A new multimedia classification approach: Bayesian of inductive cognition algorithm based on Dirichlet process

L-C Jin*, W-G Wan, B Cui and **X-Q Yu**
School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China

Abstract: In this paper, we propose a Bayesian of inductive cognition algorithm based on Dirichlet process used in virtual reality multimedia information data classification. We present a Bayesian of inductive cognition algorithm framework model for classifying scenes in virtual reality multimedia data. The multimedia can switch between different shots, the unknown objects can leave or enter the scene at multiple times, and the scenes can be classified. The proposed algorithm consists of Bayesian inductive cognition part and Dirichlet process part. This algorithm has several advantages over traditional distance-based agglomerative classifying algorithms. Bayesian of inductive cognition algorithm based on Dirichlet process hypothesis testing is used to decide which merges are advantageous and to output the recommended depth of the scenes. The algorithm can be interpreted as a novel fast bottom-up approximate inference method for a Dirichlet process mixture model. We describe procedures for learning the model hyperparameters, computing the predictive distribution and extensions to the Bayesian of inductive cognition algorithm. Experimental results on virtual reality multimedia datasets demonstrate useful properties of the Bayesian of inductive cognition algorithm.

Keywords: pattern classification, virtual reality multimedia, Bayesian model, Dirichlet process

1 INTRODUCTION

Classification algorithms usually operate on a fixed set of data. When classification is applied to perform multimedia data, the input data might be digital virtual reality multimedia data. In this paper, we consider a different problem arising from the formalisation of virtual reality multimedia data classification as a classification problem. Classification problems of this type have been actively studied in multimedia classification.[1–4] For example,[5] a parametric mixture model is proposed for optical flow features with neighbourhood constraints. The number of clusters is selected by a likelihood heuristic object. Temporal context is modelled implicitly using differential motion features. Explicit context models include designs based on hidden Markov model (HMMs)[6,7] or frame-to-frame model adaptation.[8,9] A method which approaches the problem's time series structure in a manner similar to Bayesian forecasting has recently been suggested in Ref. 10. The authors propose a Gaussian mixture model to represent image rather than motion features. Temporal context is incorporated using the estimate obtained on a given frame in the sequence as prior information for the following frame.

We propose a Bayesian method capable of addressing both temporal context and estimation of the number of classes by a single model. The distribution of each class in feature space is described by an exponential family model, which is estimated under its respective conjugate prior.[11,12] The components are combined in a Bayesian mixture model to represent a classification. For each component, the

The MS was accepted for publication on 28 May 2010.
* *Corresponding author: L.-C. Jin, School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China; email: longcunjin@shu.edu.cn*

prior is defined by the component's posterior estimated during the previous time step. Owing to the 'chaining' properties of conjugate pairs, this results in closed model formulation for the entire time series. The mixture proportions and number of components are controlled by a Dirichlet process prior.[13–15] As we will argue, the conjugate nature of Dirichlet process leads to a chaining property analogous to the exponential family case. This property is used to propagate classification structure along the time series in a similar manner to that in which the conjugate component distributions propagate parameter information. Inference of the model is conducted by an adaptation of the Gibbs sampler for Dirichlet process mixture models[16,17] to the time series model. To facilitate application of our model to the large amounts of data arising in multimedia classification, we show how the efficiency of the Gibbs sampler can be substantially increased by exploiting temporal smoothness and introduce a multi-scale sampling method to speed up processing of individual frames. Just as the model, the multi-scale algorithm is based on the properties of exponential family distributions.

The remainder of the paper is organized as follows. An overview of the related work is given in Section 2. The proposed algorithm is discussed in Section 3. Section 4 presents the experimental settings and performance evaluation. Section 5 concludes this paper.

2 RELATED WORK

The work in this paper is related to and inspired by several previously probabilistic approaches to classification, which we briefly review here. There has also been a considerable amount of decision tree-based work on Bayesian tree structures for classification and regression,[18–20] but this is not closely related to the present work. An agglomerative model is described for merging based on marginal likelihoods in the context of HMM structure induction.[7,21–24] Gaussian and diffusion-based hierarchical generative models are described, respectively, for which inference can be done using Markov chain Monte Carlo (MCMC) methods. Similarly, a hierarchical generative model is presented based on a mutation for multimedia data and uses a hierarchical fusion of contexts based on marginal likelihoods in a Dirichlet language model.[25]

An approximate approach is presented based on the likelihood ratio test statistics to compute the marginal likelihood for c and $c-1$ classes and use this in an agglomerative algorithm.[26,27] Hierarchical classifying of multinomial data consisting of a vector of features is performed. The classes are specified in terms of which subset of features have common distributions. Their classes can have different parameters for some features and the same parameters to model other features, and their method is based on merged finding that maximize marginal likelihood under a Dirichlet multinomial model.[28] A Bayesian hierarchical classifying algorithm is defined, which attempts to agglomeratively find the maximum posterior probability classifying but makes strong independent assumptions and does not use the marginal likelihood. Probabilistic abstraction hierarchies are presented from which a hierarchical model is learned in which each node contains a probabilistic model and the hierarchy favours placing similar models at neighbouring nodes in the tree as measured by a distance function between probabilistic models. The training multimedia data are assigned to leaves of this tree. An agglomerative algorithm is presented for merging time series based on greedily maximizing marginal likelihood.[29,30] A greedy agglomerative algorithm based on marginal likelihood, which simultaneously classifies rows and columns of multimedia expression data, has also recently proposed.

The proposed algorithm is different from the above algorithms in several ways. First, unlike Refs. 14 and 15, in fact it is not a hierarchical generative model of the data, but a hierarchical way of organizing nested classes. Second, the proposed algorithm is derived from Dirichlet process mixtures. Third, the hypothesis test at the core of the proposed algorithm tests between a single merged hypothesis and the alternative is exponentially many other classifying methods of the same multimedia data, not one versus two classes at each stage. Lastly, the proposed algorithm does not use any iterative approach, such as expectation maximisation (EM), or require sampling, such as MCMC, and is therefore significantly faster than most of the above algorithms. The proposed algorithm for virtual reality multimedia feature data is most closely related to methods for Bayesian phylogenetics. These methods typically assume that features are generated directly by stochastic process over a tree. The proposed algorithm adds an intervening layer of abstraction by assuming that

partitions are generated by a stochastic process over a tree, and that features are generated from these partitions. By introducing a partition for each feature, we gain the ability to annotate a hierarchy with the levels most relevant to each feature. The proposed algorithm is an extension of the block model that discovers a nested set of categories as well as which categories are useful for understanding each relation in the virtual reality multimedia dataset.

3 THE PROPOSED ALGORITHM

The proposed algorithm is based Dirichlet process mixture models by an approximate inference approach. Since a Dirichlet process mixture models with concentration hyperparameter α defines a prior on all partitions of the n_k virtual reality data points D_k (the value of α is directly related to the expected number of classes), the prior on the merged hypothesis is the relative mass of all n_k points belonging to one class versus all the other partitions of those n_k virtual reality data points consistent with the tree structure. This can be computed bottom-up as the tree built in Fig. 1.

The marginal likelihood of a virtual reality data point is

$$p(D_k)=\sum_{v\in V}\frac{\alpha^{m_v}\prod_{l=1}^{m_v}\Gamma(n_l^v)}{[\Gamma(n_k+\alpha)/\Gamma(\alpha)]}\prod_{l=1}^{m_v}p(D_l^v) \tag{1}$$

where V is the set of all possible partitioning of D_k, and m_v is the number of virtual reality data points in class l of partitioning v.

This is easily shown since

$$p(D_k)=\sum_v p(v)p(D^v)$$

where $p(v)=\dfrac{\alpha^{m_v}\prod_{l=1}^{m_v}\Gamma(n_l^v)}{[\Gamma(n_k+\alpha)/\Gamma(\alpha)]}$ and

initialize each leaf i to have $d_i=\alpha, \pi_i=1$
for each internal node k do
$d_k=\alpha\Gamma(n_k)+d_{left}\,d_{right_k}$
$\pi_k=\frac{\alpha\Gamma(n_k)}{d_k}$
end for

1 To compute prior on merging, where $right_k(left_k)$ indexes the right (left) subtree of T_k and $d_{right_k}(d_{left_k})$ is the value of d computed for the right (left) child of internal nodes k

$$p(D^v)=\prod_{l=1}^{m_v}p(D_l^v)$$

where $p(D_k)$ is a sum over all partitioning v, where the first fractional term in the sum is the prior on partitioning v and the second product term is the likelihood of partitioning v under the data. To compute by the hierarchical classifying is

$$p(\cdot D_k|T_k)=\sum_{v\in V_T}\frac{\alpha^{m_v}\prod_{l=1}^{m_v}\Gamma(n_l^v)}{d_k}\prod_{l=1}^{m_v}p(D_l^v) \tag{2}$$

where V_T is the set of all tree-consistent partitioning of D_k.

For a fixed tree, we can optimize over the hyperparameters by taking gradients. In the case of all data belonging to a single class

$$p(D|H_1)=\int p(D|\theta)p(\theta|\beta)\mathrm{d}\theta \tag{3}$$

We can compute $p(D|H_1)/\beta$. Using this, we can compute gradients for the component model hyperparameters bottom-up as the tree is being built

$$\begin{aligned}\frac{\partial p(D_k|T_k)}{\partial\beta}=&\pi_k\frac{\partial p(D|H_1)}{\partial\beta}+\\&(1-\pi_k)\frac{\partial p(D_i|T_i)}{\partial\beta}p(D_j|T_j)+\\&(1-\pi_k)p(D_i|T_i)\frac{\partial p(D_j|T_j)}{\partial\beta}\end{aligned} \tag{4}$$

Similarly, we can compute

$$\begin{aligned}\frac{\partial p(D_k|T_k)}{\partial\alpha}=&\frac{\partial\pi_k}{\partial\alpha}p(D_k|H_1)-\\&\frac{\partial\pi_k}{\partial\alpha}p(D_i|T_i)p(D_j|T_j)+\\&(1-\pi_k)\frac{\partial p(D_i|T_i)}{\partial\alpha}p(D_j|T_j)+\\&(1-\pi_k)p(D_i|T_i)\frac{\partial p(D_j|T_j)}{\partial\alpha}\end{aligned} \tag{5}$$

where $\dfrac{\partial\pi_k}{\partial\alpha}=\dfrac{\pi_k}{\alpha}-\dfrac{\pi_k}{d_k}\left(\dfrac{\partial d_k}{\partial\alpha}\right)$ and

$$\frac{\partial d_k}{\partial\alpha}=\Gamma(n_k)+\left(\frac{\partial d_{left_k}}{\partial\alpha}\right)d_{right_k}+\left(\frac{\partial d_{left_k}}{\partial\alpha}\right)d_{left_k}$$

These gradients can be computed bottom-up by additionally propagating d/α. This allows us to construct an EM-like algorithm where we find the best tree structure in the (Viterbi-like) E step and then optimizes over the hyperparameters in the M step. In our experiments, we have only optimized one of the hyperparameters with a simple line search for

Gaussian components. A simple experiments approach is to set the hyperparameters β by fitting a single model to the whole dataset.

In our proposed algorithm, we assume that virtual reality multimedia data are generated by a Dirichlet process mixture model. Virtual reality multimedia data d_j are represented as a random vector which consists of Dirichlet random variables X_{ij}, where X_{ij} has the value of within-data-frequency f_{ij} for the ith term t_i as follows

$$p(d_j)=p(X_{1j}=f_{1j}, X_{2j}=f_{2j}, \ldots, X_{|V|j}=f_{|V|j}) \quad (6)$$

If we assume that each of the variables X_{ij} is independent of one another, using an independence assumption, the probability of d_j is calculated as follows

$$p(d_j)=\prod_{i=1}^{|V|} p(X_{ij}=f_{ij})$$

where $|V|$ is the virtual reality data size and each $p(X_{ij}=f_{ij})$ is given by $p(X_{ij}=f_{ij})=\dfrac{\exp(-\lambda_{ic})\lambda_{ic}^{f_{ij}}}{f_{ij}}$

In our proposed virtual reality multimedia classification algorithm approach to Bayesian of inductive cognition model, the proposed Bayesian of inductive cognition model tests the following three measures to weight each term feature: virtual reality multimedia information gain, χ^2 statistics and an extended version of risk ratio.

Virtual reality multimedia information gain is an information-theoretic measure defined by the amount of reduced uncertainty given a piece of information. Virtual reality multimedia information gain for a term given a class, which becomes the weight of the term, is calculated using a virtual reality multimedia point data event model as follows

$$\begin{aligned} f_{w_{ic}} &= H(C)-H(\cdot C|W_i) \\ &= \sum_{C_s\in\{C,\overline{C}\}}\sum_{W_t\in\{W_i,\overline{W}_i\}} p(C_s, W_s)\log\frac{p(C_s, W_t)}{p(C_s, W_t)} \end{aligned} \quad (7)$$

where $p(c)$ is the number of virtual reality multimedia point data belonging to the class c divided by the total number of virtual reality multimedia point data and $p(\overline{w})$ is the number of virtual reality multimedia data point without the term w divided by the total number of virtual reality multimedia data point.

The second measure we used is χ^2 statistics developed for the statistical test of the hypothesis. In virtual reality multimedia data classification, given a two-way contingency table for each term t_i and the class c as represented shown in Table 1, $f_{w_{iC}}$ is calculated as follows

$$f_{w_{iC}}=\frac{(wz-xy)^2}{(w+x)(w+y)(y+z)} \quad (8)$$

where w, x, y and z indicate the numbers of virtual reality multimedia data point for each cell in Table 1.

While the above two measures have been widely tested in virtual reality multimedia data categorisation domain, we have tested additional measure: a extend version of risk ratio (ExtRR) as follows

$$f_{w_{iC}}=\frac{\lambda_{iC}}{\mu_{iC}}+\frac{\mu_{iC}}{\lambda_{iC}}$$

With this ExtRR measure, our z_{jC} is finally defined as follows

$$z_{jC}=\sum_{i=1}^{|V|}\frac{1}{FW_C}\text{if}\left(\frac{\lambda_{iC}}{\mu_{iC}}+\frac{\mu_{iC}}{\lambda_{iC}}\right)\log\frac{\lambda_{iC}}{\mu_{iC}} \quad (9)$$

This measure indicates the sum of the ratio of two Dirichlet process parameters and their reciprocal. The first term represents how term t_i is more likely to be presented in the class c compared to outside of the class c and the second term represents the reverse. With this measure, $f_{w_{ic}}$ has the minimum value of 2.0 for the term which does not have any virtual reality multimedia information data to predict the annotated class.

In our proposed Bayesian of inductive cognition model based on Dirichlet process, for Dirichlet process probability density functions, $p_i=N(x;\mu_i,\Sigma_i)$, where μ_i is the mean vector of the ith class measurements, and Σ_i is the within-class covariance matrix of the ith class, and its divergence is

$$\begin{aligned} D(p_i\|p_j)=\int \mathrm{d}x N\left(x;\ \mu_i,\ \sum\nolimits_i\right)\ln\frac{N(x,\ \mu_i,\ \sum_i)}{N\left(x;\ \mu_j,\ \sum_j\right)}\times \\ \frac{1}{2}\left[\ln\left|\sum\nolimits_j\right|-\ln\left|\sum\nolimits_i\right|+\mathrm{tr}\left(\sum\nolimits_j^{-1}\sum\nolimits_i\right)+ \\ \mathrm{tr}\left(\sum\nolimits_j^{-1}D_{ij}\right)\right] \end{aligned}$$

where $D_{ij}=(\mu_i+\mu_j)(\mu_i+\mu_j)^{\mathrm{T}}$ and $|\Sigma|=\det(\Sigma)$. To simplify the notation, we denote the divergence

Table 1 Two-way contingency

	Presence of t_i	Absence of t_i
Annotated as C	w	x
Not annotated as C	y	z

between the projected densities $p(W^{\mathrm{T}}x|y=i)$ and $p(W^{\mathrm{T}}x|y=j)$

$$D_{\mathrm{W}}(p_{\mathrm{i}}\|p_{\mathrm{j}}) = D\left[p\left(W^{\mathrm{T}}x\middle|y=i\right)\middle\|p\left(W^{\mathrm{T}}x|y=j\right)\right] = \frac{1}{2}\left[\ln\left|W^{\mathrm{T}}\sum\nolimits_{\mathrm{j}}W\right| - \ln\left|W^{\mathrm{T}}\sum\nolimits_{\mathrm{i}}W\right|\right] + \mathrm{tr}\left\{\left(W^{\mathrm{T}}\sum\nolimits_{\mathrm{j}}W\right)^{-1}\left[W^{\mathrm{T}}\left(\sum\nolimits_{\mathrm{i}}+D_{\mathrm{ij}}\right)W\right]\right\}$$

To improve the proposed Bayesian of inductive cognition model based on Dirichlet process classifying virtual reality multimedia data, we need to maximize the linear combination of the log of the Dirichlet mean of the divergences and the log of the normalized divergences

$$W^* = \arg\max_{\mathrm{W}}\left\{\alpha\log\left[\prod_{1\le \mathrm{i}\ne \mathrm{j}\le \mathrm{c}} E_{\mathrm{W}}\left(p_{\mathrm{i}}\|p_{\mathrm{j}}\right)\right]^{1/\mathrm{c}(\mathrm{c}-1)} + (1-\alpha)\log\sum_{1\le \mathrm{i}\ne \mathrm{j}\le \mathrm{c}}\left[D_{\mathrm{W}}(p_{\mathrm{i}}\|p_{\mathrm{j}})\right]^{\mathrm{q_i q_j}\Big/\sum_{1\le \mathrm{m}\ne \mathrm{n}\le \mathrm{c}}\mathrm{q_m q_n}}\right\} \quad (10)$$

where the supremum of α is 1 and the infimum of α is 0. When $\alpha=0$, the above equation reduces to

$$W^* = \arg\max_{\mathrm{W}}\prod_{1\le \mathrm{i}\ne \mathrm{j}\le \mathrm{c}}\left[D_{\mathrm{W}}\left(p_{\mathrm{i}}\|p_{\mathrm{j}}\right)\right]^{\mathrm{q_i q_j}\Big/\sum_{1\le \mathrm{m}\ne \mathrm{n}\le \mathrm{c}}\mathrm{q_m q_n}} \quad (11)$$

and when $\alpha=1$, the above equation reduces to

$$W^* = \arg\max_{\mathrm{W}}\left[\prod_{1\le \mathrm{i}\ne \mathrm{j}\le \mathrm{c}} E_{\mathrm{W}}\left(p_{\mathrm{i}}\|p_{\mathrm{j}}\right)\right]^{1/\mathrm{c}(\mathrm{c}-1)} \quad (12)$$

By setting $q_{\mathrm{i}}=1/c$, we can simplify the above formula as

$$W^* = \arg\max_{\mathrm{W}}\left\{\sum_{1\le \mathrm{i}\ne \mathrm{j}\le \mathrm{c}}\log D_{\mathrm{W}}\left(p_{\mathrm{i}}\|p_{\mathrm{j}}\right) - \alpha c(c-1)\log\left[\sum_{1\le \mathrm{i}\ne \mathrm{j}\le \mathrm{c}} D_{\mathrm{W}}\left(p_{\mathrm{i}}\|p_{\mathrm{j}}\right)\right]\right\} \quad (13)$$

Based on above equations, we define the value of the virtual reality multimedia objective function as

$$L(W) = \frac{1}{c(c-1)}\sum_{1\le \mathrm{i}\ne \mathrm{j}\le \mathrm{c}}\log D_{\mathrm{W}}\left(p_{\mathrm{i}}\|p_{\mathrm{j}}\right) - \log\left[\sum_{1\le \mathrm{i}\ne \mathrm{j}\le \mathrm{c}} q_{\mathrm{i}}q_{\mathrm{j}}D_{\mathrm{W}}\left(p_{\mathrm{i}}\|p_{\mathrm{j}}\right)\right] + \frac{(1-\alpha)}{\alpha\sum_{1\le \mathrm{m}\ne \mathrm{n}\le \mathrm{c}} q_{\mathrm{m}}q_{\mathrm{m}}}\sum_{1\le \mathrm{m}\ne \mathrm{n}\le \mathrm{c}} q_{\mathrm{i}}q_{\mathrm{j}}\log D_{\mathrm{W}}\left(p_{\mathrm{i}}\|p_{\mathrm{j}}\right) \quad (14)$$

Therefore, $W^*=\mathrm{argmax}_{\mathrm{W}}L(W)$. The virtual reality multimedia objective function $L(W)$ depends only on the sub-objective by the columns of W or $L(W)=L(WQ)$ when Q is an orthogonal $r\times r$ matrix.

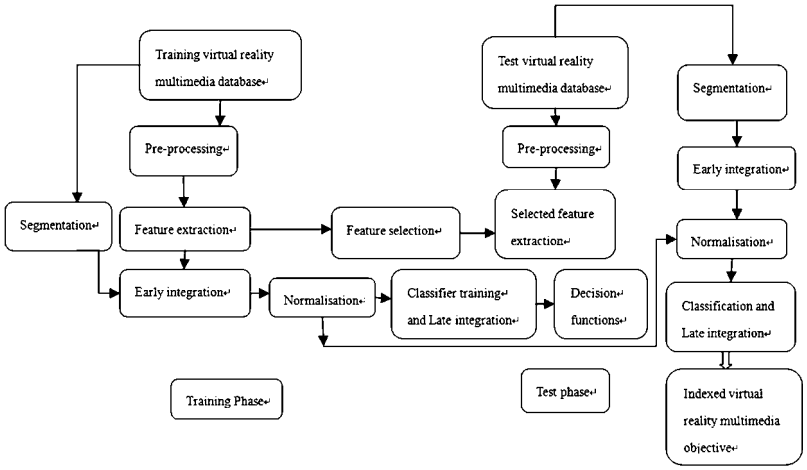
The proposed Bayesian of inductive cognition classification algorithm procedure architecture is shown in Fig. 2. Figure 2 shows how to utilize Bayesian inference to abstract feature of scenes. It also shows an extension of the block model that discovers a nested set of categories. The parameters of the algorithm procedure have been discussed in discussed in Section 3. The proposed Bayesian of inductive cognition model based on Dirichlet process classifying virtual reality multimedia data can be denoted as shown in Fig. 3. The proposed scheme consists of training phase and test phase. The training phase using Bayesian of inductive cognition algorithm to train virtual reality multimedia's database includes segmentation, pre-processing, feature extraction, feature selection and so on. The test phase utilizes the training phase to classify scenes of CORA.[31] The test phase consists of classifier training, late integration, decision functions, selected feature extraction and so on. Then, the scheme is implemented to classify scenes in our database.

Input: Training measurements $X_{i;j}$,where i denotes the i^{th} class ($1 \le i \le c$) and j denotes the j^{th} measure in the i^{th} class ($1 \le j \le n_i$), the dimensionality of selected features $k < n$ (n is the dimensionality of $x_{i;j}$), the maximum number M of different initial values for the projection matrix W, the learning rate k (a small value), the combination factor η, and a small value ε as the convergence condition.

Output: An estimate W^* of the class.

```
for m=1:M {
    Initialize W_t^m (t = 1) randomly
    while |L(W_t^m)| is defined in W* .
    do{
        conduct the gradient steepest algorithm:
        W_t^m ← W_{t-1}^m + k · ∂_W L(W_{t-1}^m), where
        ∂_{W_t^m} L(W_{t-1}^m) is given.
    t ← t + 1.
    }//while on line 3
      W^m ← W_t^m
}//for on line 1
    W ← argmax_m L(W^m)
    Orth-normalization Step: W* ←orth-normalize (W).
```

2 Architecture of the virtual reality Bayesian of inductive cognition model based on Dirichlet process classification algorithm system



3 Architecture of the virtual reality Bayesian of inductive cognition model based on Dirichlet process classification algorithm system

4 EXPERIMENTS

The experiments are conducted on several virtual reality multimedia datasets to validate our proposed Bayesian of inductive cognition model based on Dirichlet process. The datasets are more than 10 GB with resolution 320×240 and are sampled at one frame per second. These virtual reality multimedia data are named CORA.[31] We have collected a database called the CORA with more than 10 GB. In summary, these multimedia sequences pose the following challenges: the object of interest can have wild changes in appearances, including pose and lighting variations; the background can be highly cluttered and non-stationary; the object can leave and re-enter the scene multiple times, which may occur due to large multimedia motion or post-editing of the multimedia sequence.

In the virtual reality multimedia data classification literature, an itemset refers to a set of items, which in our application refers to a candidate set of regions that could represent an object of interest. A frequent itemset is an itemset that occurs at least a certain number of times, and hence more likely corresponds to an object of interest. A recent multimedia data classification algorithm[32] using semantic classification discovers frequent closed itemset, such that for each discovered frequent itemset, there exists no superset of equal frequency. This helps in reducing the final number of itemsets to be considered. The algorithm requires the minimum itemset frequency as an input parameter. Setting the minimum frequency too small will result in too many frequent closed itemsets, and many of them might not correspond to the object of interest. Hence, we start from largest possible minimum frequency, which is equal to the number of frames, and gradually decrease it until M frequent closed itemsets are found. We found $M=16$ to give the best results.

Our proposed Bayesian of inductive cognition model based on Dirichlet process provides a natural way for object-oriented scene classification, and is also able to point out 'what' is exactly the factor that separates the frames. In Table 1, we compare the proposed framework to two baseline methods. Each multimedia sequence has a natural object of interest, e.g. the BOY and BOY-HOUSE. The BOY and BOY-HOUSE sequences used in the localisation experiment are not used here because they did not contain transitions from one object to another. The proposed Bayesian of inductive cognition model based on Dirichlet process is one frame per second and the motions of both the object of interest and the background are fast, making it non-trivial to apply optical flow or layer extraction methods for discovering objects. In addition, all sequences frequently transit between different shots. The average duration

Table 2 Virtual reality multimedia object-oriented scene classification performance

Sequence	No. of Frames	The proposed algorithm (%)	Baseline (NM) (%)	Baseline (FREQ) (%)
BOY	15 844 (5248)	95.3	86.3	82.5
BOY-HOUSE	18 956 (6844)	93.5	78.9	883.6

4 Results of object-oriented scene classifying using our proposed Bayesian of inductive cognition model based on Dirichlet process. Each scene in the top three rows contains the virtual reality BOY; in the three bottom rows, each scene contains the virtual reality BOY-HOUSE. This provides a Bayesian of inductive cognition model based on Dirichlet process algorithm by object-oriented scene overview of the whole multimedia sequence in a different way from traditional key frame extraction

of a shot is compared to the multimedia length. This also demonstrates the difficulty of using optical flow-based methods. This also demonstrates the difficulty of using optical flow-based methods. The ground-truth data labels the presence or absence of the object of interest in each frame. We evaluate the object mining performance as a detection problem. The classification rates are shown in Table 2. Numbers in parenthesis indicate the number of frames containing the object of interest.

Virtual reality multimedia data in which the object-oriented classification is switching among a number of scenes are considered, for example, in the test drive scene in Fig. 4, the object-oriented switches between the houses, the frontal view of the house, the side view, and so on. We would like to classify the framework into semantically meaningful groups. In classical temporal segmentation methods, the similarity between two frames is assessed using global image characteristics. For example, all pixels are used to build

a colour histogram for each frame, and a distance measure such as the chi-square distance is used to measure the similarity between two histograms.

5 CONCLUSION

The virtual reality multimedia information data classification and objected-oriented nature of our algorithm provide promising new directions for multimedia scene classifying. At present, our proposed Bayesian of inductive cognition model based on Dirichlet process algorithm only provides a rough position estimate of the object of interest. For key frame classifying, this can be enough, but in some other areas such as high-quality editing, it might be of interest to obtain a clearer contour mining of the multimedia shots. This might require sophisticated feature detectors in addition to virtual reality multimedia information data classification.

ACKNOWLEDGEMENTS

This research work is supported by the National High Technology Research and Development Program of China (863 Program, no. 2007AA01Z319), the National Natural Science Foundation of China (nos. 60873130 and 60872115) and the Shanghai's Key Discipline Development Program (no. J50104).

REFERENCES

1 Song, D. J. and Tao, D. C. Biologically inspired feature manifold for scene classification. *IEEE Trans. Image Process.,* 2010, **19**, 174–184.
2 Brezeale, D. and Cook, D. J. Automatic video classification: a survey of the literature. *IEEE Trans. Syst. Man Cybern. C,* 2008, **38C**, 416–433.
3 Bosch, A., Zisserman, A. and Munoz, X. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. Patt. Anal. Mach. Intell.,* 2008, **30**, 712–727.
4 Gao, Y., Wang, W.-B. and Yong, J.-H. A video summarization tool using two-level redundancy detection for personal video recorders. *IEEE Trans. Consum. Electron.,* 2008, **54**, 521–526.
5 Weiss, Y. and Adelson, E. H. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models, Proc. IEEE Computer Society Conf. on Computer vision and pattern recognition: CVPR '96, San Francisco, CA, USA, June 1996, IEEE Computer Society, pp. 321–326.
6 Bregler, C. Learning and recognizing human dynamics in video sequences, Proc. IEEE Computer Society Conf. on Computer vision and pattern recognition: CVPR '97, San Juan, Puerto Rico, June 1997, IEEE Computer Society, pp. 568–574.
7 Huang, J. C., Liu, Z. and Wang, Y. Joint scene classification and segmentation based on hidden Markov model. *IEEE Trans. Multimedia,* 2005, **7**, 538–550.
8 Khan, S. and Shah, M. Object based segmentation of video, using color, motion and spatial information, Proc. IEEE Computer Society Conf. on Computer vision and pattern recognition: CVPR 2001, IEEE Computer Society, Vol. 2, pp. 746–751.
9 Chen, D.-Y., Cannons, K., Tyan, H.-R., Shih, S.-W. and Mark Liao, H.-Y. Spatiotemporal motion analysis for the detection and classification of moving targets. *IEEE Trans. Multimedia,* 2008, **10**, 1578–1591.
10 Goldberger, J. and Greenspan, H. Context-based segmentation of image sequences. *IEEE Trans. Patt. Anal. Mach. Intell.,* 2006, **28**, 463–468.
11 Bernardo J. M. and Smith A. F. M. Bayesian Theory, 1994 (Wiley, Chichester).
12 Duan, L.-Y., Jin, J. S., Tian, Q. and Xu, C.-S. Nonparametric motion characterization for robust classification of camera motion patterns. *IEEE Trans. Multimedia,* 2006, **8**, 323–340.
13 Ferguson T. S. A Bayesian analysis of some nonparametric problems. *Ann. Stat.,* 1973, **1**, 209–230.
14 Bouguila, N. and Ziou, D. A Dirichlet process mixture of Dirichlet distributions for classification and prediction, Proc. IEEE Workshop on Machine learning for signal processing: MLSP 2008, Cancún, Mexico, October 2008, IEEE, pp. 297–302.
15 Cao, L.-L., Luo, J. B., Kautz, H. and Huang, T. S. Image annotation within the context of personal photo collections using. *IEEE Trans. Multimedia,* 2009, **11**, 208–219.
16 Neal, R. M. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.,* 2000, **9**, 249–265.
17 Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Trans. Signal Process.,* 2008, **56**, 71–84.
18 Chipman, H., George, E. and McCulloch, R. Bayesian cart model search with discussion. *J. Am. Stat. Assoc.,* 1998, **93**, 935–960.
19 Denison, D., Holmes, C., Mallick, B. and Smith, A. Bayesian Methods for Nonlinear Classification and Regression, 2002 (Wiley, West Sussex).

20 Wang, X. G., Ma, X. X. and Grimson, W. E. L. Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Trans. Patt. Anal. Mach. Intell.,* 2009, **31**, 539–555.

21 Stolcke, A. and Omohundro, S. Hidden Markov model induction by Bayesian model merging. *Adv. Neural Inform. Process. Syst.,* 1993, **5**, 11–18.

22 Nikseresht, A. and Gelgon, M. Gossip-based computation of a Gaussian mixture model for distributed multimedia indexing. *IEEE Trans. Multimedia,* 2008, **10**, 385–392.

23 Williams, C. A MCMC approach to hierarchical mixture modeling. *Adv. Neural Inform. Process. Syst.,* 2000, **12**, 680–686.

24 Neal, R. M. Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Stat.,* 2003, **7**, 619–629.

25 Jensen, C., S., Lin, D. and Ooi, B. C. Continuous clustering of moving objects. *IEEE Trans. Knowl. Data Eng.,* 2007, **19**, 1161–1173.

26 Banfield, J. D. and Raftery, A. E. Model-based Gaussian and non-Gaussian clustering. *Biometrics,* 1993, **49**, 803–821.

27 Vaithyanathan S. and Dom, B. Model-based hierarchical clustering, Proc. 16th Conf. on Uncertainty in artificial intelligence, Stanford, CA, USA, June 2000, Stanford University, pp. 599–608.

28 Iwayama, M. and Tokunaga, T. Hierarchical Bayesian clustering for automatic text classification, Proc. 14th Int. Joint Conf. on Artificial intelligence: IJCAI-95, Montreal, Que., Canada, August 1995, CSCSI, pp. 1322–1327.

29 Zhang, Z. F., Masseglia, F., Jain, R. and Bimbo, A. D. Editorial: Introduction to the special issue on multimedia data mining. *IEEE Trans. Multimedia,* 2008, **10**, 165–166.

30 Ekin, A., Tekalp, A. M. and Mehrotra, R. Automatic soccer video analysis and summarization. *IEEE Trans. Image Process.,* 2003, **12**, 796–807.

31 Jin, L. C., Wan, W. G., Cui, B., Yu, X. Q. and Xu, H. W. A new multimedia information data mining method, Proc. ACM 2009 World Summit on Genetic and evolutionary computation, Shanghai, China, June 2009, ACM, pp. 899–902.

32 Duan, L.-Y., Xu, M., Tian, Q., Xu, C.-S. and Jin, J. S. A unified framework for semantic shot classification in sports video. *IEEE Trans. Multimedia,* 2005, **7**, 1066–1083.