An Indoor Sign Dataset (ISD): An Overview and Baseline Evaluation

João L. R. Almeida, Franklin C. Flores, Max N. Roecker,
Marco A. K. Braga and Yandre M. G. Costa
Departament of Informatics, State University of Maringá, Maringá, Paraná, Brazil

Keywords: Indoor signs, Visually Impairment, Indoor Signs Dataset, Convolutional Neural Networks.

Abstract: Visually impaired people need help from others when they need to find specific destinations and cannot guide themselves in indoor environments using signs. Computer Vision Systems can help them with this kind of tasks. In this paper, we present to the research community an Indoor Sign Dataset (ISD), a novel dataset composed of 1,200 samples of indoor signs images labeled into one of the following classes: accessibility, emergency exit, men's toilets, women's toilets, wifi and no smoking. The ISD dataset consists of images in different environments conditions, perspectives, and appearance that turns the recognition task quite challenging. A data augmentation technique was applied, generating 69,120 images. We also present baseline results obtained using handcrafted features, like LBP, Color Histogram, HOG, and DAISY applied on SVM, k-NN, and MLP classifiers. We further make non-handcrafted features learned using convolutional neural networks (CNN). The best result was obtained using a CNN model, with an accuracy of 90.33%. This dataset and techniques can be applied to design a wearable device able to help visually impaired people.

1 INTRODUCTION

Approximately of 285 million people have some visual impairment in the world, of whom 39 million are completely blind (Prajapati and Shah, 2016). Visual impairment is a severe condition and can turn daily tasks into challenging ones. Indoor signs are marks with symbols and/or text that communicates essential social rules of the environment: whether it is to display information, to call attention or even to show local prohibitions (Wang et al., 2013). In public places, some typical examples of indoor signs are men's and women's toilet signs, guiding signs to exit door or stairs ahead and signs to inform local wifi connection. Visually impaired people are not able to receive this information and they need to use some specialized equipment to help them to move and to interact with the environment. In the last years, the decreasing costs of hardware and the raising attention of the computer vision research field, pattern recognition, and machine learning brought new perspectives on how the technology can help visually impaired people.

Image object recognition, in the last years, has been gaining attention in the computer vision research field. Nowadays, there is a wide range of datasets in different object recognition problems which includes human face, vehicle, food, alphanumeric character and transit signs. These datasets have the main goal to present themselves as a benchmark to compare different techniques and methods in the specific problem. With the rising awareness to research in the field of autonomous vehicles, the number of datasets of recognition problems regarding public and traffic signs has significantly increased in the last years.

Despite the many datasets available, only a few of them address the indoor environment sign recognition problem, and many of them are in early stage of development (Ni et al., 2014). One can find in the literature some works presenting the use of technology to aid visually impaired people to recognize indoor signs (Ni et al., 2014) (Wang et al., 2013), (Kunene et al., 2016).

The recognition of indoor signs is not analogous to the recognition of traffic signs for two significant concerns: the highlight from the background, and appearance standardization (Ni et al., 2014). Traffic signs are heavily highlighted in the background and usually located in higher spots with good sight. Traffic signs also typically have a standardized appearance, with low or no difference in the dimensions, form and color. In opposition, indoor signs are often located

Almeida, J., Flores, F., Roecker, M., Braga, M. and Costa, Y.
An Indoor Sign Dataset (ISD): An Overview and Baseline Evaluation.
DOI: 10.5220/0007375705050512
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 505-512
ISBN: 978-989-758-354-4

in neglected spots and do not have the goal of catch much attention. Another problem is the absence of standardization, with very few cases of joint international use of symbols and colors. Some classes of signs are represented commonly with the same color, such as exit signs (green) and information sign (blue), but the vast range of forms and symbols turn the recognition into a challenging task.

In this paper, we introduce an Indoor Sign Dataset (ISD), a dataset created to support the development of researches in the indoor signs recognition task. The dataset is composed of digital images with a wide range of different sizes, forms, colors, perspectives and environmental conditions, such as illumination, occlusion and noise.

To this end, we have created classification models by using both feature engineering and representation learning approaches. Regarding the feature engineering approach, we have assessed the following well-known descriptors taken from the image processing literature: Local Binary Patterns (LBP) (Dalal and Triggs, 2005), the Histogram of Oriented Gradients (HOG) (Gonzalez and Woods, 2006), the DAISY Local Descriptor (Tola et al., 2010) and Color Frequency of the image sample. The feature vectors obtained with these methods were submitted to Support Vector Machine (SVM) (Vapnik, 1995), k-Nearest Neighbors (k-NN) (Mitchell, 1997) and Multilayer Perceptron (MLP) (Mitchell, 1997) classifiers. Regarding the feature learning approach, we have created one model using a Convolutional Neural Network (CNN) (Lecun, 1989) in such a way that it performs both feature extraction, without the need for human intervention, and classification.

This paper is organized as follows. Section 2 presents some related works with its highlights and results. The developed dataset is presented in detail in Section 3. Section 4 introduces theoretical foundations of the methods used in the models developed as a baseline of performance results in the dataset, exhibited in Section 5. Section 6 discusses the achieved results and concludes the paper.

2 RELATED WORKS

By analyzing the literature, one can find objects recognition applied to a wide range of problems, but few of them address the indoor signs recognition task. In this section we are going to describe a brief review of these works.

Ni et al. (Ni et al., 2014) present a dataset of indoor signs constituted of artificial images obtained in searches on Google Images. The dataset contains over a thousand samples unevenly distributed between 21 classes, each of those having from 40 to 80 samples. Besides, the authors presented a baseline of comparison between nine models associating different feature extraction and classification methods. Three feature extraction methods were used: Principal Component Analysis (PCA) (Jolliffe, 2005), HOG and Dense SIFT (DSIFT) (Lowe, 2004). For the classification, the authors used a MLP, SVM and kNN. The best performance was obtained when DSIFT and SVM were used in association, achieving an accuracy of 80.5%. The authors concluded that the indoor signs recognition is a challenging due to the absence of standardization, which, in some cases, as the toilet signs, may even have hundreds of different representations. The authors also reported that realistic conditions may cause the degradation of the detection and classification performance. The dataset used in that work is not publicly available.

Wang et al. (Wang et al., 2013) address the problem of recognition signs placed on doors. The roughly geometric forms as the principal feature to detect the doors and the indoor signs are recognized by combining the salience of the detected door's image region and a matching-based bipartite graph. Hardware was also developed to evaluate the method, which consists of a camera, a microphone, a portable computer and audio speakers. To evaluate the method, the authors created a dataset, which is not publicly available, with 146 samples of digital images of toilet signs, open/close lift signs and directions signs. In the first instance, the authors evaluated the method with four classes (men's toilet, women's toilet, open and close), achieving an accuracy of 86%. In the second instance, with four more classes added (up, down, left and right directions), the method achieved an accuracy of 81.6%.

Kunene et al. (Kunene et al., 2016) present a real-time system that can recognize indoor navigational signs placed over plain backgrounds. The method has four steps: detection of the sign from the background using appearance features, enhanced segmentation by masking out the background, extraction of Speeded-up Robust Features (SURF) (Bay et al., 2008) and classification using a three-search structure. The dataset used to evaluate the method, which is not publicly available, is composed of seven videoclips in which one can see eleven different signs. The method achieved an average accuracy of 67.14%. The authors also performed qualitative tests. Among ten volunteers, seven reputed the model as suitable for sign recognition, rating the usability with 3.9 on a scale from 1 to 5.

Lastly, Bashiri et al. (Bashiri et al., 2018) pre-

Table 1: Summary of Related Works.

Work	Number of classes	Features	Classifiers	Freely Available	Recognition Rate (%)
(Ni et al., 2014)	21	PCA, HOG, and DSIFT	MLP, k-NN, and SVM	No	80.5 %
(Wang et al., 2013)	4 7	Saliency Map	Matching-based bipartite graph	No	86 % 81.6 %
(Kunene et al., 2016)	4	SURF	Three-Search Structure	No	67.14 %
(Bashiri et al., 2018)	3 3	CNN with Transfer Learning		Yes	90.4[1]% 99.8[2] %

[1] Original images [2] Augmented images

Table 2: Summary of samples per classes.

Class	Number of samples
Men's Toilet	320
Women's Toilet	320
Accessibility	190
Exit	130
No Smoking	120
Wifi	120
Total	**1200**

sent the MCIndoor20000 dataset, a large-scale fully-labeled image dataset to support the development of indoor objects detection of indoor signs for hospitals and healthcare institutions. The dataset is composed of 2,055 digital images from three different indoor object categories, including 754 images of doors, 599 stairs and 702 hospital signs (with specific environment context attributes such as Clinics, Pharmacy and Ambulatory Surgery Center). A data-augmentation was employed in the training subset, resulting in more than 20,000 samples. The authors used the pre-trained CNN model AlexNet (Krizhevsky et al., 2012) to estimate the quality and quantity of attributes of the MCIndoor20000 dataset. The accuracy results were 90.4% and 99.8% for the original dataset and the augmented dataset, respectively.

Comparing all these works is not straightforward, because they are not necessarily developed using the same number of classes and not even on the same datasets. Anyway, we summarize in Table 1 some information about the related works mentioned in this section.

3 DATASET

In this work, we present to the research community an Indoor Sign Dataset (ISD)[1]. The creation of this dataset was motivated by the lack of publicly available datasets to the development of scientific works aiming to address image classification in this application domain, including images from different environments. The dataset is composed of 1,200 samples distributed in 6 classes: accessibility, emergency exit, no smoking, wifi, men's toilet and women's toilet. We choose these classes because of their high availability in public environments. Table 2 summarizes the distribution of samples per classes.

Around 90% of the samples are digital images of indoor signs at public environments in Argentina, Brazil, Bulgaria, Japan, Paraguay and United States of America captured during 2017 and 2018. The remaining 10% of the samples are digital images of indoor signs acquired by searches on public domain image datasets or e-commerce's catalog images. The dataset's samples have a wide range of appearances and capture conditions, such as illumination, perspective, sizes and occlusions. All the samples were cropped accordingly to their squared bounding boxes to minimize the background proportion in the image and stored in JPEG format. Fig. 1 illustrates some samples of our dataset. The sizes of the sample images are variable between 50×50 and 2000×2000, the dataset's size mode being 354×354.

4 FEATURE EXTRACTION AND CLASSIFICATION

In this section, we detail the process of feature extraction and classification. The feature extraction is usually made using multiple manual or statistical selection of features, commonly known as handcrafted-features. Handcrafted-features are good for a wide range of cases, but sometimes they cannot efficiently describe the complexity of the patterns of the classification with digital images. Another approach is the usage of CNN, a specialized kind of neural network for processing data that have spatial interactions, to acquire meaningful features based on a training stage. CNN methods have gained prominence recently due to its high capacity to generalize patterns in images, but it is usually costly and requires powerful hardware.

Figure 1: Some examples of samples in dataset. From top to bottom, from left to right; the class of each samples are: Emergency Exit, Accessibility, Men's toilet, No Smoking, Wifi and Women's Toilet.

[1] https://sites.google.com/view/indoorsigndatasetisd

4.1 Feature Extraction

The handcrafted features assessed in this work are the following:

- **LBP:** This method adapts a local contrast correction as a structural texture descriptor (Ojala et al., 1996). The rationale behind this method is that binary patterns in the neighborhood of a pixel are the basic properties considering the texture of an image (Costa et al., 2012). For each pixel in the image, the method differentiates gray intensities of P neighbors at a R distance. We used $P = 8$ and $R = 2$.
- **Color Histogram:** In some cases, the distribution of colors in an image can be descriptive enough to classify a sample (Gonzalez and Woods, 2006). The color histogram reveals the frequency of each element of the color space in an image. Since the RGB color space uses 8 bits per channel (red, green and blue), a channel concatenated histogram has 768 units. To reduce this dimensionality, we grouped these units in 64 buckets for each channel, thus reaching a descriptor of 196 dimensions. Since the images of the dataset have different resolutions, we resized them to 128×128 pixels resolution.
- **HOG:** This technique uses the frequency of gradients orientations of an image as a descriptor. It analyzes the distribution of pixels intensities gradients or edges directions to describe the shape and appearance of an object. It is a popular descriptor to detect objects (Dalal and Triggs, 2005). Before computing the HOG of the images in the dataset, we resized them to 128×128 pixels resolution and transformed them to a grayscale color space. The cell has the size of 16×16 pixels and is applied 1×1 in each block. The orientation parameter is set to 8.
- **DAISY:** DAISY is a feature descriptor based on oriented gradient similar to Scale-invariant feature transform (SIFT) (Lowe, 2004). This technique utilizes Gaussian pondering and one circularly symmetric kernel allowing speed and efficiency in dense calculations (Tola et al., 2010). The images were converted to grayscale and then resized to 128×128 resolution. The distance between the sample points of the descriptor was set to 16, the radius of the external ring was set to 16 pixels, and we have extracted two rings with eight samples of the histogram from each using the 8-neighborhood.

4.2 Classification

To perform classification of the images, we chose Support Vector Machine (SVM) k-Nearest Neighbors (k-NN) and MultiLayer Perceptron classifiers. We applied these classifiers with all the feature extraction methods previously mentioned in Subsection 4.1. We also collected results using CNN to perform feature learning and classification. The following subsections detail how these methods work.

4.2.1 SVM

Support Vector Machine is a supervised learning model that aims to find an n-dimensional hyperplane that maximizes the distance between elements of distinct classes (Vapnik, 1995). By definition, SVM acts as a binary linear classifier, but some alternatives allow the classification of objects that are not linearly separated or have two or more classes.

As a formal definition of SVM we have: given a set with n elements $(\vec{x_1}, y_1), (\vec{x_2}, y_2), \ldots, (\vec{x_n}, y_n)$ where $y \in \{-1, 1\}$ representing the two classes and $\vec{x_i} \in \mathbb{R}^p$ representing the features vector, the goal is to find a hyperplane n-dimensional that maximizes the distance between the two classes. The hyperplane can be expressed as a set of points $\vec{x}$ satisfying the equation

$$\vec{w} \cdot \vec{x} - b = 0 \tag{1}$$

where $\vec{w} \in \mathbb{R}^p$ is the normal vector to the hyperplane and $b \in \mathbb{R}$ is the distance parameter to the origin.

We defined these values during the training of the classifier, and the inference's equation is as

$$y_i = \vec{w_i} \cdot \vec{x_i} + b \tag{2}$$

We chose Radial Basis Function (RBF) as the kernel. The parameters of cost (c) and γ were optimized using grid-search.

4.2.2 k-NN

k-Nearest Neighbors is a supervised learning model that classifies one instance according to its k nearest neighbors (Mitchell, 1997). The central advantage of this method is the not required offline training stage. Nevertheless, the model needs to calculate the distance between all stored samples when classifying a new instance. Thus, the classification step duration is directly proportional to the dataset size.

The algorithm considers that all the problem instances match to points in a n-dimensional space ($\mathbb{R}^n$). Euclidean distance is the commonly used method to calculate the distances and formally defined

as let x be an instance and a feature set of x be $(\vec{a_1}(x), \vec{a_2}(x), \ldots, \vec{a_n}, (x))$, where $a_i(x)$ indicates the value of i-th attribute of instance x. The Euclidean distance between two instances x_j and x_k is defined by the equation 3.

$$d(x_j, x_k) = \sqrt{\sum_{i=1}^{n} (a_i(x_j) - a_i(x_k))^2} \tag{3}$$

The most voted class according to the k nearest neighbors defines the classification of the new instance. The equation 4 demonstrates how to perform it.

$$x_i \leftarrow c : \underset{c}{arg\ max}\ k_c \tag{4}$$

where k_c is the number of neighbors belonging to the class c in the k nearest neighbors.

During the evaluation step using k-NN, we used $k = 5$, different values of k were used, like $1, 3, 7$, and 9, but did not present better results.

4.2.3 Multilayer Perceptron

The MLP is used as general mapping between the inputs and outputs variables (Mitchell, 1997). It is a supervised learning model that uses the back-propagation error. The MLP has one or more hidden layers that are fully-connected. We used a MLP with one hidden layer to connect the inputs x_i, where i the index of a sample is, with the output $\widehat{y_i}$ predicted using the logistic function presented on equation 5.

$$\widehat{y_i} = \frac{1}{1 + e^{-x_i}} \tag{5}$$

The error of a prediction is defined as

$$E = \sum \frac{1}{2} (\widehat{y_i} - y_i)^2 \tag{6}$$

where y_i is the true class of i-th sample. We update the weights using the stochastic gradient-based optimizer (Kingma and Ba, 2015), the learning rate used is 10^{-3}.

4.2.4 Convolutional Networks

Convolutional networks (Lecun, 1989), also known as Convolutional Neural Networks (CNN), are a kind of neural network that employs the mathematical convolution operation instead of matrix multiplication. Due to its characteristics of sparse connections, parameter sharing and equivariant representations, Convolutional Networks leverage advantages when working with spatially related data, such as digital images (Goodfellow et al., 2016).

The basic unit of a Convolutional Network is the convolutional layer, where the input is convoluted with the kernels at some stride and outputs a feature map. This feature map can also pass through an activation function and is subsampled into pooling steps. Usually, in a single convolutional network, there are many convolutional layers connected in series or parallel until the network output.

Convolutional networks are commonly associated with standard fully-connected neural networks, such as multilayer perceptrons (MLPs) (Mitchell, 1997). In those cases, a convolutional network usually receives a "raw" digital image and outputs a feature map. Then, the MLP takes this feature map as an input and performs the classification (Lecun, 1989). The main advantage of using Convolutional Networks is its purpose as a trainable mechanism of dimensionality reduction. Thus, it can adjust its parameters to create a more generalizable feature map of the input it receives, decreasing the necessity of manual feature extraction.

To simplify the model and minimize setup of hyperparameters, we designed all the stages of the model with the same principles, as it can also be seen employed in the works of (Krizhevsky et al., 2012) and (Simonyan and Zisserman, 2015). The architecture of the model can be summarized in Table 3.

The model's input has a $64 \times 64 \times 3$ shape as it receives as input an RGB image with 64 pixels in width and 64 pixels in height. The input passes through a series of convolutional layers with tiny 3×3 filters. We fixed the convolution stride at one unit in each axis. The output of the convolution also has the same size as the input by adding a zero-valued border in the input.

A leaky rectifier activation function (LReLU) is set up in the output of the convolution. We chose the LReLU instead of the traditional rectifier (ReLU) as demonstrated by (Maas et al., 2013), which in some cases ReLU activation could "kill" some neurons and cannot activate anymore. Each stack of convolutional layers is finalized with a spatial pooling, performing

Table 3: Architecture configuration of the model.

#	Type	Input Units	Parameters	Stride (x, y)
1	Convolutional	$64 \times 64 \times 3$	$3 \times 3 \times 64$	$(1,1)$
2	Convolutional	$64 \times 64 \times 64$	$3 \times 3 \times 64$	$(1,1)$
3	Max-Pooling	$64 \times 64 \times 64$	2×2	$(2,2)$
4	Convolutional	$32 \times 32 \times 64$	$3 \times 3 \times 128$	$(1,1)$
5	Convolutional	$32 \times 32 \times 128$	$3 \times 3 \times 128$	$(1,1)$
6	Max-Pooling	$32 \times 32 \times 128$	2×2	$(2,2)$
7	Convolutional	$16 \times 16 \times 128$	$3 \times 3 \times 256$	$(1,1)$
8	Convolutional	$16 \times 16 \times 256$	$3 \times 3 \times 256$	$(1,1)$
9	Max-Pooling	$16 \times 16 \times 256$	2×2	$(2,2)$
10	Flat	$16 \times 16 \times 256$	16,384	—
11	Fully-connected	16,384	256	—
12	Fully-connected	256	128	—
13	Fully-connected	128	6	—
14	Softmax	6	6	—

a maximum-value subsample over a 2×2 unit square with a stride of 2.

In addition, a stack of two 3×3 convolutional layers (without spatial pooling) has the same effect of one 5×5 convolutional layer, but it includes two nonlinear rectifications instead of one, which makes the decision function more discriminative (Simonyan and Zisserman, 2015).

Following the convolutional stage, there is a stage of three fully-connected layers receiving as input the result of the convolutional stage. The fully connected layers have the structure similar to multi-layer perceptron (MLP). The first layer has 256 units, the second layer has 128 units and the third one, as it performs the classification, has 6 units. An LReLU activation with $\alpha = 0.01$ is also equipped in all the fully connected layers. The last stage of the model takes the fully-connected stage's output, applies a normalized exponential function (softmax) and "squashes" a vector of arbitrary real values into probabilities that add up to one.

The training process follows, in general, the works of (Krizhevsky et al., 2012) and (Simonyan and Zisserman, 2015). The training consists of optimizing a multinomial logistic regression (softmax regression) since all the classes are exclusive. The error of the model was defined as the cross-entropy of the prediction and the label of the sample. The cross-entropy is defined as

$$H(\hat{\mathbf{y}}, \mathbf{y}) = -\mathbf{y} \cdot \log(\hat{\mathbf{y}}) \tag{7}$$

where $\mathbf{y}$ is the sample label, $\hat{\mathbf{y}}$ is the sample model's prediction and $\cdot$ is a dot product.

The training also employs a dropout regularization (Srivastava et al., 2014) in the standard neural network layers. Consider a neural network with L hidden layers and let $l \in \{1, 2, \ldots, L\}$ index the hidden layers of the network. Let also $\mathbf{z}^{(l)}$, $\mathbf{y}^{(l)}$, W^l and $\mathbf{b}^{(l)}$ denote the inputs, the outputs, weights and biases of layer l, respectively; where $\mathbf{y}^{(0)} = \mathbf{x}$. The feed-forward operation of a standard neural network can be described as

$$\begin{aligned} z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)} \mathbf{y}^{(l)} + b_i^{(l+1)}, \\ y_i^{(l+1)} &= f(z_i^{(l+1)}) \end{aligned} \tag{8}$$

for $l \in \{0, 1, \ldots, L-1\}$ and for any i hidden unit where f is an activation function. With dropout regularization, the feed-forward operation becomes

$$\begin{aligned} r_j^{(l)} &\sim \text{Bernoulli}(p) \\ \tilde{\mathbf{y}}^{(l)} &= \mathbf{r}^{(l)} \circ \mathbf{y}^{(l)} \\ z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)} \tilde{\mathbf{y}}^{(l)} + b_i^{(l+1)} \end{aligned}$$

$$y_i^{(l+1)} = f(z_i^{(l+1)}) \tag{9}$$

where $\circ$ denotes the entrywise product of the operands. For any layer l, $\mathbf{r}^{(l)}$ is a vector of Bernoulli independent random variables each of which has a probability p of being 1. This vector is sampled and multiplied entrywise with the outputs of that layer, $\mathbf{y}^{(l)}$, to create an output $\tilde{\mathbf{y}}^{(l)}$ that the next layer uses as input. We set $p = 0.5$ during the training phase.

Mini-batch gradient descent with the Adam Algorithm (Kingma and Ba, 2015) was employed as the optimization approach. The batch size was set to 256 and the exponential decay rates β_1 and β_2 were set to 0.9 and 0.999. The learning rate was set at 10^{-3} and do not decay in the training procedure. We trained the model for fixed 32 epochs.

Since inaccurate initialization of the parameters in a neural network model can lead to stall the optimizer due to the instability of the gradient in training nets, we employed an optimized heuristic initialization schemes. We followed the initialization scheme proposed by (He et al., 2015) for the convolutional layers and the initialization proposed by Glorot and Bengio (Glorot and Bengio, 2010) for all the fully-connected layers. All the initial values of bias parameters were set to zero.

To prevent over fitting in CNN many images are necessary for the training subset, so we employed a data augmentation technique in the training set which is mainly composed of random transformations in the color (brightness, contrast and saturation) and space (translation, rotation, blur and sharpening) of the image. Figure 2 exhibits some examples of samples after the augmentation process. We performed the augmentation at a rate of 64, i.e., each sample in the training set generates 63 augmented samples. Consequently, each fold increased from 120 to 7.680 samples. Each cross-validation contains 69.120 trai-

Figure 2: From top to bottom, some augmented examples for each sign class: men's toilet, women's toilet, accessibility, emergency exit, no smoking and wi-fi connection.

ning images and 120 original test images. All augmented images are available in this dataset.

We choose not to apply a transfer learning technique with well-known models (such as the VGG-net (Simonyan and Zisserman, 2015) or the AlexNet (Krizhevsky et al., 2012)), because these models usually are designed for its usage in environments with abundance of computational resources with dedicated hardware for graphical or tensor processing. We designed our model considering the computational resources' constraints of an embed device. We also considered the model inference time as a main factor, which can be as close as possible of a real-time.

5 RESULTS

In this section, we present the results of the experiments using our dataset. To evaluate the models that execute the features extraction and classification in two steps, we used a stratified cross-validation technique with 1,200 images partitioned into ten folds.

The results are summarized in Table 4. The 90.33% accuracy by model #13 states the capability of the Convolutional Networks to generalize the most descriptive features of the samples. The models that employed the Color Histogram and LBP as feature extraction presented the lowest accuracy performance using any classifier. We conjecture that the absence of standardization of indoors signs is the primary cause for these models to achieve such accuracy. Another intriguing characteristic is that models with k-NN and MLP classifiers always have low accuracy hit when comparing the same feature extraction method.

The models #7 and #10 that employed the HOG and DAISY with an SVM classifier scored a significant accuracy. Although they do not achieve results such as model #13, these methods do not require such computational effort as convolutional networks to train or perform and are suitable for applications in embedded and wearable devices.

The confusion matrix of the best cross-validation using model #13 is summarized in Table 5. We perceived that the most common errors are in the classification of the class of toilet signs. We suppose this behavior is due to the similarity of appearances of these two signs. The Figure 3 illustrates some very similar samples of the men's and women's toilet sign class present in our dataset.

Figure 3: Samples of the classes Men and Women Toilet signs.

Table 4: Experimental Results.

#	Feature	Classifier	Mean Accuracy	Standard Deviation
1	LBP	SVM	0.4500	0.0306
2	LBP	k-NN	0.3541	0.0388
3	LBP	MLP	0.2833	0.0271
4	Color Histogram	SVM	0.3458	0.0408
5	Color Histogram	k-NN	0.3333	0.0554
6	Color Histogram	MLP	0.3411	0.0341
7	HOG	SVM	0.7333	0.0294
8	HOG	k-NN	0.5708	0.0442
9	HOG	MLP	0.7166	0.0349
10	DAISY	SVM	0.7125	0.0407
11	DAISY	k-NN	0.5125	0.0572
12	DAISY	MLP	0.7083	0.0344
13	CNN		0.9033	0.0163

Table 5: Confusion Matrix using CNN.

	Men's Toilet	Women's Toilet	Accessibility	Exit	No Smoking	WiFi	Total
Men's Toilet	27	4	1	0	0	0	32
Women's Toilet	2	29	0	1	0	0	32
Accessibility	0	0	19	0	0	0	19
Exit	0	0	0	13	0	0	10
No Smoking	0	1	0	0	11	0	12
WiFi	0	0	0	1	0	11	12
Total	29	34	20	15	11	11	120

Using the Table 5 we can summarizes the values of precision, recall and f-measure of each category using model #9. The "men's toilet" sign presents the lowest recall value, i.e., the model often classifies this class of sign as others classes. The macro-f, which denotes the mean of the f-measure, is evaluated as 0.930. This value shows that, in general, the classifier is stable and does not favors some classes over others.

6 DISCUSSION AND CONCLUSIONS

This paper presents an Indoor Sign Dataset (ISD) with some of the principal classes of indoor environment signs. We performed experiments on this dataset using different methods of feature extraction and classification algorithms, both in handcrafted features mode and by using representation learning as well. The collected results were summarized and presented according to the principal evaluation metrics regarding pattern recognition systems.

We emphasize the importance of having a public dataset to improve the results about this problem. It is not possible to compare the results with other techniques, because they are the first using this dataset, but the results are encouraging, because they present higher accuracy than works related to it and use real pho-

tos of different environments.

As a future work, we consider the application of different feature extraction methods, such as co-occurrence matrix and SURF, and different models of CNNs to the classification task to improve the results. Furthermore, we intend to design a wearable device to incorporate the software.

ACKNOWLEDGEMENTS

The authors would like to thank all the people who contributed with images for the elaboration of this dataset, the Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil (Capes) and the National Council for Scientific and Technological Development (CNPq) for its financial support on this work.

REFERENCES

Bashiri, F. S., LaRose, E., Peissig, P., and Tafti, A. P. (2018). Mcindoor20000: A fully-labeled image dataset to advance indoor objects detection. *Data in Brief*, 17:71 – 75.

Bay, H., Ess, A., Tuytelaars, T., and Gool, L. V. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 – 359. Similarity Matching in Computer Vision and Multimedia.

Costa, Y., Oliveira, L., Koerich, A., Gouyon, F., and Martins, J. (2012). Music genre classification using lbp textural features. *Signal Processing*, 92(11):2723 – 2737.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Gonzalez, R. C. and Woods, R. E. (2006). *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

Jolliffe, I. (2005). *Principal Component Analysis*. American Cancer Society.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

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 - Volume 1*, NIPS'12, pages 1097–1105, USA. Curran Associates Inc.

Kunene, D., Vadapalli, H., and Cronje, J. (2016). Indoor sign recognition for the blind. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, SAICSIT '16, pages 19:1–19:9, New York, NY, USA. ACM.

Lecun, Y. (1989). *Generalization and network design strategies*. Elsevier, Zurich, Switzerland.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

Ni, Z., Fu, S., Tang, B., He, H., and Huang, X. (2014). Experimental studies on indoor sign recognition and classification. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 489–494.

Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59.

Prajapati, R. and Shah, P. (2016). Design and testing algorithm for real time text images: Rehabilitation aid for blind. *International Journal of Science Technology & Engineering*, 2(11):275 – 278.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

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

Tola, E., Lepetit, V., and Fua, P. (2010). Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Wang, S., Yang, X., and Tian, Y. (2013). Detecting signage and doors for blind navigation and wayfinding. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 2(2):81–93.