



# A novel feature selection method using generalized inverted Dirichlet-based HMMs for image categorization

Rim Nasfi<sup>1</sup> · Nizar Bouguila<sup>1</sup>

Received: 7 August 2021 / Accepted: 19 February 2022 / Published online: 14 March 2022  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

## Abstract

Hidden Markov Models (HMMs) have consistently been a powerful tool for performing numerous challenging machine learning tasks such as automatic recognition. The latter perceives all objects of the universe through information carried by their characteristics or features. However, not all available data is always valuable for distinguishing between the different objects, scenes, scenarios; referring analogically to states. More often than not, automatic recognition is accompanied by a feature selection to reduce the number of collected features to a relevant subset. Although sparse, the majority of literature resources available on feature selection for HMMs, presuppose either a single Gaussian or employ a Gaussian mixture model (GMM) as emission distribution. The proposed method builds upon the feature saliency model introduced by Adams, Cogill, and Beling (in *IEEE Access* 4:1642–1657), and is adjusted to handle complex multidimensional data by using as a novel experiment, GID (Generalized Inverted Dirichlet) mixture models) as emission probabilities. We make use of an Expectation-Maximization (EM) algorithm (Dempster et al. in *J R Stat Soc* 39(1):1–22) to compute maximum a posteriori (MAP) [Gauvain and Lee in *IEEE Transact Speech Audio Process* 2(2):291–298] estimates for model parameters. The complete inference and parameter estimation of our GID-FSHMM (GID Feature Selection-based HMM) are detailed in this work. Automatic recognition applications such as facial expression recognition and scenes categorization demonstrate comparable to higher performance compared to the extensively used Gaussian mixture-based HMM (GHMM), the Dirichlet-based (DHMM) and the inverted Dirichlet-based HMM (IDHMM) without feature selection and also when the latter is embedded in all of the aforementioned models.

**Keywords** Hidden Markov models · Generalized inverted Dirichlet · Feature selection · Automatic recognition · Facial expressions recognition · Scene categorization

## 1 Introduction

The successful application of HMMs to a great number of areas ranging from speech recognition to image categorization broke new grounds by bringing many extensions and novelties not only in terms of the methods used along with HMMs to better their performance but also in the volume and diversity of data collected for analysis using these methods. There is no doubt that this expansion of data, types of information, and features contributed enormously to refining and improving machine learning tasks and methods. Nevertheless, it has triggered a considerable amount of problems

and challenges namely the formidable curse of dimensionality often resulting from the manipulation of high-dimensional data. For example, in clustering tasks, it is a widely held view that the more information, data, and features we manipulate, the better an algorithm is expected to perform [48]. However, this is not the case in practice. Many features can be just “noise” and may cause the finest pattern recognition and machine learning techniques to struggle as a result of irrelevancy and thus degrade the modelling performance [18]. Thereby, feature selection is used to increase modelling performance since it allows eliminating noise in the data, speeding up the models’ training and prediction, decreasing overfitting odds and most importantly reducing the computational cost after disregarding many features.

We intend by feature selection, the process of decreasing the number of gathered features to a relevant subset of features and is usually used to counter the curse of

✉ Rim Nasfi  
r\_nasfi@encs.concordia.ca

<sup>1</sup> Concordia Institute for Information Systems Engineering,  
1455 De Maisonneuve Blvd. West, Montreal, QC, Canada

dimensionality [3]. Aside from feature extraction, which is a separate problem, feature selection determines relevant features from a given set of features, whereas feature extraction generates new features from a given set. Unlike feature extraction, feature selection does not come up with new features nor does it amend the primary features.

The primary inducement for adopting feature selection strategies is their important potential to improve modelling and generalization capabilities if performed reliably and properly [4]. Applying feature selection permits taking into consideration the significant contribution of feature screening to the classification process. In fact, each different feature contributes differently to the classification structure based on its degree of relevance [21, 68]. The latter is intended to be determined to improve our models' performance, in particular using simultaneous feature selection and classification in the course of an unsupervised process which is considered to be one of the most challenging problems in data mining and machine learning. In practice, the said case implies selecting features without a priori knowledge about data labels.

In most applications of HMM, features are pre-selected based on domain knowledge, and the feature selection procedure is completely omitted. Usually, to train HMMs, even in the case where feature selection is considered features are selected traditionally. That is, features are selected in advance either based on already available data or relying on experts' knowledge. These practices are the result of the scarcity of literature in terms of unsupervised feature selection methods specific for HMMs [3] [30], not to mention the high computational cost of wrapping methods. Despite the extensive research and investigations that are made on feature selection in their general case, methods specific for HMMs are lacking. Feature selection methods for HMMs and mixture models are seldom treated as a joint topic. Most importantly, the use of generalized inverted Dirichlet mixtures to model the emission probabilities within the HMM framework together with feature selection as an embedded process is unprecedented. In this work, we propose a fully customized feature selection methodology with a complete empirical and experimental study of Feature Saliency embedding into the GID-based HMM.

Feature selection plays a major role in speeding the learning process and refining the models' interpretation. It can drastically minimize the risk of overfitting and mitigates the effects of the curse of dimensionality [39]. Above-mentioned, the feature selection process is embedded in the training of the HMM, which represents the main takeaway from this research work.

Our vision of integrating feature selection in the HMM framework holds beyond the simple procedure of solely combining state of the art feature selection methods such as in the case of [55], where several ranking methods like Bhattacharyya distance [43], entropy and Wilcoxon [33],

have been used to reduce the number of features fed to the HMM. As far as we are concerned, we are resolute to use the feature saliency as in [1], thoroughly embedded in the HMM framework making only one core method ready for use directly to treat any set of features.

The work presented in this manuscript can also be viewed intellectually at two different levels. First, it allows the integration of the non-conventional feature selection techniques into the framework of HMM, second, it allows the use of GID mixtures as a premiere to model data fed to HMM specifically emission distributions.

The remainder of this paper is organized as follows: In section 2 we summarize the previous works adopting HMMs. Then, we outline some of the applications using general feature selection methods along with HMMs as a predictive model. In section 3, we present our GID-FSHMM and explain all the corresponding integration steps. The subsequent section 4 showcases real-life problems experimentation and analyses obtained results. Finally, the paper closes with a summary of work and concluding remarks.

## 2 Related work

### 2.1 Hidden Markov models

In this section, we recall a handful of background information on HMMs, while focusing on previous related work using HMMs conjointly with feature selection. Hidden Markov models are a ubiquitous tool commonly utilized to model time series data [34] [37] with applications across numerous areas. Used for decades in speech recognition [62], text classification [15, 44], face recognition [58] and fMRI data analysis [26], HMMs represent a powerful statistical tool that have proven to be not only useful but also efficient in various machine learning-based applications.

An HMM consists mainly of two distinct sequences of states. The first is a sequence of hidden states modelled by a Markov chain [7], the second is a sequence of observed events or features related to the hidden states. The typical purpose behind using HMMs is to represent probability distributions over sequences of observations, with the assumption that the observations are discrete. Therefore, the hidden states sequence can be estimated from the sequence of correlated observations. It is possible to specify an HMM by an initial probability, a matrix of transition probabilities between the states, and a set of parameters of the emission probability distribution which will be more focused on later in this paper. Most importantly, an HMM is outlined by two fundamental properties. Firstly, it assumes that an observation at time  $t$  is generated by some process whose state  $h_t$  is hidden from the observer. Second, it implied that the state of the said hidden process fulfills the Markov property

[31]; that is, given the value of  $h_{t-1}$ , the current state  $h_t$  is independent of all the states before the time  $t - 1$ . Thus, the observed features are modelled meeting the property of conditional independence given the state sequence. At the application level, the learning of parameters is simply finding the best set of state transitions and emission probabilities amid the states of the model. Consequently, an output sequence or a set of sequences is specified. At each time a state sequence is handled, there is a corresponding vector of observations composed of features collected from various sources. However, not all features are likely to be useful to the model. That is why, to build a rigorous model, we ought to remove all features that do not contribute to its usefulness, without degrading its accuracy.

## 2.2 Feature selection and its application with HMMs

Feature selection is a wide research area and many methods to reduce a given set of features have been implemented in both supervised and unsupervised contexts [47].

Typically, feature selection techniques are present in the state of the art under three main categories, namely, filters, wrappers, metaheuristic methods and embedded [3]. While filter methods such as information gain [49, 71], Pearson's correlation coefficient [35] and variance threshold [72], treat the evaluation of all features and return a relevant subset out of them apart from the model building process, wrapper methods tend to optimize the classifier's performance for the most part. Wrappers, which commonly adopt either forward selection [45], backward elimination [20] or recursive feature elimination [22, 53, 78], identify the relevant features depending on the learning algorithm. That is, when using wrappers, the model itself is built depending on a certain subset of features and its performance is measured upon particular criteria. Methods relying on metaheuristic algorithms tackle feature selection as an optimization problem. Composed but not limited to evolution-based algorithms such as Genetic Algorithm [42], these methods obtain the optimal solution thanks to their simplicity, flexibility and their capability to avoid local optima [59]. They start their feature selection process by generating random solutions that do not require heavy derivatives calculations and carry on an exploration phase to thoroughly investigate search space and identify promising areas. Embedded methods namely  $L_1$  regularization and decision trees [9, 36] aspire to simultaneously select the features and build the model. Although filter methods exhibit a significant low time complexity, they are usually criticized for ignoring certain informative features [63]. On the other hand, metaheuristic and wrapper-based methods evaluate the usefulness of selected features using learner's performance and can thus be more complex but still not very time-consuming. However, it has been proved that other optimization algorithms, namely embedded

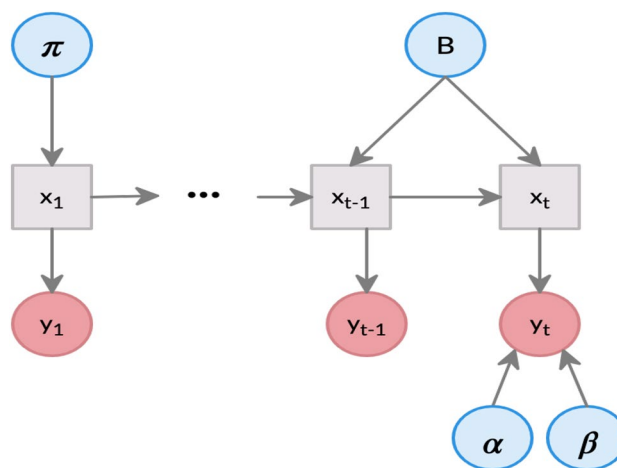
methods, can be more efficient given the fact that they not only improve the performance of the model but also facilitate results analysis. Indeed, there is a significant complexity compromise when it comes to using embedded methods, but these methods succeeded in adapting to several types of data and can be used with the majority of machine learning models. Embedded methods are also very useful when investigating relationships between features, which is an arising challenge nowadays.

In particular, feature selection for HMMs is driven by a crucial need to determine which feature to use in the model. Despite being investigated in numerous general and mixture models-based studies [25] [48], feature selection methods dedicated to HMMs are particularly limited. In fact, in the majority of applications, features are selected beforehand based on domain knowledge, and a consonant feature selection procedure is fully lacking [56] [74]. Clearly, transformation methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) do reduce the number of features in the model and for this same reason, they have been integrated into HMMs in [6, 80]. However, the mentioned methods do not really act as feature selection techniques as they are not able to eliminate data streams, and hence they merely are considered feature extraction techniques.

Embedded or integrated feature selection approaches, which are the main focus in this manuscript, ought to consider the whole set of features at once. These features serve as an input to the maximizing learning algorithm that is deployed to optimize the models' performance. As an output, the reduced set of features, as well as the models' parameters, are generated. Hence, an embedded feature selection method is identified as a simultaneous selection of features and model construction. This combines both the wrappers and filters advantages of respectively selecting feature subsets concerning a specific learning algorithm and the computational efficiency [1]. As previously indicated, one of the embedded methods for feature selection is the classification and regression trees [57]. The latter applies a recursive splitting of the feature space to generate a classification model. Features identified as the ones being involved in improving the model will exclusively be included in the learning algorithm. In contrast to the mixture models, context [11, 24, 48], literature about feature selection integration into the HMM framework is somewhat narrow. Nearly all HMM-based adaptations of feature selection were based on what is also known as the concept of feature saliency, which has been defined by [48] as a metric associated with a given feature, that is the probability that the said feature is relevant. Zhu et al. [81] are among the first to use a jointly embedded estimation and feature selection method, where they apply a variational Bayesian framework to the end of salient features inference.

They use the implemented method to simultaneously infer the number of hidden states as well as the models' parameters. The adopted approach showed interesting results, however, the use of the variational Bayesian sometimes manifested a significant underestimation of the variance for the approximate distribution. Adams et al. [1] put forward a feature saliency model using hidden Markov models. The main idea is to use feature saliency variables to represent the probability that a given feature is relevant, by drawing a distinction between state-dependent and state-independent distributions. The said model operates in the case where the number of hidden states is known. For the matter, it provides a maximum a posteriori based estimation that selects the most relevant features using an Expectation-Maximization (EM) algorithm. This approach takes advantage of the already specified number of states to provide maximum a posteriori estimates and save the most relevant features by applying an Expectation-Maximization algorithm [14]. Moreover, Zheng et al. [79] adopted a strategy that combines a hidden Markov model, a localized feature saliency measure, and two *t* Student distributions for the purpose of distinguishing between relevant and non-relevant features. This strategy made it possible to accurately model emission parameters for each hidden state. Similarly to [81], the parameter estimation was operated using a variational Bayes framework. More recently, Fons et al. [30] incorporated Adams' feature saliency HMM (FSHMM) [2] into a dynamic asset allocation system. The authors applied their HMM-based feature selection method to train their systematic trading system by testing its performance on real-life data. It showed that even without a financial expert involvement, the results reached a decent accuracy allowing the model to objectively contribute to portfolio construction and to prevent biases in the feature selection process. From their side, authors in [12] proposed a feature selection algorithm embedded in an HMM applied to gene expression time-course data, and they succeeded in reducing the feature domain by up to 90% leaving only a few but relevant features. The notable drawback of the mentioned work is that features deemed as irrelevant are eliminated and hence can drastically affect the models' accuracy in the case the aforementioned features seem to be relevant after treatment.

There is indubitably a significant challenge when analyzing dense data, that is dealing with the saliency parameters besides those imperative for the model itself. As a consequence, the parameter estimation can sometimes be a sensitive task, not to mention the huge impact that the number of needed hidden states has on the said estimation. For this particular reason, we need to adapt the model in a way that it can handle the modelling of the data using a lower number of parameters to come up with the most relevant features from the candidate sets.



**Fig. 1** The Hidden Markov model: Grey squares represent latent variable, pink circles are observations, and blue circles represent model parameters where  $\alpha$  and  $\beta$  are GID parameters (colour figure online)

## 3 The proposed GID-FSHMM model

### 3.1 Feature selection integration in Hidden Markov model

In this section, we start by presenting the Hidden Markov Model and we recall the feature saliency concept that we will embed in the HMM. Then, we

#### 3.1.1 The Hidden Markov model

We consider a HMM with continuous emissions and  $K$  states. We put  $y = \{y_0, y_1, \dots, y_T\}$  the sequence of observed data with  $y_t \in \mathbb{R}^L$ , where  $T$  designates the time factor and  $L$  is the number of features. The observation for the  $l$ -th feature at time  $t$ , which is represented by the  $l$ -th component of  $y_t$ , is denoted by  $y_{lt}$ .

Let  $x = \{x_0, x_1, \dots, x_T\}$  be the sequence of hidden data. The transition matrix of the Markov chain associated to this sequence is denoted as  $B = \{b_{ij} = P(x_t = j | x_{t-1} = i)\}$  and  $\pi$  is the initial state probability. Thus the complete data likelihood can be expressed as:

$$p(x, y | \Lambda) = \pi_{x_0} c_{x_0}(y_0) \prod_{t=1}^T b_{x_{t-1} x_t} c_{x_t}(y_t) \quad (1)$$

where  $\Lambda$  is the set of model parameters, and  $c_{x_t}(y_t)$  is the emission probability given state  $x_t$ .

In our feature selection hidden Markov model (FSHMM) we apply a feature saliency approach over the emission probability distribution in order to select the relevant features and to estimate our parameters [48]

The graphical model of the GID Hidden Markov model can be seen in Fig. 1.

### 3.1.2 Feature saliency-based Hidden Markov model

The feature-saliency based HMM measures the relevancy of a certain feature as follows; if the latter’s distribution is dependent on the underlying state, the feature is believed to be relevant. In the case where its distribution is independent of the state, the feature is considered irrelevant [1].

Thus, we put a set of binary variables  $z = \{z_1, \dots, z_L\}$  indicating the relevancy of features, that is  $z_l = 1$  if the  $l$ -th feature is relevant and  $z_l = 0$  if it’s irrelevant. The feature saliency  $\rho_l$  is the probability that the  $l$ -th feature is relevant.

In this work we assume that all features are conditionally independent given the state. Hence, the conditional distribution of  $y_t$  given  $z$  and  $x$  can be written as follows:

$$p(y|z, x_t = i, \Lambda) = \prod_{l=1}^L r(y_{it}|\theta_{il})^{z_l} q(y_{it}|\lambda_l)^{1-z_l} \tag{2}$$

where  $r(y_{it}|\theta_{il})$  is the conditional feature distribution for the  $l$ -th feature with state-dependent parameters  $\theta_{il}$  which later will be detailed with depending on the adopted type of mixture, and  $q(y_{it}|\lambda_l)$  is the state independent feature distribution with parameters  $\lambda_l$ .

$\Lambda = \{\theta, \rho, \lambda\}$  is the set of all our FSHMM model parameters. The marginal probability of  $z$  is:

$$p(z|\Lambda) = \prod_{l=1}^L \rho_l^{z_l} (1 - \rho_l)^{1-z_l} \tag{3}$$

The joint distribution of  $y_t$  and  $z$  given  $x$  can be expressed as:

$$p(y_t, z|x_t = i, \Lambda) = \prod_{l=1}^L [\rho_l r(y_{it}|\theta_{il})]^{z_l} [(1 - \rho_l) q(y_{it}|\lambda_l)]^{1-z_l} \tag{4}$$

The marginal distribution for  $y_t$  given  $x$  over all values of  $z$  is:

$$c_{x_t}(y_t) = p(y_t|x_t = i, \Lambda) = \prod_{l=1}^L (\rho_l r(y_{it}|\theta_{il}) + (1 - \rho_l) q(y_{it}|\lambda_l)) \tag{5}$$

The complete data likelihood of the FSHMM can thus be written as:

$$p(x, y, z|\Lambda) = \pi_{x_0} p(y_0, z|x_0, \Lambda) \prod_{t=1}^T b_{x_{t-1}, x_t} p(y_t, z|x_t, \Lambda) \tag{6}$$

The form of  $q(\cdot)$  indicates our prior knowledge about the distribution of the non-salient features. We put  $q(\cdot)$  and  $r(\cdot)$  to follow an inverted generalized Dirichlet distribution, as this can lead to better results for the reasons explained earlier in this paper.

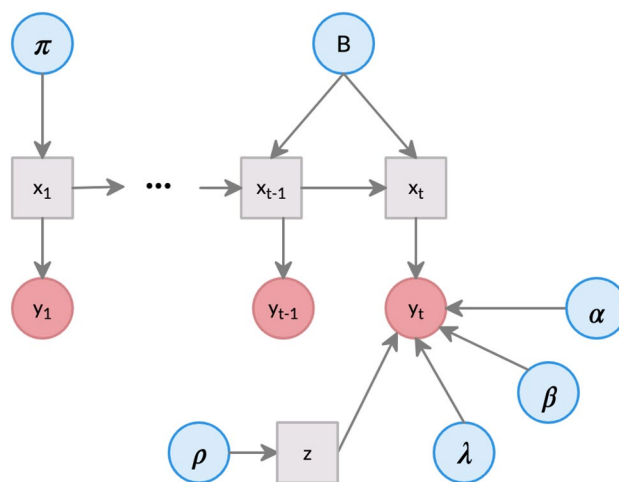


Fig. 2 The feature saliency GID-based Hidden Markov Model: Grey squares represent latent variable, pink circles are observations, and blue circles represent model parameters (colour figure online)

In this work, the state-dependent and the state-independent distributions are assumed to be GID mixtures. Accordingly, the set of model parameters for the GID-FSHMM is  $\Lambda = \{\pi, B, \alpha, \beta, \rho, \lambda\}$ . Figure 2 shows the feature saliency GID-based HMM.

## 3.2 GID mixtures and integration into the FSHMM framework

### 3.2.1 Generalized Inverted Dirichlet

The choice of GID is backed by the several interesting mathematical properties that this distribution has. These properties allow for a representation of GID samples in a transformed space where features are independent and follow inverted Beta distributions. Adopting this distribution lets us take advantage of conditional independence among features. This interesting strength is used in this paper to develop a statistical model that handles not only positive data but also feature selection.

Let  $\vec{X}$  a  $D$ -dimensional positive vector following a GID distribution. The joint density function is given by Lingappaiah [52] as:

$$p(\vec{X}|\vec{\alpha}, \vec{\beta}) = \prod_{d=1}^D \frac{\Gamma(\alpha_d + \beta_d)}{\Gamma(\alpha_d)\Gamma(\beta_d)} \frac{X_d^{\alpha_d-1}}{\left(1 + \sum_{l=1}^d X_l\right)^{\eta_d}} \tag{7}$$

where  $\vec{\alpha} = [\alpha_1, \dots, \alpha_D]$ ,  $\vec{\beta} = [\beta_1, \dots, \beta_D]$ .  $\eta$  is defined such that  $\eta_d = \alpha_d + \beta_d - \beta_{d+1}$  for  $d = 0, \dots, D$  with  $\beta_{D+1} = 0$ .

The GID estimation is made simple thanks to an essential property, that is if there exists a vector  $\vec{X}$  that follows a

GID distribution, then we can come up with another vector  $\vec{W}_n = [\vec{W}_{n1}, \dots, \vec{W}_{nD}]$  where each element follows an inverted Beta (IB) distribution following the transformation:

$$W_{nd} = f(X_{nd}) = \begin{cases} X_{nd}, & d=1 \\ \frac{X_{nd}}{1+X_{n1}+\dots+X_{nd-1}}, & d=2, \dots, D \end{cases} \quad (8)$$

Then, the multivariate extension of the 2-parameters inverted Beta distribution is given by:

$$p_{IBeta}(W_{nd}|\alpha_{jd}, \beta_{jd}) = \frac{\Gamma(\alpha_{jd} + \beta_{jd})}{\Gamma(\alpha_{jd})\Gamma(\beta_{jd})} \frac{W_{nd}^{\alpha_{jd}-1}}{(1 + W_{jd})^{(\alpha_{jd}+\beta_{jd})}} \quad (9)$$

The mean of IB is given by:

$$E(W_d) = \frac{\alpha_d}{\beta_d - 1} \quad (10)$$

The variance of IB is given by:

$$Var(W_d) = \frac{\alpha_d(\alpha_d + \beta_d - 1)}{(\beta_d - 2)(\beta_d - 1)^2} \quad (11)$$

### 3.2.2 GID mixture model

Let us consider a data set  $\mathcal{X}$  of  $N$   $D$ -dimensional positive vectors,  $\mathcal{X} = (\vec{X}_1, \vec{X}_2, \dots, \vec{X}_N)$ . We assume that  $\mathcal{X}$  is governed by a weighted sum of  $M$  GID component densities with parameters  $\Theta = (\vec{\theta}_1, \vec{\theta}_1, \dots, \vec{\theta}_M, p_1, p_2, \dots, p_M)$  with  $\vec{\theta}_j$  is the vector of parameters of the  $j$ -th component and  $p_j$  are the mixing weights which are positive and sum to one [4]:

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^M p_j p(\vec{X}_i|\vec{\Theta}_j) \quad (12)$$

where  $p(\vec{X}_i|\vec{\Theta}_j)$  is the GID distribution with  $\vec{\Theta}_j = (\alpha_{j1}, \beta_{j1}, \alpha_{j2}, \beta_{j2}, \dots, \alpha_{jD}, \beta_{jD})$  is the set of parameters defining the  $j$ -th component. Furthermore, in mixture-based clustering, each data point  $\vec{X}_i$  can be assigned to all classes with different posterior probabilities  $p(j|\vec{X}_i)$ . Therefore, a factorization of the posterior probability can simply be expressed as:

$$p(j|\vec{Y}_i) \propto p_j \prod_{l=1}^D p_{IBeta}(X_{il}|\theta_{jl}) \quad (13)$$

where  $X_{i1} = Y_{i1}$  et  $X_{il} = \frac{Y_{il}}{1+\sum_{l=1}^D Y_{il}}$  for  $l > 1$ ,  $p_{IBeta}(X_{il}|\theta_{jl})$  is an inverted Beta distribution with  $\theta_{jl} = (\alpha_{jl}, \beta_{jl}), l = 1, \dots, D$

In this fashion, the clustering structure underlying  $\mathcal{X}$  is the same as that underlying  $\mathcal{Y} = (\vec{Y}_1, \dots, \vec{Y}_N)$ , and it can be described by the following mixture model with conditionally independent features:

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^M p_j \prod_{l=1}^D p_{IBeta}(X_{il}|\theta_{jl}) \quad (14)$$

### 3.2.3 GID mixture-based FSHMM

As a first attempt in the context of feature saliency-driven HMMs, to the extent of our knowledge, we are using a mixture of GID as emission probabilities of our FSHMM. Gaussian mixtures, in particular, have seldom been tested previously and applied successfully [48] [81]. Assuming the relevant feature distribution is represented by a mixture of  $M$  GID distributions, we let  $\Phi = \{\phi_{1t}, \dots, \phi_{Mt}\}$  be the set of variables indicating the mixture component, where  $\phi_m = 1$  if observation  $t$  comes from the  $m^{\text{th}}$  mixture and  $\phi_{mt} = 0$  otherwise. To indicate the probability that the observation comes from the  $m^{\text{th}}$  mixture, given the state, we put  $\omega_{im}$ . In this regard, the set of model parameters  $\Lambda$  becomes  $\{\pi, B, \alpha, \beta, \rho, \lambda, \omega\}$ . The idea behind the GID-based FSHMM is to suppose that a given feature  $y_{lt}$  is generated from a mixture of two univariate distributions. The first one is supposed to generate relevant features and is distinct for each cluster. The second is common to all clusters in a way that it is independent of class labels, and generates irrelevant features. This purpose can be formulated as follows.

The marginal probability of  $\phi_i$  can be expressed as

$$p(\Phi|\Lambda) = \prod_{m=1}^M \omega_{im}^{\phi_{im}} \quad (15)$$

In the same manner as in (3), we assume the features are conditionally independent given the state. Thus, the conditional distribution of  $y_t$  given  $x, y$  and  $\Phi$  can be formulated as

$$p(y_t|\Phi, z, x_t = i, \Lambda) = \prod_{l=1}^L [r(y_{lt}|\alpha_{ilm}, \beta_{ilm})^{z_l} q(y_{lt}|\alpha_{\lambda,ilm}, \beta_{\lambda,ilm})^{1-z_l}]^{\phi_{mt}} \quad (16)$$

The joint distribution of  $y_t, \Phi$ , and  $z$  given  $x$  is

$$\begin{aligned} p(y_t, \Phi, z|x_t = i, \Lambda) &= p(y_t|\Phi, z, x_t = i, \Lambda)p(\Phi|\Lambda)p(z|\Lambda) \\ &= \prod_{m=1}^M \left[ \omega_{im} \prod_{l=1}^L [\rho_l r(y_{lt}|\alpha_{ilm}, \beta_{ilm})^{z_l}] \right. \\ &\quad \left. [(1 - \rho_l)q(y_{lt}|\alpha_{\lambda,ilm}, \beta_{\lambda,ilm})^{1-z_l}]^{\phi_{mt}} \right] \end{aligned} \quad (17)$$

The marginal distribution for  $y_t$  given  $x$  is obtained by summing (17) over  $z$  and  $\Phi$  such as

$$\begin{aligned}
 c_{x_t}(y_t) &= p(y_t|x_t = i, \Lambda) \\
 &= \sum_{m=1}^M \omega_{im} \prod_{l=1}^L (\rho_l r(y_{tl}|\alpha_{ilm}, \beta_{ilm}) + (1 - \rho_l)q(y_{tl}|\alpha_{\lambda,ilm}, \beta_{\lambda,ilm}))
 \end{aligned}
 \tag{18}$$

The complete data likelihood for the FSHMM with GID emissions is

$$p(x, y, z, \Phi|\Lambda) = \pi_{x_1} p_{IBeta}(y_1, \Phi, z|x_1, \Lambda) \prod_{t=2}^T b_{x_{t-1}, x_t} p_{IBeta}(y_t, \Phi, z|x_t, \Lambda)
 \tag{19}$$

### 3.3 Parameter estimation of the GID-FSHMM

#### 3.3.1 Update equations for FSHMM parameters

In order to perform the estimation of parameters, we opt for using the EM algorithm, referred to as Baum-Welch when applied in the context of HMMs [8, 62]. We use this algorithm to calculate the maximum-likelihood (ML) estimates for the model parameters. For the part where we evaluate the features, we are bound to place priors on the parameters to compute the maximum a posteriori (MAP) estimates [32]. We need to go over the two steps of the Baum-Welch algorithm. First, in the E-step we need to find the expected value of the complete log-likelihood taking into consideration the data and the underlying model parameters. Second, in the M-step we proceed to maximize the expectation computed in the previous step in order to figure the next set of model parameters out. The Baum-Welch is iterated until an experimentally determined stopping threshold is met. The  $Q$  function designates the expectation of the complete log-likelihood and is given by:

$$\begin{aligned}
 Q(\Lambda, \Lambda') &= \mathbb{E}[\log p(x, y, z, \Phi|\Lambda)|y, \Lambda'] \\
 &= \sum_{x, z, \Phi} \log(p(x, y, z, \Phi|\Lambda)|\Lambda') p(x, z, \Phi|y, \Lambda')
 \end{aligned}
 \tag{20}$$

where  $\Lambda$  and  $\Lambda'$  represent the set of model parameters for the current iteration and the set of parameters from the previous iteration respectively. We place priors on the parameters and calculate the MAP estimates with an eye toward the automatic feature assessment and selection. Hence the  $Q$  is changed by adding  $G(\Lambda)$  the prior on the model parameters such as:

$$Q(\Lambda, \Lambda') + \log G(\Lambda)
 \tag{21}$$

By analogy to the previously explained EM procedure, the complete log-likelihood  $Q$  is calculated in the E-step (20), then the  $\log G(\Lambda)$  is added up and equation (21) is maximized in the M-step. For this matter, several probabilities are needed for the FSHMM, the E-step takes in charge the computation of the following quantities:

$$\zeta_t(i) = \mathbb{P}(x_t = i|y, \Lambda),
 \tag{22}$$

$$\xi_t(i, j) = \mathbb{P}(x_t = i, x_{t+1} = j|y, \Lambda)
 \tag{23}$$

where  $\zeta_t(i)$  et  $\xi_t(i, j)$  are respectively the conditional state probabilities and the conditional transition probabilities. These quantities are calculated using the forward-backward algorithm. As a result, the following quantities are what the E-step probabilities turn out to be after iterating the forward-backward algorithm

$$\begin{aligned}
 \delta_{ilm_t} &= p(y_{lt}, z_l = 1|\phi_{mt} = 1, x_t = i, \Lambda') \\
 &= \rho_l p(y_{lt}|\alpha_{ilm}, \beta_{ilm}),
 \end{aligned}
 \tag{24}$$

$$\begin{aligned}
 \epsilon_{ilm_t} &= p(y_{lt}, z_l = 0|\phi_{mt} = 1, x_t = i, \Lambda') \\
 &= (1 - \rho_l)q(y_{lt}|\alpha_{\lambda,ilm}, \beta_{\lambda,ilm}),
 \end{aligned}
 \tag{25}$$

$$\begin{aligned}
 \tau_{ilm_t} &= p(y_{lt}|\phi_{mt} = 1, x_t = i, \Lambda') \\
 &= \delta_{ilm_t} + \epsilon_{ilm_t},
 \end{aligned}
 \tag{26}$$

$$\begin{aligned}
 u_{ilm_t} &= p(z_l = 1, x_t = i, \phi_{mt} = 1|y, \Lambda') \\
 &= \zeta_t(i) \left( \frac{\delta_{ilm_t}}{\tau_{ilm_t}} \right) \left( \frac{\omega_{im} \prod_{l=1}^L \tau_{ilm_t}}{\sum_m^M \omega_{im} \prod_{l=1}^L \tau_{ilm_t}} \right),
 \end{aligned}
 \tag{27}$$

and

$$\begin{aligned}
 v_{ilm_t} &= \mathbb{P}(z_l = 0, x_t = i, \phi_{mt} = 1|y, \Lambda') \\
 &= \zeta_t(i) \left( \frac{\epsilon_{ilm_t}}{\tau_{ilm_t}} \right) \left( \frac{\omega_{im} \prod_{l=1}^L \tau_{ilm_t}}{\sum_m^M \omega_{im} \prod_{l=1}^L \tau_{ilm_t}} \right),
 \end{aligned}
 \tag{28}$$

The  $Q$  function is expanded into a sum of terms where each term can be maximized independently. These terms are the  $Q$  function applied to the initial state  $\pi$ , the state-transition  $b$ , and the parameters for the emission distribution  $\theta = \{\alpha, \beta, \lambda, \rho\}$ . Consequently, for all parameters, except for the GID distribution ones, the maximization step gives, as a result, the following parameters and their updates

$$\hat{\pi}_i = \zeta_t(i),
 \tag{29}$$

$$\hat{b}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \zeta_t(i)},
 \tag{30}$$

$$\hat{\omega}_{im} = \frac{\sum_{t=1}^{T-1} \sum_{l=1}^L u_{ilm_t}}{\sum_{t=1}^{T-1} \sum_{l=1}^L \sum_{m=1}^M u_{ilm_t}}
 \tag{31}$$

$$\begin{aligned}\hat{\rho}_l &= \frac{\sum_{t=1}^{T-1} \sum_{l=1}^L \sum_{m=1}^M u_{ilm}}{\sum_{t=1}^{T-1} \sum_{l=1}^L \sum_{m=1}^M u_{ilm} + \sum_{t=1}^{T-1} \sum_{l=1}^L \sum_{m=1}^M v_{ilm}} \\ &= \frac{\sum_{t=1}^{T-1} \sum_{l=1}^L \sum_{m=1}^M u_{ilm}}{T}\end{aligned}\quad (32)$$

Complete estimation of GID parameters, as well as the MAP estimation, can be consulted respectively in Appendix sections A and B.

## 4 Experiments and results

In this section, extensive experiments are conducted and we have implemented several real-world topical yet challenging applications using the FSHMM with GID emission probabilities. We are mainly comparing our new approach to its classical FSHMM competitors and other new adaptations that we executed for the sake of comparison and testing, e.g., inverted Dirichlet-based FSHMM (ID-FSHMM) and Dirichlet-based FSHMM (Dir-FSHMM), not to mention the widely used GMM-FSHMM. It is noteworthy that the learning of the mentioned adaptations has been based on the same methodology described in the previous section to learn the GID mixture-based FSHMM. Two real-world applications, facial expressions recognition, and scene categorization are here tested and explained. Experimentation and results presented in this section have been yielded on a macOS environment over a 2.3 GHz Dual-Core Intel Core i5 MacBook Pro, using Python.

### 4.1 Facial expressions recognition

Facial expression recognition is a powerful process that usually commends the way we interact with other people. It is one of the non-verbal communication media that humans naturally use in everyday interactions. Besides its role in supporting humans' understanding of people's intentions and feelings, facial expression recognition plays a major role in making decisions about relationships or situations. For all these reasons, substantial efforts have been devoted to automating this recognition [19, 23] and using it as a fundamental step within multiple decision-making systems. A human being is naturally empowered to interpret these expressions and make his decisions in a real-time matter. Nonetheless, this task is still approached as a complex and challenging process in the field of machine learning [27, 38]. Facial Expression Recognition is applied in a wide range of contexts and is used in numerous applications such as Human-Computer Interaction [65], student automatic E-learning [54], Behavioural Science [46], psychological studies [50], image understanding, and synthetic face

animation. The principal purpose of researchers working on these applications is to produce automated systems capable of automatically recognizing the emotional state of a person and further draw an analysis or take a decision based on a specific context [16].

#### 4.1.1 HMM-based facial expression recognition

Classification is the most significant part of a facial expression recognition system [69]. Methods applied to classify this type of images are generally sorted into static or dynamic [77]. Static methods are based on the information acquired from the input image, they take benefit from the use of support vector machine, neural network, Bayesian network to perform the assigned task. HMMs are dynamic classifiers that exploit temporal records to analyze facial expressions. Hence, they are highly recommended by psychological experiments carried out as indicated in [5].

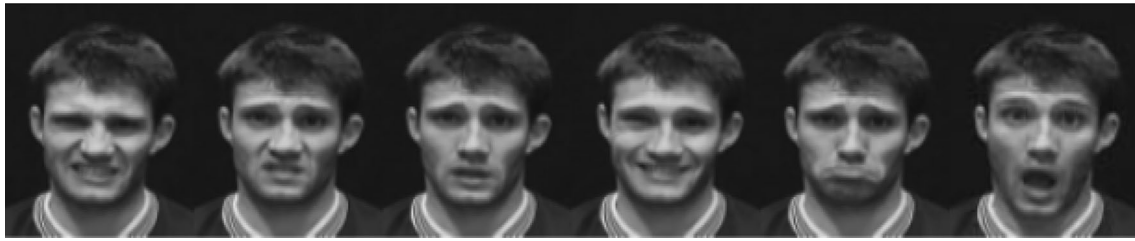
As early as 1990, [66] used HMMs to come up with a solution for the challenging task of automating facial expressions recognition. Authors in [66], used an HMM along with the integration of a priori structural knowledge with statistical information. HMMs offer a perfect analogous representation to the experience of observing a particular feeling through the way they statistically handle the behaviour of an observable symbol sequence. These models provide a specification of the probability distribution over all hidden events that are behind a certain symbol sequence. The performances of HMMs when dealing with such challenges are promising, especially in the case when they learn through an entire sequence of images describing a group of actions taken by a person when undergoing a certain feeling. The learning process is conducted in a much smoother way thanks to HMMs capability of handling the Spatio-temporal nature of the debated application. In fact, there is a metaphorical resemblance between human performance when naturally processing the recognition task, and the stochastic nature of the HMM process inasmuch as it analyses the measurable (observable) actions in order to infer the immeasurable (hidden) feelings of the person.

In this particular context, we choose to apply our model on the challenging Dollar facial expression database [17] 3. This application is unprecedented as it uses for the first time and embedded model-based feature selection into the HMM structure.

#### 4.1.2 Experimental trials and results

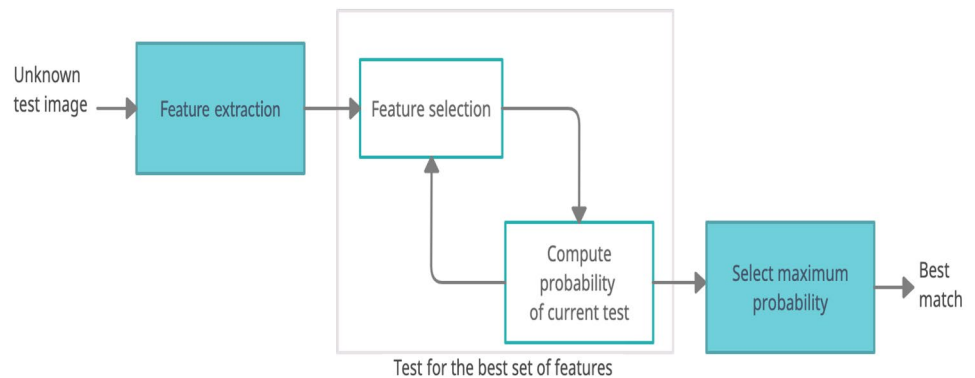
The Dollar database is composed of 192 sequences performed by 2 individuals, each expressing 6 different basic emotions 8 times under 2 lighting setups. Each subject starts with a neutral expression, then expresses emotion, and returns to a





**Fig. 3** Samples of facial frames from the Dollar facial expressions dataset

**Fig. 4** Block diagram for FSHMM-based face recognizer

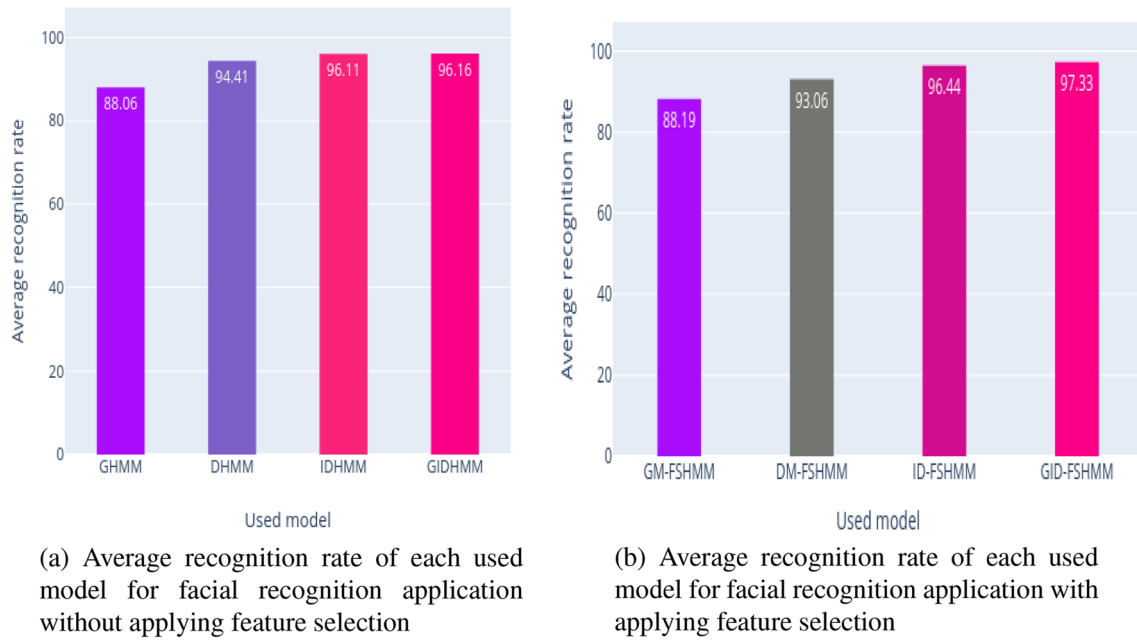


neutral expression. For our simulations, we follow the experimental setting considered in [58], which consists of using three peak frames of each sequence for 6-class expression recognition (576 images: anger, disgust, fear, joy, sadness, and surprise). The pre-processing steps are also the same and consist of extracting features from the whole face region by cropping original face images into  $110 \times 150$  pixels, keeping only the central part of facial extraction. A Local Binary Pattern (LBP) descriptor [60], is used for feature extraction. More specifically, each cropped face image is first divided into small regions from which LBP histograms are extracted and then concatenated altogether into a single feature histogram representing the face image. We use a 59-bin LBP operator in the (8, 2) neighbourhood. This means 8 sampling points on a circle of radius 2, then we divide each image ( $110 \times 150$ ) into  $18 \times 21$  pixels regions. Therefore, each face image is divided into 42 ( $6 \times 7$ ) regions and is then represented by LBP histograms with a length of 2478 ( $59 \times 42$ ). After that, these histograms are normalized. The procedure is applied as the one originally used in [67]. We figured that if we reduce the feature vector the algorithm tends to early diverge, however since features will later be reduced by the model itself, and in order to give the algorithm the time to learn we will not reduce the feature vector ourselves and will leave it as it is. The obtained feature vector is actually handled with our GIDHMM where the feature saliency is considered. Hence, recognition is carried out via a single HMM recognizer. A collection of HMMs each representing a different subject is matched against the test image and the highest match is selected as explained

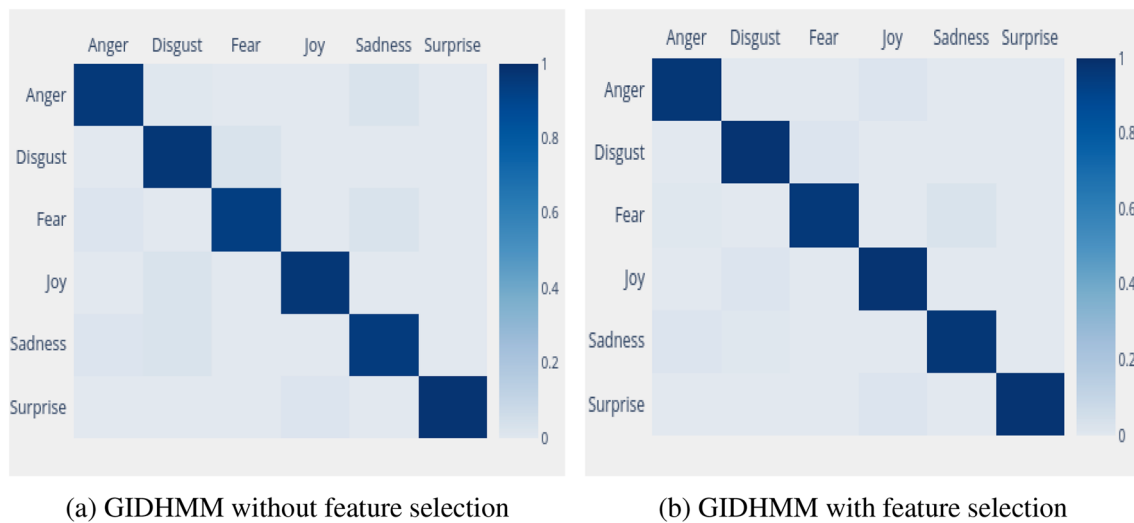
in figure 4. For the sake of comparison, we conduct several experiments with different used models, with and without taking into consideration of feature relevancy, then we report the results including features relevancies and the confusion matrix for each experiment.

In order to evince the advantages of the proposed approach and most importantly underline the crucial role of feature selection integration in improving results, we compare the latter with other emotion recognition approaches that are mainly based on mixture models. These approaches, have been personally implemented, and include inverted Dirichlet-based HMM without feature selection (IDHMM) [58], inverted Dirichlet-based HMM with feature selection (ID-FSHMM), generalized inverted Dirichlet-based HMM without feature selection (GIDHMM), and generalized inverted Dirichlet-based HMM with feature selection (GID-FSHMM). On top of that, we put a special emphasis on the improvements noticed from the use of GID mixtures measured against the Gaussian mixtures-based HMM with feature selection (GM-FSHMM). Results obtained are displayed in Fig. 5, where we present the average recognition rates for the different used methods.

There is an interesting observation to make after nailing these results, that is the amelioration in average recognition results after using the GIDHMM as a model. Initially, using only the latter allowed for a slight but worth mentioning amelioration in the average recognition rate from 96.11% to 96.16%, this itself shows that the GIDHMM is better in modelling our data than the IDHMM. Further, we



**Fig. 5** Average recognition rates for facial expressions recognition with and without applying feature selection (colour figure online)



**Fig. 6** Confusion matrices for facial expressions recognition with and without applying feature selection for GID-FSHMM

noticed when incorporating feature selection into our previously established IDHMM model, the average recognition rate improved from 96.11 to 96.44%. On top of that, we plainly succeeded to bear out our theoretically anticipated projections regarding recognition rates. In fact, the feature selection-based GIDHMM executed the task of recognizing each facial expression considerably better than GIDHMM without FS. This conclusion comes after several trials on each emotion type separately. The individual recognition

rates per category and confusion matrices for the mentioned trials are displayed in Table 1 and Fig. 6.

As indicated in Fig. 5, average recognition rates for GIDHMM with and without feature selection integration are respectively 97.33% and 96.16% with the corresponding average misclassified images of 22.11 and 15.32 per dataset. There is also a significant variation in the run time when using each of the cited methods, 36.4 min for GIDHMM, and 41.6 min for GIDHMM-FS.

**Table 1** Detailed recognition rates in the case of facial recognition application with and without applying feature selection for GIDHMM

	Anger	Disgust	Fear	Joy	Sadness	Surprise
Recognition rates without applying feature selection						
Anger	0.96	0.01	0	0	0.03	0
Disgust	0	0.97	0.03	0	0	0
Fear	0.02	0.01	0.94	0	0.03	0
Joy	0	0.03	0	0.97	0	0
Sadness	0.02	0.03	0	0	0.95	0
Surprise	0	0	0	0.02	0	0.98
Recognition rates when feature selection is applied						
Anger	0.97	0	0	0	0.03	0
Disgust	0	0.98	0.02	0	0	0
Fear	0.01	0	0.96	0	0.03	0
Joy	0	0.02	0	0.98	0	0
Sadness	0.02	0.01	0	0	0.97	0
Surprise	0	0	0	0.02	0	0.98

Needless to say, the integration of feature selection in our models brought an obvious amelioration to the yielded results for all adopted distributions, this shows the important role of taking into consideration the feature saliency when dealing with image classification tasks. Further investigations with respect to features relevancy are conducted in the following applications to emphasize this role.

## 4.2 Scene categorization

Recently, there has been an abundance of research works and experimental trials aiming to bridge the semantic gap between the perceptual ability of human vision and the capacity of automated systems when performing the same related tasks. This challenge is prompted by the impressive trait of the human visual system to rapidly, accurately, and comprehensively recognize and understand a complex scene [28, 29, 76]. Thus, it would be worthwhile if each image in a studied collection could be annotated with semantic descriptions allowing for a better automatic interpretation and hence an improved visual recognition ability. In this section, we work on a challenging problem related to the mentioned area of research, which is recognizing scene categories. Visual scenes classification has many applications in robot navigation and robot path planning [70], video analysis [73], content-based image retrieval [13].

Inasmuch as this application is complex due to the variety of scenes and variations of viewing angles and changing backgrounds, choosing efficient features plays a major role in the accuracy level of the recognition task.

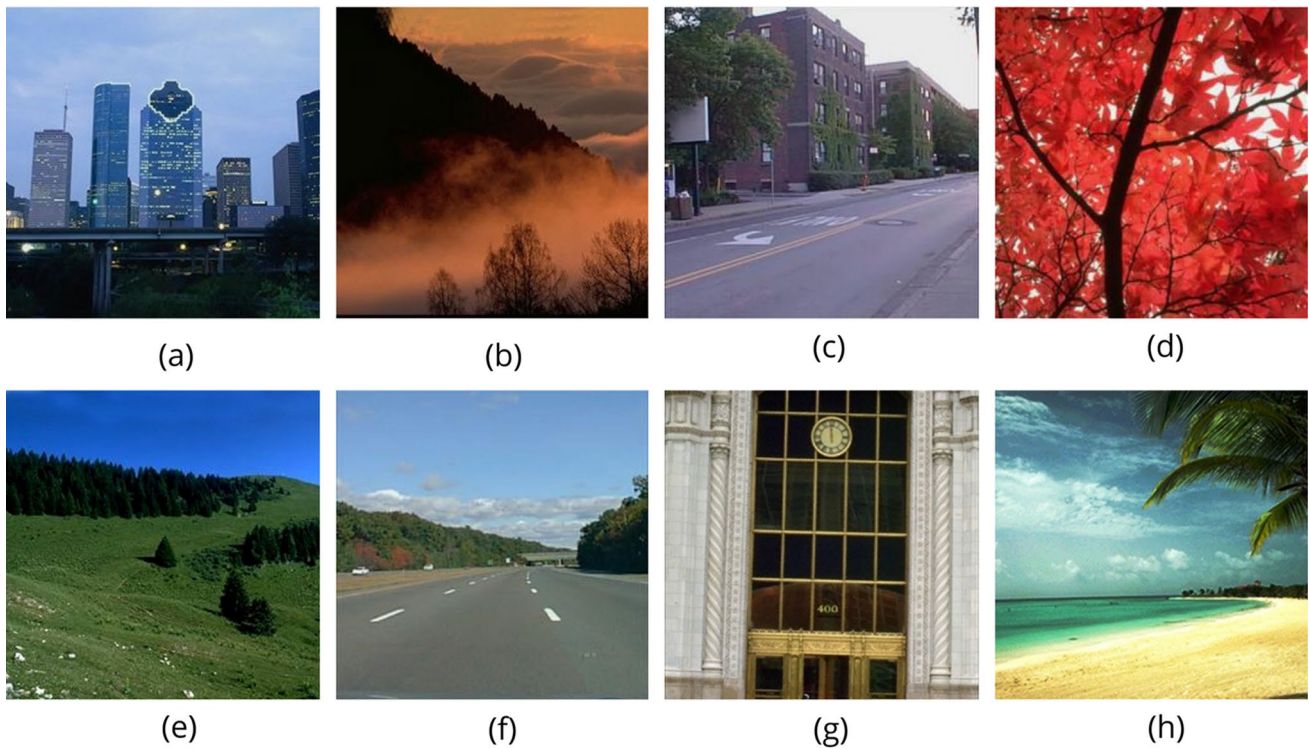
In this section, we test the effectiveness of our proposed feature-selection-based GIDHMM, in categorizing images of real-world scenes from the notorious MIT benchmark

[61].<sup>1</sup> The indicated database contains about 2688 diverse outdoor scene images in colours from 8 categories: coast (360 images), mountain (374 images), forest (328 images), open country (410 images), inside city (308 images), street (292 images), tall building (356 images) and highways (260 images). Images come in  $256 \times 256$  pixels resolution. We choose to randomly select 200 images from each category for training and leave the rest for testing purposes. Figure 7, shows example images from the MIT outdoor data set.

A crucial step for the scene categorization task is feature extraction. For this matter, we adopt a process where we normalize images which will afterward be represented each by a collection of local image patches. These patches are scanned and low-level feature vectors are thereafter extracted. We then use the bag of words approach (BOW) method, adapted likewise by [75] for scene classification, in which Yang et al. mapped the key points of an image into visual words. Hence, each image could be represented as a “*bag of visual words*” BoVW, and in this instance as a vector of counts of each visual word in that image. This will allow for an overall representation for each image through a feature vector, upon which the task of image classification is built. Following [10], and after obtaining the intended histograms, we apply a probabilistic Latent Semantic Analysis (pLSA) [40, 41] in order to represent each image by a D-dimensional vector with D being the number of latent aspects (hidden aspects, features, or hidden states in our analogy). Ultimately, our objective is to identify the right category for each image by applying our previously developed model.

In this work, we use dense SIFT  $16 \times 16$ -pixel patches calculated over a grid of 8 pixels. Besides, we build a bag of words dictionary using a K-means algorithm [41] to cluster our descriptors in a V visual words vocabulary. For each

<sup>1</sup> <http://people.csail.mit.edu/torralba/code/spatialenvelope/>



**Fig. 7** Sample images from the 8 categories MIT data set: (a) Tall buildings, (b) Mountain, (c) Street, (d) Forest, (e) Open country, (f) Highway, (g) Inside city, (h) Coast

**Table 2** The confusion matrix in the case of MIT scene recognition problem when applying GIDHMM without feature selection

	Tall building	Mountain	Street	Forest	Open country	Highway	Inside city	Coast
Tall building	0.94	0.02	0	0	0	0	0.04	0
Mountain	0.01	0.92	0	0	0.07	0	0	0
Street	0	0	0.92	0	0	0.06	0.02	0
Forest	0	0.02	0	0.95	0.03	0	0	0
Open country	0	0.03	0	0.01	0.87	0	0	0.09
Highway	0	0.01	0	0	0.03	0.88	0	0.08
Inside city	0.01	0.01	0.05	0	0.03	0	0.90	0
Coast	0	0.01	0.01	0	0.05	0	0	0.93

SIFT point in a candidate image, the nearest neighbour within the vocabulary is computed, and thereby a feature vector with dimension  $V$  is built. Hence, each image can be represented as a frequency histogram over the  $V$  visual words. As previously explained, in this work we apply pLSA to allow for a description through a  $D$ -dimensional vector where  $D$  is the number of aspects. We employ our GIDHMM to model the set of images designated for training. We compute the class-relationship likelihood of each input image and classify it to the class that maximizes more its likelihood. In our approach, each image class is characterized by its own behaviour, therefore each class is described by its own HMM. That being the case, for each scenery type, a distinct 8-state HMM is trained. Experiments are

carried out 30 times with the average accuracy reported for both feature-saliency-based and non-feature-saliency-based methods.

Through these experiments, we aim to evaluate not only the effectiveness of GIDHMM measured against IDHMM and GHMM but also the effectiveness of embedding the process of feature selection in the core of each of the aforementioned models. Experiments are chosen to be conducted in the following order: first, we compare the performance of GIDHMM, IDHMM, and GHMM without taking into consideration the relevancy of features. Then we reproduce the same experiments by taking into account feature relevancy. Table 2 presents the confusion matrix when GIDHMM is applied without feature selection. According to this table,

**Table 3** The confusion matrix in the case of MIT scene recognition problem when applying GIDHMM with feature selection

	Tall building	Mountain	Street	Forest	Open country	Highway	Inside city	Coast
Tall building	0.96	0.02	0	0	0	0	0.02	0
Mountain	0.01	0.95	0	0	0.04	0	0	0
Street	0	0	0.93	0	0	0.06	0.01	0
Forest	0	0.02	0	0.96	0.02	0	0	0
Open country	0	0.03	0	0.01	0.90	0	0	0.06
Highway	0	0	0	0	0.02	0.91	0	0.07
Inside city	0.02	0.01	0.05	0	0.03	0	0.89	0
Coast	0	0.01	0.01	0	0.03	0	0	0.95

**Table 4** Average recognition rates for different used HMMs in the context of natural scenes recognition, with and without feature selection

Method	Average recognition rate (%)	Integrating feature selection
GHMM	87.60	88.14
DHMM	88.79	89.05
IDHMM	90.01	90.66
GIDHMM	<b>91.37</b>	<b>93.12</b>

Bold values are the values achieved by the model proposed in this paper

we get an average accuracy of 91.37%. On the other hand, Table 3 shows the confusion matrix when GIDHMM is used along with feature selection: the average accuracy is 93.12%.

Results of other experimentation on the different used models are presented in Table 4 and confirm our previous assumptions about the role of feature selection in improving recognition rates. Our algorithm analyzed all extracted features and succeeded to determine their saliency, hence the use of the better features yielded better results. Figure 8 shows the feature saliencies obtained by our GID-FSHMM.

### 5 Conclusion

While there are multiple general techniques of applying feature selection, and despite the buildup of standardized procedures for features and dimensionality reduction, the literature reveals time and again that custom methods keep outperforming the general methods. Besides there is an overwhelming need for some sort of supervised data and knowledge when applying general feature selection models. In our context, this supervised data can take the form of information about the class, a label of each observation, or even a piece of knowledge about the latent variable. This additional

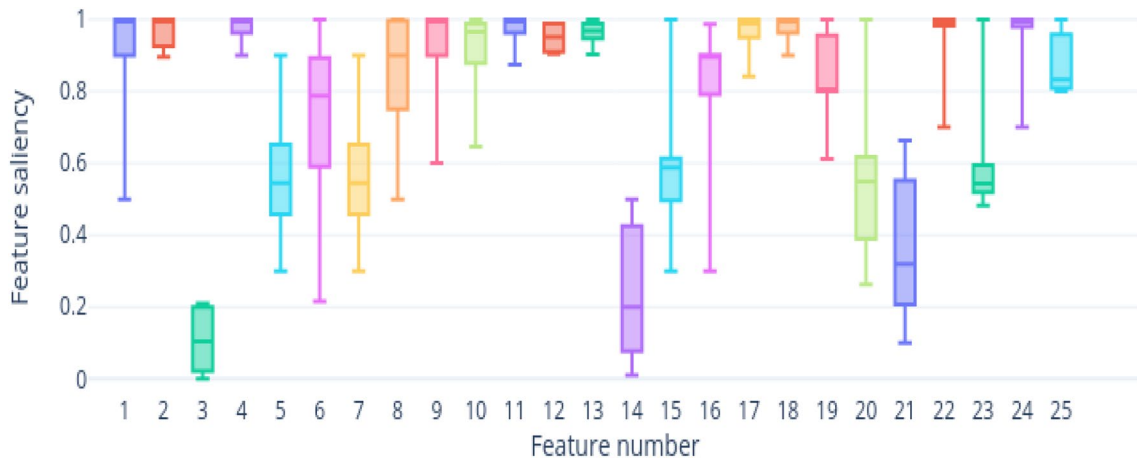
information is often not readily accessible especially in areas where mixture models and HMMs are applied, considering that those models account for the fact that supervised data is unavailable. Therefore, unsupervised feature selection methods are essentially needed when using HMMs, allowing for significantly better performing models compared to those based upon general feature selection methods. Further, the interest in adopting the GID for modelling our data arose from the limitations encountered when inverted Dirichlet is adopted, in particular its restraining strictly positive covariance. In this paper, we proposed a framework in which all the aforementioned problems are addressed simultaneously in the case of automatic recognition. The developed approach applies feature selection to a GID-based HMM. Parameters are learned via a MAP method adding a huge advantage in raising both accuracies of parameter estimates and feature saliencies. Experimental results involving challenging real-life applications such as facial expressions recognition and natural outdoor scene recognition showed that the proposed approach is highly promising. Future works are intended to be done in the near future extending this work to different flexible distributions and considering online learning for more precise results.

### Estimation of GID parameters

Here, the estimation of the GID parameters is equivalent to maximization of the following log-likelihood function

$$\begin{aligned}
 \log(p(\vec{X}_i|\Theta)) &= \sum_{i=1}^N \sum_{j=1}^M p_j \prod_{l=1}^D p_{l\text{Beta}}(X_{il}|\theta_{jl}) \\
 &= \sum_{i=1}^N \sum_{j=1}^M \sum_{l=1}^D (\log(p_j) + \log p_{l\text{Beta}}(X_{il}|\theta_{jl}))
 \end{aligned}
 \tag{33}$$

In the E-step, we compute the conditional expectation of log-likelihood, which is reduced to the computation of the posterior probabilities meaning the probability that a vector  $\vec{X}_i$  is assigned to a cluster  $j$ , such as following



**Fig. 8** Feature saliencies obtained in the case of fnatural scenes recognition problem when performing feature selection-based GIDHMM (colour figure online)

$$p(j|\vec{X}_i, \Theta, \vec{p}_j) = \frac{p_j p(\vec{X}_i|\theta_j)}{\sum_{j=1}^M p_j p(\vec{X}_i|\theta_j)} = \frac{p_j \prod_{l=1}^D p_{lBeta}(X_{il}|\theta_{jl})}{\sum_{j=1}^M p_j \prod_{l=1}^D p_{lBeta}(X_{il}|\theta_{jl})} \tag{34}$$

where  $\vec{p}_j = (p_1, \dots, p_M), p_j > 0$  and  $\sum_{j=1}^M p_j = 1$   
Hence, we have

$$\log p(\mathcal{X}|\Theta, \vec{p}_j) = \sum_{i=1}^N \sum_{j=1}^M \sum_{l=1}^D p(j|\vec{X}_i, \Theta, \vec{p}_j) (\log(p_j) + \log p_{lBeta}(\vec{X}_i|\theta_{jl})) \tag{35}$$

Whence, the conditional expectation of the complete-data log likelihood

$$\mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon) = \log p(\mathcal{X}|\Theta, \vec{p}_j) + \Upsilon (1 - \sum_{j=1}^M p_j) \tag{36}$$

where  $\Upsilon$  is the Lagrange multiplier.

We move forward in maximizing the log-likelihood function by finding the roots to its derivations with respect to the set of parameters. The mixture weights can be easily estimated as follows

$$p_j = \frac{\sum_{i=1}^N p(j|\vec{X}_i, \Theta, \vec{p}_j)}{N} \tag{37}$$

Regarding the derivatives with respect to  $\alpha_{jl}$  and  $\beta_{jl}$ , we have

$$\begin{aligned} \frac{\partial \mathbb{Q}}{\partial \alpha_{jl}} &= \sum_{i=1}^N p(j|\vec{X}_i, \Theta, \vec{p}_j) \frac{\partial \log p_{lBeta}(\vec{X}_i|\theta_{jl})}{\partial \alpha_{jl}} \\ &= \sum_{i=1}^N p(j|\vec{X}_i, \Theta, \vec{p}_j) (\Psi(\alpha_{jl} + \beta_{jl}) - \Psi(\alpha_{jl}) + \log(\frac{X_{il}}{1 + X_{il}})) \end{aligned} \tag{38}$$

$$\begin{aligned} \frac{\partial \mathbb{Q}}{\partial \beta_{jl}} &= \sum_{i=1}^N p(j|\vec{X}_i, \Theta, \vec{p}_j) \frac{\partial \log p_{lBeta}(\vec{X}_i|\theta_{jl})}{\partial \beta_{jl}} \\ &= \sum_{i=1}^N p(j|\vec{X}_i, \Theta, \vec{p}_j) (\Psi(\alpha_{jl} + \beta_{jl}) - \Psi(\beta_{jl}) + \log(\frac{1}{1 + X_{il}})) \end{aligned} \tag{39}$$

where  $\Psi(\cdot)$  is the digamma function. We can clearly see that a closed form solution to estimate  $\theta_{jl}$  does not exist. Therefore, we ought to use the Newton-Raphson method [51] such as

$$\theta_{jl}^{old} = \theta_{jl}^{new} - H_{jl}^{-1} \tag{40}$$

where  $H_{jl}$  is the Hessian matrix associated with  $\mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)$  with first derivatives vector  $G_{jl} = (\frac{\partial \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \alpha_{jl}}, \frac{\partial \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \beta_{jl}})$

$$H_{jl} = \begin{pmatrix} \frac{\partial^2 \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \alpha_{jl}^2} & \frac{\partial^2 \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \alpha_{jl} \partial \beta_{jl}} \\ \frac{\partial^2 \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \alpha_{jl} \partial \beta_{jl}} & \frac{\partial^2 \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \beta_{jl}^2} \end{pmatrix} \tag{41}$$

with the following second and mixed derivatives

$$\frac{\partial^2 \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \alpha_{jl}^2} = \sum_{i=1}^N p(j|\vec{X}_i, \Theta, \vec{p}_j) (\Psi'(\alpha_{jl} + \beta_{jl}) - \Psi'(\alpha_{jl})) \tag{42}$$

$$\frac{\partial^2 \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \beta_{jl}^2} = \sum_{i=1}^N p(j|\vec{X}_i, \Theta, \vec{p}_j) (\Psi'(\alpha_{jl} + \beta_{jl}) - \Psi'(\beta_{jl})) \tag{43}$$

$$\frac{\partial^2 \mathbb{Q}(\mathcal{X}, \Theta, \vec{p}_j, \Upsilon)}{\partial \alpha_{jl} \partial \beta_{jl}} = \sum_{i=1}^N p(j|\vec{X}_i, \Theta, \vec{p}_j) (\Psi'(\alpha_{jl} + \beta_{jl})) \tag{44}$$

## MAP estimation

A standard choice for the mixing parameters vector  $\vec{p}_j$  is the Dirichlet distribution (Dir), given its definition on the simplex  $\{(p_1, \dots, p_M) : \sum_{j=1}^{M-1} p_j < 1\}$  [64]. We pick the same distribution for both initial and transition probabilities  $\pi$  and  $B$  as well as for the observation weights such that

$$\begin{aligned}\pi &\sim \text{Dir}(\pi|p) \\ B_i &\sim \text{Dir}(B_i|b_i) \\ \omega_i &\sim \text{Dir}(\omega_i|\Omega_i) \\ \rho_l &\sim \frac{1}{\mathbf{Z}} e^{-k_l \rho_l}\end{aligned}$$

where  $\mathbf{Z}$  is the normalizing constant and  $\Omega_i$  is the hyperparameter vector of the observation weights.

The prior for the mixing parameters vector  $\vec{p}_j$  can be written as follows

$$p(\vec{p}_j|M, \vec{\Delta}) = \frac{\Gamma(\sum_{j=1}^M \Delta_j)}{\prod_{j=1}^M \Gamma \Delta_j} \prod_{j=1}^M p_j^{\Delta_j-1} \quad (45)$$

where  $\vec{\Delta} = \{\Delta_1, \dots, \Delta_M\}$  is the Dirichlet parameters vector.

For the GID parameters, the Gamma (G) function is chosen as a prior given its exponential nature under the assumption of parameters independence. Thus we have the priors

$$p(\alpha_{jl}) = G(\alpha_{jl}|\nu_{jl}, \vartheta_{jl}) = \frac{\vartheta_{jl}^{\nu_{jl}}}{\Gamma(\nu_{jl})} \alpha_{jl}^{\nu_{jl}-1} e^{-\nu_{jl} \alpha_{jl}} \quad (46)$$

$$p(\beta_{jl}) = G(\beta_{jl}|\kappa_{jl}, \varsigma_{jl}) = \frac{\varsigma_{jl}^{\kappa_{jl}}}{\Gamma(\kappa_{jl})} \beta_{jl}^{\kappa_{jl}-1} e^{-\kappa_{jl} \beta_{jl}} \quad (47)$$

where  $\nu$ ,  $\vartheta$ ,  $\kappa$  and  $\varsigma$  are positive hyperparameters.

**Acknowledgements** The authors would like to thank the editor and the anonymous referees for their valuable comments. The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Adams S (2015) Simultaneous feature selection and parameter estimation for hidden Markov models. PhD thesis, Dissertation, University of Virginia
- Adams S, Beling PA (2019) A survey of feature selection methods for gaussian mixture models and hidden markov models. *Artificial Intelligence Rev* 52(3):1739–1779
- Adams S, Beling PA, Cogill R (2016) Feature selection for hidden markov models and hidden semi-markov models. *IEEE Access* 4:1642–1657
- Al Mashrgy M, Bdiri T, Bouguila Nizar R (2014) simultaneous positive data clustering and unsupervised feature selection using generalized inverted dirichlet mixture models. *Knowl-Based Syst* 59:182–195
- Ambadar Z, Schooler Jonathan W, Cohn Jeffrey F (2005) Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions. *Psychol Sci* 16(5):403–410 (( PMID: 15869701))
- Bashir FI, Khokhar AA, Schonfeld D (2007) Object trajectory-based activity classification and recognition using hidden markov models. *IEEE Transact Image Process* 16(7):1912–1919
- Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite state markov chains. *Ann Math Statist* 37(6):1554–1563
- Bilmes JA et al (1998) A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *Int Comput Sci Institut* 4(510):126
- Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97(1–2):245–271
- Bosch A, Zisserman A, Muñoz X (2006) Scene classification via pls. In: Leonardis Aleš, Bischof Horst, Pinz Axel (eds) *Computer vision - ECCV 2006*. Berlin, Heidelberg, Springer, Berlin Heidelberg, pp 517–530
- Boutsidis Ch, Mahoney MW, Drineas P (2009) Unsupervised feature selection for the k-means clustering problem. In: *Proceedings of the 22nd International Conference on Neural Information Processing Systems, NIPS'09*, page 153–161, Red Hook, NY, USA, Curran Associates Inc
- Cárdenas-Ovando RA, Fernández-Figueroa EA, Rueda-Zárate Héctor A, Julieta N, Rangel-Escareño Claudia A (2019) Feature selection strategy for gene expression time series experiments with hidden markov models. *Plos One* 14(10)
- Dacheng T, Xiaoou T, Xuelong L, Xindong W (2006) Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transact Pattern Anal Mach Intelligence* 28(7):1088–1099
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society* 39(1):1–22
- Denoyer L, Zaragoza H, Gallinari P (2001) HMM-based passage models for document classification and ranking. *ECIR'01 - 23rd European Colloquium on Information Retrieval Research*. Darmstadt, Germany, pp 126–135
- Dewan MAA, Mahbub M, Lin F (2019) Engagement detection in online learning: a review. *Smart Learn Environ* 6(1):1
- Dollár P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. *VS-PETS Beijing, China*
- Dy JG, Brodley Carla E (2004) Feature selection for unsupervised learning. *J Mach Learn Res* 5:845–889
- Edwards GJ, Lanitis A, Taylor CJ, Cootes TF (1998) Statistical models of face images - improving specificity. *Image Vision Comput* 16(3):203–211
- Eom H, Son Y, Choi S (2020) Feature-selective ensemble learning-based long-term regional pv generation forecasting. *IEEE Access* 8:54620–54630
- Esfandian N, Razzazi F, Behrad A (2012) A clustering based feature selection method in spectro-temporal domain for speech recognition. *Eng App Artificial Intelligence* 25(6):1194–1202

22. Ezenkwu Chinedu P, Akpan Uduak I, Stephen Bliss U-A (2021) A class-specific metaheuristic technique for explainable relevant feature selection. *Mach Learn App* 6:100142
23. Fan W, Bouguila N (2013) Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE Transact Neural Netw Learn Syst* 24(11):1850–1862
24. Fan W, Bouguila N (2019) Simultaneous clustering and feature selection via nonparametric pitman-yor process mixture models. *Int J Mach Learn Cybern* 10(10):2753–2766
25. Fan W, Sallay H, Bouguila N, Bourouis S (2015) A hierarchical dirichlet process mixture of generalized dirichlet distributions for feature selection. *Comput Electrical Eng* 43:48–65
26. Fan W, Yang L, Bouguila N, Chen Y (2020) Sequentially spherical data modeling with hidden markov models and its application to fmri data analysis. *Knowledge-Based Syst* 206:106341
27. Fathima A, Vaidehi K (2020) Review on facial expression recognition system using machine learning techniques. In: *Advances in Decision Sciences, Image Processing, Security and Computer Vision*, pages 608–618. Springer
28. Fei-Fei L, Koch C, Iyer A, Perona P (2004) What do we see when we glance at a scene? *J Vision* 4(8):863–863
29. Fei-Fei L, Iyer A, Koch C, Perona P (2007) What do we perceive in a glance of a real-world scene? *J Vision* 7(1):10–10
30. Fons E, Dawson P, Yau J, Zeng Xiao J, Keane J (2019) A novel dynamic asset allocation system using feature saliency hidden Markov models for smart beta investing. *Papers* 1902.10849, arXiv.org
31. Frydenberg M (1990) The chain graph Markov property. *Scand J Stat* 17:333–353
32. Gauvain J-L, Chin-Hui L (1994) Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transact Speech Audio Proces* 2(2):291–298
33. Gehan Edmund A (1965) A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52(1–2):203–224
34. Ghahramani Z (2001) An introduction to hidden markov models and bayesian networks. In *Hidden Markov models: applications in computer vision*, pages 9–41. World Scientific
35. Graña M, Termon M, Savio A, Gonzalez-Pinto A, Echeveste J, Pérez JM, Besga A (2011) Computer aided diagnosis system for alzheimer disease using brain diffusion tensor imaging features selected by pearson’s correlation. *Neurosci Lett* 502(3):225–229
36. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182
37. Hajirahimi Z, Khashei M (2019) Hybrid structures in time series modeling and forecasting: a review. *Eng App Artif Intelligence* 86:83–106
38. Harms Madeline B, Alex M, Wallace Gregory L (2010) Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. *Neuropsychol Rev* 20(3):290–322
39. Hegde S, Achary KK, Shetty S (2015) Feature selection using fisher’s ratio technique for automatic speech recognition. arXiv preprint [arXiv:1505.03239](https://arxiv.org/abs/1505.03239)
40. Hofmann T (1999) Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’99*, page 50–57, New York, NY, USA, Association for Computing Machinery
41. Hofmann T (2013) Probabilistic latent semantic analysis
42. Holland JH (1992) Genetic algorithms. *Sci Am* 267(1):66–73
43. Kailath T (1967) The divergence and bhattacharyya distance measures in signal selection. *IEEE Transact Commun Technol* 15(1):52–60
44. Kang M, Ahn J, Lee K (2018) Opinion mining using ensemble text hidden markov models for text classification. *Expert Syst App* 94:218–227
45. Khan Naseer A, Waheeb Samer A, Riaz A, Shang X (2020) A three-stage teacher, student neural networks and sequential feed forward selection-based feature selection approach for the classification of autism spectrum disorder. *Brain Sci* 10(10):754
46. Kwang Hyeon K, Kyeongyun P, Haksoo K, Byungdu J, Sang Hee A, Chankyu K, Myeongsoo K, Tae Ho K, Se Byeong L, Dongho S et al (2020) Facial expression monitoring system for predicting patient’s sudden movement during radiotherapy using deep learning. *J Appl Clin Med Phys*
47. Kittler J, Pudil P, Somol P (2001) Advances in statistical feature selection. In: *Proceedings of the Second International Conference on Advances in Pattern Recognition, ICAPR ’01*, page 425–434, Berlin, Heidelberg, Springer-Verlag
48. Law MHC, Figueiredo MAT, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. *IEEE Transact Pattern Anal Mach Intelligence* 26(9):1154–1166
49. Lee C, Geunbae Lee G (2006) Information gain and divergence-based feature selection for machine learning-based text categorization. *Inform Process Manag* 42(1):155–165
50. Lee W-C, Yoon D (2019) A study on facial expression and first impression through machine learning. In: *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pages 298–301. IEEE
51. Lindstrom Mary J, Bates Douglas M (1988) Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *J Am Stat Assoc* 83(404):1014–1022
52. Lingappaiah GS (1976) On the generalised inverted dirichlet distribution. *Demonstratio Mathematica* 9(3):423–433
53. Maldonado S, Weber R (2009) A wrapper method for feature selection using support vector machines. *Inform Sci* 179(13):2208–2217
54. Mayer Richard E (2020) Searching for the role of emotions in e-learning. *Learn Instr* 70:101213
55. Mohammadreza M, Mohammadreza S, Hossein R (2019) A novel feature selection method for microarray data classification based on hidden markov model. *J Biomed Inform* 95:103213
56. Montero JAV, Sucar LES (2004) Feature selection for visual gesture recognition using hidden markov models. In: *Proceedings of the Fifth Mexican International Conference in Computer Science, 2004. ENC 2004.*, pages 196–203. IEEE
57. Murphy Kevin P (2012) *Machine learning: a probabilistic perspective*. MIT press
58. Nasfi R, Soui M (2014) Extraction of interesting adaptation rules. *Procedia Comput Sci* 34:607–612. <https://doi.org/10.1016/j.procs.2014.07.081>
59. Nasfi R, Amayri M, Bouguila N (2019) A novel approach for modeling positive vectors with inverted dirichlet-based hidden markov models. *Knowledge-Based Syst* 105335
60. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 24:971–987
61. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J Comput Vis* 42(3):145–175
62. Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proc IEEE* 77(2):257–286
63. Rahmaninia M, Moradi P (2018) Osfsmi: online stream feature selection method based on mutual information. *Appl Soft Comput* 68:733–746
64. Robert C, Casella G (2013) *Monte Carlo statistical methods*. Springer Science & Business Media



65. Samara A, Galway L, Bond R, Wang H (2019) Affective state detection via facial expression analysis within a human-computer interaction context. *J Ambient Intelligence Humanized Comput* 10(6):2175–2184
66. Samaria F, Fallside F (1993) Automated face identification using hidden markov models. Olivetti Research Limited
67. Shan C, Gong S, McOwan Peter W (2009) Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vision Comput* 27:803–816
68. Shang R, Song J, Jiao L, Li Y (2020) Double feature selection algorithm based on low-rank sparse non-negative matrix factorization. *Int J Mach Learn Cybern* 11:1891–1908
69. Sun Y, Akansu AN (2014) Facial expression recognition with regional hidden markov models. *Electron Lett* 50(9):671–673
70. Tian X, Tao D, Rui Y (2012) Sparse transfer learning for interactive video search reranking. *ACM Trans Multimedia Comput Commun Appl* 8(3):26. <https://doi.org/10.1145/2240136.2240139>
71. Uğuz H (2011) A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Syst* 24(7):1024–1032
72. Wang J, Chen X, Gao W (2005) Online selecting discriminative tracking features using particle filter. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) 2:1037–1042 (IEEE)
73. Xiang T, Gong S (2008) Activity based surveillance video content modelling. *Pattern Recogn* 41(7):2309–2326
74. Xie L, Xu P, Chang S-F, Divakaran A, Sun H (2004) Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recogn Lett* 25(7):767–775
75. Yang J, Jiang Y-G, Hauptmann Alexander G, Ngo C-W (2007) Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, pages 197–206
76. Yu Y, Zhu H, Wang L, Pedrycz W (2021) Dense crowd counting based on adaptive scene division. *Int J Mach Learn Cybern* 12(4):931–942
77. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transact Pattern Anal Mach Intelligence* 31(1):39–58
78. Zhang W, Yin Z (2020) Eeg feature selection for emotion recognition based on cross-subject recursive feature elimination. In 2020 39th Chinese Control Conference (CCC), pages 6256–6261. IEEE
79. Zheng Y, Jeon B, Sun L, Zhang J, Zhang H (2018) Student's t-hidden markov model for unsupervised learning using localized feature selection. *IEEE Transact Circuits Syst Video Technol* 28(10):2586–2598
80. Zhou J, Zhang X (2008) An ica mixture hidden markov model for video content analysis. *IEEE Transact Circuits Syst Video Technol* 18(11):1576–1586
81. Zhu H, He Z, Leung H (2012) Simultaneous feature and model selection for continuous hidden markov models. *IEEE Signal Process Lett* 19(5):279–282

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.