Scene Classification Using Hidden Markov Models

Benrais Lamine
Dept. of computer science
University of Science and Technology Houari Boumedien
Algiers, Algeria
mbenrais@usthb.dz

Baha Nadia
Dept. of computer science
University of Science and Technology Houari Boumedien
Algiers, Algeria
nbaha@usthb.dz

***Abstract*—In common multiclass classification problem, the main difficulties occur when classes are not mutually exclusive. In order to solve problems such as document classification, medical diagnosis or scene classifications we need to use robust and reliable tools. In this paper, we consider the problem of scene classification treated by hidden Markov models (HMMs) using a novel and intuitive classification process. We introduce a modeling system that scales the parameters of the HMM (observations and hidden states) into the variables of the scene classification problem (scene categories and objects belonging to the scene). The HMM is constructed with the support of object's weight ranking functions. Inference algorithms are developed to extract the most suitable scene category from the generated discrete Markov chain. In order to approve the efficiency of the proposed method, we used the MIT Indoor dataset (*2700* scenes distributed into *67* scenes categories) to evaluate the classification accuracy. We also compared the obtained results with the current state of the art's methods. Our approach distinguishes itself by obtaining results going until *76%* of well classified scenes.**

Keywords*—Scene classification; object's weight; hidden Markov models

I. INTRODUCTION

Having information about the surrounding environment is a major asset in achieving tasks or taking decision for any exciting agent (human or robot). The ability to build such an instantaneous concept guides the agent for a better accommodation and efficiency. For this purpose, the general problem of scene classification has received considerable attention in the recent past and turned out to be a major field in computer vision. In this paper, we consider the scene classification problem with objects as attributes [1] formally defined as the following: Given a set of finite scene categories $SC = \{\mathrm{SC}_1, \mathrm{SC}_2, \dots, \mathrm{SC}_\mathrm{n}\}$ an input scene S containing a set of finite properties $P = \{\mathrm{P}_1, \mathrm{P}_2, \dots, \mathrm{P}_\mathrm{n}\}$; we are not going to define rigorously what a property is, but we can simply say that it contains semantic information about S, e.g. Objects, Actions, Size, relationships, etc. We wish to assign the most suitable scene category SC_i to S knowing P. For convenience, let us assume the compact notation (1).

$$\mu = (SC, P) \tag{1}$$

We introduce an innovative new approach in scene classification problem relying on a recognized and strong mathematic tool of prediction and classification: The hidden Markov models (HMM). The first challenge consists on finding the right modeling of the scene classification problem so it can be solved by the HMMs architecture. Analogies between the inputs and outputs parameters of both entities (HMMs and scene classification) need to match. In parallel, properly ordered input parameters in the HMM are very critical to the final result accuracy which made us develop weight functions that assign a weight measure to each object of the dataset. This weight measure is first used to quantify the saliency and importance of an object as a single entity then to distinguish the most suitable object knowing the current scene category. Once elaborated, the process generates a discrete Markov chain containing the scene categories that represent the most the selected objects. Afterwards, an inference algorithm is developed to extract the most suitable scene category from the discrete Markov chain. This way of approach is not common [2], [32] for the classification using an HMM and is going to be explained and tested throughout this paper.

The remainder of the paper is organized as follow. Section II will introduce the reader to an overview of existing approaches and methods treating the scene classification problem. Section III presents the formal definition of hidden Markov models (HMM) and the construction of the discrete Markov chain. Section IV introduces the proposed method and explains with details all the stated contributions. Finally, Section V experiments the proposed method showing the obtained accuracies while varying the different input parameters. A comparison with some existing method in the literature is also presented. We conclude by summarizing our results and outlining steps to improve the scene classification accuracy using hidden Markov models.

II. RELATED WORKS

In recent years, several image descriptors have been developed in order to increase the ability of computer vision system towards a higher level of interpretation. Arens and Ottlik [4] are one of the first to implement a scene classification experiment in concrete street traffic application retrieving textual description of videos sequences. Later on, their work has been improved by Dejan and Rok [5] where a scene interpretation based a Description Logic (DL) and a top down guided 3D CAD model-based vision algorithm were developed aiming to bring more autonomous activity to robot on objects and scenes. Such as [5] logical languages for scene classifications have been widely studied [6], [7] where predicates represent the different properties of the scene (objects, size, positions) and the inference system is used to identify the associated scene categories. The Description Logic (DL) was the most successful for representing a real

world state, Neumann and *al* [8] introduced the DL as knowledge reasoning and representation system for scene classification with temporal and special relationships. Their proposed approach exploits relations between objects, occurrences, events and episodes joining at the same time visual evidence and contextual information. A more specific contribution has been made by [9] applying the DL for road scenes classification and intersections geometries. The formalism of the DL being similar to the problem of scene classification, the approach shows successful and promising results. The complexity of scene classification increases relatively to the number and size of scenes. To face this issue, first approaches were to reduce the choice of scene categories to a binary perception: Indoor/outdoor scene classification [10], [11] and very satisfying results were obtained. Nevertheless, the approaches were not extendable to multiclass classifications. Another approach consists on predicting the location of salient area in the scene [12], [13]. In the same perspective, the classification process is isolated to a "Focus of attention" analogously to human vision activities [14]. Agnes and Sven [15] proposed an indoor scene classification using a 3D approach mixed with Gist scene features, while [16] recorded better results using Gist features in outdoor scene classification. We can find in the literature several methods of scene classification using low level approaches [17]-[19] even if [19] was able to get quality results by adopting an approach that shares discriminative feature between the different scene categories, however, based on [20], [21], [11], scene classification depending on low level approaches works poorly. In contrast, high level approaches of scene classification were developed. In this case, a scene is represented with high level information such as objects, actions, etc. [0]. Quattoni and Torbralba [3] have proposed model of indoor scene classification where a comparison between scenes is made using a set of ROI to find the right scene category of the given image. The main idea is the fact that scenes containing the same objects tend to have similar scene category. In the same perspective, [22], [23] proposed a deformable part-based models (DPM's) using SVM's as training models. The originality of [22], [23] is the introduction of an open-ended learning of latent structures for scene classification problems. [24] proposed an SVM classification model using maximization likelihood and margin, this approach is made possible by the fact that the optimization problem was efficiently solved. [25] introduced a new visual descriptor for recognizing scene categories based on a holistic representation and has a strong generalizability for category recognition. It's mainly based on encoding the structural properties within an image and suppresses detailed textural information. Representing an image as a bag of objects has recently demonstrated impressive results [20], [26], [27], [40]. Herranz et al. [15] explored the path of scene classification using conventional neural networks (CNNs) exploring the way to combine effectively scene centric and object centric knowledge into a CNN architecture. Scene classification state of the art based on CNNs becomes very successful [27] principally due to the impressive obtained results. However CNNs are known for two main inconveniences: 1) the huge amount of data needed for the training part; 2) the high computational cost. In our case, we are not going to compare our results with the CNNs architecture due to the differences in terms of environment's preparation (Amount of data and hardware prerequisites). The proposed method requires much less data and computational cost. In the same perspective of high level scene classification, Biederman et al. [13] assume that relations between an object and its environment can be reduced to five classes in order to characterize the organization of objects into real-world scenes. These classes have the ability to reduce the anomalies that can occur in scene classification problem. Further investigations have been introduced later on by [28] integrating other classes of relationship. Fuzzy logic has also been used widely for scene classification [29], [30], Baiget et al. [31] were one of the first who computerized the geometrical construction of scenes studying human behavior, the learning was done using a derivation of fuzzy logic called FMTHL (fuzzy metric temporal horn logic). In the same idea, Zitnick et al. [40] adopted a statistic approach to extract semantic information and identify the scene categories. Their approach and results are influenced by the assumption that abstract images can accurately represent world real scenes. While all the approaches reviewed in the literature differ in many features, they share the same aspect of using a learning part known as background knowledge to assist the identification of the scene category.

III. Hidden Markov Models (HMMs)

A. Definition of HMMs

The hidden Markov model is a probabilistic signal processing approach that aims to extract the maximum likelihood model from a sequence of observable events [2]. Robust and efficient, it has been known to mathematicians since a long time but has only been applied recently on numerous modern applications such as speech recognition, computational molecular biology and other areas of artificial intelligence and pattern recognition [2]. In theory, the HMMs are presented as a finite number N of states and M of observations symbols. Each state is assigned to a clock time t and possesses a measurable property. Transitions from different states are based upon a transition probability distribution that depends on the previous state (The Markovian property). After each transition made, an observation output symbol is yield based upon an emission probability specific to the current state. There are thus N emission probabilities for each observation. Formally, the HMM are defined as follow [2]:

- T: Observation sequence length (total number of clock times t)
- N: Number of hidden states $\{S_1 \ldots S_n\}$
- M: Number of observations symbols $\{o_1 \ldots o_n\}$
- A: state transition probability $\{ a_{ij} \}$ where $a_{ij} = P[q_{t+1}=S_i \mid q_t=S_j)]$ j,i in $[1,N]$
- B : observation emission probability $\{ b_i(o_k) \}$ where $b_i(o_k)= P[\text{o}_\text{k}$ at $t[q_t=S_i]$ i in $[1,N]$, k in $[1,M]$
- π: The initial state distribution $\{\pi_\text{i}\}$ where $\pi_\text{i}=P[q_1=S_i]$ i in$[1,N]$

Having the appropriate values of *N, M, A, B* and π an observation sequence $O= \{o_1, o_2, o_3 \ldots o_T\}$ is generated following Algorithm 1:

Algorithm 1

1- *Choose an initial state q1 according to the initial state distribution* π
2- *Set t=1.*
3- *Choose* o_t *according to* $B_i(o_t)$ *the symbol probability distribution in state t*
4- *Choose t+1 according to* $a_{i,i+1}$. *The state transition probability distribution for state t;*
5- *Set t=t+1*
6- *If t <T go to 3 else terminate the process*

A compact notation λ is used to represent in (2) for a given HMM.

$$\lambda=(A, B, \pi) \quad (2)$$

B. Inference of Hidden Markov Model and Dynamic Programming

Given a model λ=(*A*, *B*, π) and an observation sequence $O=\{O_1,O_2...O_n\}$ the most basic approach to estimate the probability of *O* knowing λ i.e $P(O|\lambda)$ is by computing the probabilities of all possible sequences of hidden states having a length of *T* (*T=Card(O)*) that are eligible to emit *O*. The probability of such a sequence can be compute as follow: We first start by computing the probability of a fixed set of hidden states *I* knowing a model λ using (3).

$$P(I|\lambda) = \pi_{1i}\, a_{i1i2}a_{i2i3}a_{i3i4} \ldots a_{iT-1iT} \quad (3)$$

Next, we compute the probability of a given observation *O* knowing the hidden states *I* and the model λ using (4).

$$P(O|I,\lambda) = b_{i1}(O_1)b_{i2}(O_2)b_{i3}(O_3) \ldots b_{iT}(O_T) \quad (4)$$

The probability where *O* and *I* occur at the same time (i.e *O* is emitted by *I)* is simply the product of (3) and (4) as illustrated in (5).

$$P(O,I|T) = P(O|I,\lambda)(P(I|\lambda) \quad (5)$$

Finally, the probability of *O* knowing λ is obtain by summing the probability computed in (5) over all the possible hidden states *I* as represented in (6).

$$P(O|\lambda) = \sum_{i=1}^{T} P(O_i|I_i,\lambda) * P(I_i|\lambda) \quad (6)$$

An explanation of (6) can be seen as the following: At time *t=1*, we are in the hidden state *i1* with an initial probability of π_i and emit the symbol o_1 with the probability $b_{i1}(o_1)$. At time *t=2*, we will make a transition to the hidden state i_2 with the transition probability a_{i1i2} (note that the transition can be reflexive) and emitting the symbol o_2 with probability $b_{i2}(o_2)$ and so on until *t=T*.

The reader can easily notice that computing the probability (6) requires a lot of computation time, exactly up to $(2T-1)N^T$ multiplications and N^T-1 additions. Another approach is proposed with dynamic programming.

In our case, we are more interested in finding the most likely sequence of hidden states that can emit a sequence of given observations. A dynamic programming algorithm for finding such a sequence is widely known as the Viterbi algorithm [2]. The key idea of the Viterbi Algorithm is to keep only the max probability path of the hidden states -not all the paths- that can emit the current sequence of observation.

Given a model $\lambda=(A,B,\pi)$ a set of observation $O=\{o_1, o_2..o_T\}$ The Viterbi algorithm, presented in algorithm 2, introduces the dynamic programing method [2].

Algorithm 2 : Viterbi (λ,O) return BEST_PATH

1- *Creates a path probability matrix VITEBI[N+2,T]*
2- *For each state I do*
3- *VITERBI[S,1]:=pi1*bi(O1)*
4- *BackPointer[s,1]:=0*
5- *End for*
6- *For each time step t from 2 to T do*
7- *For each state I*
8- *Viterbi[S,t]:=MAX{s'=1:N} viterbi[s',t-1]*as's*bi(Ot)*
9- *Backbpointer[s,t] :=argmax{s'=1,N} viterbi[s',t-1]*as',s*
10- *End for*
11- *End for*
12- *ZT= argmax{s'=1,N} viterbi[s',T]*as',s*

The complexity of the Viterbi algorithm is on the order of *o(MN)* where *M* Number of observations symbols and *N* number of hidden states [2]. This complexity is significantly better than the previous method. The obtained result "BEST_PATH" is generally named as "Discrete Markov Chain".

IV. Proposed Method

In this section, we're going to explain the different contributions made in this paper. We describe the weight functions which calculate the saliency of a given object. Then, we explain the investigation made to solve the scene classification problem using an HMM architecture. The aim is to ensure a perfect analogy between their formal definitions.

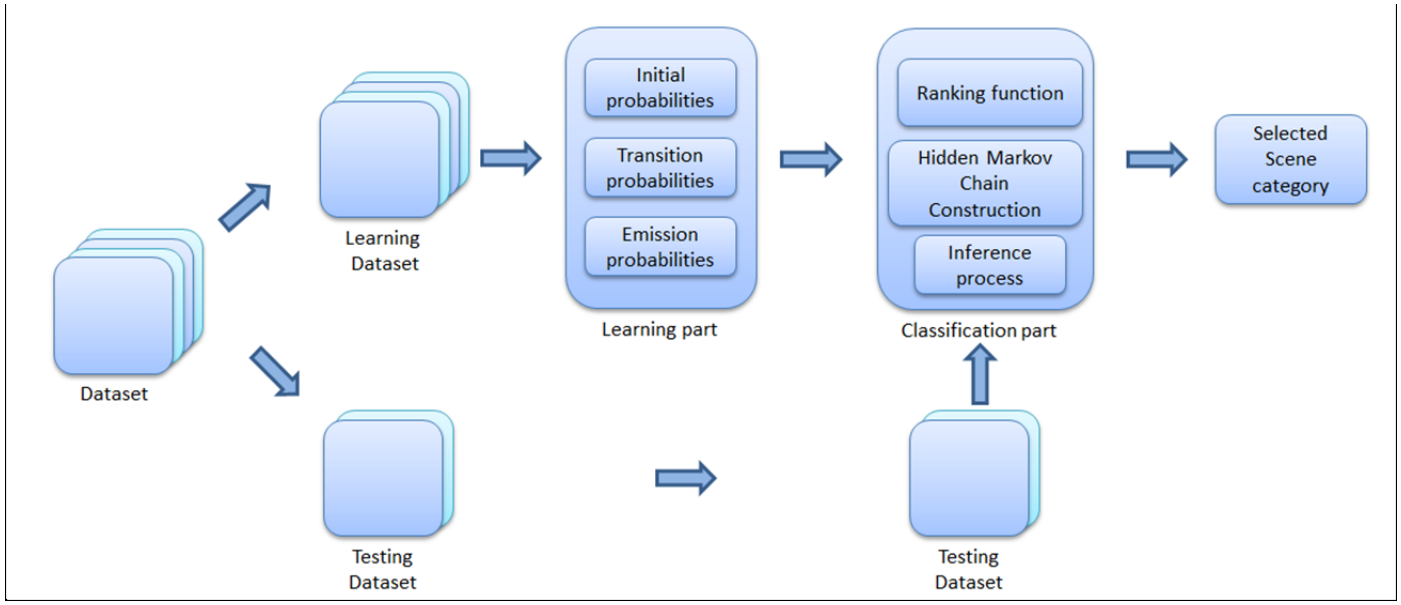

Fig. 1. Workflow of the proposed classification process based on hidden Markov model.

Finally, an inference algorithm is presented in order to extract the most suitable scene category from the discrete Markov chain. To illustrate the given contributions, Fig. 1 summarizes the complete workflow of the proposed method. First, The Dataset [3] is divided into a learning dataset (80%) and test dataset (20%) as recommended by [3]. The learning dataset is used to construct the necessary entities to the proposed classification process composed of initial probabilities distribution, transition and emission probabilities computation. This part of the workflow is called the *"learning part"*. The test dataset is used to certify the reliability of the proposed method's classification. A ranking function selects the most salient objects to represent the input scene. The next step consists on going through the selected object while constructing the discrete Markov chain. Finally, an inference algorithm is developed in order to extract the most suitable scene category from the discrete Markov chain. Each of these steps will be deeply explained in the upcoming subsections.

A. Object's Weight Measure and Scene Similarities

In this subsection we're going to present the weight measures developed to quantify the saliency of given objects. After that the similarity measure between two different scene categories is introduced.

1) Object's Weight Measure Computation

In order to determine if an object has an impact on the scene category classification process, we need to develop a weight function to quantify its saliency.

We introduce the following definitions to clarify the content of equations.

Let,

- $FO(SC_i, O_i)$: A function that returns the frequency of appearance (all the occurrences) of object O_i in all the different scenes that are labeled as scene category SC_i
- $NO(SC_i, O_i)$: A function that returns the number of times (without counting doubles) object O_i appears in all scenes labeled in current scene category SC_i
- $FO_{all}(O_i)$: A function that returns the frequency of appearance (all the occurrences) of object O_i in all the dataset
- $NO_{all}(O_i)$: A function that returns the number of times (without counting doubles) object O_i appears in all the dataset.

Note: The FO (and FO_{all}) returns all the occurrences of appearance of the object O_i in current scene, conversely, the NO (and NO_{all}) returns the number of time an object O_i exists in the current scene. The correlation between FO (respectively FO_{all}) and NO (respectively NO_{all}) is defined in (7)

Let O_{all} be the set of all objects in the dataset

$$\forall\, o_i \in O_{all}, FO(o_i) \geq NO(o_i) \tag{7}$$

First, we take into consideration the fact that an object o_i belongs to a particular scene category SC_i. (8) demonstrates how the weight measure $\acute{W}$ is calculated.

$$\acute{W}(SC_i, o_i) = \begin{cases} 0 \text{ if } \mathrm{NO}(\mathrm{SC_i}, \mathrm{o_i}) = 0 \\ \left(\frac{\mathrm{NO}(\mathrm{SC_i}, \mathrm{o_i})^{\mathrm{FO}(\mathrm{SC_i}, \mathrm{o_i})}}{\mathrm{NO_{all}}(\mathrm{o_i}) * \mathrm{FO_{all}}(\mathrm{o_i})}\right) \quad else \end{cases} \tag{8}$$

We generalized (8) to get an equation independent of any scene category as presented in (9).

Nevertheless, in order to generate appropriate calculations processes, a normalized version of (9) is elaborated in (10) to ensure that the weight values of objects o_i are held between 0 and 1.

Since (10) has no upper bound, its value expends as the occurrences of the object raises, we associate the value of the variable "*MaxValue*" according to the current dataset.

$$W(o_i) = \begin{cases} 0 \text{ if } \max_{SC_i}(\mathrm{NO}(\mathrm{SC_i}, \mathrm{o_i})) = 0 \\ \dfrac{\max_{SC_i}\left(NO(SC_i,o_i)\right)^{\left(\max_{SC_i}\left(FO(SC_i,o_i)\right)\right)}}{No_{all}(SC_i,o_i) * FO_{all}(SC_i,o_i)} \quad \text{else} \\ \mathrm{FO}(\mathrm{SC_i}, \mathrm{o_i})\ \mathrm{SC_i} = \mathrm{SC_1}, \ldots, \mathrm{SC_N} \\ \text{and } \mathrm{NO}(\mathrm{SC_i}, \mathrm{o_i})\ \mathrm{SC_i} = \mathrm{SC_1}, \ldots, \mathrm{SC_N} \end{cases} \quad (9)$$

$$\widetilde{W}(O_i) \begin{cases} 0 \text{ if } \max_{SC_i}(\mathrm{NO}(\mathrm{SC_i}, \mathrm{o_i})) = 0 \\ \dfrac{\max_{SC_i}\left(NO(SC_i,o_i)\right)^{\left(\max_{SC_i}\left(FO(SC_i,o_i)\right)\right)}}{No_{all}(SC_i,o_i) * FO_{all}(SC_i,o_i)} \%\mathrm{MaxValue} \quad \text{else} \\ \mathrm{FO}(\mathrm{SC_i}, \mathrm{o_i})\ \mathrm{SC_i} = \mathrm{SC_1}, \ldots, \mathrm{SC_N} \\ \text{and } \mathrm{NO}(\mathrm{SC_i}, \mathrm{o_i})\ \mathrm{SC_i} = \mathrm{SC_1}, \ldots, \mathrm{SC_N} \end{cases} \quad (10)$$

Experimentation made us assume that the weight functions $\acute{W}$ and $\widetilde{W}$ represent more faithfully the saliency of a given object o_i than simple probability measures. Nevertheless, the results are biased by the experimented dataset.

2) Scene Categories Similarities Computation

The aim of quantifying the similarity measure between two scenes categories SC_i and SC_j is to grant the classification process the possibility to switch to the most suitable scene category in a given clock time "*t*". (11) shows how to calculate the similarity measure α between two scene categories SC_i and SC_j (*i* can be equal to *j*).

Let SC_i and SC_j be two scene categories from the given dataset and $SC_iO_i=\{o_{i1},o_{i2}..o_{ik}\}$ be the set of objects belonging to all the scenes in the dataset labeled as SC_i and $SC_jO_j=\{o_{j1},o_{j2}..o_{jk'}\}$ be the set of objects belonging to all the scenes in the dataset labeled as SC_j.

We introduce the function α ($SC_i\_O_i$,$SC_j\_O_j$) which returns the similarity of SC_i toward SC_j as in (11).

$$\alpha(SC_iO_i, SC_jO_j) = \frac{Card(SC_iO_i \cap SC_jO_j)}{Card(SC_iO_i)} \quad (11)$$

The similarity function is non-commutative $\alpha(SC_io_i, SC_jo_j) \neq \alpha(SC_io_i, SC_jo_j)$.

B. Object's Ranking Function

The first step consists on providing a ranking function which sorts the set of objects o_i before their submission to the hidden Markov chain construction. It is very important and crucial to have the most significant and finest ranking function since the promoted scene categories depends deeply on it. Additionally, a truncation of insignificant (less salient) objects is made in order to reduce the length of the hidden Markov chain and thus the combinatory computation and also to protect the classification process to get lost. The ranking function relies exclusively on the weight measure as presented in (10). Fig. 2 shows how an input scene containing a set of objects will be ranked and truncated to a smaller and more salient set.

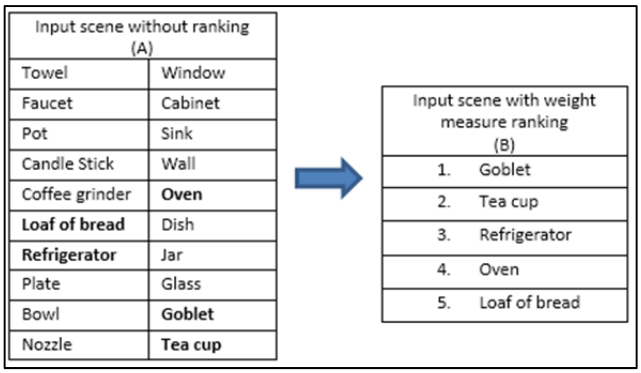

Fig. 2. Example of ranked and trancated objects from a given input scene.

From Fig. 2 we can see that some objects of the input scene in state (A) were deleted in state (B) ex: "Towel", "Pot", "Jar" …etc. while the most salient objects according to the weight measure calculated in (10) are ordered as presented in state (B). This ranking method will, in most of the time, directly guide the hidden Markov chain construction toward the most suitable scene category.

C. Analogy Between the Scene Classification Problem and the HMMs Architecture

In this subsection, we are going to demonstrate how the high level scene classification problem μ can be represented using an HMM λ model. Based on the definition given in Section III let be λ = (A, B, π) and on the definition given in the introduction, let be $\mu = (SC, P)$. In order to achieve the analogy, each component of μ will get its correspondent in λ. Table 1 shows the different correspondences.

TABLE I. ANALOGY BETWEEN THE HIGH LEVEL SCENE CLASSIFICATION PROBLEM AND THE HIDDEN MARKOV MODEL FORMAL DEFINITIONS

Hidden Markov model λ	**High level scene classification problem μ**
T: Observation sequence length (total number of clock times t)	T' : Cardinality of the set of properties P in a given the input scene S
N: set of hidden states $\{S_1, S_2 \ldots S_n\}$	SC : Set of scene categories $\{ SC_1, SC_2, \ldots SC_n\}$
M: set of observation symbols $\{o_1, o_2, \ldots o_n\}$	P: Set of properties $\{p_1, p_2 \ldots p_n\}$
Transition probabilities	Similarity between two SC as in (11)
Emission probabilities	Weight measure $\acute{W}$ based on a given Scene category SC_i as in (8)
Initial probabilities distribution π	Absolute weight measure $\widetilde{W}$ independent from any scene category SC_i as in (10)

Based on the comparison made in Table 1, we can easily see that the analogy between λ and μ is conceivable and indeed the scene classification problem μ can be represented by an HMM architecture λ. The same steps and algorithm as defined in Section III will be used to construct the hidden Markov model but in this case, for the purpose of scene

classification problem. Fig. 3 presents a theoretical example of an HMM while Table 3 shows the corresponding hidden Markov chain for observations generated in Table 2.

Let *S* be an input scene represented by an unsorted set of objects *O* as follow:

O = {Toothbrush, Phone, Bed, Book, TV}

Given a theoretical HMM λ containing just two scene categories: "Bedroom" and "Bathroom" are presented in Fig. 3.

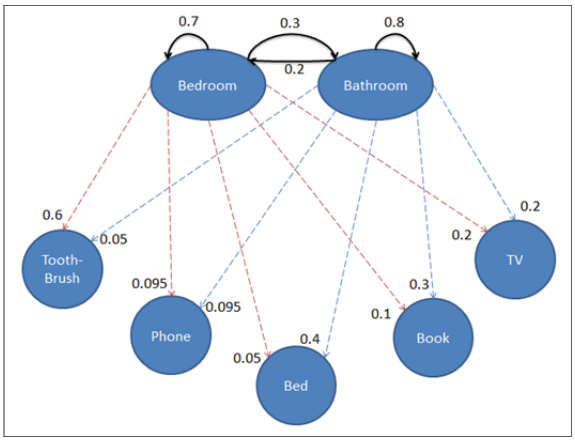

Fig. 3. Example of a hidden Markov model that contains 2 hidden states : "Bedroom" and "Bathroom" and 5 observations : "Toothbrush", "Phone", "Bed", "Book", "TV".

The object's ranking based on the absolute weight measure $\widetilde{W}$ calculated in (10) of the set *O* is presented in Table 2.

TABLE II. EXAMPLE OF ABSOLUTE WEIGHT MEASURE DISTRIBUTION EQUIVALENT TO INITIAL PROBABILITIES DISTRIBUTION

N°	Object	Absolute weight measure
O1	Toothbrush	0.6
O2	Bed	0.4
O3	Book	0.3
O4	TV	0.2
O5	Phone	0.095

In this example, we took only 3 objects (in green) from the set *O* to represent the input scene *S*. Thus, the objects O_4 and O_5 (in orange) won't be taken under consideration when constructing the discrete Markov chain.

In the second step, we construct the discrete Markov chain. It consists on applying the Viterbi algorithm presented in Section III (Algorithm 2) based on λ and following the top 3 ranked object in Table 2.

TABLE III. EXAMPLE OF DISCRETE MARKOV CHAIN

Observation (Objects)	O_1	O_2	O_3
Bathroom	.6	.6*.7*.05=.021	.072*.2*.1=.00144
Bedroom	.05	.6*.3*.4=.072	.072*.7*.3=.01512

To the hidden Markov chain to be constructed, the process chooses the most suitable hidden state (scene category) based on the upcoming and the current observation. The next hidden state can either stay the same or switch to another more salient (as describe in the Viterbi algorithm). At the end of the process, the hidden Markov chain can contain as much hidden states as existing observations (worst case scenario) or only one hidden state (best case scenario). From Table 3 we can see that the extracted hidden states are: "Bathroom", "Bedroom". In practice, to avoid useless calculations, we only go thought the max path (green in Table 3) omitting the other paths (orange in Table 3). In the following, the final inference step is presented which extracts the most suitable scene category from the discrete Markov chain.

D. Inference Algorithm

The common way to handle multiclass classification problems using hidden Markov models is by adopting the "one Vs all" approach [32]. In this paper we introduce a novel approach of multiclass classification that models the scene classification problem represented by the hidden Markov models such as only one HMM is used to classify all the classes represented by the different scene categories SC_i.

The inference process is developed to select the most suitable hidden state from the set generated by the hidden Markov chain. The method used to extract the most suitable scene category is simply by counting their frequency of appearance in the hidden Markov chain. If two scene categories get the same frequency of appearance, the priority goes to the scene category appearing first in the hidden Markov chain since the object are initially ordered by weight measure. Algorithm 3 will show how the process is executed.

Algorithm 3
Input : Discrete Markov chain DMC
Output : Chosen scene category : S

1- *Extract the priorities between scene categories existing in the DMC. (first to appear gets the highest priority) .*
2- *Count the number of occurrences of each scene category mentioned in the DMC.*
3- *If Tow (or more) scene categories get the same number of occurrence in the DMC, split the conflict with the priority calculated in Step 1.*

Algorithm 3 outputs the most suitable hidden state having as an input a constructed hidden Markov chain.

V. TEST AND RESULTS

In this section, we perform experiments of scene classification over the CVPR09 dataset [3]. First, we evaluate the accuracy of the proposed method varying its own input parameters, then, a comparison with the existing state of art's methods, which uses the same dataset for the scene classification problem, is provided.

A. Variying the objects taken

To avoid combinatory explanation, and to have a hidden Markov chain relatively small and exploitable, we varied the amount of objects taken into consideration in each input scene from *3, 5, 7* and *9* objects for all scenes categories. The truncation of chosen objects is made based on the weight measure $\widetilde{W}$ as calculated in (10). Fig. 4 shows the different obtained results.

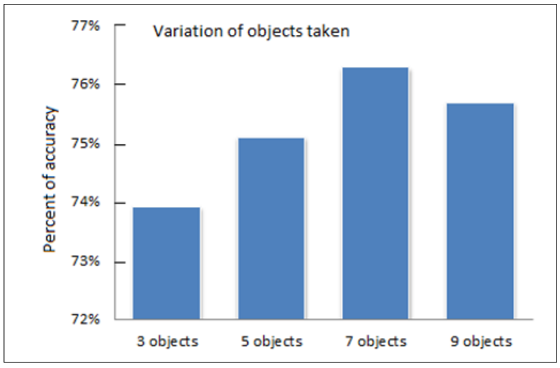

Fig. 4. Proposed method accuracy calculation variying the number of objects taken into consideration.

Fig. 4 shows a rise of accuracy in the classification process when objects are added (from *3* to *7* objects). In this part, the classification process is gaining practical information going until *76.28%* of accuracy for *7* objects. Above *7* objects, the accuracy drops down. The classification process is misled for getting useless information.

B. Varying the number of results

Illustrated in Fig. 5 is the rate of well-classified input scenes when *1, 2* then *3* scene categories are suggested by the classifier.

We notice from Fig. 5 that the accuracy increases when the suggested scene categories increase. This result claims that the hidden Markov chain holds, for most of the time, the right scene category but the inference algorithm fails to extract it. This discordance is sanctioned with a gap of *17%*. Nevertheless, this gap is contained in the *3* scene categories and proves that ranking made by the inference algorithm is reliable.

C. Summary of Proposed Method's Results

Table 4 presents all the results obtained by the proposed method while varying the different parameters: "Number of objects taken" and "Number of scene categories suggested". The minimum result is obtained when *9* objects are taken and *1* suggested scene category is made getting only *54.94%*, while the best result obtained is when *7* objects are taken and *3* scene categories are suggested getting *76.28 %*.

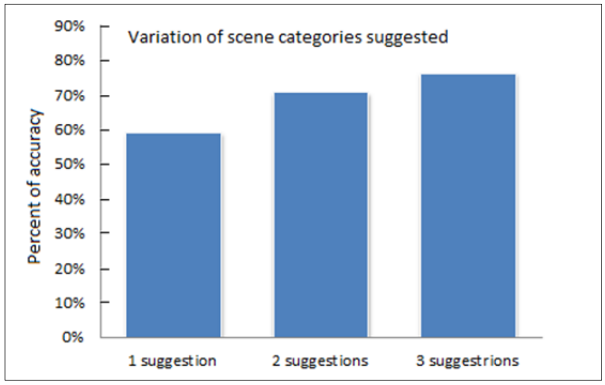

Fig. 5. Accuracy of the proposed method variying the number of scene category suggested.

TABLE IV. PRESENTATION OF ALL THE RESULTS OBTAINED BY THE PROPOSED METHOD VARIYING ALL THE PARAMETERS

Number of objects taken	Number of scene categories suggested	Obtained results
3	1	57.90 %
	2	68.57 %
	3	73.91 %
5	1	59.09%
	2	70.75 %
	3	75.09 %
7	1	58.69 %
	2	69.56 %
	3	**76.28%**
9	1	54.94 %
	2	68.97 %
	3	75.69 %

TABLE V. PRESENTATION OF A COMPARISION BETWEEN THE PROPOSED METHOD'S BEST RESULTS AND EXHISTING METHODS IN THE LITERATURE

Methods	Accuracy
ROI+GIST [3]	26.50 %
MM-SCENE [24]	28.00 %
DPM [23]	30.40 %
CENTRIST [25]	36.90 %
Object Bank [17]	37.60 %
DPM+GIST-Color [23]	39.00 %
DPM+SP [23]	40.50 %
DPM+SP+GIST-Color [23]	43.10 %
Singh et al. [33]	49.40 %
Zuo et al. [19]	52.24 %
Method in [34]	59.50 %
Juneja et al. [35]	63.18 %
Doersch et al. [36]	66.87 %
Method in [37]	68.20 %
Fc8-FV [38]	72.86 %
MPP[39]	75.67 %
Proposed method	**76.28** %

D. Comparaison to the State of the Art

In this subsection we are going to compare the proposed method's best result with the existing methods in the literature that uses the same dataset (CVPR09 [3]). Table 5 summarizes the comparison results.

From Table 5 we can see that the proposed method performs better in terms of scene classification accuracy compared to the other methods getting a rate of 76.28 % of accuracy.

VI. CONCLUSION AND PERSPECTIVES

In this paper was introduced a novel approach of classification using the hidden Markov model applied on the scene classification problem. After going through the learning process which computes all the entities of the hidden Markov model (HMM), the classification process starts by ranking the input objects called observations putting the most salient ahead. The construction of the discrete Markov chain starts by generating scene categories (hidden states) while examine the ranked objects one by one. At the end, the discrete Markov chain contains a set of scene categories. The final step consists on extracting the most suitable scene category from the discrete Markov chain. The obtained results are very satisfying - *76%* of well classified scene categories - while some improvements are still possible by changing the ranking

function and make it dynamic to the current hidden state or providing parallelism in the construction of the hidden Markov chain (more than one chain is constructed in the same time).

REFERENCES

[1] L. Li, H. Su, Y. Lim and F. Li, "Objects as Attributes for Scene Classification," Paper presented at the meeting of the ECCV Workshops (1), pp. 57-69, (2010).

[2] Z. Ghahramani, "An introduction to hidden Markov models and Bayesian networks," International journal of pattern recognition and artificial intelligence, pp.9-42, 2001.

[3] A. Quattoni and A.Torralba, "Recognizing Indoor Scenes," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[4] M. Arens and A. Ottlik, "Using behavioral knowledge for situated prediction of movements," In: Proc. 27th German Conference on Artificial Intelligence, pp.141–155, 2004.

[5] D. Pangercic, R. Tavcar, M. Tenorth and M. Beetz, "Visual scene detection and interpretation using encyclopedic knowledge and formal description logic," In Proceedings of the International Conference on Advanced Robotics (ICAR), pp. 605-610, 2009.

[6] L. Hotz and B. Neumann, "Scene interpretation as a configuration task," Künstliche Intelligenz 3, pp. 59-65, 2005.

[7] F. Baader and P. Hanschke, "A schema for integrating concrete domains into concept languages," In Proc 12th International Joint Conference on Artificial Intelligence (IJCAI), pp. 452–457, 1991.

[8] Neumann, Bernd and R. Möller, "On Scene Interpretation with Description Logics," Cognitive Vision Systems Lecture Notes in Computer Science, pp. 247-275, 2006.

[9] B. Hummel W. Thiemann and I. Lulcheva, "Description logic for vision-based intersection understanding," Proc. Cognitive Systems with Interactive Sensors (COGIS), 2007.

[10] A.N. Ghomsheh and A. Talebpour. "A new method for indoor-outdoor image classification using color correlated temperature," Int. J. Image Process 6, pp. 167-181. 2012.

[11] M. Szummer and R. W. Picard, "Indoor-outdoor image classification, " In IEEE International Workshop on Content-Based Access of Image and Video Database Proceedings, pp. 42-51,1998.

[12] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence 20, pp. 1254-1259, 1998.

[13] I. Biederman, R.J. Mezzanotte and J.C. Rabinowitz. "Scene Perception: Detecting and Judging Objects Undergoing Relational Violations," Cognitive Psychology, 1982.

[14] Hwang, S. Ju, and K. Grauman, "Learning the Relative Importance of Objects from Tagged Images for Retrieval and Cross-Modal Search," International Journal of Computer Vision, pp. 134-153, 2011.

[15] A. Swadzba, and S. Wachsmuth, "Indoor scene classification using combined 3D and gist features," Paper presented at the meeting of the ACCV, pp.201-215, 2011.

[16] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp. 273-280, 2003.

[17] L. Li, H. Su, E. Xing, and F. Li, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," In NIPS, pp.1378-1386, 2010. 3, 5

[18] K. Grauman, and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," Paper presented at the meeting of the ICCV Tenth IEEE International Conference on, pp. 1458-1465, 2005.

[19] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang and X. Jiang, "Learning discriminative and shareable features for scene classification," In Proceedings of European Conference on Computer Vision (ECCV), pp. 552-568, 2015.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," Proceeding CVPR '06 Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2169-2178 , 2006.

[21] A. Oliva, and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of computer vision, pp. 145-175, (2001).

[22] Li, Lu, and Siripat Sumanaphan. "Indoor Scene Recognition.", unpublished, 2011. [online]. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.374.9006&rep=rep1&type=pdf

[23] M. Pandey, and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," In Computer Vision (ICCV) 2011 IEEE International Conference on, pp. 1307-1314, (2011).

[24] J. Zhu, L. Li, F. Li, and E. Xing, "Large margin learning of upstream scene understanding models," In Advances in Neural Information Processing Systems, pp. 2586-2594, (2010).

[25] J. Wu, and J. Rehg, "CENTRIST: A visual descriptor for scene categorization," IEEE transactions on pattern analysis and machine intelligence, pp. 1489-1501, 2011.

[26] L. Nanni, and A. Lumini, "Heterogeneous bag-of-features for object/scene recognition," Applied Soft Computing, pp. 2171-2178, (2013).

[27] L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNs: objects, scales and dataset bias," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 571-579, (2016).

[28] M. Sadeghi, and A. Farhadi, "Recognition using visual phrases." In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1745-1752, (2011).

[29] S. Wei, and H. Hagras, "A Big-Bang Big-Crunch fuzzy logic based system for sports video scene classification." In IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 642-649, 2016.

[30] E. Elbaşı, "Fuzzy logic-based scenario recognition from video sequences," Journal of applied research and technology, pp. 702-707, (2013).

[31] P. Baiget, C. Tena, F. Roca and J. Gonzàlez., "Automatic learning of conceptual knowledge in image sequences for human behavior interpretation," In Iberian Conference on Pattern Recognition and Image Analysis, pp. 507-514, (2007).

[32] G. E. Hinton and A. D. Brown, "Training many small hidden markov models." Proc. of the Workshop on Innovation in Speech Processing, (2001).

[33] S. Singh, A. Gupta, and A. Efros, "Unsupervised discovery of mid-level discriminative patches," In Proceedings of European Conference on Computer Vision (ECCV), (2012).

[34] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," In ICML, pp. 647-655, (2014).

[35] M. Juneja, A. Vedaldi, C. Jawahar and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification,". In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 923-930, (2013).

[36] C. Doersch, A. Gupta and A. Efros, "Mid-level visual element discovery as discriminative mode seeking," In Advances in neural information processing systems, pp. 494-502, (2013).

[37] L. Liu, C. Shen, L. Wang, A. Hengel and C. Wang. "Encoding high dimensional local features by sparse coding based fisher vectors," In Neural Information Processing Systems, pp. 1143–1151, 2014.

[38] M. Dixit, S. Chen, D. Gao, N. Rasiwasia and N. Vasconcelos, "Scene classification with semantic fisher vectors." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2974-2983, (2015).

[39] D. Yoo, S. Park, J. Lee, and I. Kweon, "Multi-scale pyramid pooling for deep convolutional representation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 71-80, (2015).

[40] C. Lawrence, R. Vedantam, and D. Parikh, "Adopting abstract images for semantic scene understanding, " IEEE transactions on pattern analysis and machine intelligence, pp.627-638, (2016).