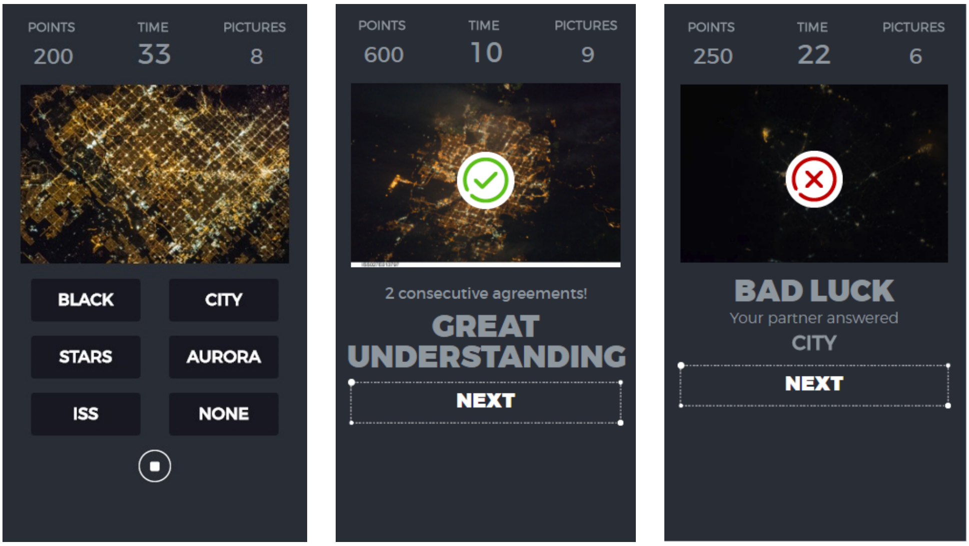
In the rest of the paper, we illustrate how we addressed the classification objective by both Human Computation and Machine Learning approaches. The ambition is to understand to what extent the two approaches are complementary or overlapping. We aim to classify several thousands of images obtained during various ISS missions and made available on the Web by NASA[1].

[1] Cf. https://eol.jsc.nasa.gov/.

4. HUMAN COMPUTATION EXPERIMENT

To solve the classification task by involving people, we developed the Night Knights game[2], a GWAP designed for the light pollution purpose.

The human participant plays the role of an astronaut with the mission of classifying images taken from the ISS. At the beginning of the game, each player is randomly coupled with another player and a sequence of images is shown to them simultaneously. Each image must be classified into 6 predefined categories and players can gain points only if they agree on the same classification. Each game lasts one minute and so people have to be quick to classify as many images as possible. Figure 2 shows some screenshots of the game user interface.



2 consecutive agreements!
GREAT UNDERSTANDING

BAD LUCK
Your partner answered
CITY

(a) Classify an image (b) Agreement (c) Disagreement

Figure 2. Night Knights: the game play

Night Knights takes inspiration from one of the most famous GWAP for image classification, the ESP Game by Luis von Ahn (Law and Ahn, 2011), but it implements different game mechanics and aggregation algorithm. Differently from the ESP Game that aimed to add free-text tags to images (*annotation* or *labelling* purpose), the goal of Night Knights is to classify images into a closed set of categories (*classification* purpose). As regards the algorithm used to aggregate players' answers, the ESP game is based on simple agreement, which means that only two coherent contributions are required to consider the annotation task completed (i.e., if two users type the same tag, the tag is considered valid). On the other hand, Night Knights is still based on an output-agreement strategy (Von Ahn and Dabbish, 2008), but the algorithm is more articulated: each image is offered to multiple players, whose contributions are weighted according to their reliability (measured on a gold standard) and aggregated to form a classification score; once the score overcomes a specific

[2] Cf. http://www.nightknights.eu/.

threshold, the image is classified and removed from the game (cf. also (Celino et al., 2012)). By design, a minimum of four agreeing users are required to reach the classification threshold; more details on the design and implementation of Night Knights are provided in (Celino et al., 2016).

The game was released in October 2016 in beta version and was then more extensively advertised for a related competition[3] whose winner joined the 2017 Summer Expedition to observe the Solar Eclipse in USA. The competition lasted about one month, from June 12th to July 7th 2017, and was addressed to all EU University students. At the time of writing (October 2017), the global results obtained by Night Knights are summarized in Table 1.

***Table 1.* Night Knights experimental results**

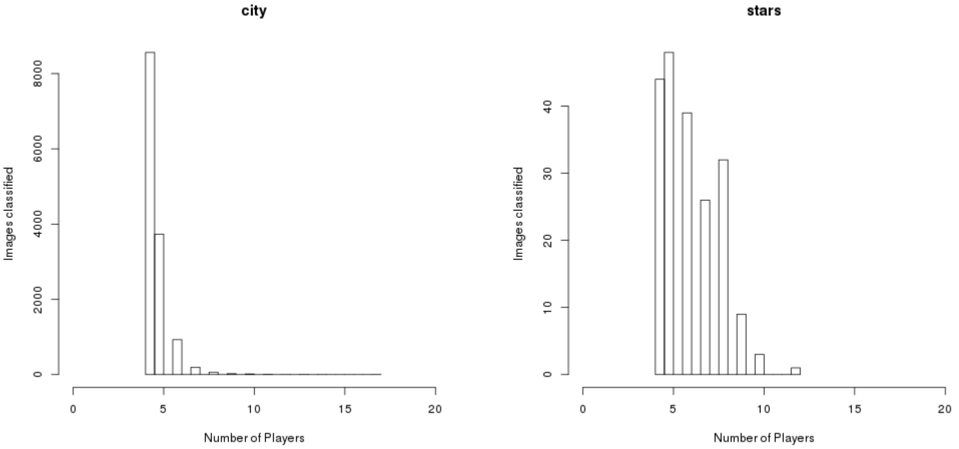
classified images	27,699
players	633
total played time	~226 hours
throughput	122.44 task/hour
average life-play (ALP)	21.44 min time/user
expected contribution (EC)	43.76 task/user

To evaluate efficiency, attractiveness and effectiveness of Night Knights, we measure the main GWAP metrics (Law and Ahn, 2011) which are the *throughput* (number of solved task per unit of time), the *average life play* (ALP, average time spent by each user), and the *expected contribution* (EC, average number of tasks solved by a user). In our case, a "solved task" is a classified image (with aggregated score over the threshold).

The numbers in Table 1 show that the game managed to engage users to play a substantial amount of time, thus classifying almost 28,000 photos. Monitoring the evolution over time of the main metrics, we noticed a significant increase of players' participation during the competition period (e.g. ALP reached values over 100 minute/player). This means that providing a tangible reward to players can make them contribute more efficiently, speeding up the classification process (higher throughput), engaging them for a longer time (higher ALP), and ensuring a larger contribution rate to the human computation task (higher EC). As a global result, more images get classified.

With specific reference to the number of contributions required to reach an agreement on classification, on average 4.6 players are needed, with a number of users ranging between 4 and 17. This result is coherent with the minimum number of contributions required to classify an image that we set at design time. Making this analysis separately for each category, we discover that there are categories that are easier to classify than others. CITY is one of the easiest classes and it is evident by looking at the histogram in Figure 3, which shows a peak in correspondence to 4 players, quickly decreasing towards low frequencies for higher number of users. On the other hand, the histogram of the STARS class is almost constant between 5 and 8 players, indicating that these images are intrinsically harder to classify and thus an higher number of contributions is needed to reach an agreement.

[3] Cf. http://stars4all.eu/index.php/lpis-usa-competition/.



(a) City

(b) Stars

Figure 3. ***Number of users required to reach an agreement on classification for different categories***

However, while the competition results seem encouraging, it is very challenging (if not impossible) to guarantee prizes and tangible rewards until the completion of the classification task over *all* images taken on board of the ISS: this kind of incentive works well within a limited period of time with clear competition objectives. Ensuring an indefinite and continuous engagement of players requires different strategies and incentives that cannot be ensured by a “casual game” like Night Knights. For this reason, we set up different experiments to evaluate the possibility to rely on automatic classification means.

5. MACHINE LEARNING EXPERIMENTS

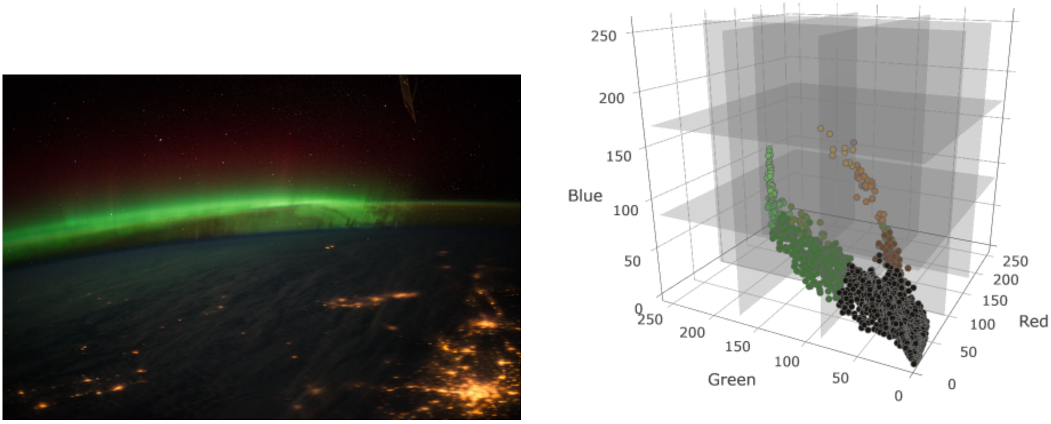
Automatic classification methods do not suffer from the typical problems of Human Computation activities, such as users engagement, varying classification time, incentives to motivate participation and so on. On the other hand, the main requirement to set up a supervised learning classifier is to have an already-labelled dataset.

In our experiments, we decided to use the $\sim$26,000 images classified through Night Knights from the game launch to the end of the competition; while this dataset does not exactly constitute a “ground truth” (since we do not know the *true* classification of those images), we use it as such, dividing it in a training set and a test set, so to enable a direct comparison between the human and machine classifications. The distribution of the images in the dataset across the 6 classes is not balanced: most of the images represent cities (62%) and only few of them (1-5%) belong to BLACK, STARS and ISS classes. Different experiments are required to understand how to suitably handle a dataset which is so unbalanced and heterogeneous.

As described in Section 2, different methods can be adopted for image classification; we decided to follow two parallel approaches: a Random Forest classifier using as input a set of manually-chosen color-based features and a more complex deep learning methodology using Convolutional Neural Networks. In both cases, we tuned the model parameters to find the optimal configuration in terms of precision and recall of the classes of interest with respect to the Night Knights results. For both models, we also studied the effect of the size and composition of the training set (i.e., number of images and class distribution with balanced vs. stratified data). Additionally, for each of the two approaches, we also tuned their specific distinguishing parameters, as explained in detail in the following paragraphs.

5.1. Color-based features and Random Forest Classifier

This first approach requires two phases: feature extraction and model training. With respect to the former, we define a set of color-related features, based on the intuitive idea of creating a 3D histogram of each image colors. This phase is as follows: firstly, the image colors are defined in a mathematical way, by representing each pixel in a RGB three-dimensional color space (red, green and blue); then, the pixels' distribution is analyzed with respect to a color palette to detect the most prominent colors. We define our color palette as a set of fixed colors by partitioning the RGB color space in a uniform way: each axis is divided in *n* splits, generating a number of boxes (*buckets*) equal to the cube of the number of splits (i.e. with 3 splits we have 9 boxes, with 4 splits 64 boxes, with 5 splits 125 boxes and so on). Figure 4 shows the original image and the same image represented in an RGB color space partitioned in 9 buckets. Finally, we count how many pixels fall in each bucket and we compute the relative count, with respect to the total number of pixels in the image. Each bucket becomes a feature of our classification model, indicating the percentage of pixels that fall into that bucket in the RGB color space.



(a) Original image

(b) Partition of the RGB space in 9 buckets

***Figure 4.* Preprocessing an image by representing it in the RGB color space.**

The number of splits of the RGB space is a parameter that needs to be carefully tuned in order to obtain the best classification performances. It is worth noting that, while increasing the number of buckets makes the description of the image more accurate from a color composition perspective, on the other hand, it increases the model complexity, with a consequent worsening of the time performance. Therefore an accurate evaluation of the best number of splits is essential. We made different experiments with 3, 4 and 5 splits: the subsequent model evaluation improves with the number of buckets, but while the difference between 3 and 4 split was sensible, the improvement from 4 to 5 splits was less relevant; thus, we stopped at 5 splits (i.e. 125 features) to avoid a too complex and computationally-expensive modelling phase.

As regards the model building phase, we adopt a Random Forest (Breiman, 2001) classifier, based on the common-sense intuition that our images can be distinguished based on their colors with a tree-based algorithm like "Are more than 50% of pixels black? Then advance along a certain branch. Less? Take this other branch". While the simplest model implementing this classification approach is the decision-tree, we choose the more robust Random Forest implementation, which is an ensemble of trees (a forest, indeed) in which the classifications of all the trees are taken into account to generate the aggregated result.

The best model is obtained after tuning the different parameters as follows: 85%/15% data split between training/test set, the training set conserving the original stratified distribution of images across classes and, as already mentioned, 5 splits and 125 color-based features. Table 2 presents the confusion matrix between the color-based classifier and the Human Computation experiment.

The results are generally positive, with a global accuracy of almost 84%. Also the agreement indexes indicate a good accordance between the color-based classifier and the Human Computation system: the Rand Index is 0.87 and its adjusted version is 0.68[4].

***Table 2.* Confusion matrix of the Night Knights and the best-performing Color-based model**

		Night Knights					
		ISS	**Aurora**	**Black**	**City**	**None**	**Stars**
Color-model	**ISS**	**0.98**	0.00	0.00	0.00	0.02	0.00
	Aurora	0.05	**0.92**	0.00	0.00	0.00	0.03
	Black	0.05	0.04	**0.80**	0.03	0.03	0.04
	City	0.07	0.01	0.00	**0.85**	0.04	0.03
	None	0.11	0.03	0.02	0.01	**0.80**	0.03
	Stars	0.12	0.00	0.09	0.03	0.18	**0.58**

The precision obtained on the different classes is quite high (in the 80-98% range) but for the STARS class which is less that 60%. In terms of recall, again in most cases the model gets good results (78-98%) except for the ISS class (23%) and again for the STARS class (14%). This means that the color-based model is a good classifier, therefore confirming our intuition that colors are a powerful discriminant characteristics of our target images; nonetheless, for the light pollution research objective, an improvement is needed at least with respect to the STARS pictures, which are

[4] The Rand Index (Rand, 1971) is a metric used to evaluate the similarity between data clusters by counting the elements that are classified in concordant and discordant ways; the Adjusted Rand Index is the Rand Index adjusted for the chance grouping of elements. Both indexes range between 0 (no accordance) and 1 (perfect accordance).

used for calibration purposes.

To improve the prediction results, especially on the worse performing classes, we could change the feature selection strategy (e.g., by splitting the color space in a non-uniform way or by adding additional predictors), also on the basis of a qualitative evaluation of the mistakes made by our model. Enriching the feature selection, however, leads to more complex models; since the above-mentioned experiment was already quite computationally expensive (cf. Section 5.3), we decided to try a different approach, by adopting a deep learning method.

5.2. Deep Convolutional Neural Network

Building and training a Deep Convolutional Neural Network from scratch is a long process that involves manually sizing and tuning the architecture of the DCNN. Consequently, we decided to follow another course of action, by re-training an existing, well-performing model (which was originally trained on a similar task) and to make use of its prior knowledge while adapting it to our dataset: this is a technique known as Transfer Learning (Donahue et al., 2014). This approach enables us to cut the training time of a model to a few hours on a normal laptop, whereas deep learning from scratch could take days or weeks and must rely on powerful GPUs for processing. Moreover, this process hides most of the complexity normally associated with the full-scale building of a Deep Convolutional Neural Network, allowing us to focus on just a few hyper parameters and settings.

On a technical level, the neural network library we use is Google's Tensorflow[5], one of the most popular and appreciated in the deep learning community. The recommended network model for the image recognition task suggested by the Tensorflow platform is the open-source Inception-v3 model (Szegedy et al., 2015), which is a network trained on the ImageNet dataset[6], a common benchmark challenge for academic studies in the field of image recognition. The 2014 ImageNet challenge asked to classify 150,000 images into 1,000 different classes, where each class was a common object, plant, animal, etc. The Inception-v3 model achieved 21.2% top-1 and 5.6% top-5 error rate on the ImageNet dataset[7].

We apply the Transfer Learning approach by re-training the model as follows: we load the pre-trained Inception-v3 model and remove the old top layer (which is the one responsible for making the actual prediction based on the features analyzed by the underlying hidden layers); we create a new top layer and train it on our dataset (even if our classes weren't part of the 1,000 originally present in ImageNet); the strong assumption here is that the information that allowed Inception-v3 to categorize over 1,000 classes in ImageNet are also useful for our task, even in a decidedly different (and very peculiar) dataset.

The re-training process actually consists of two phases: a first one in which "bottlenecks" are created, and a second one when the actual training takes place. Bottleneck is an informal term indicating the penultimate layer of the neural network, which produces a summary of the image features for the final classification layer. After the bottlenecks creation, the actual training of the new top layer begins, by repeating a series of learning steps in which a different subset of training images is

[5] Cf. https://www.tensorflow.org.

[6] Cf. http://image-net.org/.

[7] Cf. https://github.com/tensorflow/models/tree/master/inception.

analyzed and used to train the model; after each step, a training accuracy and a validation accuracy values are computed to evaluate the balance between improving accuracy and avoiding overfitting. The number of learning steps is a model configuration parameter. Additionally, to create a more robust model, training images can be modified by adding distortion, scaling, cropping and adjusting brightness.

The best model is obtained after tuning the different parameters as follows: 85%/15% data split between training/test set and 4,000 learning steps; we experimentally evaluated that changing between balanced and stratified did not impact on results, thus we used the same stratified training set employed in the color-based model to enable a direct comparison; moreover, the addition of image distortion did not help either, so the best model we illustrate makes use of the original images, to avoid further processing time. Table 3 presents the confusion matrix between the DCNN classifier and the Human Computation experiment.

Table 3. Confusion matrix of the Night Knights and the best-performing DCNN-based model

		Night Knights					
		ISS	**Aurora**	**Black**	**City**	**None**	**Stars**
DCNN-model	**ISS**	**0.96**	0.00	0.00	0.00	0.03	0.00
	Aurora	0.01	**0.97**	0.00	0.00	0.01	0.01
	Black	0.01	0.00	**0.81**	0.07	0.08	0.04
	City	0.00	0.00	0.00	**0.99**	0.01	0.00
	None	0.01	0.00	0.00	0.01	**0.97**	0.01
	Stars	0.13	0.05	0.09	0.07	0.05	**0.61**

The results are again positive and even better than the color-based model, with a global accuracy of 95%. Also the agreement indexes improve: the Rand Index is 0.97 and its adjusted version is 0.91. The precision obtained on the different classes is high (in the 81-99% range) but for the STARS class which, also in this case, achieves the lowest precision of 61%. In terms of recall, the DCNN model gets good results on all classes (83-98%) with the lowest value again for the STARS class.

The deep learning approach seems therefore a very promising solution for our specific case of image classification. Further considerations are offered hereafter.

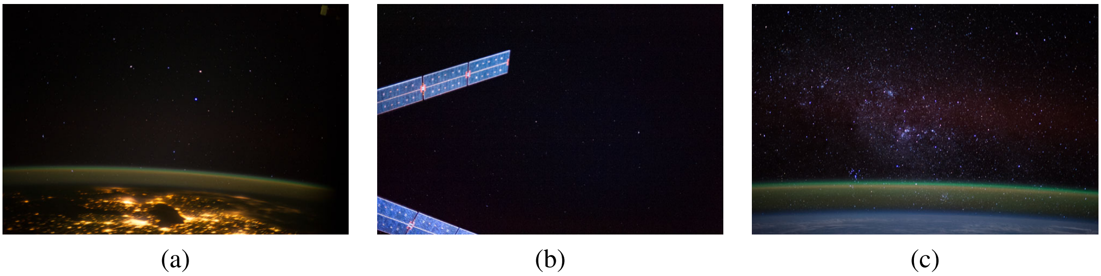
5.3. Comparison between the two Machine Learning approaches

Both classification models show very high values of precision and recall in almost all the classes. Overall, the DCNN model performs better than the color-model: this appears reasonable, since the deep learning methodology builds a model that is more precise and can generalize better than the color-based one, because more different features are taken into account and thus the description of each picture is more accurate and thorough.

A result, which is very relevant for our purposes, is the prediction of the CITY class with extremely high precision and recall (99% and 96% respectively for the DCNN model), which guarantees an almost perfect recognition of the pictures of cities, which are the first and foremost goal of our light pollution evaluation. We attribute this result to at least a couple of factors: firstly, the uniqueness of the city images with respect to the others classes, as they have more peculiar features and are less likely to include elements from other classes; secondly, city images are the most prevalent in our corpus, so a more refined training can have occurred on them.

As regards the other two classes we are interested in – BLACK and STARS – further considerations are needed. The BLACK class is predicted with 81% precision and 88% recall in the DCNN model: since they are used for calibration only, we can be quite satisfied with the precision and give up with respect to a higher recall, thus paying the price of discarding some of the photos.

STARS, on the other hand, seems to be the most difficult class to predict, since both models achieve the lowest precision of 58% and 61% (and in the color-based model, recall is also very low at 14%). Having a look at the confusion matrices in Tables 2 and 3, this is also evident, since they are often mistaken with all the other categories, with almost equal percentages. This apparently negative result does not completely surprise us: within the photos taken from the ISS, it is quite difficult to find images depicting *only* a starry sky, since in most cases the astronauts take mixed pictures, putting in the composition also parts of the Earth's surface or the ISS, as the ones shown in Figure 5.



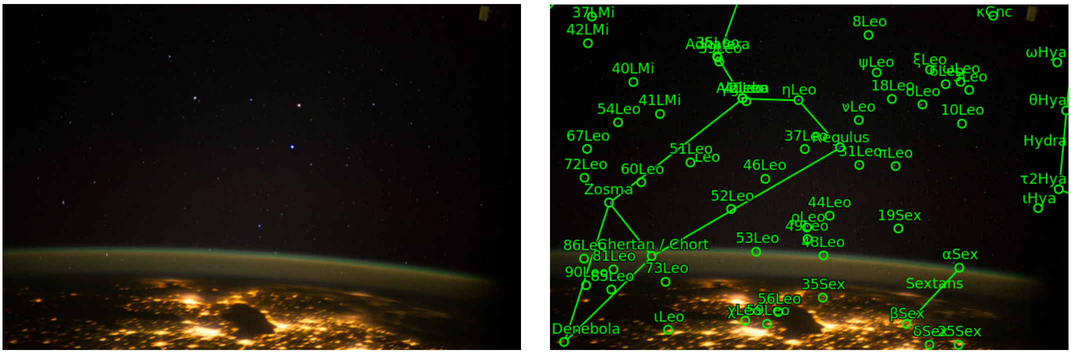
(a) (b) (c)

Figure 5. ***Example of images depicting stars mixed with other categories (CITY, ISS and AURORA)***

As explained before, the starry sky pictures are useful for calibration purposes: the brightness of well-known stars can be compared with the city illumination, thus allowing for a quantitative interpretation of the cities' light pollution phenomenon. Therefore, if "mixed" pictures can be analyzed to identify the starry part, they can be considered valuable. As an example, we took a mixed picture and submitted it to an online service that automatically identifies the visible stars[8]: the result shown in Figure 6 demonstrates that the "poor" results on the STARS class can anyway lead to a positive outcome for our specific purpose.

Given the fact that we are interested in only 3 classes (CITY, STARS and BLACK) out of the 6 proposed categories, the choice of training a 6-class classifier may appear questionable. Indeed, we tried also to train both color-based models and DCNN models on 4 classes (the 3 we are interested in and a fourth encompassing all other cases) to check if precision on the relevant classes could increase. Experimentally, however, we observed that both families of classifiers, trained on 4 classes, returned worse results than the aforementioned experiments; we interpret this outcome as a confirmation of the value of our original 6-class approach: the artificial fourth class was probably

[8] Cf. http://nova.astrometry.net/.



(a) Original image

(b) Identification of stars and constellations

***Figure 6.** Automatic recognition of stars and constellations in a mixed picture*

too mixed in terms of image content, so that the models were unable to learn to correctly distinguish it from the 3 relevant classes.

Finally, the two machine learning approaches can also be compared with respect to their computational performance. All the illustrated experiments were run on a simple laptop, using only CPU processing, with a 16GB 1600Mhz RAM and a 2,8GHz Intel Core i7 processor.

In analyzing the color-based model, we need to take into account the image pre-processing time to extract the color features (1.2 second/image), the model building time (10 minutes for the whole training set) and the classification time for new images of the test set. The training phase (pre-processing + model building) globally takes 7 hour and 30 minutes for 22,000 images; since the time to predict a single image is negligible (0.0003 seconds/image), we can say that the classification phase requires only the time to extract the color features, which is 1.2 second/image.

For the DCNN model, the total training time is around 2 hours, including both the creation of bottlenecks and the learning steps, while the prediction time is 0.6 seconds/image. Those number show that, also from a performance point of view, the DCNN model is to be preferred with respect to the color-based approach, both for the training phase (in which we "save" the time required by Google to build the Inception-v3 model on the ImageNet dataset) and for the prediction phase (which is roughly one half of its competitor's).

6. HUMAN VS. MACHINE COMPARISON

Given the results obtained in the previous sections, we deepen our investigation by having a closer look at the classification confidence measures. On the one hand, we further analyse the similarities between the different approaches, to understand if they show the same "behaviour", so to answer the question "do humans and machines find the same images easy/hard to classify?". On the other hand, we evaluate a technique to address the scenario-specific objectives outlined in Section 3, so to answer the question "can we improve precision (at the expense of recall) on some of the relevant classes?".

6.1. Comparing Human-Machine Classifications on Confidence Measures

When applying a machine learning model to classify new data, it usually not only outputs the predicted class, but it also returns a value between 0 and 1 that represents the "confidence" of the algorithm about the given classification. This confidence score can be interpreted as the "difficulty" to classify the input data. Both the color-based model and the DCNN model give us this confidence score. On the other hand, Night Knights aggregates the contributions from the different players to output the image classification; given our aggregation algorithm, which depends on both the level of agreements between players and the reliability of each contributor, we can say that the higher the number of players required to classify an image, the less "confident" the human classifiers, hence the more "difficult" the classification (cf. also Section 4 and Figure 3). Therefore we analyse the collected data to compare the machine learning approaches to the human computation experiment: do humans and machines find the same images easy/hard to classify?

To find an answer, we take the test set and we divide it in two groups: an *agreement set*, which contains the images that were classified in the same way by Night Knights players and by the two Machine Learning classifiers, and a *disagreement set*, which contains all other images which therefore were either classified in the same way by only 2 out of 3 methods, or were attributed to 3 different classes by the 3 methods. The agreement set contains 79% of the images, thus the disagreement set includes the remaining 21%. Then we apply the t-test[9] on the difference of means between the "confidence measures" of the 3 methods in the agreement/disagreement sets: if a difference is statistically significant, it means that that classifier "distinguishes" between easy- and hard-to-classify images.

***Table 4.* Statistically relevant difference in means on confidence measures between agreement and disagreement sets**

	agreement set	**disagreement set**	**t-test p-value**
DCNN confidence	0.96	0.81	$<10^{-16}$
Color confidence	0.90	0.59	$<10^{-16}$
# of NK players	5.6	8.0	$<10^{-16}$

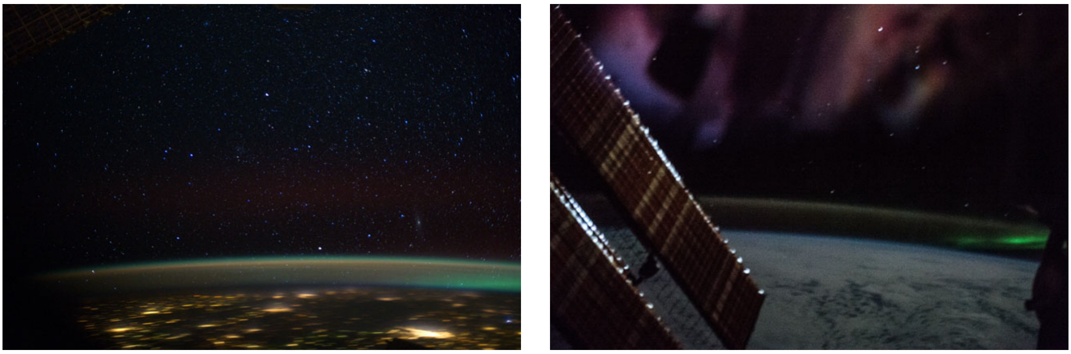
The results are presented in Table 4. All the 3 mean differences are statistically significant as testified by the p-values of the corresponding t-tests. This means that humans and machines "agree" on the difficulty of the image classification: when the machine learning classifiers give a high confidence score, the human computation method requires a smaller number of players to reach an agreement, and vice versa.

Analysing the disagreement set, we notice that the number of images that are classified the same by the two machine learning algorithms and differently by the human computation approach is negligible (less that 1%); considering only the images classified equally by the two automatic classifiers, the Rand/Adjusted Rand Index computed with the respective Night Knights classification

[9] The t-test usually assumes the populations to be normally distributed, which is not our case; however, the t-test is reliable in case of large samples, as per our populations; in any case, the Wilcoxon and the Kolmogorov-Smirnov tests confirm the statistically significant difference in means.

is 0.98/0.96. Therefore we can conclude that, when the two proposed machine learning classifiers agree, it is highly probable that humans would also agree on the same classification.

As regards the other images of the disagreement set, Figure 7 show some pictures that were differently classified by the three approaches (and the respective confidence measures); indeed, we can say that it is difficult or even impossible to tell which should be the "true" classification, since they are all mixed images. As explained in Section 3, for the purpose of measuring the light pollution effect, those images are are not of interest, exactly because they are mixed in content; therefore, we can safely discard all images that were classified differently by the two automatic approaches.



(a) *Night Knights*: STARS 7 users.
DCNN: AURORA 0.54 confidence.
COLOR: CITY 0.70 confidence

(b) *Night Knights*: AURORA 8 users.
DCNN: ISS 0.66 confidence.
COLOR: NONE 0.37 confidence

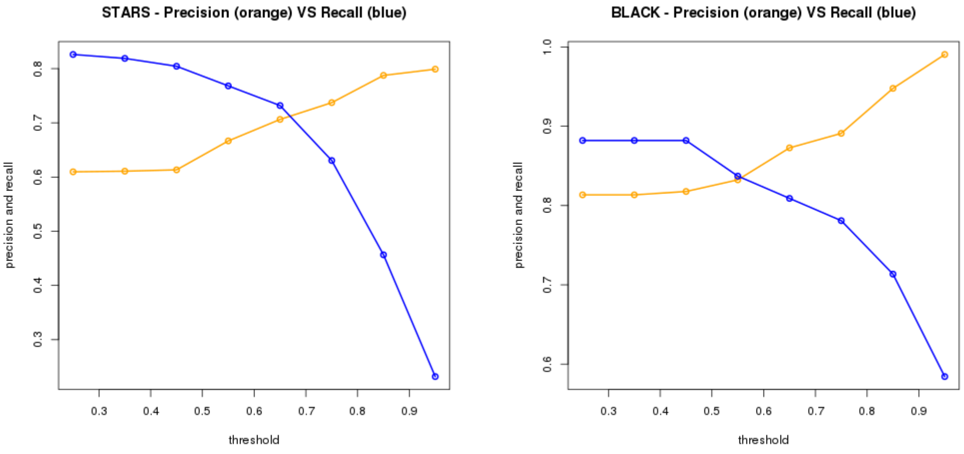
Figure 7. Examples of images classified in different classes by the three methods with low confidences and high number of players

6.2. Improving Precision at the expense of Recall with Confidence Scores

Given the previous analyses, we test if the confidence score of machine learning classifiers can be further exploited to improve classification precision, especially in the classes we are more interested in. In other word, we observe how precision (and recall) vary, by putting a minimum threshold on the confidence score and discarding all classifications with a confidence score smaller than the threshold.

Figure 8 shows the case of STARS and BLACK classes by varying the confidence score threshold of the DCNN model between 0.25 and 0.95. We observe that precision monotonically grows with the confidence threshold; of course, the higher the threshold, the more images get discarded and the lower the recall (which indeed monotonically decreases). We don't make this analysis on the CITY class, since we already got very high levels of precision (cf. Section 5).

To increase the precision on STARS images to 80%, we can accept the classifications with a confidence score greater than 0.85, decreasing recall to 45%. Using the same threshold for BLACK images, we can increase the precision to over 90%, correspondingly reducing recall to 75%. The recall cut can seem too high, but that should be interpreted in context: the images we are leaving out



(a) (b)

***Figure 8.** Precision and recall values varying the threshold on DCNN-model confidence score*

are the most confusing ones, which may not be of particular importance anyway, as they probably contain several contrasting elements at the same time.

The same analysis conducted on the color-based model leads to the same results, with precision/recall monotonically increasing/decreasing with the confidence score threshold; however, since recall on STARS images was already very low in the color-based model, the gain in precision would come at the expense of only a few classified images.

7. CONCLUSIONS

In this paper, we compared a Human Computation approach and two Machine Learning methods to classify images taken on board of the ISS, in order to employ the relevant part of them for a measurement and quantification of the light pollution issue.

The Human Computation approach was implemented through a Game with a Purpose and a related competition, which allowed us to collect a set of ∼26,000 labelled images; those images were then exploited to train two different Machine Learning classifiers, one based on the Random Forest algorithm with manually-identified color-based features and one based on Deep Convolutional Neural Networks, previously trained on a large image dataset and adapted through a transfer learning approach to our scenario (see upper part of Figure 9).

Once trained on the Human Computation results, the Machine Learning approaches worked well and showed a classification "behaviour" quite similar to the one of the GWAP players. On the other hand, the Human Computation system was very effective in collecting labels from human classifiers in quite a short time, even if its long-term sustainability is unknown.

From the point of view of our light pollution objective, CITY images – useful to monitor artificial illumination levels on Earth – were automatically classified with high precision, while BLACK and STARS pictures – used for light pollution measurement calibration – were more difficult to predict; nonetheless, we demonstrated that further analysis based on confidence scores can improve precision and that even mixed pictures can be successfully employed to identify stars and constellation of known brightness.

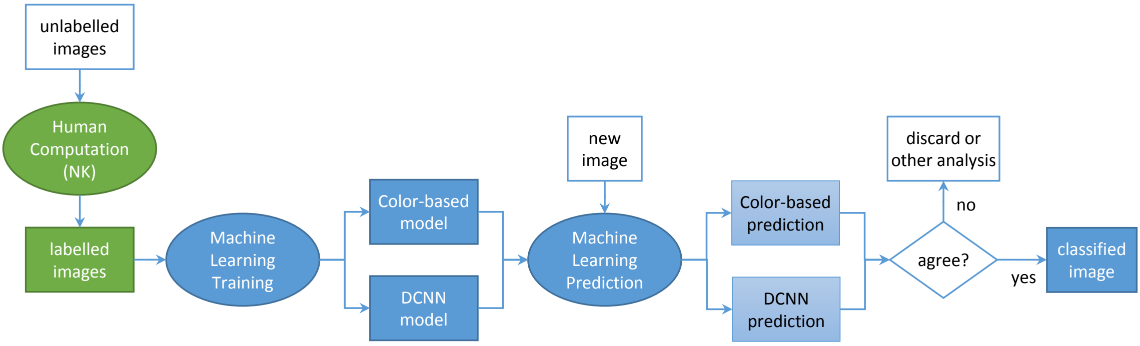


Figure 9. ***Proposed general approach for image classification***

Therefore, we think that, in the long term, a solution to our light pollution-related image classification problem can follow the process illustrated in Figure 9, mixing and combining the different approaches: the Human Computation can be only used to create the training set for Machine Learning approaches; the trained models can then be run in parallel on new unlabelled images to get their respective classifications (and related confidence scores). When the two Machine Learning classifiers agree on the image classification, this can be quite safely considered the "true" classification result; when they disagree, the image can be either completely discarded or further analyses can be performed. In the latter case, we think it useless to put those images again in our Human Computation workflow – both because of the unpredictable time to obtain a human classification and because of the concordance between the three methods on prediction "difficulty", as shown in Section 6; the classification of those images can come either from a further elaboration on confidence scores or from a human expert or from another "oracle", as in an active learning setting (Settles, 2012).

The above conclusions, however, should not be considered as applicable to any other classification case, because they reflect both the specificity of our photos and the light pollution research goal: as we explained in Section 3, the objective was to get high precision/recall for CITY pictures and high precision for STARS and BLACK images, while mixed images were of no value.

In other cases, it could be worth to adopt an *iterative process* in which all images with different automatic classifications are resubmitted in a feedback loop to humans, possibly experts. In this way, the expert-classified images could be used to retrain the classifiers, so to improve their classification power. In our case, such an iterative solution would not bring any substantial benefit to the classification result for the above-mentioned reasons.

Moreover, with respect to mixed images, in some cases it could be worth to keep *multiple classifications*, i.e. allowing two or more categories for each picture. In those cases, alternative approaches could be adopted, like keeping all classifications with a confidence value higher than a given threshold or defining additional mixed categories to merge two or more basic categories.

Acknowledgements

This work is partially supported by the STARS4ALL project (H2020-688135), co-funded by the European Commission. We thank Andrea Fiano for the implementation of the Night Knights game, Alejandro Sánchez de Miguel and Lucía García for their support in the interpretation of our experimental results from a light pollution research point of view and Esteban González Guardia for the retrieval and provision of images from the NASA repository.

8. REFERENCES

Breiman, L. (2001). Random forests. *Machine learning* 45, 1 (2001), 5–32.

Celino, I, Contessa, S, Corubolo, M, Dell'Aglio, D, Della Valle, E, Fumeo, S, and Krüger, T. (2012). Linking Smart Cities Datasets with Human Computation: the case of UrbanMatch. In *Proceedings of the 11th international conference on The Semantic Web*. Springer-Verlag, 34–49.

Celino, I, Fiano, A, and Re Calegari, G. (2016). *Games Release (initial release)*. Technical Report. STARS4ALL project deliverable, https://figshare.com/s/5bd9c9f96c8dcee121b5.

Debevec, P. (2008). A median cut algorithm for light probe sampling. In *ACM SIGGRAPH 2008 classes*. 33.

Donahue, J, Jia, Y, Vinyals, O, Hoffman, J, Zhang, N, Tzeng, E, and Darrell, T. (2014). DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. 647–655.

Hollan, J. (2009). *What is light pollution, and how do we quantify it?* Technical Report. N. Copernicus Observatory and Planetarium, Brno.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Irwin, A. (1995). *Citizen science: A study of people, expertise and sustainable development*. Psychology Press.

Kotsiantis, S. B, Zaharakis, I, and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. (2007).

Krizhevsky, A, Sutskever, I, and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*. 1097–1105.

Law, E and Ahn, L. v. (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 3 (2011), 1–121.

Lin, Y, Lv, F, Zhu, S, Yang, M, Cour, T, Yu, K, Cao, L, and Huang, T. (2011). Large-scale image classification: fast feature extraction and svm training. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. 1689–1696.

Nixon, M and Aguado, A. S. (2008). *Feature Extraction & Image Processing, Second Edition* (2nd ed.). Academic Press.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66, 336 (1971), 846–850.

Russakovsky, O, Deng, J, Su, H, Krause, J, Satheesh, S, Ma, S, Huang, Z, Karpathy, A, Khosla, A, Bernstein, M, Berg, A. C, and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. DOI:http://dx.doi.org/10.1007/s11263-015-0816-y

Sánchez de Miguel, A. (2015). *Variación espacial, temporal y espectral de la contaminación lumínica y sus fuentes: Metodología y resultados*. Ph.D. Dissertation. Universidad Complutense de Madrid.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–114.

Smith, J. R and Chang, S.-F. (1995). Single color extraction and image query. In *Image processing, 1995. Proceedings., International conference on*, Vol. 3. IEEE, 528–531.

Szegedy, C, Liu, W, Jia, Y, Sermanet, P, Reed, S, Anguelov, D, Erhan, D, Vanhoucke, V, and Rabinovich, A. (2015). Going Deeper with Convolutions. In *Computer Vision and Pattern Recognition (CVPR)*.

Von Ahn, L. (2006). Games with a purpose. *Computer* 39, 6 (2006), 92–94.

Von Ahn, L and Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.