Global image features for scene recognition invariant to symmetrical reflections in robotics

D. Santos-Saavedra, X. M. Pardo, R. Iglesias, V. Álvarez-Santos, A. Canedo-Rodríguez, C. V. Regueiro

***Abstract*—Scene understanding is still an important challenge in robotics. Robots must be aware of the kind of the environment where they move. In our case we plan to combine scene recognition with a multisensor localization system to allow the retrieval of knowledge (as robot controllers), according to *where the robot is*. In this paper we analyse the impact of several global image representations to solve the task of scene recognition. The performances of the different alternatives were compared using a benchmark of images taken in the Centro Singular de Investigacion en Tecnoloxias da Informacion (CITIUS), at the University of Santiago de Compostela, since this is the environment where the robot moves. The results are promising not only regarding the accuracy achieved, but mostly because we have found a holistic representation that allows the correct classification of images corresponding to rooms that are symmetrical (we increased the size of the test set including images that are obtained by specular reflection from other images also included in the same set).**

***Index Terms*—Scene recognition, holistic representations, invariance to symmetries, CENTRIST, spatial pyramid, Local Difference Binary Patterns.**

D. Santos-Saavedra, X. M. Pardo, R. Iglesias, A. Canedo-Rodríguez and V. Álvarez-Santos are with the CITIUS, University of Santiago de Compostela, Spain. E-mails: roberto.iglesias.rodriguez@usc.es, zeugin18@gmail.com, xose.pardo@usc.es, adrian.canedo@usc.es, victor.alvarez@usc.es

C. V. Regueiro is with the Department of Electronic and Systems, University of A Coruña, Spain. E-mail: cvazquez@udc.es.

I. Introduction

ALTHOUGH the industrial sector has been the main user of robots for many years, nowadays there is a clear shift towards the service sector. This expansion of service robots responds to a strong demand of an ageing society, in which the urban centres continue to grow in size and density. The growth of service robots has not progressed more due to technological limitations that make their progress difficult. One of the limitations of these robots is *scene understanding*. Today's robots are unable to understand their environments, they are not aware if they are moving in a room that is similar to another one where they have been moving previously. Robots are not able to distinguish whether they are moving for instance in a kitchen or living room. Scene based robot navigation is still a very difficult and open task. Knowing "where am I" has always being an important research topic in robotics and computer vision [1]. Thus, the automatic detection of representative situations – a room without people that can be tidied up, people sitting in a sofa or children playing, people that have just entered home, amongst others— would represent an important qualitative leap forward, as robots would stop from being passive and transform into robots with "initiative". On the other hand, the identification of similarities amongst different working spaces would

allow the retrieval of controllers. These are some simple examples of the benefits of scene understanding to mention but a few.

The retrieval of robot controllers considering the location of the robot is, in fact, one of the aspects we are interested in. As part of the project TIN2012-32262, we are interested in the development of learning algorithms able to link *what* the robot is learning to *where* the robot learns it. If we consider the main scenarios where future robots are expected to move, or the tasks they are expected to carry out (assisting with the house work, security and vigilance, rehabilitation, collaborating in the care-entertainment, etc), we immediately realize that this new generation of robots must be able to learn on their own. Robots should be able to learn from what the user does, but also from their interaction with the physical and social environment. Furthermore, this adaptation should not be constrained to a time interval, but on the contrary it should be continuous, i.e., during the life of the robot. Nevertheless, this makes necessary the achievement of *knowledge retrieval from scene recognition and robot localization*. Like people, the behavior of the robots must change not only as a consequence of the time goes by, but also according to where the robots are. All the information about the environment will be used to retrieve robot controllers that have been previously learnt.

In the past, we already developed a robust multi-sensor system for mobile robot localization [3]: a localization system that can combine data supplied by a 2D laser range finder, a Wi-Fi positioning system, and a magnetic compass to estimate the pose of the robot relative to a map. In this paper we describe the work we have done to get a different kind of information about the localization of the robot, the recognition of the scene. In this case we assume that the robot will take one observation at some location and it will use a classifier to identify this observation. The different classes represent general categories of places and not particular instances. That means that the robot assigns the same label to different places that pertain to the same category.

There are two main options to address the problem of scene recognition: using local or global descriptors. Regarding the first option (local descriptors), the usual approach consists on selecting some salient points in the image (using SURF, SIFT or similar strategies), and then building a description from this unstructured set of points. A sufficient number of such key points have to be detected and, in addition, they should be distinguishable and stable features that can be accurately localized. In contrast, global descriptors summarize the whole image in a single descriptor, being GIST and CENTRIST the most representative strategies. Several studies in scene perception have shown that humans are able to understand the general context of novel scenes even when the presentation time is very short ($< 100msec$), when images are not fully attended to, or are presented blurred. Oliva and Torralba [4], [5] showed that scenes which belong to the same category, normally have the same spatial layout properties, and proposed a holistic approach to build the *gist* of the scene. This *gist* of a scene represents all that information that an observer can comprehend after a single glimpse of an image, and it is most commonly associated with low-level global features such as color, spatial frequencies and spatial organization. This experience of understanding everything in one glimpse, is similar to what we experience watching television and flipping rapidly through the channels. Torralba's representation provides low resolution information about the frequency contents of the scene. Image descriptors characterize the global view of the image without the use of localized information. The low spatial resolu-

tion of Torralba's representation guarantees certain robustness regarding the change of viewpoint. Nevertheless, the image regions that are considered to calculate Torralba's descriptors are pre-established and do not depend on the contents of the image. Wu and Regh [7] also used a holistic approach and proposed *CENTRIST* (Census Transform Histogram), a representation that captures properties, such as, rough geometry and generalizability by modeling the distribution of local structures. CENTRIST is easy to implement, has nearly no parameters to tune, and is invariant to uniform illumination variations. Besides, it is extremely fast (from the computational point of view), this is a very important characteristic in robotics, specially if we consider that we want the robot to identify the environment while it is interacting with it. Because of this, in this paper we analyse the use of image descriptors inspired by CENTRIST to carry out the classification of the environments where the robot moves.

II. Classification based on global features

As we mentioned in the introduction, in this work we want to analyse the use of different global visual descriptors to classify scenes from images taken with the robot camera. The goal is to assign the same category to places with the same structures.

One of the first descriptors we have considered is the CENsus Transform hISTogram (CENTRIST) [7]. The Census Transform (CT) is a non-parametric local transform based on the comparison amongst the intensity value of each pixel of the image with its eight neighboring pixels, as illustrated in Figure 1. As we can see in this figure, if the center pixel is bigger than (or equal to) one of its neighbors, a bit 1 is set in the corresponding location. Otherwise a bit 0 is set. The eight bits generated after all the comparisons have to be put together following always the same order, and then they are converted to a base-10 number in the interval $[0, 255]$. This process maps a 3×3 image patch to one of 256 cases, each corresponding to a special type of local structure, and it is repeated for every pixel of the original image. The Census Transform (CT) is also referred to as the *Local Difference Sign Binary Pattern* (LSBP). This name is due to the fact that the way the Census Transform is obtained is equivalent to the *Local binary pattern code* $LBP_{8,1}$, while the *Sign* word in the name reflects the fact that this code only considers the sign of the comparisons (1 if the sign of the comparison is positive, cero otherwise).

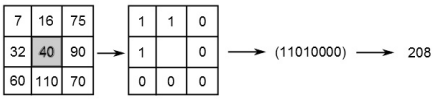

Fig. 1. Illustration of the Census Transform Process on a 3×3 image patch

Obviously the information about the magnitude of the intensities is lost in the $LSBP$. This is the reason why X. Meng et al. [6] suggested the use of the so-called *Local Difference Magnitude Binary Pattern* (LMBP) as a further piece of information to complete the representation of the image. In this case, the LMBP is computed as the intensity difference between the center pixel and its neighboring pixels. In this case, if the difference in intensity amongst the center pixel and one of its neighbors is higher than a threshold T, a bit 1 is set, otherwise a bit 0 is set. Like in the case of the Census Transform, the eight bits generated after all the comparisons have to be put together following always the same order, and then they are converted to a base-10 number.

Thus, for every image we can compute the holistic representation given by the combination of the *LMBP* and the *LSBP* (Figure 2). Once this process is over, we can compute the histogram of LSBP and LMBP as a feature representation. Both the LSBP and the LMBP histograms are 256 dimensions (the bins of the histograms are each one of the

Fig. 2. One of the holistic representations that will be used to classify the images consists on the combination of the Local Difference Sign Binary Pattern and the Local Magnitude Binary Pattern.

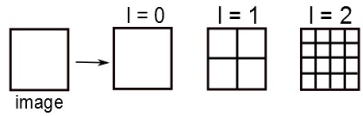

Fig. 3. Spatial Pyramid Splitting with non-overlapping blocks. This figure shows the partitioning of the image in different blocks for the first three levels $l = 0, 1, 2$.

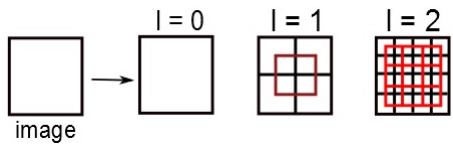

Fig. 4. Spatial Pyramid Splitting with overlapping blocks. This figure shows the partitioning of the image in different blocks for the first three levels $l = 0, 1, 2$.

values that the LMBP and LSBP can take), therefore, the new feature representation is 512 dimensional (256×2). It is a common practice to suppress the first and the last bins of these histograms, due to noise cancellation and the removal of not significant information, that is the reason why the dimension that appears in Figure 2 is 508.

As we will see in the experimental results, in this paper we have analyse the performance of classifier when both alternatives are used, i.e., the CT histogram, and the combination of the LSBP and LMBP histograms.

III. Spatial Pyramid Splitting

Lazebnik et al. [8] suggested the use of *spatial pyramids* when holistic approaches are used for image categorization. This technique works by partitioning the image into increasingly fine sub-regions and computing histograms of features inside each region. The resulting *spatial pyramid* is a simple and computationally efficient extension of an order less *bag of words*. The operation of partitioning an image into blocks and compute histograms in these blocks has been carried out very often in computer vision, both for global image description and for local description of interest regions. Nevertheless, the spatial pyramid method addresses the issue of determining what is the right subdivision scheme by working with multiple resolutions. A schematic illustration of the spatial pyramid splitting is provided in Figure 3. This method partitions an image into a sequence of spatial blocks at resolutions: 0,...,L, such that the grid at level 0, has just a single block (and the whole image is treated as unique entity), at level 1, the image is subdivided into four disjoint quadrants, yielding to four feature histograms, and in general the grid at level l has $(2^l)^2$ blocks.

There are some researchers who have suggested the usage of overlapping spatial regions, instead of breaking the image into a set of disjoint blocks, actually several combinations have been tried (half size, circular and rectangular overlapping areas, etc.) [9]. In our case we will analyse the partition shown in Figure 4. With this overlapping partition, the number of blocks at each level is determined as $(2^l)^2 + (2^l - 1)^2$.

As we will see in the experimental results, we have analysed the performance of the classifier with both alternatives, overlapping and non-overlapping blocks.

Due to the different size of the blocks obtained in the different partitions, the number of features in each one of them would be different. This makes necessary the normalization of all histograms by the total weight of the number of features. An efficient way of doing it is by enforcing the total number of features in all images to be the same, i.e., resizing the image between different levels so that all blocks contain the same number

of pixels. Finally, the feature representations obtained for all blocks are concatenated to form an overall feature vector for each image (Figure 5) . This means that at each level the size of the image descriptor is $(508 \times number\_of\_blocks)$ (where 508 is the number of dimensions of the vector with the concatenated LSBP and LMPB histograms). If we consider that this size would now be multiplied by the number of levels of the spatial pyramid, it is easy to understand that we would end with a very high dimensional descriptor. Because of this, and considering the fact that the histogram representation of LSBP and LMBP are correlated with each other [6], it is possible to use principal component analysis (PCA) to reduce their size. In our case, we have carried out this principal component analysis (using a set of images taken in the environment where the robot moves), and as a result of it, we could project each 508 descriptor to a new space of 80 dimensions (Figure 5).

IV. Experimental Setup

We have created an image set obtained in the Centro Singular de Investigacion en Tecnologias de la Informacion (CITIUS), at the University of Santiago de Compostela, to analyse the performance of a classifier that categorizes each image using the different holistic representations described in the previous sections: CT, LSBP+LMBP, Spatial Pyramid with and without overlapping areas, etc.

There are two reasons why we chose to create a new set of images: First of all, there are many published results where the performance of global image representations has been analysed in outdoor images, but this is not what we want. Second, this is the environment where our robot (a Pioneer 3DX) moves, so in this case the accuracy we get with these images is the real accuracy the robot will have to deal with. We will use this set of images as a benchmark to compare the performance of the classifier using the different holistic representations mentioned before.

TABLE I
DESCRIPTION OF THE SET OF IMAGES TAKEN AT THE CITIUS RESEARCH CENTRE, UNIVERSITY OF SANTAGO DE COMPOSTELA, SPAIN. THIS SET OF IMAGES HAS BEEN USED AS BENCHMARK FOR THE DIFFERENT TESTS DESCRIBED IN THIS PAPER

Type of room \Class	Number of img.	Class
Common staff areas (ground floor)	52	0
Kitchen	28	1
Assembly Hall	29	2
Common staff areas (S1 floor)	53	3
Entrepreneurship Laboratory	34	4
Laboratories 1, 2 and 3	106	5
Instrumentation Laboratory	18	6
Office	18	7
Common staff areas (first and second floors)	73	8
Staircase	16	9
Robotics Laboratory	21	10

As we can see in Table I, our benchmark is made up of images taken in different areas within the research centre. According to this table, there are common staff areas that are identified as different classes, this is because these staff areas are very different from each other and therefore we considered that they should be labeled differently. Something similar happens with the laboratories, their size and furniture is so different that we considered that they should be identified as different scenarios.

V. Experimental results

Using the benchmark of images described in the previous section, we have analysed the performance of a Support Vector Machine (SVM) as classifier, using different image descriptors (Table II).

As we can see in Table II, using the combination of the LSBP and the LMBP is better than using only the Census Transform. We have analysed the performance when different values of the parameter T (described in section II) has been used to get the LMBP. The best performance is obtained with $T = 50$. On the other hand we have also analysed the performance of the

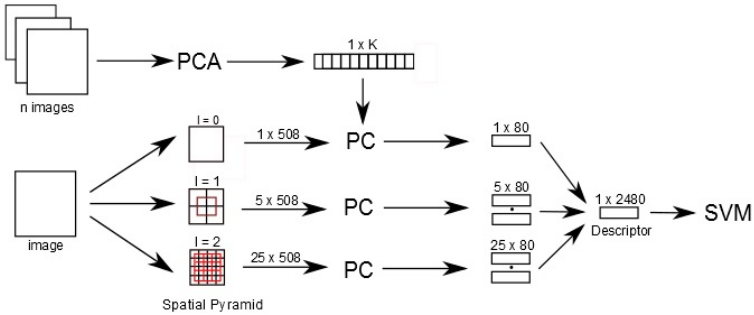

Fig. 5. Schematic representation of the combined use of the global features described in section II and the Spatial Pyramid Splitting. To classify any image the spatial pyramid partitions the image into different blocks, being necessary to get the feature description (CT, or the combination LSBP+LMBP) for every block. The vector that describes every block is projected into a new space with fewer dimensions (using the results of a principal component analysis). Finally, the transformed vectors are combined together and used as input to a SVM that classifies the image.

TABLE II
PERFORMANCE OF A SVM CLASSIFIER USING DIFFERENT FEATURE REPRESENTATIONS. THESE RESULTS ARE THE AVERAGE PERFORMANCE OBTAINED AFTER USING FOUR TEST SETS (4-FOLD CROSS-VALIDATION)

Feature representation	Performance
Census Transform	64.74%
LSBP+LMBP (T=50)	70.01%
LSBP+LMBP (T=100)	67.47%
LSBP+LMBP (T=150)	67.28%
Census Transform+S. Pyramid	76.5%
LSBP+LMBP (T=50)+S. Pyramid	78.32%
LSBP+LMBP+S. Pyramid (non-overlpping blocks)	78.96%

classifier when we include different levels of abstraction in the descriptor of the image (Spatial Pyramid described in section III). In this case we have used the Spatial Pyramid Splitting in three levels ($l = 0, 1, 2$). We have analysed the Spatial pyramid when the Census Transform is computed in every block in which the image is divided, or when the LSBP+LMB ($T = 50$) is used instead. Considering the results we obtained, we can conclude that the best performance is obtained with the Spatial Pyramid splitting and using the LSBP+LMBP as feature representation of every block. Finally, we also analysed whether the performance is better when the splitting of the spatial pyramid is done using overlapping or not overlapping blocks, in this case the use of non-overlapping blocks seems to be slightly better.

To carry out the experiments we have used 4-fold cross-validation. The best parameters of the SVM (using grid search) are $C = 0.1$ and $\gamma = 1$. When the image descriptor was the combination LSBP+LMBP (without spatial pyramid), the best value of C used is 0.5.

Something that is important to be aware of, is the fact that the size (number of dimensions) of the Census Transform representation, or the representations that combine the LSBP and LMBP histograms could be reduced by applying Principal Component Analysis. For this kind of representations in particular we have not carried out such analysis and possible dimension reduction, since we only wanted to compare which of the two strategies seemed to be the best alternative.

Finally, table III shows the average confusion matrix obtained when the spatial pyramid splitting with non-overlapping blocks was used, and the combined LSBP and LMBP histograms are used to describe every block.

TABLE III
CONFUSION MATRIX: EVERY LINE SHOWS THE AVERAGE RECOGNITION OF THE 100% IMAGES OF EACH CLASS. THESE RESULTS HAVE BEEN COMPUTED AS THE AVERAGE OF THE RESULTS OBTAINED USING THE 4 TEST SETS (4-FOLD CROSS-VALIDATION). THE IMAGE REPRESENTATION WAS OBTAINED USING LSBP+LMBP (T=50)+S. PYRAMID

Class\Class	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	60.8%		1.9%	12.9%					24.3%		
1	9.4%	64.6				9.4%			16.7%		
2			93.3%	3.1%			3.6%				
3	20.5%			45.1%	19.6%	3.6%	1.9%		7.4%		1.9%
4					73.0%	27.0%					
5		3.8%		2.7%	6.3%	83.5%		3.7%			
6				5.0%			88.8%		6.3%		
7		5.0%			14.6%	26.7%		39.2%	14.6%		
8	4.2%	0.5%		3.2%					92.0%		
9	11.3%	10.0%		5.0%					6.3%	67.5%	
10		4.2%		4.2%							91,7

A. Robustness to symmetries

We selected holistic representations that capture the structural properties within an image an suppresses detailed textural information. Nevertheless, there are many factors that might affect the performance of the classifier. Thus, viewpoint, illumination, occlusion, cast shadows, might alter the identification of the scene from a set of images. In this work, the set of images that are in our benchmark have been obtained under different illumination conditions. Regarding the viewpoint, we have analised the robustness of the SVM classifier on recognizing mirror images. In particular we have doubled the size of the test set, by adding images that are got from the original ones by specular reflection (Figure 6). Obviously we have analized the performance of the SVM classifier when the training and validation data set are still the same, but the test set has been altered by including the extra images obtained from the original ones by specular reflection. This kind of analysis is very useful, since a class like *office*, represents environments which are the same (regarding type of furniture, size, etc), but it can happen that the arrangement of the furniture in one of the offices is symmetrical to the arrangement of the same furniture but in other office. In this case, it would be desirable that both rooms are identified as "offices" by the robot (no matter the distribution of the furniture).

TABLE IV
PERFORMANCE OF A SVM CLASSIFIER WITH DIFFERENT FEATURE REPRESENTATIONS AND USING AUGMENTED TEST SETS WITH SYMMETRICAL IMAGES

Feature representation	Perfomance
Census Transform	53.88%
LSBP+LMBP (T=50)	59.23%
LSBP+LMBP (T=100)	56.32%
LSBP+LMBP (T=150)	54.11%
Census Transform+S. Pyramid	69.25%
LSBP+LMBP (T=50)+S. Pyramid	71.69%
LSBP+LMBP+S. Pyramid (non-overlapping blocks)	70.31%

Table IV shows the experimental results we got in this case. It is clearly noticeable that although there is a drop of performance due to the new reflected images in the test set, this drop is much higher in those representations which do not use the spatial pyramid splitting. Therefore, we can conclude that Spatial Pyramid not only help to get a better performance, but mostly helps to improve robustness.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an analysis of different holistic representations inspired by Centrist, and we have analysed the performance of a SVM that classifies every observation (image taken from a robot). With this aim we have collected images from the environment (CITIUS research centre at the University of Santiago de Compostela),

Fig. 6. Example showing how the test set is modified. The figure shown in the top (original image) was always in the test set. In this case, a new image (bottom image) obtained from the original one by specular reflection, is also included in the test set. The performance of the SVM classifier is analysed using this augmented test sets, but the SVM is not trained again (the training and validation sets do not change).

where our Pioneer 3DX robot moves. We have used this set of images as a benchmark for all the tests described in this paper. According to the results we can conclude that working with a representation that combines the Local Difference Magnitude Binary Pattern (LMBP) with the Local Difference Sign Binary Pattern (LSBP) allows the achievement of a better performance than simply using the Census Transform. On the other hand, the use of this representation (LMBP+LSBP), together with the Spatial Pyramid Splitting, allows not only a further improvement in the performance of the classifier, but also obtaining a scene classification that is robust and invariant to symmetrical images, a problem that is very common in a robot moving in an structured environment (office-type environment with many symmetrical rooms). This result, and hence the importance of the spatial pyramid, is achieved regardless the overlapping or non overlapping nature of the blocks into which the hierarchical splitting of the image is done. We consider that the performance achieved with the global image descriptors described in this paper is very promising, especially if we consider that these are the results of classifying individual images, but the confidence can be improved if we consider that the final identification of the environment can be the result of the categorization of a sequence of images taken by the robot when it moves in the environment, and not only a single image.

Regarding our future work, we plan to analyse other global image representations based on gist, but what we consider especially interesting is the merge of this holistic representations with local descriptors, that can provide further robustness or even other benefits. Finally, we also plan to use this scene recognition together with the multisensor localization system developed by our research group for the retrieval of robot controllers.

ACKNOWLEDGMENT

This work was supported by the grant TIN2012-32262, "Consolidation of Competitive Research Groups, Xunta de Galicia ref. 2010/6", and the scholarhip BES-2010-040813 FPI-MICINN.

REFERENCES

[1] J. Wu, H. I. Christensen, J. M. Rehg, *Visual Place Categorization: Problem, Dataset, and Algorithm*, The 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4763-4770, 2009.

[2] A. Canedo-Rodriguez, V. Alvarez-Santos, D. Santos-Saavedra, C. Gamallo, M. Fernandez-Delgado, R. Iglesias, C.V. Regueiro, *Robust multisensor system for mobile robot localization*, Natural and Artificial Computation in Engineering and Medical Applications, LNCS 7931, 5th International Work-Conference on the Interplay between Natural and Artificial Computation, pp. 92-101. 2013

[3] A. Canedo-Rodriguez, V. Alvarez-Santos, D. Santos-Saavedra, C. Gamallo, M. Fernandez-Delgado, R. Iglesias, C.V. Regueiro, *Robust multi-sensor system for mobile robot localization*, Natural and Artificial Computation in Engineering and Medical Applications, LNCS 7931, 5th International Work-Conference on the Interplay between Natural and Artificial Computation, pp. 92-101. 2013

[4] A. Oliva, *Gist of the scene* Encyclopedia of Neurobiology of Attention. L. Itti, G. Rees, and J.K.Tsotsos (Eds), Elsevier, San Diego, CA, pp 251-256.

[5] A. Oliva, A. Torralba, *Modeling the shape of the scene: a holistic representation of the spatial envelope.* Int. J. Comput. Vis. 42 (3), pp 145-175. 2001

[6] X. Meng,Z. Wang, L. Wu, *Building global image features for scene recognition*. Pattern Recognition 45 (1), pp 373-380. 2012.

[7] J. Wu, J. M. Rehg, *CENTRIST: A Visual Descriptor for Scene Categorization*. IEEE Trans. Pattern Analysis and Machine Intelligence. 33(8), pp 1489-1501. 2011.

[8] S. Lazebnik, C. Schmid, J. Ponce *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*, Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), pp 2169-2178, 2006.

[9] K. Kristo, Ch. S. Chua *Image representation for object recognition: utilizing overlapping windows in Spatial Pyramid Matching*, 20th IEEE International Conference on Image Processing (ICIP). 2013