Mapping sports interest with social network

Gabriel Sousa Ferreira[1], **Flávio Luis Cardeal Pádua**[2],
William Robson Schwartz[3], **Marco Túlio Alves Nolasco Rodrigues**[1]

[1] Universidade de Itaúna (UIT) – Itaúna, MG – Brazil

[2]Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Belo Horizonte, MG – Brazil

[3]Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

gabriel.sofe@gmail.com, cardeal@cefetmg.br,

william@dcc.ufmg.br, tulio@uit.br

Abstract. *Discovering regions that have sports interest in a set of images acquired from a scene at different times and possibly from different viewpoints and cameras is a crucial step for many applications. Physical activity can be effective at all stages of chronic disease, therefore, finding regions with the presence of physical activities might contribute to is important for the elaboration of public policies to minimize the presence of diseases such as obesity. This work addresses the problem of sport/non-sport image classification. We combine Convolutional Neural Network (CNN), traditional classifiers and geographical information to provide robust training and testing stages. As result, we achieved a high area under the curve (AUC) in a social network dataset. The experimental results show the feasibility of our proposed model. These results can be used and applied to develop public health policies based on statistics of sports interest.*

1. Introduction

Physical activity is increasingly being debated as a solution to global health problems [Committee et al. 2008, Sallis et al. 2004, Blair 2009] The World Health Organization (WHO) estimates that 3.3 million people worldwide die each year due to physical inactivity, the fourth major risk factor for global mortality [organization (WHO et al. 2017, Pratt et al. 2014].

The power of physical activities presents several benefits, such as functional health (quality of life, functional independence, balance, pain and falls prevention), cardiopulmonary health (dyslipidemias, blood pressure, coagulation, asthma, cardiac and pulmonary function), metabolic health (resistance and sensitivity to insulin, glucose uptake, metabolic syndrome, overweight, constipation, hormonal influences and sleep quality), mental health (anxiety, depression, self-concept, sleep quality and cognitive function), musculoskeletal health (bone mineralization, flexibility, strength, balance, growth, development of motor skills, muscle fiber) and cancer prevention (intestinal transit time, hormonal factors, immune function, and connection with other behaviors). On the other hand, physical inactivity is a factor that causes several complications: diabetes, cancer, obesity, hypertension, bone, cardiovascular and joint diseases and may even affect our mental health, lead to depression, stress and anxiety [Bouchard et al. 1994, Blair et al. 1999, Chodzko-Zajko et al. 2009, McAuley 1994].

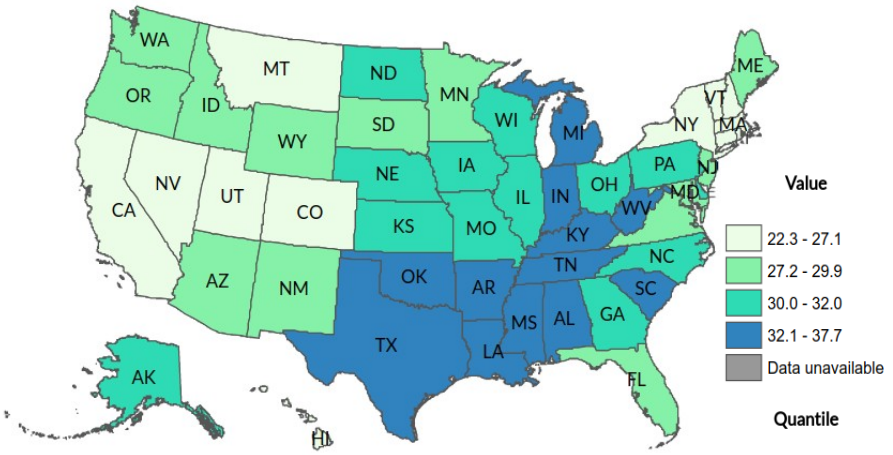

Figura 1. Percent of adults aged 18 years and older who have obesity in US. Obese is defined as body mass index (BMI) > 30.0; BMI was calculated from self-reported weight and height ($weight\ [kg]/height\ m^2$) [CDC 2018].

The Centers for Disease Control and Prevention (CDC) has created a map that provides information on the health of Americans state-by-state to learn more about the state of health and behaviors of residents of a particular state. Figure 1 shows the percentage of adults with obesity in several American states. Besides the information regarding the state of health, it is also necessary to know the behavior of the population regarding the sports interest. Therefore, a map with the social interest in sports is an important tool for several governmental actions to foster public health.

Recently, several techniques using social network have been proposed to prevent diseases. These methods, although useful, usually do not reveal the available geographical information. [Kagaya and Aizawa 2015], for instance, investigate a problem of food/non-food binary classification, without making geolocation, which can introduce considerable bias in public health policies. In another example, [Ferwerda et al. 2016] proposes to predict the personality of the user through the images posted on Instagram. Accurate public health policies are only possible when the geographical information is available, and, therefore, the strength of any public policy regarding health is strictly related to the geographical information of the available inputs.

This paper proposes a set of vision-based algorithms to map the sporting interest in a social network by image classification. We adopt convolutional neural network (CNN) as feature extractor and representation [Krizhevsky et al. 2012]. A CNN is trained using images extracted from social networks on the task of sport vs non-sport image classification. Then, the features in the last layer the CNN framework are flattened to form the feature vectors. Afterward, the classification is performed by classifier [Boland and Murphy 2001], support vector machine (SVM) [Rejani and Selvi 2009]

or logistic regression [Ciocca et al. 2015], which exploit the characteristic of CNN features to detect which group shares more similarities with an input image. Finally, with the geographic information present in the image, it is possible to determine which city of the map the image of the sport was posted [Jose and Hernandez 2017].

To the extent that by posting an image they reflect your interests, people who are positioned close to each other can be seen as reflecting the common interests. Our main contribution is a method for the generation of a map with sporting interest based on the classification of social network images. Thus, measures can be adopted by public authorities to increase sports activities, with the aim of minimizing diseases resulting from the non-practice of physical activities.

The main idea of these classification schemes is to propose flexible and general tools for performing classification sports/non-sports images in social networks. This research aims to develop an intelligent method capable of analyzing a range of data and returning quantitative information that may be of interest to individuals, health professionals, companies or governments.

2. Related Work

Our method is related to numerous works on convolution neural network, social network and image classification which we briefly discuss below.

Convolution Neural Network. Several methods based on neural network [Cherkassky and Mulier 2007] (consists of layers of interconnected compute units - neurons) are triggering umpteen applications using supervised and non-supervised methods. Weighting factors, summation function, transfer function, learning function, error Function and back-propagated value make up an artificial neuron which is arranged in layers. Features are presented to the network via the input layer, which communicates to one or more hidden layers, where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer where the answer is output. Convolutional Neural Networks [LeCun et al. 1998, Kalchbrenner et al. 2014] are a biologically-inspired class that replaces this three steps with a single neural network that is trained end to end from raw pixel values to classifier outputs.

[Jose and Hernandez 2017] performs the mapping of the activity of cats and dogs using social media and a set of georeferenced images in San Francisco. A trained CNN was used to recognize numerous visual classes. Applied to more than one million images collected from Flickr, the CNN used for detection of dogs and cats was the Inception-v3 [Szegedy et al. 2016], which proved to be very effective with an error rate of 3,46%.

Social Network. Service that allows individuals to construct a public or semi-public profile within a bounded system, articulate a list of other users with whom they share a connection and a view [Wasserman and Faust 1994]. Although the classification of sports-related images is convenient for real applications, the set of current works have the sole purpose of classifying and differentiating different types of sports[GUGULOTHU and RAO 2016]. The geolocation of sports practices to public measures using the classification of images obtained through social media was not explored up to then.

The use of social networks is increasing and results in a wide range of data that

can be explored and studied. Barbier et at. [Barbier and Liu 2011] show the main points and challenges related to information through social networks and how they use data mining methods. The data available on social networks can transcend the boundaries of the physical world in scale and extension that was previously not possible. Data mining can help researchers and professionals overcome challenges by searching for new information that until then would be hidden in a massive set of data.

The work of [Kagaya and Aizawa 2015] approaches the classification of images collected by Instagram, given an image the objective is to find out if the image contains food or not, a CNN-based approach was realized through images labeled as food / non-food.

Image Classification. Image classification is a way to gain knowledge of a large collection of images [Rejani and Selvi 2009, Boland and Murphy 2001, Jose and Hernandez 2017]. It consists of performing the grouping of images in different classes with semantic characteristics. A group of labeled images for feature extraction is provided in a supervised manner which will be used later to predict the respective class of each image. The classification system is adapted based on the user's objective and includes data balancing, data preprocessing, extraction and selection of characteristics and choice of a classification method [Lu and Weng 2007].

In [GUGULOTHU and RAO 2016], the author proposes a Bayesian method for classification of sports images, that combines four different characteristics to represent an image: Color coherence vector (CCV), Edge direction histogram (EDH), edge direction coherence vector (EDCV) e color Histogram (CH). An image database composed of several sports classes was used, in which the images were classified according to the different characteristics, dividing primarily into indoor sports and outdoor sports, after that the images belonging to the indoor category were subdivided into 3 groups, table tennis, boxing e snooker. The images belonging to the outdoor category were subdivided into water games (swimming and windsurfing), snow games (ski images and snowboard) e ground sports (cricket, kabaddi, and hockey).

Compared to image data domains, there is relatively little work on applying CNN's to sports in social network classification. Since all successful applications of CNN's in image domains share the availability of a large training set, we speculate that this is partly due to lack of large-scale sport in social network classification.

3. Proposed Aproach

The approach proposed in this work is composed of the three main steps: (1) training step; (2) prediction step, and (3) analysis step. These steps are illustrated in Figure 2 and will be described in details in the next sections.

Our method receives social network images as input(either sports or not sports images). Its output is a score that indicates if the image belongs to the sports class. Therefore, the goal of our algorithm is to decide whether there is significant social interest in sports in a region, then such information is placed into the map.

3.1. Training Phase

In the first step, the database is built with the extraction of resources and the data are submitted to cross-validation in conjunction with the machine learning algorithms. The

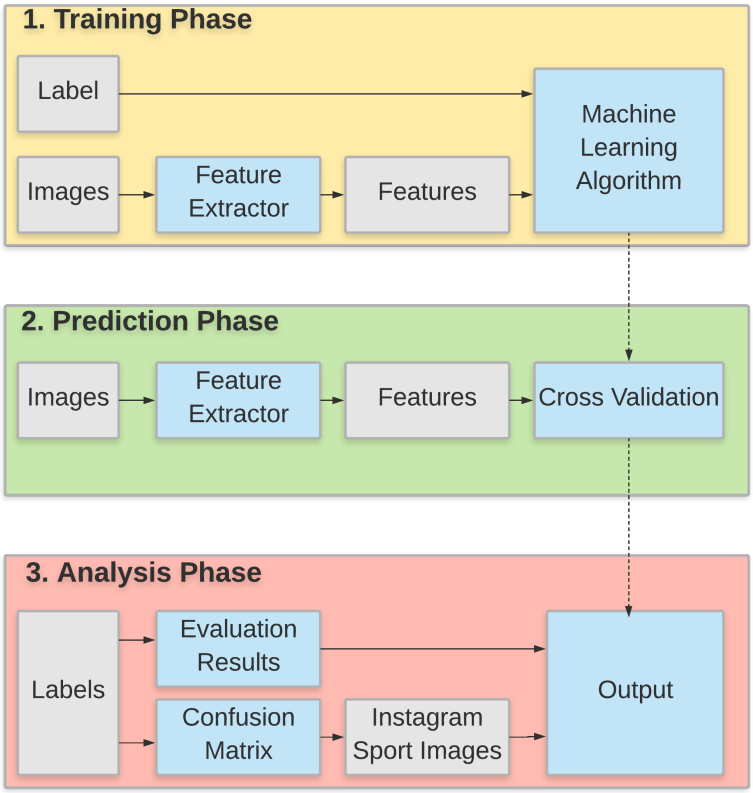

Figura 2. Diagram showing sport/non-sport classification approach employed in this method. Each block is detailed in the section with similar name.

database is collected through Flickr and Instagram. The data collected from Flickr has only sports images. On the other hand, the data collected from Instagram contain sport and non-sport images. The data collected by Flickr belong to the sports class containing images related to sports, such as athletics, basketball, cycling, football, futsal, handball, judo, fighting, bodybuilding, swimming, skateboarding, surfing, tennis and volleyball. The Instagram images were collected by region, selecting 128 cities in the state of São Paulo, Brazil.

The images were labeled identifying whether they belong to the sports group or Instagram, also specifying the different types of sports. The model responsible for extracting the feature from the images will be inception-v3, a Google's neural network model. After extracting the features from the images, three machine learning models are used to train and recognize the patterns present in the features: Support vector machine(SVM), neural network and logistic regression.

SVM. The model basically builds a hyperplane in a space which can be used for classification. The mapping in the hyperplane is performed by a kernel function, responsible for finding a linear decision rule in space looking for an ideal separation boundary, making the widest margin between the decision boundary and the vector mapped on the hyperplane [Hearst et al. 1998, Charfi et al. 2012]. The SVM model can cope well with large sets of examples and the classification process is fast.

Neural Network. Neural Network is based on a collection of artificial neurons, which can transmit a signal from one to another. The neuron that receives the signal processes the information and transmits it to the neurons connected to it [Hagan et al. 1996, Yadav et al. 2014]. The Neural Network develops its own set of important features from the data processed, without the need for a prior knowledge about the data.

Logistic Regression. Logistic Regression is a statistical model used to analyze a dataset. The Logistic Regression have the objective to find the best model to describe the relationship between the result and a set of variables. The model coefficients are generated to predict the probability of the presence of the characteristic of interest [Ciocca et al. 2015, Harrell 2001]. It provides results in terms of probability and has high reliability.

3.2. Prediction

In the second stage, the prediction of the images is performed, training the algorithms, the generalization capability will be evaluated through the set of images collected, using the features extracted by the inception-v3 model and cross-validation, using 10 folds. The cross-validation technique will receive as input, features extracted from the images and machine learning algorithms, generating as output results of testing classification algorithms.

3.3. Analysis Phase

In the last step, the data from previous steps are represented in different ways and analyzed. As a result of cross-validation and machine learning algorithms, new labels will be determined, identifying the respective groups of each image, sport or not.

Comparing the labels of the training phase with the labels generated in this last step a matrix will be constructed, which will be responsible for selecting the initial images of the Instagram that were labeled in different sports. The classification models will then be evaluated by comparing the rates of false positive and true positive. As output, our method presents a map containing the intensity of the social interest in sports in a geolocation to improve public health.

4. Experimental Evaluation

In this section, we evaluate convolution neural network (CNN) combined with support vector machine (SVM), logistic regression, neural network.

4.1. Datasets and Evaluation Protocol

Flickr Dataset. This dataset is built by forming links between images sharing common metadata from Flickr. This dataset has 487 images of 14 sports classes. The classes are athletics, basketball, cycling, football, futsal, handball, judo, MMA, tennis, swimming, skateboarding, surfing, bodybuilding, and volleyball. The images of this dataset were labeled as a sport and were manually analyzed for ensure that served as ground truth.

Instagram Dataset. This dataset only has one class and it was built with images of cities with more than 50 thousand people. It contains 26,390 sport/non-sport images collected from 128 cities in the state of São Paulo in Brazil. There is no previous knowledge about the images collected in this dataset, they are current images collected by Instagram. It

will be of responsibility of the machine learning algorithms to detect and identify sports images in the midst of these data.

Feature Extraction. The features of the images were extracted through Google's Inception v3 model trained on ImageNet, representing the images with a numeric vector with 2048 dimensions.

Machine Learning Algorithms. Three different machine learning algorithms were used to perform the experiment.

- Support Vector Machine (SVM): The model was configured using the kernel radial basis function (RBF) and its two parameters (Cost = 1 and Regression loss epsilon 0,1) [Hsu et al. 2003];
- Neural Network: The model was configured using the hidden layer activation function, ReLu (the rectified linear unit function) [Krizhevsky et al. 2012]. To optimize the weights, was used SGD (stochastic gradient descent) [Maas et al. 2013];
- Logistic Regression: The model uses the type of regularization L2 with force of cost 1 [Srivastava et al. 2014].

Cross-Validation. Cross-validation was used with 10 folds to compare and evaluate machine learning algorithms [Refaeilzadeh et al. 2009]. The models trained in the training phase represent the extracted knowledge. With the model obtained, it is applied to a test database. The test base is previously labeled, making it possible to measure the correctness rate of the model by comparing the results obtained with the labeling available on the test basis. The technique was used dividing the database into 10 folds. Of these 10 folds, 9 are used for training and one serves as a test base. The process is repeated 10 times, so each part is used once as a set of tests. At the end of the process is calculated by the average of the results obtained in each step an estimate of the quality of the model.

Metrics Evaluation. The methods used were evaluated by calculating the area under ROC, classification acuraccy, F-1, precision and recall of each method [Zheng 2015]:

- AUC: The ROC curve shows the sensitivity of the classifier by plotting the rate of true positives to the rate of false positives;
- CA: It's the ratio between the number of correct predictions and the total number of predictions;
- F-1 is a weighted harmonic mean of precision and recall;
- Precision is the proportion of true positives among instances classified as positive;
- Recall is the proportion of true positives among all positive instances in the data.

4.2. Experimental Results

In this section, it is reported results and comparisons.

Tabela 1. Evaluation Results.

	AUC	CA	F-1	Precision	Recall
Neural Network	0.990	0.994	0.993	0.993	0.994
Logistic Regression	0.994	0.994	0.994	0.993	0.994
SVM	0.996	0.993	0.991	0.991	0.993

(a) (b) (c)

Figura 3. Instagram images that was missclassified according to the confusion matrix (in red are false positives). (a) Instagram images predicted as sport by Logistic Regression; (b) Instagram images predicted as sport by SVM; (c) Instagram images predicted as sport by Neural Network.

Evaluation Results. The values computed by the displayed metrics range from 1 (perfect) and worst to 0. All models presented showed scores higher than 0.9. These results show that the combination of the CNN model and the classifiers produced good results. The modeling of the dataset with sports images with very descriptive features makes it easier for the classifiers to generalize and distinguish between sports and non-sport images, justifying the good results presented by the Table 1.

Confusion Matrix. The confusion matrix results in the number of instances between the predicted class and the current one. In this way, we can visualize the Instagram class images that were predicted as some kind of sport. The results obtained are shown in Figures 3(a), (b) and (c). The images highlighted in red are false positives, non-sport images erroneously classified as a sport.

The model Logistic Regression returned 28 images classified as a sport, in 28 images 4 are false positives (non-sport images classified as a sport). The model Neural Network returned as a result 34 images classified as a sport, in 34 images only one is considered as false positive. The model SVM returned as a result 24 images classified as a sport, all classified correctly. It is possible to observe that there is a difference between the number of images obtained and the precision of each model. The Neural Network was the model capable of returning the largest number of images and with good accuracy. The SVM application returned the smallest number of images but obtained the best accuracy among the models. The results obtained through Logistic Regression based on the images resulting from the confusion matrix were not satisfactory when compared to the other methods, despite having returned more images than the SVM was the model that most presented false positives.

Social Sport Map. The developed model makes it possible to create a heat map by relating the images obtained by the confusion matrix to its geotags. With the help of a Google API, we plotted the points belonging to each geotag of Instagram sports images on the map.

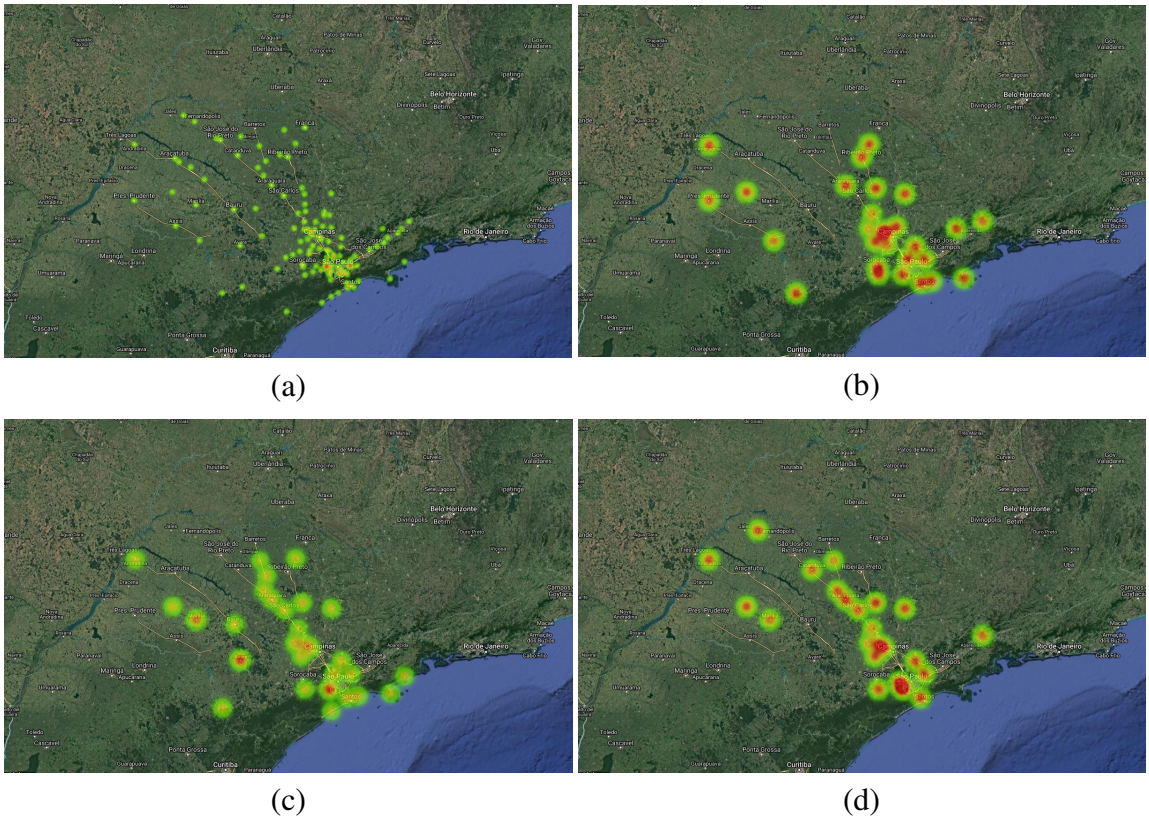

Figura 4. Social Sport map. (a) Cities considered in the model. (b) Social Sport map by Logistic Regression; (c) Social Sport map by Neural Network; (d) Social Sport map by SVM.

We computed the sports activity for each city by normalizing the number of Instagram images classified as a sport by the total number of Instagram images collected in that city.

The distributions of each model are distinct due to different results obtained by the model but present some similarity regarding the spatial distribution. Spatial distributions make sense based on the cities collected. The highest level of heat is centralized where there is a greater density of cities.

The sports Figure 4(a) maps the cities where Instagram images were collected. The social sports maps presented in the Figures 4(b), (c) and (d) demonstrate the mapping of georeferenced images classified as a sport by Logistic Regression, Neural Network, and SVM models. Cities, where sports images were not found, were not pointed by the social sports map. The intensity of each point on the map is proportional to the number of sports images found at that point, keeping the proportion of normalization computed previously. The cities indicated in the map of Figure 4(a) and not present in Figures 4(b), (c) and (d) denote more attention to public actions. On the other hand, in regions with higher sports interest intensity, attention should be paid, because it may indicate the need to distribute the sports practice.

It is notable that there is a certain similarity between the main points of interest based on the result of the proposed model and the distribution of the soccer stadium in the cities considered in the model as shown in the Figure 5. With this, it is possible to

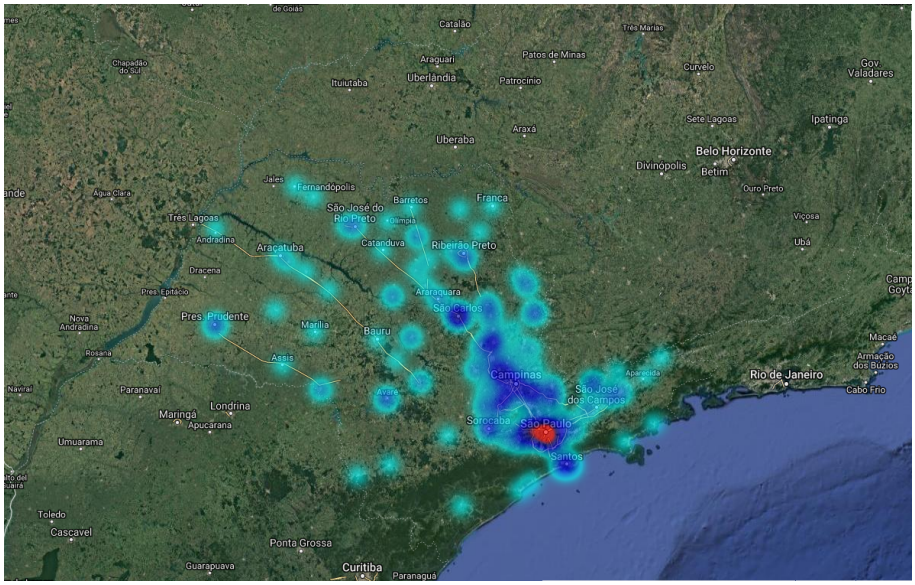

Figura 5. Soccer stadiums in the cities considered in the model [Guillermo Rivera and RSSSF 2017].

conclude that in a certain way, the social interest can reflect in the practice of sports. It is important to analyze the points where there is a strong interest in sports but there is not a great density of stadiums. On the other hand, there are many stadiums where you do not have much interest.

5. Conclusions

Sport and non-sport classification have been exploited using CNN representation and employing different classification approaches. We reported high results on two well-known social networks with visual representations. In this paper, we have CNN representation and classification techniques to design a robust and efficient sport and non-sport classifier. Results show that the combination CNN model and classifiers produce good results and demonstrate the effectiveness of our method with the mapping of interest in physical activities. In other words, the method can help in the elaboration of public policies regarding physical activity. Future works will investigate how the proposed method can be optimized to reduce the required memory and computational resources.

Referências

(2018). Centers for disease control and prevention (cdc).

Barbier, G. and Liu, H. (2011). Data mining in social media. In *Social network data analytics*, pages 327–352. Springer.

Blair, S. N. (2009). Physical inactivity: the biggest public health problem of the 21st century. *British journal of sports medicine*, 43(1):1–2.

Blair, S. N., Brodney, S., et al. (1999). Effects of physical inactivity and obesity on morbidity and mortality: current evidence and research issues. *Medicine and science in sports and exercise*, 31:S646–S662.

Boland, M. V. and Murphy, R. F. (2001). A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells. *Bioinformatics*, 17(12):1213–1223.

Bouchard, C. E., Shephard, R. J., and Stephens, T. E. (1994). Physical activity, fitness, and health: International proceedings and consensus statement. In *International Consensus Symposium on Physical Activity, Fitness, and Health, 2nd, May, 1992, Toronto, ON, Canada*. Human Kinetics Publishers.

Charfi, I., Miteran, J., Dubois, J., Atri, M., and Tourki, R. (2012). Definition and performance evaluation of a robust svm based fall detection solution. In *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, pages 218–224. IEEE.

Cherkassky, V. and Mulier, F. M. (2007). *Learning from data: concepts, theory, and methods*. John Wiley & Sons.

Chodzko-Zajko, W. J., Proctor, D. N., Singh, M. A. F., Minson, C. T., Nigg, C. R., Salem, G. J., and Skinner, J. S. (2009). Exercise and physical activity for older adults. *Medicine & science in sports & exercise*, 41(7):1510–1530.

Ciocca, G., Cusano, C., and Schettini, R. (2015). Image orientation detection using lbp-based features and logistic regression. *Multimedia Tools and Applications*, 74(9):3013–3034.

Committee, P. A. G. A. et al. (2008). Physical activity guidelines advisory committee report, 2008. *Washington, DC: US Department of Health and Human Services*, 2008:A1–H14.

Ferwerda, B., Schedl, M., and Tkalcic, M. (2016). Using instagram picture features to predict users' personality. In *International Conference on Multimedia Modeling*, pages 850–861. Springer.

GUGULOTHU, V. K. and RAO, S. M. (2016). Classification of sports images using naive bayesian classifier. *International Journal of Engineering Technology and Computer Research*, 4(4).

Guillermo Rivera, M. L. d. A. and RSSSF (2017). Sao paulo state stadia - rsssf brazil.

Hagan, M. T., Demuth, H. B., Beale, M. H., et al. (1996). *Neural network design*, volume 20. Pws Pub. Boston.

Harrell, F. E. (2001). Ordinal logistic regression. In *Regression modeling strategies*, pages 331–343. Springer.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Jose, A. S. and Hernandez, E. (2017). City-scale mapping of pets using georeferenced images. *SIGSPATIAL Special*, 8(3):5–6.

Kagaya, H. and Aizawa, K. (2015). Highly accurate food/non-food image classification based on a deep convolutional neural network. In *International Conference on Image Analysis and Processing*, pages 350–357. Springer.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lu, D. and Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3.

McAuley, E. (1994). Physical activity, fitness, and health: The consensus knowledge. *Champaign, IL: Human Kinetics*.

organization (WHO, W. H. et al. (2017). Global health risks-mortality and burden of disease attributable to selected major risks. *Cancer*.

Pratt, M., Norris, J., Lobelo, F., Roux, L., and Wang, G. (2014). The cost of physical inactivity: moving into the 21st century. *Br J Sports Med*, 48(3):171–173.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems*, pages 532–538. Springer.

Rejani, Y. and Selvi, S. T. (2009). Early detection of breast cancer using svm classifier technique. *arXiv preprint arXiv:0912.2314*.

Sallis, J. F., Frank, L. D., Saelens, B. E., and Kraft, M. K. (2004). Active transportation and physical activity: opportunities for collaboration on transportation and public health research. *Transportation Research Part A: Policy and Practice*, 38(4):249–268.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Yadav, J. S., Yadav, M., and Jain, A. (2014). Artificial neural network. *International Journal of Scientific Research and Education*, 1(6):108–117.

Zheng, A. (2015). *Evaluating Machine Learning Models A Beginner's Guide to Key Concepts and Pitfalls*.