Indoor environment recognition based on ambient classification and people detection

Topic: *Research and Development* Environment and person recognition through vision.

Marco Pirrone, Massimo Romano, Fiora Pirri, Valentina Capone,
Autonomous Agent Laboratory for Cognitive Robotics (ALCOR)
Dipartimento di Informatica e Sistemistica
Universitá di Roma “La Sapienza”
Via Salaria 113, 00198, Roma, Italy
{pirrone, romano, pirri, capone}@dis.uniroma1.it

***Abstract*— In this paper we present a new approach to indoor environment recognition, that exploits a classification of indoor features, human detection, and uncertainty reasoning. The motivation of our work relies on the fact that, for an aware indoor localization, an intelligent autonomous robot needs to show not only an ability not to get lost, but also a sort of awareness of the specific ambient it is stepping into, that is, an ability to identifying the specific ambient, and whether the house inhabitants are present or not, in order to achieve particular tasks. In other words, our goal consists in building a symbolic house ambient map, i.e. one in which each ambient can be suitably labeled by hypotheses of the kind “bedroom, with probability p”. The basic ideas exploit a suitable combination of methodologies leading to a reasoning process that elaborates over the hypotheses emerging from the features analysis, in so establishing probabilistic causal relations between observations and current states. Several methods based on low level features analysis have been investigated so far to solve the indoor environment classification problem. Several disciplines have been, in fact, involved in the study of indoor classification. Just to mention some of them, let us recall the studies in the field of *cognitive psychology*, such as categorization [3] [26], visual similarity [28], and the analysis of the distinction between perceptual and conceptual structures [3] [24]. On the other hand, in the field of *information sciences* , the focus is mainly on the analysis of the image subject [20], on issues related to image indexing [21] [27], on the attributes to be used for describing an image classification [4] [12], on query analysis [10] [11], and on the role of indexing schemata [4].**

Early methods were based on low level features analysis; these techniques, however, hardly lead to a suitable generalization. So, even if low-level features have been largely used in this area, however, they cannot solve, by their own, the underlying classification problem, without appealing to semantic concepts. Indeed, the scene classification problem is often approached from two different points of view: 1. is based on the computation of low-level features (e.g. color, texture, etc.), that are processed with classification methods leading to high-level image properties [25]; 2. is based on the combination of low-level with high-level features, in order to improve the whole classification performance.

Both approaches have arguments for and against. Low-level image features such as color, textures, shape, motion, etc., can be computed automatically and efficiently but the semantics of the image is seldom captured by low-level features. On the other hand, there is no effective method yet to automatically generate good semantic features of an image. In general, a common compromise is to obtain further image information with methods requiring human interaction, e.g. supervised learning methods or manual annotation processes. However, we have to consider that a close human interaction can introduce ambiguousness and biased opinions that could compromise the whole classification process.

This paper gives a contribution to the indoor classification problem using an *active visual attention* approach. In particular, our efforts are focused on the development of an active vision system able to extract useful information from a set of indoor ambient images, and infer the indoor environmental classes. To this end we also face the problem of people detection. In this case our approach is inspired by the work of Poggio [16], that exploits the hierarchical structure of human body to obtain a very efficient classifier. So, given an image we perform a search by components (face, hands, feet), then we collect and combine the results to verify that the human body geometric constraints are complied with. The search of the different parts is performed by distinct classifiers, obtained using the boosting algorithm [13], a general method that makes it possible to yield extremely efficient classifiers from any given learning algorithm. The geometric constraints that we verify are not very strict, so that a partially occluded human body, also in different postures, can be found. Further, the approach we propose seems to perform nicely also in noisy scenes. The whole classification system, is thus based on a combination of context free and context dependent analysis. The first one uses image features independent from the context, like color and intensity, to obtain a set of homogeneous regions, through a clustering procedure. The second one is a two phases context dependent classification process. At the first stage, the system gives a probabilistic classification of the textures extracted from the operating environment images. At the second stage, on the basis of a probabilistic region growing approach, a classification of the textured area is provided. The output of the previous steps is used to find a probabilistic hypothesis concerning the indoor environments. The results of both analyses are combined using suitable weights depending on three factors: the entropy of the features, the correlation among the resulting areas obtained from the context free and context dependent image analysis and the results drawn by an off line learning procedure. The upshot of the weighted combination allows for the construction of a symbolic structure which is the agent inner state representation. A probabilistic reasoning deduction system, based on Hidden Markov Models (HMM), serves to update the agent indoor environment degrees of belief. We provide also some example and experiments.

I. The Background and the Test

The whole classification process, that we illustrate in the next Sections, is supported by a texture DB, used in the texture classification, and a set of both positive e negative example images used in the people detection algorithm.

The texture DB stores a set of possible textures of indoor objects (beds, doors, carpets, windows, etc.) and their clusters, classified and labeled according to the rooms where they can be found (bedroom, bathroom, corridor, etc.). In other words for each texture, we keep a confidence vector, i.e. a set of probabilistic values, having one element for each indoor object [1]. This set defines an *object belonging probability distribution* associated to the selected texture. A texture DB, according to a suitable learning procedure will store data of the form $f_i(X) = \{P(X\widetilde{\in}a_1), \ldots, P(X\widetilde{\in}a_n)\}$ where $P(X\widetilde{\in}a_i)$ is the learned probability that, in a generic ambient scene, the texture X belongs to the indoor entity a_i. By the classification each sub block is labeled with the probability distribution taken from the texture DB having smallest distance from the texture contained in the sub block.

Regarding the people detection training set, we use 754 hand images in the set of positive examples and 816 hand images in the set of negative example, subdivided in the following way:
- 194 positive examples and 240 negative examples for the palm (right and left hand);
- 560 positive examples and 576 negative examples for the back of the hand (right and left hand);

These sets are used to classify hands in different position and angle ($0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, 315^0$ degree). An analogous set of images (580 images) is stored for the feet. This set is subdivided in right foot images, left foot image, and each of this subset is further subdivided bottom foot image alto and down. For the face classification we have used the examples supplied by Intel OpenCV library.

The system has been tested on different robots, both with a single camera and with a pan-tilt head endowed with a couple of ccd cameras. Experiments have been performed in the simulated arena constructed inside the Alcor Laboratory, further at the AAAI Robot Exhibition and Competition, in Edmonton (2002), at the ENEA-Casaccia, for the first national conference on autonomous intelligent systems and advanced robotics, in the Padua rescue Arena, where however the agent could not compete because her right frontal pinion get broken, and finally in Milan at IST2003.

II. Indoor Classification

To face the indoor classification problems we have proposed an active vision system based on a combination of context free and context dependent analysis. The first one uses image features independent from the context, like color and intensity, to obtain a set of homogeneous regions, through a clustering procedure. The second one is a two phases context dependent classification process. At the first stage, the system gives a probabilistic classification of the textures extracted from the operating environment images. At the second stage, on the basis of a probabilistic region growing approach, a classification of the textured area is provided. The output of the previous steps is used to find a probabilistic hypothesis concerning the indoor environments. The results of both analyses are combined using suitable weights depending on two factors: the entropy of the features, the correlation among the resulting areas obtained from the context free and context dependent image analysis. The weighted combination allows for the construction of a symbolic structure which is the agent inner state representation. To reach our goal we have followed the multilevel features analysis described below.

As first step of our path we compute an **early image analysis** consisting of:
- A subdivision of the acquired image in $subblock$ in order to allow local and spatial properties to improve the classification process. In order to work with different resolution, the dimension of sub blocks is variable[2].
- A symmetry characteristic image extraction using the *Generalized Symmetry Transform* introduced in [6], [2]. After applying generalized symmetry transform, a symmetry map is computed and each block is labeled with a *symmetry degree*.
- A distance block extraction: we calculate the difference between the cells central point and the point of view centered on the camera[3].

After the above preliminary operation on the acquired image, we provide a **context free** direct computation of a set of simple features (we have chosen color and intensity) in order to obtain a clusterization of sub blocks: to this end we use a region growing algorithm that uses sub blocks as base elements. Moreover, since we are interested in a region growing finalized to a visual attention procedure, the information we use to chose the seed, necessary in this kind of algorithm, are the variance of each block (accordingly to traditional techniques) and the Symmetry of Sub block. In particular, we are interested in blocks having a small variance (that is homogeneous characteristic) and an high symmetry (that is an interesting characteristic in the real world)

Parallel to the above analysis we provide **context dependent** computation based on two levels texture classification. In this phase, for each sub block, we extract the texture contained in it and, using the discrete wavelet transform [14], [5], [23], we compute its energy. Therefore, a **tree-structured wavelet transform** (TSWT) has been computed in the following way:
- for each **sub block**, the wavelet transform returns four **sub bands** image blocks;
- the energy e of each **sub band** is computed and the maximum energy e_{max} is determined;
- let $c \in [0, 1]$ a suitable constant. If the energy of a sub band is sufficiently small ($e < ce_{max}$) the decomposition is stopped; otherwise it continues to the above step.

The energy values and their ranking order are used to implement a distance classification function df i.e. the distance between two compared textures, defined as the product $de \times dr$, where de is the Euclidean distance between two energy values of the same sub bands, and dr is the Euclidean distance of ranks of sub bands. Once the distance classification function

[1] The table is filled on the basis of a semiautomatic research and classification method applied on images taken from the web and representing indoor environments

[2] The dimension of sub block depends also on the dimension of the texture stored in the DB described in Section I

[3] This distance could be obtained using standard triangulation visual techniques or using telemeter

is given, we can use its values to search into the Texture DB (see Section I) and find the object belonging probability function associated to the texture analyzed. The classification provides a labeling of each each sub block with the probability distribution taken from the texture DB having smallest distance from the texture contained in the sub block.

The results of the texture classification are then combined using a **probabilistic region growing** approach in which the texture associated to similar probability distributions are grouped in a textures cluster. To extend the region growing algorithm was necessary to define the probability distribution of a group of clusters and mean and variance of this distribution. The output of the above probabilistic region growing algorithm is a set of k clusters denoting a textured area associated with a probabilistic *complex* (denoted in the following as elC_i) i.e. an ambient element as a chair, a table and so on according to the classification given in the classification in the texture DB.

At this point, having the results of both previous analysis we have to compute a more accurate analysis of the interesting region coming from the above stages. To this end it is necessary to elaborate a **weighted combination** of the obtained clustered regions. In order to calculate the weights to associate to each result and feature, we consider both prior knowledge and information depending on the current situation. In particular we consider the entropy of the features observed and the correlation among result clustered images. Regarding the entropy, we use an approach based on a *learning process*, in which the above features analysis are performed on a set of positive examples extracted from the acquired image[4]. Since we work with tasselled images, a positive example will be composed by one or more blocks.

The learning process proceeds as follows:

- Each feature is measured on the positive examples. Regarding color and intensity, the value is computed over the sub blocks of the positive examples. Regarding texture, using the procedure described above the *DB texture index* corresponding to the texture contained in the sub blocks is extracted;
- Relevant statistical information concerning the feature distribution over the sub blocks are obtained organizing the values extracted in the previous step into a set of **histograms**, one for each feature.

Using the histogram, we can compute the entropy of each feature. This approach is justified by the assumption that if a feature is affected by a diffuse uncertainty the entropy assumes high values, so its relevance is low.

Another value considered in the feature weight definition is the **mutual overlapping degree** that is strictly connected to the correlation among features used in the whole classification process. Since we have obtained, from the image analysis, a clustered map for each feature, a possible feature weight, is obtained from a suitable combination of the clustered map and an analysis of the cluster similarity (position and dimension)

Moreover, a suitable combination of the values extracted from the above analysis allows us to define a **feature saliency function**, i.e. a function that supplies information about features importance in a particular location of the image i.e. the salience of the location. Recall, given a feature F_h, the proposed feature weight function has the form: $w_h = \mathcal{F}(Entropy, AreaOverlap)$. Since the first term is derived from an interaction with a human expert, the prior knowledge is intrinsically considered in our formalization. Regarding the saliency function, we have to consider the weight function defined above, the symmetry of subblock containing the feature, the volume classification and the center view classification described previously.

Finally, we shall gather all the above mentioned data into a **first state estimation**, subsequent estimation will be provide by the HMM that we are going to describe in Section IV. A state estimation is defined by the following set of data:

-j ($j \in [1..5]$) clustered sets of textured cells with distributions probability, already cut down by the classification. The chosen outcome is the probability estimation that each of the current clusters, under consideration by the observation, belongs to a specific element of the environment: $f(elC = r_i)$, where r_i denotes one of the k possible elements identified in the texture analysis. In the state are present only the clusters that have the j highest values of saliency. elC is a clustered set that we call the *element complex*, to emphasize that we assume that the clustered set of complexes belongs to a single element that is supposed to be in the environment.
-The 2D dimension of the element complex, in terms of its bounding box, and its position in the ambient with respect to the system current position. This data is obtained by simple rigid geometric transformation using geometric camera model, so that any *head movement* is rectified according to an absolute reference frame centered on the agent and the specific action of the head is taken as implicit in the observation.
-The particular salience of the element complex with respect to any other clustered set already processed. This value is obtained as a mean value of the sub blocks saliency.
More specifically each state estimation singles out an hypothesis of existence of a certain element in the environment returning also a set of relevant features needed to elaborate a suitable explanation. An observation state (as far as the system ambient classification is concerned) aggregates the following data:
- $f(elC_i = P_i)$ is obtained in the clusterization process
-The saliency of the complex elC_i, $Sal(elC_i)$;
-The coordinate of the element center in absolute world coordinates (x, y, z), $Center(elC_i)$;
-The distance between the center of elC_i and the system origin, $\delta(elC_i)$ is.
-The approximate absolute dimension, computed by projecting, the system dimension on the distance, of the smallest square containing the element complex elC_i,$dim(elC_i)$.
-The time of the observation, t.

To complete this set of data we also collect the information coming from the people detection procedure (explained in Section III). In particular, in case of the people detection algorithm discovers people, we store each cluster (position and dimension) contains the human body.

[4] Such a set is obtained interacting with a human *expert* who select the interesting regions on the image, i.e. regions useful to classify the indoor environment

III. People Detection

Our approach to the problem of people detection, named *S3C* (*Soft Constraints for Component Classifiers*) *system*, is mainly inspired by the work of Poggio, Mohan and Papageorgiou [1], that is very simple and efficient due to its ability to exploit the hierarchical structure of the human body using component classifiers, but can detect only standing people with arms and legs aligned with the torso. Our system keeps as much simple structure, but can recognize a wider range of configurations than [1]. In fact, its main characteristic is a new, simple way to verify the geometric constraints, based on measures and proportions supplied by Leonardo da Vinci in its Vitruvian Man.

Given an image, we first perform a search by components (face, hands, feet), then we collect and combine the results to verify that the human body geometric constraints are complied with. The search of the different parts is performed by distinct classifiers, obtained using the boosting algorithm [8], [9], [22], in particular the Adaboost version. The boosting is a general method that makes it possible to yield extremely efficient classifiers from any given learning algorithm (named *weak classifier*) whose performance is a little better than random guessing, by repeatedly running it on various distributions over training data, and then combining the classifiers produced into a single composite classifier. The most basic theoretical property of Adaboost concerns its ability to reduce the training error. In fact, Freund and Schapire prove that the training error (the fraction of mistakes on the training set) drops exponentially with respect to the number of training rounds *T*.

As set of weak classifiers we choose a family of simple features, proposed by Viola and Jones [17] and extended by Lienhart, Maydt, Kuranov e Pisarevsky [7], [19] with the 45^o rotated features. Then the task of the learning algorithm is to identify a set of features that consistently distinguishes the face (or the left hand and so on) in an image: at each stage of Adaboost we choose as weak classifier the feature that best separates the positively-labeled images from the negatively-labeled images.

The classifiers of our system are trained to recognize 5 different parts, i.e. face, left hand, right hand, left foot and right foot. Every classifier returns a list of the regions that contain the component it is trained on; all these results are the input for the next phase, that checks if and where in the image some of these parts belong to a same human body.

The geometric constraints that we verify are not very strict, so that a partially occluded human body, also in different postures, can be found. Their definition is based on the measurements and proportions and their visual synthesis supplied by Leonardo da Vinci in its innovative drawing, the Vitruvian Man. The unit of measurement is the length of the head. For example, a man is 8 heads tall, the arm is 1 head and a half, the forearm is 1 head and 1/4, the hand is 3/4 of a head and so on. So, the hand can stay in the circle with centre in the articulation of the shoulder and radius 11/4, but not more far.

In this way, we don't calculate the details of the posture of the body, but, more simply, that it isn't incorrect; with this choice we avoid an expensive off-line phase to construct a model, we obtain a very fast and efficient system, due to a control phase of geometric constraints very simple (it asks only for the calculation of a little set of inequalities) and, above all, we can detect people with arms and legs in any position, provided it is correct.

IV. HMM

Once the information coming from the image analysis has been suitably collected, the classification process goes through the focal part of the attentional process: the basic ideas exploit a suitable combination of methodologies leading to a reasoning process that elaborates over the hypotheses emerging from the features analysis, in so establishing probabilistic causal relations between observations and current states. The system will find an explanation of a sequence of observations, and this explanation will return an hypothesis of the kind "this is..".

We intertwine two inference processes in order to construct this explanation, a Hidden Markov Model analysis (or recognition) that from a sequence of observations returns a sequence of states (i.e. the possible ambient scenes), and an inference process on the states selection, so as to constrain the sequence to collapse in a single *most likely* state, with varying time, namely the current ambient scene. In particular, suppose the system is in a specific environment (e.g. a bedroom), and it begins to make observations, and each image is treated as we have explained so far, in the previous section. From each observation, at time t, a set of elements (more salient w.r.t. the specific image) are taken, and a sequence is constructed. Now the system is supposed to have no knowledge about where it is, but with some prior knowledge about what it can expects from each environment it could end up into. The prior knowledge is used in a double way: 1. to assess the probability of the elements that could be observed in a given state/ambient scene; 2. to specify the definition of each ambient/scene, together with all the necessary constraints, e.g. if there is a bed then there will be a bed spread and a pillow.

We expect that given the sequence of observations $(O_{t_1}, \dots, O_{t_n})$ a sequence of states $(S_{t_1}, \dots, S_{t_n})$ is returned, such that all the states account for the same ambient scene, in this case the bedroom. The ambient scene is the explanation of the observation sequence; furthermore, from the description associated with each state/ambient, some of the observed elements will be discarded and some will be kept with a probability, obtained by averaging on the probabilities of the elements, as given from the observation, and as given in the chosen state.

To formalize these two processes we use the Situation Calculus [18]. In this context to simplify the presentation we assume that actions are deterministic, since we are concerned only with observation actions. In fact, whenever the system changes its position the whole process restarts. Note also that actions changing the head and thus the camera position are implicit in the observation actions and are suitably rectified by the usual calibration methods. Therefore the outcome of an observation action is seen, by the reasoning process looking for an explanation, as the presupposition that a certain element has been seen in the environment, with its absolute position and dimension. In other words the current head position, that has sanctioned the presence of the specific element in the image, is definitely transparent to the observation. This fact amounts to the definition of an observation as an action having no effect in the environment (which would not be true if the head, for example, by turning, would hit somewhere). Furthermore we assume that the inference process drawing a specific explanation use a *tight rule* on the sequence of observations, i.e. that there is no intra-movement (except for the head).

In order to use the HMM in our problem we associate states with ambient scenes and observations with the observation of a specific grouping of objects in the ambient scene. Each environment is expected to have several elements that can be individuated, and many of them could be those core elements characterizing the ambient scene, e.g. a computer, a mouse, a table, etc. in a bedroom. To this end it was necessary to define axioms in SC to describe the concept of group and the way in which each group is perceived during an observation. First of all we have to establish, and mapping with a set of axioms, when two element are close one to the other, even if they have been captured from two different images but in the same observation sequence[5]: the axiom introduced imply that two element complexes are close in the ambient scene if they have been observed in the same observation sequence, meaning that either they belong to the same image or to two different images, but belonging to the same ambient scene, and furthermore their distance is such that their clusters are adjacent or there is a k (a fixed constant defined for each couple of element complexes) that weaken the adjacency. A group is a graph whose conditional tables have to be given in order to allow a suitable computation during perception. Each group is indeed a graph of probabilistic dependencies, i.e. it is a Bayes network, therefore it requires a conditional probability table to be specified, in terms of the joint probability, concerning exclusively the group:

$$\begin{aligned} &Pr(desk \circ monitor \circ mouse \circ keyboard, s) = \\ &\quad Pr(desk, s) \times Pr(monitor|desk, s) \times \\ &\quad Pr(mouse|desk, monitor, s) \times \\ &\qquad Pr(keyboard|mouse, monitor, desk, s) \end{aligned} \tag{1}$$

Given the conditional probability tables in the Bayesian graph for each group, then perception can be now regarded as an activity that can exploit the combined probabilities to increase the reliability of each element captured in the image.

The axiom that we have to introduced say that a group (or an its subgroup) can be perceived in a course of observations if each element of the group is either observed in the same observation state, hence their probabilities can be taken as independent, or they have been accumulated through previous perceptions, and in such a case the previous probability is disregarded and a new conditional probability is established, according to the conditional table given for the group. Closeness is a property required just in case there might be an opening in the ambient scene and some element not belonging to the specific ambient scene is aggregated.

Once the groups have been fixed and we know how a group is perceived in a course of observations, we introduce their **likelihood**, with respect to each ambient scene. Observe that the likelihood can be easily learned, according to the frequency of a specific group in each ambient scene. A set of rules for probability of a group in each situation, given a specific ambient, is specified. In particular, the likelihood is given only for the initial situation. In fact, in any situation successive to this, the observation is affected both by the observation state and the probability of each element detected. Therefore the observation matrix for the next situations have to be computed according to the perception and the amount of elements constituting a group that have been perceived: the values of this matrix are calculated by considering the current percept, and the previous likelihood with respect to the subgroup that is a maximal subset of the current group.

The following table illustrates the likelihood matrix for a set of a priori defined groups and the expected ambient scenes. The values read $Pr(g_i|Ambient(A, S_{t_0}))$, i.e. the probability of observing a specific group given that the ambient scene is A, in situation S_{t_0}.

	$shelves\circ$ $book\circ$ $papersheet$	$bed\circ$ $bedSpread\circ$ $pillow$	$desk \circ mouse\circ$ $monitor\circ$ $keyboard$	$stove\circ$ $table\circ$ $oven$
$Ambient(studio, S_{t_0})$	0.45	0.1	0.45	0
$Ambient(bedroom, S_{t_0})$	0.25	0.7	0.05	0
$Ambient(kitchen, S_{t_0})$	0.2	0	0.2	0.6

(2)

Analogously we can give a transition matrix in S_{t_0} accounting for the likelihood matrix. That is $A_{ij} = \sum_j Pr(g_j \;|A_i)$, hence:

	$Ambient$ $(studio, S_{t_0})$	$Ambient$ $(bedroom, S_{t_0})$	$Ambient$ $(kitchen, S_{t_0})$
$Ambient(studio, S_{t_0})$	0.9	0.1	0
$Ambient(bedroom, S_{t_0})$	0.3	0.7	0
$Ambient(kitchen, S_{t_0})$	0.4	0	0.6

(3)

At this point we have all elements to correctly define the HMM and given a course of observation actions we can compute the sequence of the most probable states. Once we have this sequence then we would choose the most probable state, or ambient scene. We shall show the computation with an example.

Example 1: Suppose we have only the three ambient scenes given in the matrices provided above. The initial probability distribution for each ambient scene is:

$$\pi_i = Pr(A_i|S_{t_0}) = 1/N$$

here N is the number of states, i.e. the number of expected ambient scenes. Let $\{a_{ij}\}$ be the transition given in 3.Let $\delta_t(i)$ be the maximum probability of all state sequences ending at state i at time t Let the Observation matrix $\{b_i(g_j)\}$, obtained through perception, be as follows:

	g_1 $shelves\circ$ $book\circ$ $papersheet$	g_2 $bed\circ$ $bedSpread\circ$ $pillow$	g_3 $desk \circ mouse\circ$ $monitor\circ$ $keyboard$	g_4 $stove\circ$ $table\circ$ $oven$
$Ambient(studio, S_{t_0})$	0.35	0.1	0.35	0
$Ambient(bedroom, S_{t_0})$	0.15	0.4	0.03	0
$Ambient(kitchen, S_{t_0})$	0.18	0	0.03	0.01

(4)

[5] An observation sequence is a sequence of image coming from the same ambient scene

From our computation, the observation matrix has already eliminated some state. And now the sequence of observations is $\{g_1, g_2, g_3\}$ with the above described probabilities for the last situation, that can easily be ordered as follows, since all the groups have been constructed through subsequent perceptions.

$$\langle g_1, s_{t_1} \rangle, \langle g_2, s_{t_2} \rangle, \langle g_3, s_{t_3} \rangle$$

So we shall compute the $\delta_j(i)$ for $i \in \{studio, bedroom, kitchen\}$

$$\begin{array}{l}
\delta_1(studio) = \pi_1 \times Pr(g_1|studio) = 0.3 \times 0.35 = 0.0825 \\
\delta_1(bedroom) = \pi_2 \times Pr(g_1|bedroom) = 0.3 \times 0.15 = 0.0495 \\
\delta_1(kitchen) = \pi_3 \times Pr(g_1|kitchen) = 0.3 \times 0.18 = 0.0594 \\
\hline
\delta_2(studio) = max_j(0.0825 \times 0.9, 0.0495 \times 0.3, 0.0594 \times 0.4) \times Pr(g_2|studio) = \\
\qquad max_j(0.0825 \times 0.9, 0.0495 \times 0.3, 0.0594 \times 0.4) \times 0.1 = 0.007425 \\
\delta_2(bedroom) = max_j(0.0825 \times 0.1, 0.0495 \times 0.7, 0.0594 \times 0.0) \times Pr(g_2|studio) = \\
\qquad max_j(0.0825 \times 0.1, 0.0495 \times 0.7, 0.0594 \times 0.0) \times 0.4 = 0.0033 \\
\delta_2(kitchen) = max_j(0.0825 \times 0.0, 0.0495 \times 0.0, 0.0594 \times 0.6) \times Pr(g_2|kitchen) = \\
\qquad max_j(0.0825 \times 0.0, 0.0495 \times 0.0, 0.0594 \times 0.6) \times 0.0 = 0.0 \\
\hline
\delta_3(studio) = max_j(0.007425 \times 0.9, 0.0033 \times 0.3, 0.0 \times 0.4) \times Pr(g_3|studio) = \\
\qquad max_j(0.007425 \times 0.9, 0.0033 \times 0.3, 0.0 \times 0.4) \times 0.35 = 0.00233 \\
\delta_3(bedroom) = max_j(0.007425 \times 0.1, 0.0033 \times 0.7, 0.0 \times 0.0) \times Pr(g_3|bedroom) = \\
\qquad max_j(0.007425 \times 0.1, 0.0033 \times 0.7, 0.0 \times 0.0) \times 0.03 = 0.00003 \\
\delta_3(kitchen) = max_j(0.007425 \times 0.0, 0.0033 \times 0.0, 0.0 \times 0.6) \times Pr(g_3|kitchen) = \\
\qquad max_j(0.007425 \times 0.0, 0.0033 \times 0.0, 0.0 \times 0.6) \times 0.03 = 0.0
\end{array} \tag{5}$$

Thus the maximal sequence is $\delta_1(studio), \delta_2(studio), \delta_3(studio)$, and therefore following the sequence of observations, the recognized ambient is a studio.

We have implemented the above given axiomatization in Golog (see [15]) and through the interface between C++ and prolog we have used an online implementation of the Viterbi algorithm to test our approach, which we have verified during an exhibition given for the EU-IST.

V. Conclusion

In this paper we have presented a new approach to indoor classification problem and people detection problem. In particular, our efforts are focused on the development of an active vision system able to extract useful information from a set of both indoor environment and human body images and infer the environment classes and position of the human body (if presents). We have introduced an innovative indoor classification system in which we use a probabilistic combination of context free and context dependent analyses to infer high level scene properties from low level image features. The approach we propose in the people detection phase, based on a search by components approach and a combination human body geometric constraints solves many cases of occlusion (or self-occlusion). In fact, the system allows to set in a parametric way the minimum number of parts that have to satisfy the geometric constraints in order to assert the presence of a human body. The system shows good results real-time also.

The goal reached consists in building a symbolic house ambient map, in which each ambient can be suitably labelled by hypotheses of the kind “This room is a roomX, with probability p”, where roomX stands for bedroom, bathroom, kitchen, etc. In case of the system discover a human body, it will be supply the position of the body w.r.t. the global reference frame.

References

[1] Anuj Mohan, Constantine Papageorgiou e Tomaso Poggio. Example-Based Object Detection in Images by Components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[2] Y. Bonneh, D. Reisfeld, and Y. Yeshurun. Quantification of local symmetry: Application for texture discrimination.

[3] B. Burns. *Percepts, Concepts, and Categories: The Representation and Processing of Information*. Elsevier Accademy Publisher, New York, 1992.

[4] Jorgensen C. Classifying images: criteria for grouping as revealed in a sorting task. In American Society for Information Science, editor, *Schwartz R.P., Beghtol C., Jacob E.K., Kwasnik B.H., Smith P.J. (ed) Proceedings of the 6th ASIS SIG/CR Classification Research Workshop*, pages 65–78, 1995.

[5] Tianhorng Chang and C.-C. Jay Kuo. Texture analysis and classification with tree-structured wavelet transform. *IEEE Trans. Image Processing*, 2(4):429–441, 1993.

[6] Reisfeld D., Wolfson H., and Yeshurun Y. Context free attentional operators: the generalized symmetry transform. *Int. J. of Computer Vision, Special Issue on Qualitative Vision*, 1994.

[7] Rainer Lienhart e Jochen Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. In *IEEE Conference on Image Processing*, volume 1, pages 900–903, September 2002.

[8] Y. Freund e R. Schapire. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[9] Yoav Freund e Robert E. Schapire. Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[10] J. P. Eakins. *Design Criteria for a Shape Retrieval System*, volume 21, pages 167–184. Computers in Industry, 1993.

[11] P. G. B. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–52, 1993.

[12] So That Others May See: Tools for Cataloguing Still Images. *Describing Archival Materials: the Use of the MARC AMC Format*. 1990.

[13] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

[14] Andrew Laine and Jian Fan. An adaptive approach for texture segmentation by multi-channel wavelet frames. Technical Report 25, 1993.

[15] H.J. Levesque, R. Reiter, Y. Lesperance, F. Lin, and R. Scherl. Golog: A logic programming language for dynamic domains. *Journal of Logic Programming*, 31:59–84, 1997.

[16] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[17] Paul Viola e Michael Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *IEEE Proceedings of Conference on Computer Vision and Pattern Recognition*, 2001.

[18] F. Pirri and R. Reiter. Some contributions to the metatheory of the situation calculus. *ACM*, 46(3):325–362, 1999.

[19] R. Lienhart, A. Kuranov e V. Pisarevsky. Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. In *25th Pattern Recognition Simposium*, September 2003.

[20] H. Roberts. Do you have any pictures of...? subject access to works of art in visual collections and book reproductions. *Art Documentation*, Fall 1998.

[21] Layne S. S. Some issues in the indexing of images. In *JASIS*, volume 45, pages 583–588, September 1994.

[22] Robert E. Schapire. A Brief Introduction to Boosting. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1401–1406, 1999.

[23] J. Smith and S. Chang. Automated binary texture feature sets for image retrieval. In *ICASSP-96*, pages 2239–2242, 1996.

[24] L.B. Smith and D.Heise. Perceptual similarity and conceptual structure. *Percepts, Concepts and Ctegories B.Burns*, 1992.

[25] Martin Szummer and Rosalind W. Picard. Indoor-outdoor image classification. In *IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*, pages 42–51, 1998.

[26] Edward R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, Connecticut, 1997.

[27] J. Turner. Cross-language transfer of indexing concepts for storage and retrieval of moving images: Preliminary results. In *ASIS Annual Conference Proceedings*, pages 19–24, October 1996.

[28] A. Tverky. Features of similarity. *Psychological Review*, 84(4), July 1997.