PROCEEDINGS

The HKBU 6th Computer Science Postgraduate Research Symposium

July 3, 2007

PG Day 2007

Department of Computer Science
Hong Kong Baptist University

The 6th HKBU-CSD Postgraduate Research Symposium (PG Day) Program

July 3 Tuesday, 2007	
Time	**Sessions**
09:15-09:30	**On-site registration** (LMC 514)
09:30-09:40	**Welcome**: Prof. Jiming Liu, Head of Computer Science Department (LMC 514)
09:40-11:10	Session A1: (Chair: JunYang Zhou) (LMC 514) *Intelligent Informatics*
	■ *Scientific Workflow Creation with Privacy Policy Compliance Checking* KaiKin Chan ■ *Automatic Semantic Classification of Images* ChunFan Wong ■ *A Model for System Functionality Definition* Di Wu
11:10-11:20	**Tea Break**
11:20-12:50	Session A2: (Chair: Xin Li) (LMC 514) *Intelligent Informatics*
	■ *Topic Detection Via Participation Using Markov Logic Network* ChiWa Cheng ■ *Iterative Feature Selection in Gaussian Mixture Clustering with Automatic Model Selection* Hong Zeng ■ *Reasoning about Decommitment in G-Negotiation Mechanism* BenYun Shi
12:50-14:00	**Noon Break**
14:00-15:00	Session B: (Chair: KaHo Chan) (LMC 514) *Networking*
	■ *Processing Continuous Spatial Queries in Wireless Sensor Networks* Yu Li ■ *Precise Modeling of Saturation Throughtput of IEEE 802.11 Point-to-Point Link* Yong Yan
15:00-15:10	**Tea Break**
15:10-17:10	Session C: (Chair: Zhili Wu) (LMC 514) *Pattern Recognition*
	■ *Object Tracking Using Information Content* Chang Liu ■ *Behavior of Virtual Human in equipment maintenance* YueSheng He ■ *Image Segmentation based on the Maximal Variance and Improved Genetic Algorithm* JianJia Pan ■ *Face Recognition Using Wavelet Packet Decomposition and Support Vector Machine* LiMin Cui
18:30	**Best Paper & Best Presentation Awards Announcement via Email**

TABLE OF CONTENTS

Session A1: Intelligent Informatics

Session A2: Intelligent Informatics

Session B: Networking

Session C: Pattern Recognition

Scientific Workflow Creation with Privacy Policy Compliance Checking

Kai-kin Chan

Abstract

There are several similarities between distributed data mining processes and scientific workflows. In this paper, we model the distribution data mining processes as scientific workflows that can provide a better creation and management of data mining discovery pipeline. As the processes may involved different parties, privacy becomes an important issue. Polices we are used which can help the privacy enforcement. As the data sets may come from different parties, we are purpose to execute the workflows in service-oriented environment in our future work.

1 Introduction

Data mining aims to extract useful information and discover patterns from large data sets. Recently, computating power and computer storage are continuously increasing, the sizes of data sets become larger and larger. Moreover, the Internet is well developed, the data sets are more distributed. Many of these data sets are stored in different physical locations. Thus, distributed data mining becomes an important issue.

There is an increasing number of scientific applications that run on distributed computing environments. Modeling complex scientific applications as scientific workflows is an effective means for presenting and managing the underlying processes of execution. In general, workflows are composed of interelated computational or data management jobs submitted to the remote hosts for execution to fulfill the goal of designing the workflow.

In distributed data mining process, a discovery pipeline involves different interactive and iterative stages. In these stages, users need to access, analyse and integrate from different distributed data sets. The workflow system enables users to compose processes to achieve the particular purpose, for example, data mining purpose. In our work, we model the distributed data mining processes as workflows, so that it is better to create and manage the discovery pipeline and processes.

The life cycle of a scientific workflow includes the steps to (1) create valid workflow description with respect to some domain independent constraints, (2) handle data sets with many elements and manage the creation of interactive substructures in the workflow that process each of those elements, and (3) submit the fully described workflows for execution. Validating the result workflow is a challenge, most of the validation steps are done by hand. With the semantic approach, data and processes are well described, the management and validation are easier.

For example, Wings [10, 7] is a workflow creation tool designed to create and validate very large scientific workflows. In Wings, workflow templates and instances are semantic objects. The workflow instances created by Wings can then be submitted to a grid-based distributed computing environment called Pegasus [6] for execution. Pegasus maps abstract workflows onto grid environment. Physical locations for both workflow components and data are automatically located. Moreover, Pegasus finds appropriate resources to execute the components. Existing data products where applicable are reused.

As the distributed data sets for distributed data mining may come from different parties, service-oriented architectures (SOAs) may be needed. Pegasus provides a stable execution environment, but it is not suitable for SOAs. Business Process Execution Language for Web Services (BPEL4WS) [1] is an industry standard for SOAs. Users can define business processes that make use of web services. These business processes can be externally treated as web services. Web Services are based on a number of standards such as WSDL [4] and SOAP [2]. Business process specifies the execution order of a collection of web services, data sharing between these web services. Moreover, different partners can be involved in the business process.

There are several challenges in modeling scientific workflow as business process. First, the scale of scientific workflow is usually larger than that of business process. Second, the data in scientific workflow are more complex than that in business process. Third, scientific workflow may need to change frequently, but business process may not.

Many data sets in distributed data mining process may come from different parties, data privacy becomes an important issue. Policies can be a constraint that controls the access of data. In our work, we add the policies into workflow that support reasoning, so that can achieve pri-

vacy management purpose.

To achieve privacy enforcement, the object would go through several difficulties. First, how to utilize the application of the analysis and methodology of data processes, the concepts of data privacy, as well as the affiliated workflow composition (ontological issue) to represent polices in a proper and significant manner. The next obstacle would be: how to be automatically enforced of the policy during the analytical process and its managerial methodology within the workflow system (policy enforcement issue). Finally, how to contribute those same categorized data from the provided provenance and data analysis history, the system can justify the function of the data (provenance issue) further.

2 Background

2.1 Distributed Data Mining

The followings are some related work on distributed data mining.

Collective Data Mining (CDM) is a framework to learn from heterogenous data sites. Heterogenous sites storing data for different set of features, but they still can have some common features.

In Distributed cooperative Bayesian learning approach, different Bayesian agents estimate the parameters of the target distribution and a global learner combines the outputs of each local model.

Similarity-based distributed data mining (SBDDM) is a framework which explicitly take the differences among distributed sources into consideration. This framework virtually integrate the data sets into groups based on their similarities and various distributed data mining techniques then apply to each resulting group.

2.2 Workflow Creation

The followings are some related work on workflow creation.

Sedna [5] is a graphical BPEL editor based on Eclipse platform. Sedna provides a number of features such as usability, automation, validation and deployment. These features aim is to further abstract away from BPEL and simplify the development of workflows. Sedna translates and exports the various language elements such as WSDL, XSD, Scientific PEL, Domain PEL etc. into standard BPEL and creates deployment descriptions for various BEPL workflow engines like Active BPEL engine, Oracle BPEL engine.

Triana is another GUI workflow creation tool. Triana allows users to drags services onto a canvas and to connect these services to each other. Triana supports a subset of BPEL and can export its workflows into BPEL.

Oracle BPEL Designer is a free BPEL editor based on Eclipse platform. It provides a one-to-one mapping to BPEL. It also offers macros that can be used to arrange sets of activities into reusable components. But it only can submit the BPEL output to Oracle's BPEL engine for execution.

Active BPEL Designer is another Eclipse based BPEL editor. It is offered by ActiveEndpoint. Most of the functions are similar to Oracle BPEL Designer, but it submits the BPEL output to Active BPEL engine for execution.

3 Creating Semantic Workflow for Data Analysis Process Management

The goal of our work is to extend the Wings workflow system for privacy relevant concepts. In this section, we show how to extend Wings that can support workflow creation in particular domain to achieve the privacy enforcement purpose. We take clinical data analysis as an example in the following sections.

3.1 Background of Wings

Wings and Pegasus provide three stages in creation of workflows. The first stage is to compose workflow templates. Workflow templates specify the abstract structure of a workflow. It is a high level structure that without identify any particular data and resources. When composing a workflow template, user can access and search the existing workflow templates and component libraries in a particular domain. Experienced users can create and validate a workflow template. Less experienced users can search the predefined workflow template and specify the input date for execution.

The second stage is to create workflow instances. Workflow instances specify the data and resources used for the workflow template, such as the input data and output data. As the workflow template can be reused by different users, so it should be an abstract structure that independent to the data and resources. Each time, users can process this stage to specify the resources for this execution.

The third stage is to create an executable workflow. It involves the resources managements and data movements in the distributed services environment. The first and second stage can be done by Wings. After creating the workflow instances, Wings outputs a DAX (DAG XML description) file and a file library file. These files are XML based files which specify the inputs and outputs files for the workflow, and then are submitted to Pegasus for execution.

In Wings, all the objects are semantic objects represented by OWL-DL, such as components, files, collections, work-

flows and workflow templates. Users can create a set of ontology representing the files, components etc. in a particular domain. Then users can compose a workflow in Wings, and semantic check for workflow creation can be done by Wings based on this set of ontology.

3.2 Workflow Representation in Wings

There are three fundamental ontologies in Wings. They are file ontology, component ontology and workflow ontology.

File ontology: Files represent data. A number of files can be grouped as a file collection. The following basic types are defined in file ontology:

- **File**: It represents the basic file class.
- **FileCollection**: It represents a collection of files.

Component ontology: Components are the processes for computation. They can process several inputs and return several outputs. In our service-oriented workflow design, components can be web services. The following basic types are defined in component ontology.

- **ComponentType**: It is the abstract classes of the component types. A component is an instance of a component type class, and this instance refers to the actual computation service or code.
- **ComponentCollection**: It is a collection of component types.

Workflow ontology: It represents the data-independent workflow templates. The following basic types are defined in workflow ontology.

- **Node**: It represents the component or component collection to be executed.
- **Link**: It represents the generic link in the workflow. The following are the subclass of link.
  - **InputLine**: It represents links do not have origin node.
  - **InOutLink**: It represents links must have an origin node and a destination node.
  - **OutputLink**: It represents the links do not have a destination node.

3.3 Ontologies for Data Privacy and Data Analysis

Fig. 1 shows how do we extends the fundamental ontologies in Wings that can support clinical data analysis. File ontology, component ontology and workflow ontology are the fundamental ontologies described. Then we creates a privacy domain which is for privacy enforcement purpose and those ontologies in this domain are colored in grey. We introduce the extended ontologies as follow.

Workflow Ontology (Extension): The following classes and properties are added.

- **Purpose** class and **for** property: It represents the data analysis purpose of the workflow template.
- **OutputQuality** class and **hasOutputQuality** property: It represents the overall output quality.

File Ontology (Extension): The following classes and properties are added.

- **DataSet** class: It extends the File class and represents the raw data sets.
- **PPDataSet** class: It extends the DataSet class and represents the privacy preserved data sets.
- **ParameterFile** class: It extends the DataSet class and represents the parameter files for the processes.
- **Clusters** class: It extends the DataSet class and represents the intermediate forms of data products for later data analysis.
- **ClustersWithDataItems** class: It extends the Clusters class and represents the clustering result stored with data items.
- **ClustersWithStatistics** class: It extends the Clusters class and represents the clustering result stored with per-cluster statistics.
- **FileLocation** class and **hasLocation** property: It represents the physical file location.
- **Attribute** class and **hasAttribute** property: It represents the attribute of data.

Component Library: It extends component ontology to represent data analysis processes. The following classes and properties are added.

- **Aggregate** class: It extends the ComponentType class. We have **GMMAggregate** and **DataSetAggregate** under Aggregate class.
- **DAComponentType** class: It extends the ComponentType class. We have **Clustering**, **ManifoldLearner**, **Classifier** and **AssociationRuleGenerator** under DAComponentType class. Each of these classes can have their subclasses such as **GMM-Basic**, **GMM-FLA**, **GTM**, **ISOMAP**, **SVM** and **C4.5**.

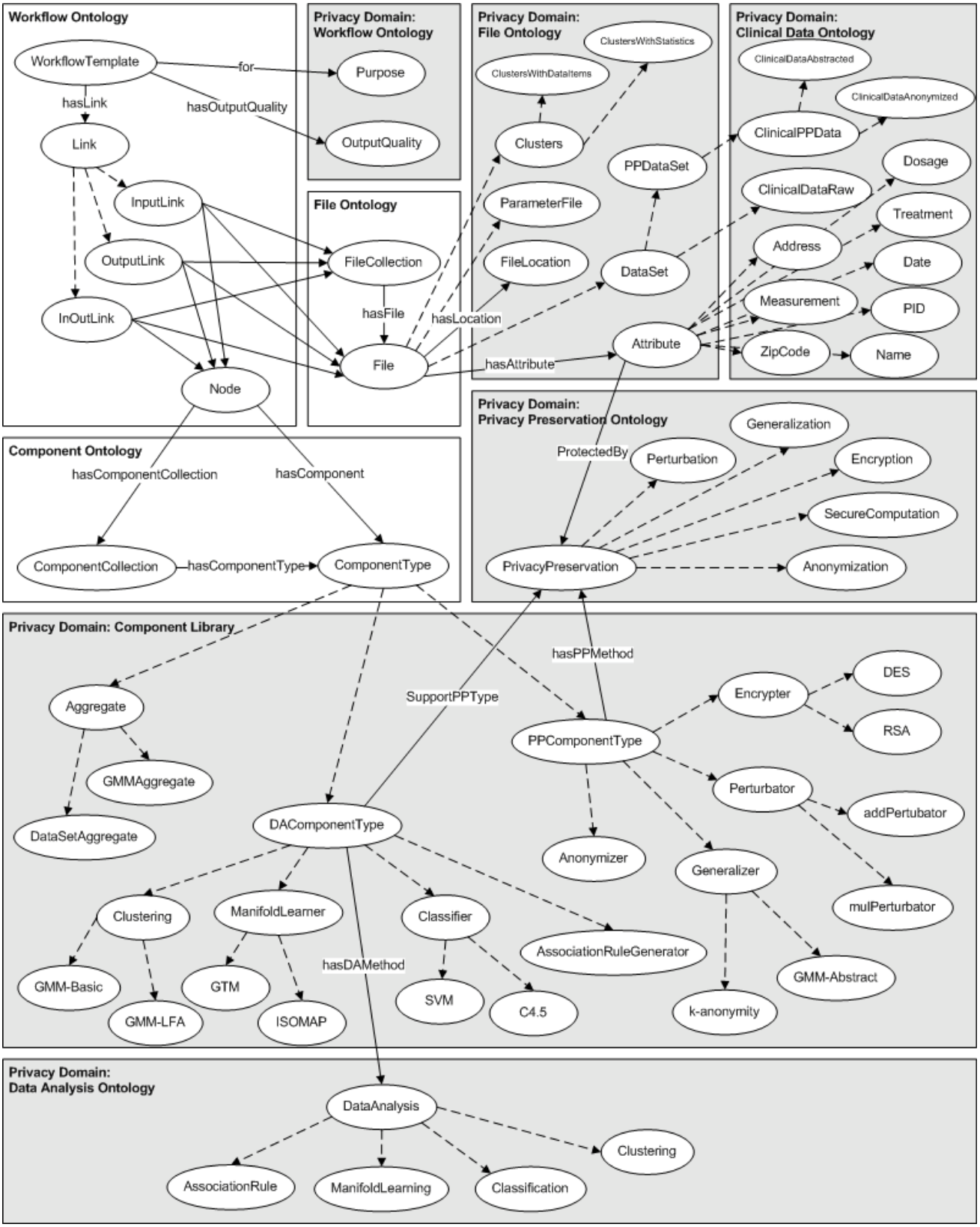


Figure 1. Ontologies for clinical data analysis

- **PPComponentType** class: It extends the ComponentType class. We have **Anonymizer**, **Generalizer**, **Encrypter** and **Perturbator** under PPComponentType class. Each of these classes can have their subclasses such as **k-anonymity**, **GMM-Abstract**, **addPertubator**, **mulPertubator**, **DES** and **RSA**.

Privacy Preservation Ontology: The following classes and properties are added.

- **PrivacyPreservation** class: It represents the privacy preservation methods. Its subclasses contains **Perturbation**, **Generalization**, **Encryption**, **SecureComputation** and **Anonymization**.

- **SupportPPType** property: It uses to specify which DAComponentType can support which privacy preservation method.

- **hasPPMethod** property: It uses to specify which PPComponentType has which privacy preservation method.

Data Analysis Ontology: The following classes and properties are added.

- **DataAnalysis** class: It represents the data analysis methods. Its subclasses contains **AssociationRule**, **ManifoldLearning**, **Classification** and **Clustering**.

- **hasDAMethod** property: It uses to specify which DAComponentType has which data analysis method.

3.4 Example - Clinical Data Analysis

The above extended ontologies are domain independent. They are the ontologies for privacy aware data analysis. We create a domain dependent ontology for clinical data analysis purpose. In this approach, we can easily reuse the domain independent part, and create the domain dependent part for different aspects. We create clinical data ontology for describing clinical data.

Clinical Data Ontology: The following classes and properties are added.

- **ClinicalDataRaw** class: It extends the DataSet class and represents the raw data sets of clinical data.

- **ClinicalPPData** class: It extends the PPDataSet class and represents the privacy preserved data sets of clinical data. We have ClinicalDataAbstracted and ClinicalDataAnonymized under ClinicalPPData Class for describing abstracted and anonymized privacy preserved clinical data.

4 Policy Representation and Reasoning

Semantic Web technology world be used in many areas, such as education, medical, government and finance. A semantic web application may involve a number of parties. Each party may involve its constraints on its data, such as security, authorization, privacy and preferences. Policy representation can help to represent and maintain the constraints on data. The following are several XML-based policy languages.

The Platform for Privacy Preferences Project (P3P) is normally used in websites. Privacy policies represented in a P3P standard format can be expressed in the websites. And then the policies can be retrieved automatically and interpreted easily by user agents. P3P user agents will allow users to be informed of site practices and to automate decision-making based on these practices when appropriate. Thus users need not read the privacy policies at every site they visit.

KAoS services and tools allow for the specification, management, conflict resolution, and enforcement of policies within the specific contexts established by complex organizational structures. While initially oriented to the dynamic and complex requirements of software agent applications, the services are also being adapted to general-purpose grid computing and web services environments as well.

Rei includes few constructs, based on deontic logic, that allow security policies, management policies and even conversation policies to be described in terms of rights, obligations, dispensations, and prohibitions. We specified the functionality of the Rei policy engine that interprets and reasons over Rei policies. This policy engine accepts policies in first order logic and RDF.We showed through examples how the policy engine can be used.

Rei represents policies using OWL-S. Authorization, Privacy and Confidentiality Policy are subclasses of Rei's Policy class. Authorization policies usually associated with services. Privacy and confidentiality policies usually associated with clients. The authorization policies grant permissions and prohibitions over attributes of the requester, service and the invocation context. The privacy policies restrict access to services satisfying I/O conditions. The confidentiality policies restrict on cryptographic characteristics of I/O parameter.

4.1 SWRL

Semantic Web Rule Language (SWRL) [3] is a rule language of the Semantic Web. SWRL is based on a combination of the OWL DL and OWL Lite sublanguages of the OWL Web Ontology Language. It allows users to write Horn-like rules to reason about OWL individuals and to infer new knowledge about those individuals. These rules are

expressed in terms of OWL concepts. SWRL is more expressive that OWL DL alone yet retains its formal semantics. As it is more expressive, it is more suitable then other rule languages. And our aim is not only detect violation, but also suggestion corrective actions, so SWRL is more suitable in our case.

4.2 Demonstrate the policy reasoning in Protege

In this section, we demonstrate the policy reasoning in Protege. Protege is a free, open source ontology editor and knowledge-base framework.

In [11], the policy representation in our current design contains four parts. They are context, usage requirement, protection requirement and corrective action. Context specifies which links, data or components the policy applies. Usage requirement and protection requirement are using for detecting policy violation. Correction actions are the suggestions for resolving the policy violation.

We have created several rules in the clinical data workflow template.

General Policy G1:"For all the inputs, it is required that the purpose of the workflow should be equal to the authorized usage of the inputs."
Context: WorkflowTemplate(?w) ∧ for(?w, ?l) ∧ hasFile(?l, ?d)
Usage: +ve: for(?w, ?pw) ∧ hasAuthorizedUse(?d, ?pdl) ∧ equal(?pw, ?pd)
Protection: NUL
Correction: prompt [workflow and data purpose mismatch]

After we created the workflow template in Wings, we modify the template from domain independent template to clinical domain specific template. We specify the input and output data set as clinical data set. Clinical data set is more specific one that provides some properties such as has AuthorizedUse and isMedSubClassOf for specific policy checking. To approach for policy reasoning, we create a data property **G1purposematch** in the workflow template.

Fig. 3 shows representation of policy in SWRL rule. Protege converts the SWRL rule to JESS, and then reasoning the rule. Then we can get the reasoning result. The G1purposematch is set to 1, it means that the workflow and data purpose is mismatch.

5 Conclusion and Future Works

In this paper, we are going to demonstrate how to create ontologies for privacy aware data analysis, and hot to represent and reason the policy. Our future work is going to extend Wings that can support (1) BPEL4WS execution and (2) policy reasoning.

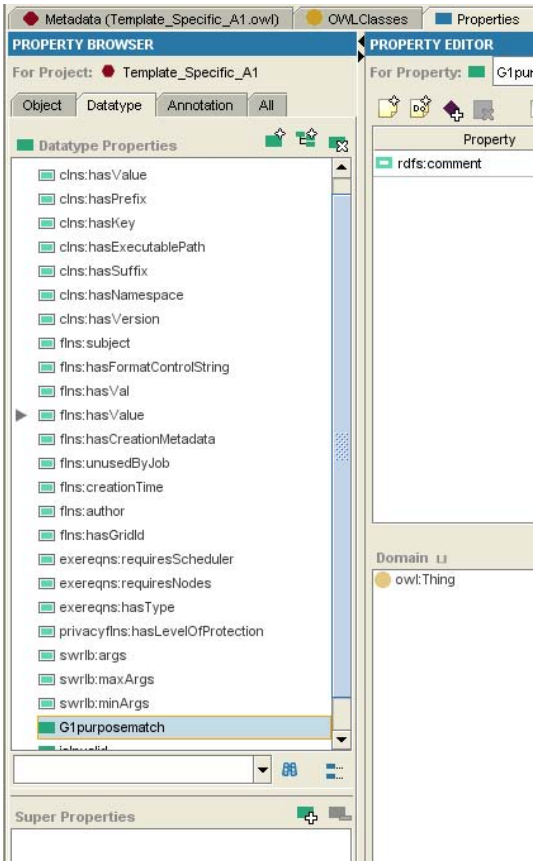


Figure 2. Data Property G1purposemarch added in Protege

There are three difficulties to submit the Wings output to BPEL4WS. First, the scale of scientific workflow is usually larger than that of business process. Second, the data in scientific workflow are more complex than that in business process. Third, scientific workflow may need to change frequently, but business process may not.

Moreover, to achieve policy reasoning, there are several difficulties. First, how to utilize the application of the analysis and methodology of data processes, the concepts of data privacy, as well as the affiliated workflow composition (ontological issue) to represent polices in a proper and significant manner. The next obstacle would be: how to be automatically enforced of the policy during the analytical process and its managerial methodology within the workflow system (policy enforcement issue). Finally, how to contribute those same categorized data from the provided provenance and data analysis history, the system can justify the function of the data (provenance issue) further.

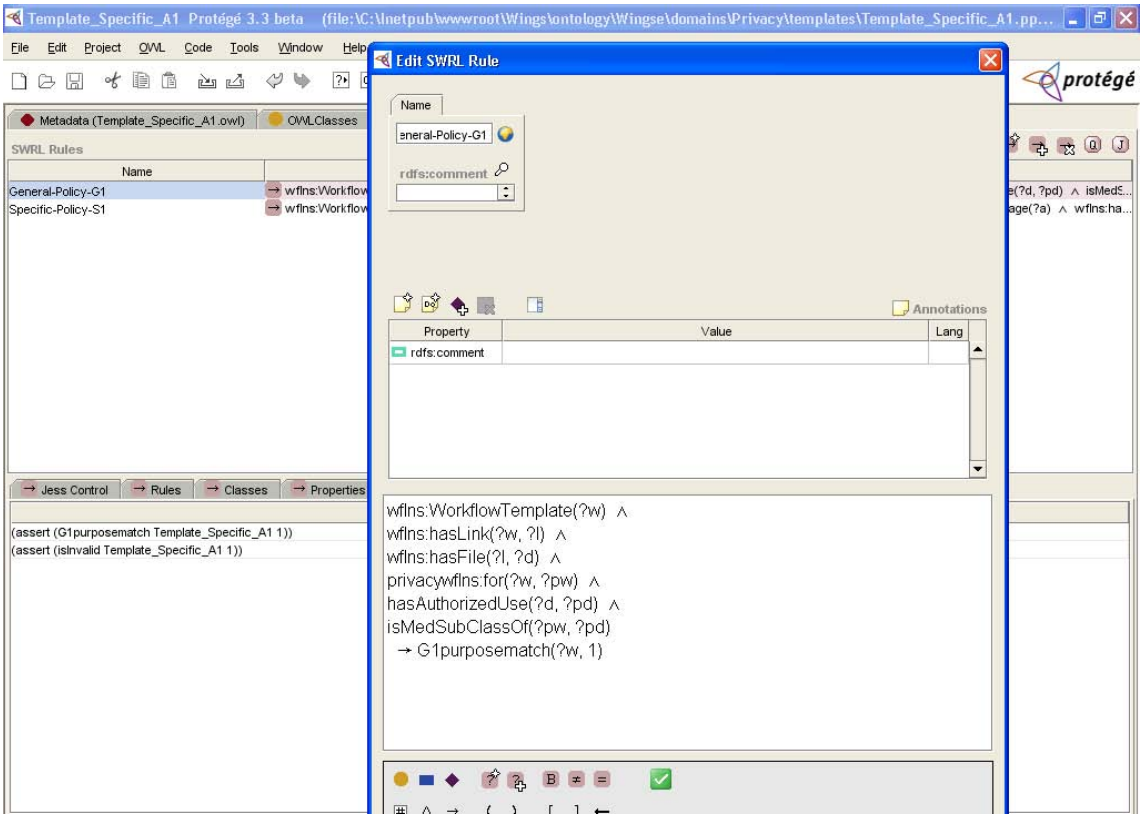


Figure 3. Policy reasoning in Protege

References

[1] Business Process Execution Language for Web Services version 1.1 http://www.ibm.com/developerworks/library/specification/ws-bpel/

[2] SOAP Version 1.2 http://www.w3.org/TR/soap/

[3] SWRL: A Semantic Web Rule Language Combining OWL and RuleML http://www.w3.org/Submission/SWRL/

[4] Web Services Description Language (WSDL) 1.1 http://www.w3.org/TR/wsdl

[5] Bruno Wassermann, Wolfgang Emmerich, Ben Butchart, Nick Cameron, Liang Chen, and Jignesh Patel: Sedna: A BPEL-Based Environment for Visual. Scientific Workflow Modeling. In Workflows for eScience - Scientific Workflows for Grids, Taylor, I.J., Deelman, E., Gannon, D., and Shields M.S. (eds.), Springer Verlag, 2006

[6] Ewa Deelman, James Blythe, Yolanda Gil, Carl Kesselman, Gaurang Mehta, Sonal Patil, Mei-Hui Su, Karan Vahi and Miron Livny: Pegasus: Mapping Scientific Workflows onto the Grid, 2004.

[7] Jihie Kim, Yolanda Gil and Varun Ratnakar: Semantic Meatdata Generation for Large Scientific Workflows. Proceeding of 5th International Semantic Web Conference, ISWC-2006, 2006.

[8] William K. Cheung, Xiao-Feng Zhang, Ho-Fai Wong, Jiming Liu, Zong-Wei Luo and Frank Tong: Service-Oriented Distributed Data Mining. IEEE Internet Computing, 10(4)(2006) 44-54, 2006.

[9] William K. Cheung and Yolanda Gil: Towards Privacy Aware Data Analysis Workflows for e-Science. To appear in 2007 Workshop on Semantic e-Science (SeS2007), held in conjunction with the Twenty-Second Conference of the Association for the Advancement of Artificial Intelligence (AAAI), 2007.

[10] Yolanda Gil, Varun Ratnakar, Ewa Deelman, Marc Spraragen and Jihie Kim: Wings for Pegasus: A Semantic Approach to Creating Very Large Scientific

Workflows. Proceeding of OWL: Experiences and Directions 2006, 2006.

[11] Yolanda Gil, William K. Cheung, Varun Ratnakar and Kai-kin Chan: Privacy Enforcement through Workflow Systems in e-Science and Beyond, 2007.

Automatic Semantic Classification of Images

Roger C. F. WONG

Abstract

For magazine editors and others, finding suitable scenes of images manually for a particular purpose is increasingly problematic. In this paper, we present a semantic query technique based on the use of image capture parameters and metadata to index and search collections of images. We develop a rule-based approach to help formulate annotations and search for specific scene of images. Experimental results indicate that this approach is able to deliver good classification performance. This approach can work in conjunction with common sense reasoning and ontology techniques to enable search embracing a high level of semantic richness.

1 Introduction

For magazine editors and others, finding suitable scenes of images for a particular purpose is increasingly problematic. In current practice we always attach tags or keywords into each image and categorizes it manually[11, 9]. Afterward, by using semi-structured indexing scheme to find the desired scenes of images. that allows a keyword search but not much more to help the user find the desired image[13].

In this paper, we explore a semantic query technique based on the use of image capture parameters and metadata to index and search collections of images. We develop a rule-based approach to help the automatic classification and search for specific scene of images.

This paper is structured as follows. First, we frame our problem statement of how to organize a semantic query. Second, we introduce some image acquisition parameters, such as scenes of images and exposure values, which is the basis of our approach to classify images. A brief description of the structure of metadata is also given. Third, we discuss how we evaluate our approach in the image retrieval. The paper concludes with a discussion of the implications of this work to the development of proposed approach and proposals for future work.

2 Problem Statement

Existing Text-based retrieval techniques for image retrieval have several limitations. Text-based methods is very powerful in matching context with high precision rate, but do not have access to image content[10]. Using textual captions to retrieve image is a possible solution, but searching captions for keywords and names will not necessarily yield the correct information, as objects mentioned in the caption are not always in picture. This results in a large number of false positives which need to be eliminated or reduced.

Furthermore, user queries always in semantic ways where textual-based retrieval may not deliver satisfy results. In this paper, we explore a semantic query technique based on the use of image acquisition parameters and the basic information attached in images without involving any human interaction and classification. As a result, we can resolve the following types of semantic queries :

- Night scene in Summer in France
- Sunset by the sea in New York in Autumn

3 Image Acquisition Technique

3.1 Scenes of Image

In the world of image acquisition many images can be broken down to few basic scenes[8], such as nature and wildlife, portrait, landscape and sports. Sub-scenes can be grouped into further categories, for example, landscape can be grouped into cloudscape photography, aerial landscape, street photography. However, this paper focuses on the four basic scenes in image retrieval system by metadata. A brief introduction to those scenes is followed.

A landscape comprises the visible features of an area of land, including physical elements such as landforms, living elements of flora and fauna, abstract elements such as lighting and weather conditions, and human elements, for instance human activity or the built environment. Landscape may also signify the objects around one in a building. Landscape Photography is the normal approach to ensure that as much of objects is in focus as possible. The simplest way to

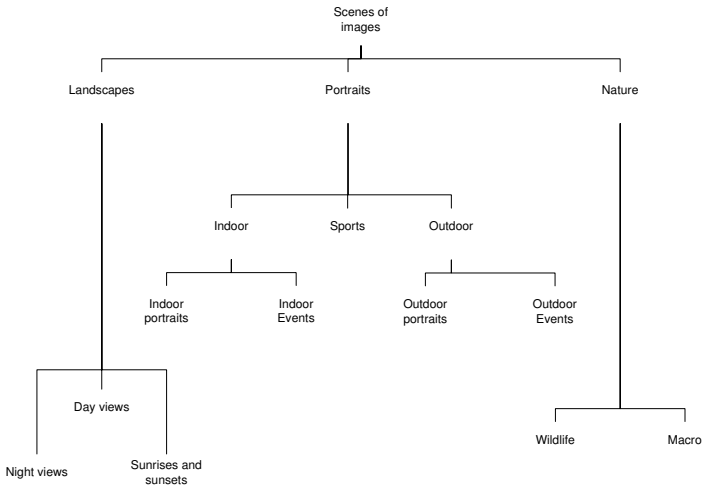


Figure 1. Tree of image scenes

do this is to choose a small Aperture setting as the smaller aperture the greater the depth of field in shots[1, 5].

The goal of portrait photography is to capture the likeness of a person or a small group of people, typically in a flattering manner. Like other types of portraiture, the focus of acquisition is the person's face, although the entire body and the background may be included. Many people enjoy having professionally made family portraits to hang in their homes, or special portraits to commemorate certain events, such as graduations or weddings[1].

Nature scenes refers to a wide range of photography taken outdoors and devoted to displaying natural elements such as landscapes, wildlife, plants, and close-ups of natural scenes and textures[5]. Nature photography tends to put a stronger emphasis on the aesthetic value of the photo than other photography genres, such as photojournalism and documentary photography[1].

Sports photography refers to the genre of photography that covers all types of sports. The equipment used by a professional photographer usually includes a fast telephoto lens and a camera that has an extremely fast shutter speed that can rapidly take pictures[1].

3.2 Exposure value

In image acquisition, exposure value (EV) denotes all combinations of camera shutter speed and relative aperture that give the same exposure. Exposure value is a base-2 logarithmic scale [8] defined by following:

$$\mathrm{EV} = \log_2 \frac{f^2}{t}$$

where

- f is the relative aperture (f-number)
- t is the exposure time (shutter speed)

EV 0 corresponds to an exposure time of 1 s and a relative aperture of f/1.0. If the EV is known, it can be used to select combinations of exposure time and f-number.

Each increment of 1 in exposure value corresponds to a change of one "step" in exposure, i.e., half as much exposure, either by halving the exposure time or halving the aperture area, or a combination of such changes. Greater exposure values are appropriate for image acquisition in more brightly lit situations, or for higher film speeds.

For a different ISO speed, increase the values by the number of exposure steps by which the speed is greater than ISO 100; formally

$$\mathrm{EV}_S = \mathrm{EV}_{100} + \log_2 \frac{S}{100}$$

For example, ISO 400 speed is two steps greater than ISO 100:

$$\mathrm{EV}_{400} = \mathrm{EV}_{100} + \log_2 \frac{400}{100} = \mathrm{EV}_{100} + 2$$

Scenes and sub-scenes	criteria 1	criteria 2	criteria 3	criteria 4	criteria 5
Landscape					
Night scenes	$EV > 8$	$d > 50m$	$h = 0$	$t > 0.6$	
Day scenes	$L < 30$	$d > 10$	$f > 8$	$EV > 8$	$h = 0$
Sunrises and Sunsets	$f > 20$	$EV > 11$	$d > 50$	$h = 0$	
Portrait					
Outdoor portraits	$30 < L < 200$	$f < 6$	$d < 10$	$EV > 10$	
Outdoor events	$30 < L < 150$	$f > 6$	$d <10$	$EV > 8$	
Indoor portraits	$30 < L < 70$	$d < 10$	$EV > 8$		
Indoor events	$30 < L < 70$	$d < 10$	$EV > 8$	$h = 1$	
Sports	$150 < L < 40$	$d > 10$	$t < 0.005$	$f < 10$	
Nature					
Macro	$d < 5$	$t < 0.03$			
Wildlife	$L > 450$	$t < 0.005$	$h = 0$	$d > 20$	$f > 4$

Table 1. Image classification scheme

To capture outdoor night sports with an ISO 400 speed imaging medium, find the tabular value of 9 and add 2 to get EV400 = 11.

For lower ISO speed, decrease the values by the number of exposure steps by which the speed is less than ISO 100. For example, ISO 50 speed is one step less than ISO 100:

$$\mathrm{EV}_{50} = \mathrm{EV}_{100} + \log_2 \frac{50}{100} = \mathrm{EV}_{100} - 1$$

4 Image Retrieval by metadata

The image acquisition techniques are essential in producing quality images, and photographer can apply each of these fundamentals each time he takes a picture. Those techniques are focused on the relationship to settings of lens and cameras, targeted subject and background, color tune and also the subject placement and composition.

The different categories of picture taking, the different combination of setting of camera and lens . For example, if the photographer wants to concentrate attention on just one part of the scene, and throw the rest out-of-focus, we should select a large aperture and a low volume of exposure time. Again, for the landscape photo, we will probably try to get in as much as possible, and thus, we need a wide-angle lens with a small aperture and long exposure time. Those acquisition techniques must be mastered before they can become an accomplished photographer.

4.1 Metadata

The digital still camera image file format standard, commonly referred to as Exif or metadata, is widely used as an international standard for digital cameras. The Standard, defining camera file system standards to enable image files to be exchanged among different recording media, was standardized in December 1998 as a companion to the Exif Standard. The most recent version, Exif Standard Version 2.2, was issued on Apr 2004 with additional tag information and recording format options [12].

The metadata tags defined in the Exif standard cover a broad spectrum including: Date and time information, Camera settings and Descriptions and copyright information. As it describe the content of images managed by digital libraries and are used for searching the documents [9], only limited tags and comments are entered manually. Some other common records about setting of camera including aperture(f), shutter speed (t), subject distance(d) and focal length (L) and fire activiation(h).

One of the typical example of Exif is to identify the geography information of the image. Location information can be included in the Exif, which could come from a GPS receiver connected to the camera. The GPS eXchange Format (GXF) format is a light-weight XML data format for the interchange of GPS data between applications and Web services on the Internet. The GPX 1.1 schema was released on August 9, 2004. GPX has been the de-facto XML standard for lightweight interchange of GPS data since the initial GPX 1.0 release in 2002. GPX is being used by dozens of software programs and Web services for GPS data exchange, mapping, and geocaching. [4]

The MakerNote tag is one of the tags in the public metadata area. Stored in this tag is very interesting information about camera settings which were in effect when taking the picture. Some of this information is repeated in normal metadata tags but most of it exists just in MakerNote and are more precise.

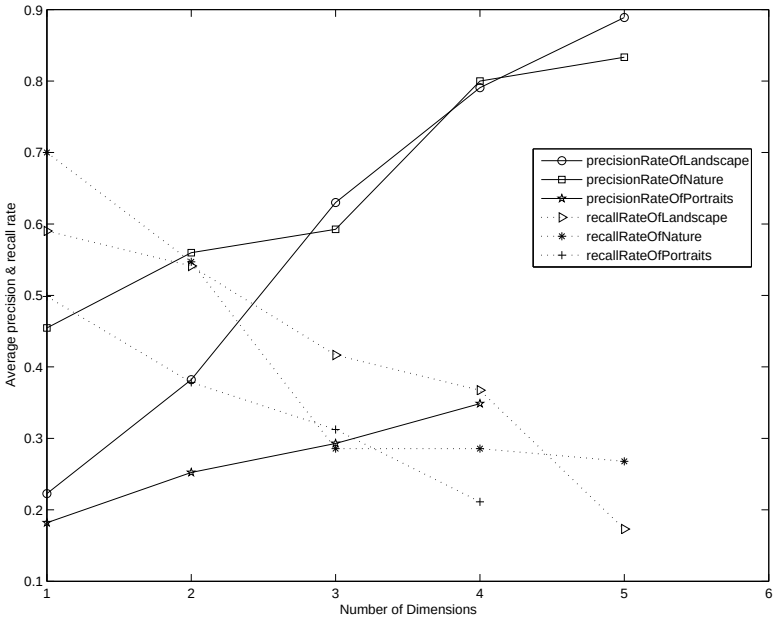


Figure 2. Average precision and recall rate of three primary scenes

4.2 Classification approach

There, we propose a novel technique for automatic image retrieval by using metadata of documents. For example, the lens aperture controls how much light per unit time reaches the sensor while shutter speed determines how long a certain amount of light will reach the sensor [3]. It annotates images with predefined semantic concepts by combining classification methods and techniques of image processing and visual feature extraction.

By proper use of camera settings of metadata, we could retrieve knowledge and acquisition techniques from photographer and consequently classify scene of images photographer intended to represent. Table 1 illustrates the proposed approach for scheme of image classification.

In table 1, there are a number of criteria(c) determines the scenes of images where weight of first criteria is higher than the fifth's one. The more criteria be satisfied, the higher precision rate is. For example, if image(i) satisfy the criteria of "Night Scenes"(s) of landscape, we consider images belongs to such scenes:

Example 1

IF $i = c1$ (ie $EV > 8$) THEN p = certain amount where p is the probability of $i \subseteq s$

Example 2

IF $(i = c1)\&(i = c2)$ (ie $EV > 8$ & $d > 50$) THEN p is higher than example 1

4.3 Semantic query manipulation

Back to the semantic query example:France's night scene in Summer, we may breakdown the query into three phases[7, 6], location, timestamp and also the image scenes. Location phase could be processed by GPS records in metadata of images. Timestamp phase could extracted and evaluated the season and the time of day based on the image timestamp. For the last image scenes phase, scenes could be classified base on the above image classification approach.

5 Experiments

In this section, our proposed approach is evaluated on a database of 882 images. After preparing images, metadata extraction and data cleaning are carried out. The classification system is able to extract the scenes of images from the image database.

5.1 Preparation and Data Cleaning

For our experiments we used images from one of the free photo album over the Internet. Images are downloaded ran-

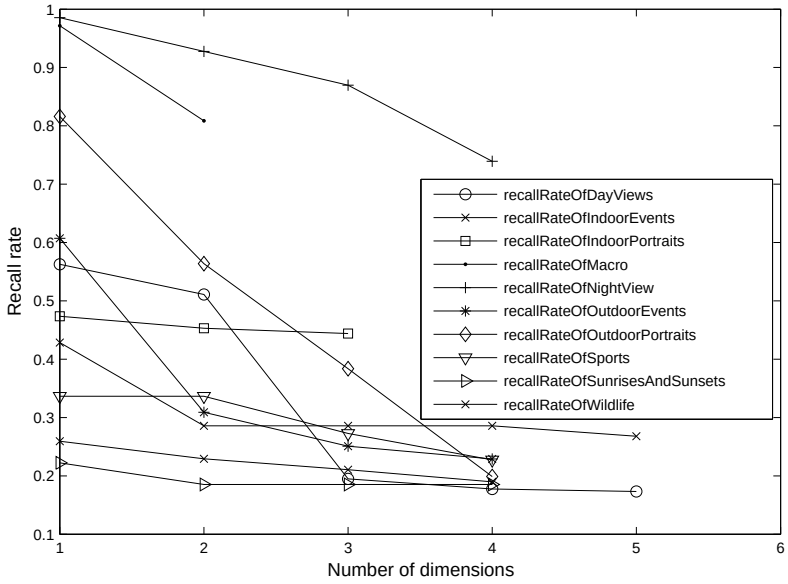


Figure 3. Recall rates of all sub-scenes

domly chosen without any bias. All images contained metadata such as brand and model of camera, aperture, exposure time etc. As Digital Single-Lens reflection (DSLR) camera are popular for most of professional photographers, we excluded images taken from consumer-level digital camera.

5.2 Experimental setup

A total of 8459 images are downloaded from the Internet. However, we focus only on images with metadata and exclude images taken from consumer-level digital camera and 3242 images satisfy our filtering Criteria .

Although standardized documentation [9] fully describe the use of metadata, most of the camera's manufacturers deliver information in their own way, especially in the tag of makernotes. For example, subject distance tag is available in the standard metadata; however, most of manufacturer store such information in their own makernotes tag. Data normalization will need to take place before the image classification process.

5.3 Precision and recall

The proportion of retrieved and relevant documents to all the documents retrieved[2]:

5.3.1 Precision

The proportion of retrieved and relevant documents to all the documents retrieved:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

In binary classification, precision is analogous to positive predictive value. Precision takes all retrieved documents into account. It can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision.

Note that the meaning and usage of "precision" in the field of Information Retrieval differs from the definition of accuracy and precision within other branches of science and technology.

5.3.2 Recall

The proportion of relevant documents that are retrieved, out of all relevant documents available[2]:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

In binary classification, recall is called sensitivity.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore recall alone is

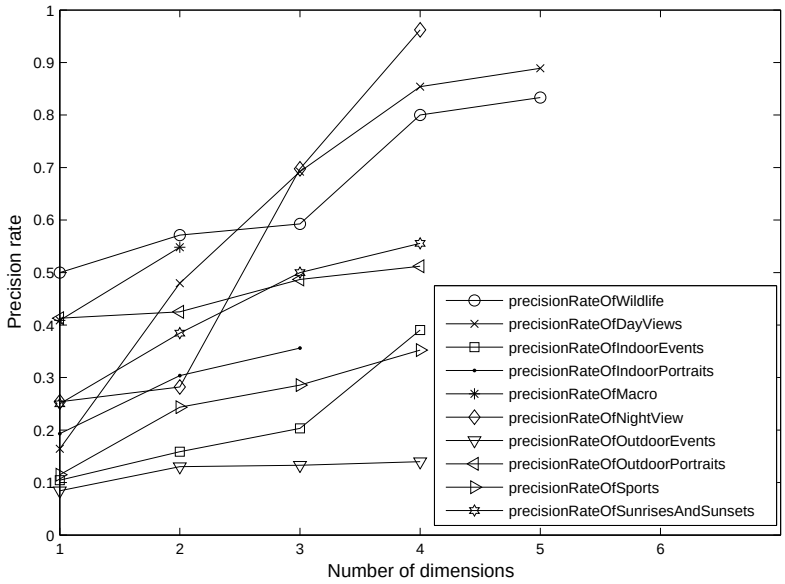


Figure 4. Precision rates of all sub-scenes

Scenes and sub-scenes	Number of images	True positive	False Positive	Recall rate	Precision rate
Landscape					
Night scenes	268	53	2	73.91%	96.23%
Day scenes	45	40	5	17.32%	88.89%
Sunrises and Sunsets	9	5	4	18.52%	55.56%
Portrait					
Outdoor portraits	414	212	202	19.89%	51.21%
Outdoor events	451	63	388	22.91%	13.97%
Indoor portraits	545	194	351	44.39%	35.60%
Indoor events	210	82	128	18.98%	39.05%
Sports	71	25	46	22.73%	35.21%
Nature					
Macro	208	114	94	80.85%	54.81%
Wildlife	18	15	3	26.79%	83.33%

Table 2. Detail experimental result

not enough but one needs to measure the number of non-relevant document also, for example by computing the precision.

5.4 Experimental Results

A better performance can be achieved by evaluating about the scene of images whether or not it could be classified correctly. In this way, the number of scenes can vary depending on the complexity of the image.

A total of 8459 images are downloaded from album and 3242 images included in this experiment. The experimental results are listed in the table 2. It is demonstrating the effectiveness of our scheme for determining the scenes of images.

Based on the results obtained using the proposed classification approach, where illustrate in figure 2,3, 4 and table 2,some of images fall into scenes while others are left untouched. Experience indicate that scenes of landscape and nature deliver better experimental results, while scenes of portraits offer lower precision and recall rate. We believe it is caused by similar image acquisition techniques used for different portrait sub-scenes. On the other hand, landscape and nature scenes always use an identical techniques and camera settings to benefit from certain environments.

6 Conclusions and future work

On the basis of the above reported approach and experiments, it is concluded that results of classification of proposed approach are competent and useful. Using our image classification approach we were able to use a simple algorithm to classify scenes of image from any unlabeled images. While the results presented in this paper are promising, they are preliminary and there is a lot of room for improvement. So far we used only a few basic scenes of images in the image classification while huge numbers of scenes present in the photographic world.

Currently, performance using text similarity is useful in the specialized topic databases. Adding more words to the text description will result a better matches. By using the proposed classification approach, additional index to images are built that enables better performance of image retrieval.

Our next goal is to build a system that combines between the proposed classification approach and the use of common sense reasoning and ontology techniques in order to deliver a better image retrieval system.

References

[1] T. Ang. *Dictionary of Photography and Digital Imaging, The Essential Reference for the Modern Photographer.* 2001.

[2] B. Baeza-Yates, R.; Ribeiro-Neto. *Modern Information Retrieval.* New York: ACM Press, Addison-Wesley, 1999.

[3] P. domain book. Basic photographic techniques pp 2-3, 2000.

[4] D. Foster. Dan foster, official gpx web site, http://www.topografix.com/gpx.asp.

[5] R. Lenman. *The Oxford Companion to the Photograph.* Oxford University Press, 2005.

[6] H. Liu and H. Lieberman. Robust photo retrieval using world semantics.in proceedings of the lrec 2002 workshop on creating and using semantics for information retrieval and filtering: State-of-the-art and future research, las palmas, canary islands, pp. 15-20.

[7] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge, 2003.

[8] S. F. Ray. *Camera Exposure Determination. In The Manual of Photography, 9th ed.* Oxford: Focal Press, 2000.

[9] B. L. Saux and G. Amato. Image recognition for digital libraries. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 91–98, New York, NY, USA, 2004. ACM Press.

[10] R. K. Srihari, Z. Zhang, and A. Rao. Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval*, 2(2/3):245–275, 2000.

[11] Y. Sun, S. Shimada, and M. Morimoto. Visual pattern discovery using web images. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 127–136, New York, NY, USA, 2006. ACM Press.

[12] Technical Standardization Committee on AV and IT Storage Systems and Equipment and Standard of Japan Electronics and Information Technology Industries Association. *Exchangeable image file format for digital still cameras: Exif Version 2.2*, jeita cp-3451 edition, April 2002.

[13] C.-F. Tsai, K. McGarry, and J. Tait. Claire: A modular support vector image indexing and classification system. *ACM Trans. Inf. Syst.*, 24(3):353–379, 2006.

A MODEL FOR SYSTEM REQUIREMENTS DEFINITION

Wu Di

Abstract

In this paper, we describe goals by two methods: identify goals by people, and abstract goals from process flows. Then we refine goals using constraints, which are picked up from goals, conflicts, functional requirements (FR) and non-functional requirements (NFR). At last, we could identify functionality of the system by getting final goals of the system.

1. Introduction

The requirements of an information system, which are usually understood as stating what a system is supposed to do,[11] are determined partly by its functionality, and partly by other non-functional requirements, such as operational costs, performance, reliability, and the like. [6] In this paper, we divided requirements definition into functional requirements definition and non-functional requirements definition accordingly.

Functional requirements definition describes the functionalities, which is the implemented function, of a system. "In computer science, functionality is a portion of code within a larger program, which performs a specific task and is relatively independent of the remaining code."[8] This indicates that if we could describe the functionalities of a system, we would get the tasks of the system. "Task is part of a set of actions which accomplish a job, problem or assignment." [8] So identifying tasks is equal to identifying the way to accomplish certain purpose. Goals are used to describe the "desirable states of the world". [7] We have illustrated this in Figure 1, which could generate the idea that using goals to formulate functions. Letier and Larnsweerde [5] said that "goals are intended outcomes to be achieved by the system under consideration". Also Anton pointed out that "goals are logical mechanism for identifying, organizing and justifying software requirements." [1] If we want to derive functions from goals, we should identify goals first. "Strategies are needed for the initial identification and construction of goals" [1] Although many researchers have done a lot on how to get goals of a system, however, they could not prove their methods could identify goals sufficiently. In this paper, we derive goals from two aspects--roles of people and process flows, which could compare with each other to improve the goal-sufficiency.

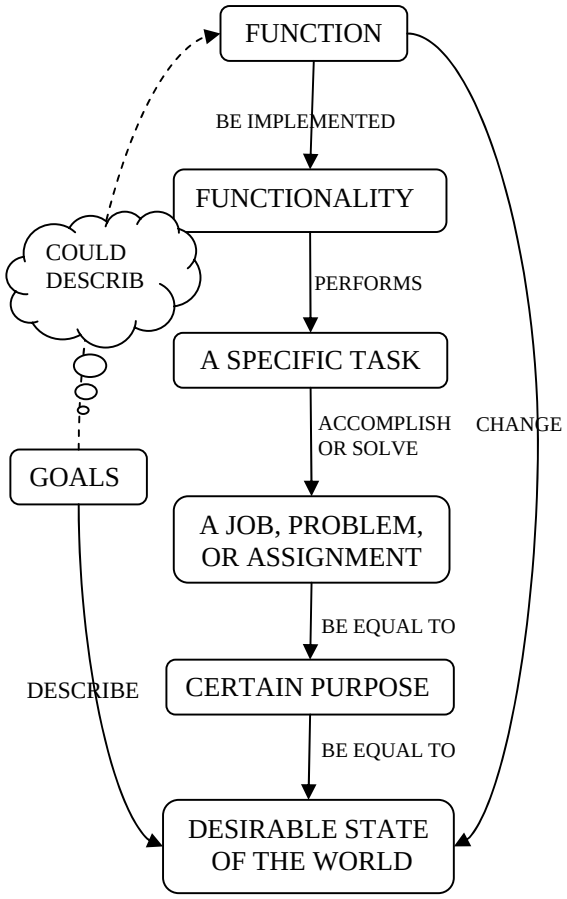


Figure 1 Goals could describe functionalities

2. Model Description

Figure 2 is the sketch map of our model

2.1 Get Goals

2.1.1 Get Goals from Role of People

The definition of a goal is "a broad statement of what the program hopes to accomplish." [10] For one program, different role of people has its own anticipation, which could not be listed one by one. In this paper, we divided roles of people into several groups according to the similarity of their goals.

CATWOE analysis and root definition in soft system methodology (SSM) [3] [4], which would be introduced

later, are applied for forming goal-groups. Three aspects-actors, owners, and clients or customers, have been identified.

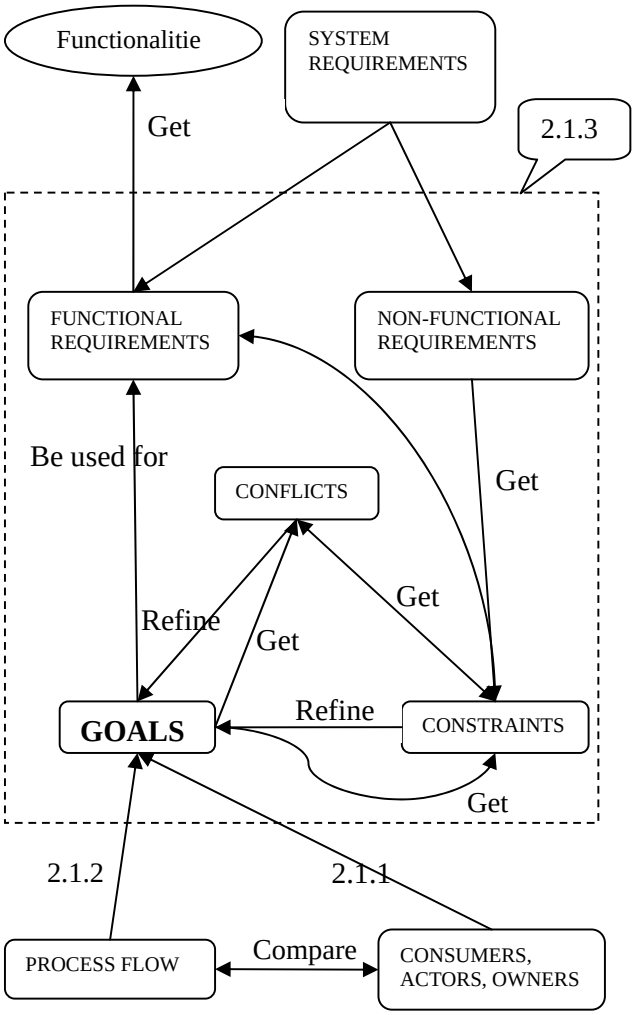


Figure 2 model of system requirements

2.1.1.1 Introduction to SSM

In soft system methodology (SSM), CATWOE analysis and root definition define the real world situation.

a) The CATWOE Analysis

In the CATWOE Analysis, we have primarily six components:

C Clients or Customers - Those who are benefited or are affected

A Actors - Those who carry out the system activities

T Transformation - Changes within or because of the system

W Worldview - How system is perceived, making T meaningful in context

O Owner-Those who could stop T

E Environment - The world that surrounds and influences the system, but the clients, actor and owner have no control over it. The environment is taken as given. [2]

b) Root definition

The root definition is a statement about the situation represented as a system. This definition is mainly formed by the results of the CATWOE analysis plus input and output components of the transformation. The following is a template of a root definition of a relevant system.

"An (...0...) owned system which, under the following environmental constraints which it takes as given; (...E...), transforms this input (...Input...) into this output (...Output..,) by means of the following major activities among others: (...T...), the transformation being carried out by these actors: (...A...) and directly affecting the following beneficiaries and/or victims (...C...), The world-image which makes the transformation meaningful contains at least the following elements among others: (.. .W...)" [2]

As CATWOE analysis and root definition could describe the real world, they should reflect every aspect of the real situation. It would be sufficient that we pick up goals of a system from clients or customers, actors, and owners.

2.1.2 Get Goals from process flow(s)

Sometimes, goals of some people who know how to do something, but do not know why they do it like that, are not clear. In this situation, it would be not sufficient if we only extract goals from roles of people. So getting goals from another way, which could help to exploit sufficient goals, becomes necessary. "Identifying goals from process descriptions by searching for statements which seem to guide design decisions at various levels within a system or organization" [1] would be accepted.

2.1.2.1 Steps for getting Goals from Process Flows [1]

- Searching for action words from process flow(s).
- Extract goals by describing the real situations using these action words.

2.1.3 Refine Goals

These initial goals would have some conflicts and the rationale issues, which could refine them. Constraints, which come from goal themselves, conflicts, and non-functional requirements, should be identified to refine the goals and solve conflicts. Functionalities are provided to satisfy these criteria. So constraints could also generate functionalities.

3. Case Study

Case Description

The Student Resident Hail is owned by ABC University. If a student wants to enjoy the hall life, he/she should apply the hall place at the starting of each semester. The staff of the Student Resident Hall then assign the available bed to the student according to his/her interest. The staff and student want to view the application results. [9]

According to our model, we have 4steps to generate goals of the system.

Step 1 Define goals from Client, Actor, and Owner. See Table 1. (2.1.1)

Aspect	Real World Character	Goals
Client	Student	1 Gets a bed according to his/her interest. 2 View the application result.
Actor	Staff of the Hall	3 Assign available bed to student according to his/her interest. 4 View the application result.
Owner	ABC University	5 Arrange students to the Hall

Table 1 Goals for people

Step 2 Define goals from process flow. Figure 3 is the process flow of Hall Resident System. (2.1.2)

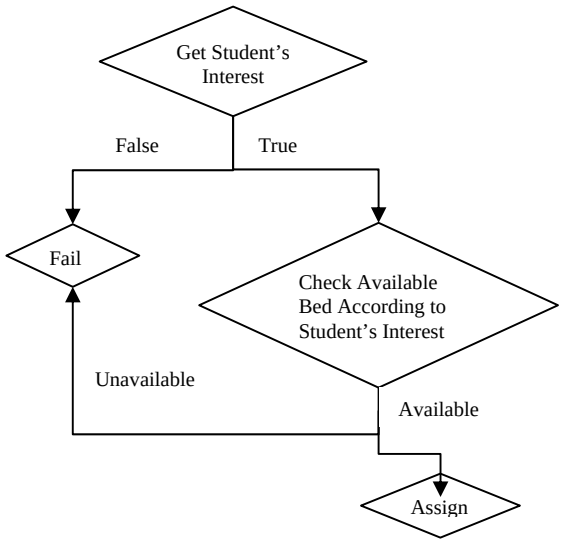


Figure 3 process flow of Hall Resident System

Action words from the process flow: Get, Check, and Assign. So we extract goals from the process flow:

6 get student's interest
7 check available bed according to student's interest
8 finish the arrangement

Step 3 Derive original Goals. (Compare)

After extracting the goals in step 1 and 2, we compare goals we have got, to remove the same goals in two steps. Finally, we get the original goals:

I. gets bed according to his/her interest (student)
II. assign available bed to student (staff)
III. view the application result (student and staff)
IV. arrange students to Hall (University)

Step 4 Refine goals (2.1.3)
Step 4.1 Get constraints

- Beds are limited.
- Students have to submit their requirements before a fixed date.
- If only one bed, which fits more than one student's requirements, is left, student who submitted his/her requirement earlier can get the bed.
- After the first round arrangement, if some beds are still available, and some students have not got beds, staff should make an announcement, and let students re-apply if they like.
- Students who apply for the Hall must be a full-time ABC University student.

Step 4.2 Get conflicts

From goals and constraints listed before, we could generate some conflicts:

- Student wants to get a bed according to his/ her interest, but there are not enough beds that could meet all the students' interests.
- Staff want to assign beds that are available to students, but students do not like that kinds of beds

We could solve these conflicts by using constraints.

Step 4.3 Get the final goals of the system.

After constraints have been listed and conflicts have been solved, the final goals of the system have been identified.

- Establish student's identity
- Get student's interest
- Check bed's availability
- Record student's requirement submit time
- Assign a bed to a student
- Make a bed unavailable after it is assigned
- Display application results
- Students who submitted their application after the deadline should not be considered

4. Conclusion and Future Work

In this paper, we construct a model that could get system functionalities from goals and constraints of the system. In order to make our goals more sufficiently, we get goals from two aspects: people and process flow. This is only the beginning of our work. Many problems, such as formulizing specific rules to identify constraints, how to refine goals, have not been proved richly. After we have got the functionality of a system, what should we do next? No suggestions have been given on this part.

References

[1] Annie I. Anton, "Goal-Based Requirements Analysis", Requirements Engineering, Proceddings of the Second International Conference, April 1996

[2] Chan S, "soft system methodology", lecture notes, Hong Kong Baptist University, 2005

[3] Checkland, P., "Soft systems methodology: an overview", J.Appl.Sys.Anal., **15**, 27-30., 1988

[4] Checkland, P., "Soft Systems Methodology: A 30-year Retrospective", John Wiley & Sons, Ltd, Chichester., 1999

[5] Emmanuel Letier and Axel van Lamsweerde, "Agent-Based Tactics for Goal-Oriented Requirements Elaboration", 24th International Conference on Software Engineering, ACM Press, May 2002

[6] John Mylopoulos, Lawrence Chung and Brian Nixon, "Representing and Using Non-Functional Requirements: A process-Oriented Approach", IEEE Transactions on Software Engineering, June 1992

[7] Shuichiro YAMAMOTO, Member, and etc., "Goal Oriented Requirements Engineering: Trends and Issues", IEICE TRANS. INF. & SYST., VOL.E89-D, NO.11, November 2006

[8] Web Dictionary, http://en.wikipedia.org/

[9] YEUNG Chung Kei, "Ontological Model for Information System Development Methodology", Mphil Thesis, Hong Kong Baptist University, Sept 2005

[10] Work Writing Tips, "The Difference between a Goal and an Objective", New Mexico State University, 2006

[11] Eric S. K. Yu, "Towards Modeling and Reasoning Support for Early- Phase Requirements Engineering", Proceedings of the Third IEEE International Symposium on, Jan 1997

Topic Detection Via Participation Using Markov Logic Network

Victor Cheng and C.H.Li

Abstract

The advent of Web2.0 enables the proliferation of online communities in which tremendous number of Internet users contribute and share enormous information. Proper exploitation of community structure help retrieving useful information and better understanding of their features. We employ Markov Logic Network to explore topic tracking by finding clusters, which represents latent topics, best fitting a set of rules. Rather than using contents in investigating discussions of a community, the user participation is used because it is believed that topics can be somehow reflected by the preferences of participation. User participation is also easier to process than text. The clustering results show this approach can reveal latent topics of a community effectively.

1. Introduction

Recent years have seen a great increased attention being given to online communities as the advent of Web2.0 grant internet users a dominant role in content contribution and sharing. For example, weblogs may be regarded as the most important representation for Web2.0. Also, Wikipedia [14], the internet encyclopedia, is not written by a small group of professionals but contributed by tremendous internet users all over the world. Study of these users centric phenomena can be beneficial to a better understanding of the psychology and sociology of the communities which is significant to their healthy development. Kollock [8] researched the motivations of users in online forums and Bishop [2] investigated the methods to encourage participation. Social network analysis [4] has also been used to analyze various online communities and study the relationship and roles of network users. Nolker and Zhou [10] applied social network theory to newsgroups to find out leader, motivators and chatters.

Apart from the human aspect, works on the content analysis of online communities are also conducted with the help of document categorization and clustering algorithms. Recently, topic detection [1] has become an increasingly important of research in information retrieval. It mainly deals with clustering of discussions of online communities and finds out the corresponding latent topics. In this paper, the online community to be studied is an online forum. A discussion here is referred as a forum thread which contains a sequences of messages/posts contributed by individual users. The messages of a thread should be highly related to its title for efficient and effective information retrieval. However, realistic situation is always not perfect, it is far from perfect. Since contents and the structure of online discussions are contributed by users, they may not be adequately placed and managed. Even the title of a discussion is sometime ambiguous. For example, the title "Sam the Record Man" in an Audio-visual forum is very unclear to anyone. Although many forums have guidelines for placing messages, a considerable of them are still misplaced. It is due to 1)users may make wrong judgement, 2)massages may be of multiple disciplines, 3)users even do not know the right place to delivery their messages, 4)users ignore the guidelines, etc. As a result, topic detection for online forums is necessary and beneficial to information retrieval.

Conventionally, clustering and categorization of text documents are usually done through the words appearing in the contents. In some systems such as systems with hypertext links, additional information can be incorporated and retrieval performance can be improved because it is believed linked objects are likely to be categorized as the same class, e.g. PageRank [11]. Cross-posting is also find useful in categorizing newsgroups [3]. A cross-posted message is a reflection that the author judges it should be interested to additional newsgroups and hence these groups should be similar, in the author's view. In addition to content categorization, collaborative filtering [12] has also been applied to newsgroups where ratings on news articles are predicted to help users locate and rank suitable news. While much success has been obtained in topic detection of news broadcast, progress in online forum has been affected by the imprecise, terse and causal communication styles. In this paper, we discuss the topic detection of a forum to be done solely with user participation. User participation can be employed in featuring a discussion because it is analogous to words, in sense of identifying a topic from the view of a forum. Since interests of a user is limited and seldom beyond dozens, the appearance of a specific user in a discussion indicates the

topic of the discussion is likely to be an interest of him. This is similar to the appearance of a specific word can give information about the topic. Some words (stop words) are very general and appear in nearly all discussions, and there are also some users they have interests in posting messages in many different discussions. Similar to words, synonymy and polysemy are also found among users.

Comparing to words, users are less specific. Number of meanings of a word is usually less than the number of interests of a user. This makes originally words are more preferable than participation in topic detection. Nevertheless, a closer look on contents of many messages reveals some problems associated with words. Firstly, many misspelled words, improperly used words, formal and informal idioms, and a range of various abbreviated words are found in almost all forums. These additional variations, sometimes can be regarded as document noise, degrade the quality of categorization. Secondly, mixed languages appearing in messages, e.g. Chinese and English, also requires additional and careful preprocessing. In contrast, participation information is relatively 'less noisy' and easier to process because it only involves in the collection of users' IDs and posting frequency in discussions. In fact, our results show that user participation can be another useful set of attributes in topic detection. It should be noted that this paper is not intended to show author participation is superior to word attributes but it can be another set of useful attributes.

In topic detection, clustering of discussions is usually done with conventional clustering algorithms such as K-means. Recent algorithms like spectral clustering or non-negative matrix factorization (NMF) [9] are also popular and widely used. These algorithms have a common feature that they cluster the objects based on the similarities or dissimilarities between object. A proper similarity/dissimiilarity measure is always helpful in getting high quality clustering. In this paper, a Markov Logic Network (MLN) [13] is employed in similarity evaluation. Rather than relying on attributes of individual objects, the relations between objects are used in evaluating similarities. Two discussions of a forum are considered to be similar if a large portion of users are common to each other. On the other hand, two users are considered similar to each other if they are found to post messages to similar discussions. MLN is a random network for studying statistical relational data. Syntactically, it can be regarded as first-order logic except that each formula has a weight attached. The advantage of using MLN is two fold. First, it can deal with different natures of information due to the expressive power of the first order logic. For example, attributes of objects and different kinds of relations among objects can be taken into account simultaneously. Secondly, even partial information can be incorporated into the network. An example is that we can incorporate similarity or dissimilarity information of a small amount of discussions of a forum into the MLN. Exploiting this partial information with conventional clustering algorithms may require considerable modification, while it is as simple as adding a formula to the MLN. Details of MLN will be given in Section 4. The remaining contents of this paper are organized as follow. Section 2 describes the Zipf's Law behavior of forums. Section 3, 4 is devoted to the introduction of Markov Networks and MLNs respectively, and a case study with a real web forum is presented in Section 5. Finally, the conclusion is given in Section 6.

—-

2. Zipf's Law Behavior of Forums

User participation is vital to the understanding of any online forums, and one important feature is the frequency of user participation. Similar to many self-organized networks, user participation of online forums often exhibits Zipf's Law behavior [15]. The Zipf's Law states that the appearing frequency of a set of words in a document corpus is inversely proportional to its rank in the frequency table. Figure 1 shows the plot (in log-log scale) of frequency of user participation versus the rank of users in the decreasingly sorted frequency table. From the plot an approximate straight line manifests user participation frequency nearly follows the Zipf's law. In other words, there is a small portion of active users posts significantly more than other users, and a majority portion of users post only rarely. In forums, not only users exhibit Zipf's law, distribution of number of messages in a discussion also has the behavior, see Figure 2. This means a small portion of discussions are very hot while a large amount of them have few messages only.

3. Markov Network

Before proceeding to Markov Logic Network (MLN) [13], this section gives an introduction on Markov network as MLN can be regarded an extension of Markov network with first order logic. A Markov network (also known as Markov random field) [5] is a model for the joint distribution of a set of random variables $X = (X_1, X_2, ..., X_n) \in \chi$. It is composed of an undirected graph G and a set of potential functions ϕ_k for cliques. A clique is a set of fully interconnected nodes, not necessary of maximum clique. Each node of G corresponds to a variable X_i and each clique of G has a potential function ϕ_k, which is a non-negative real-valued function of the state of that clique. With this configuration, the join distribution of X modeled by Markov network is given by

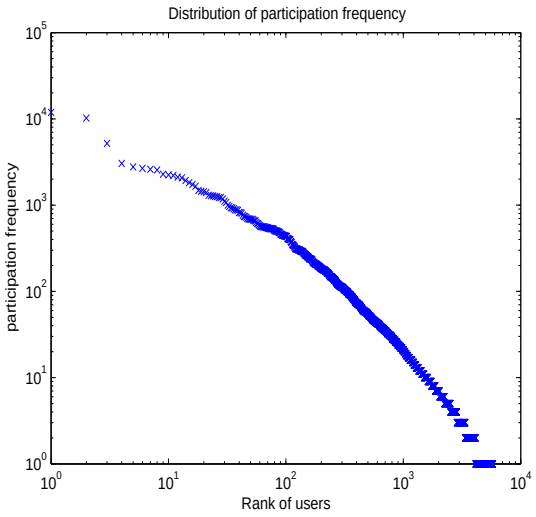


Figure 1. Distribution of participation frequency vs its rank.

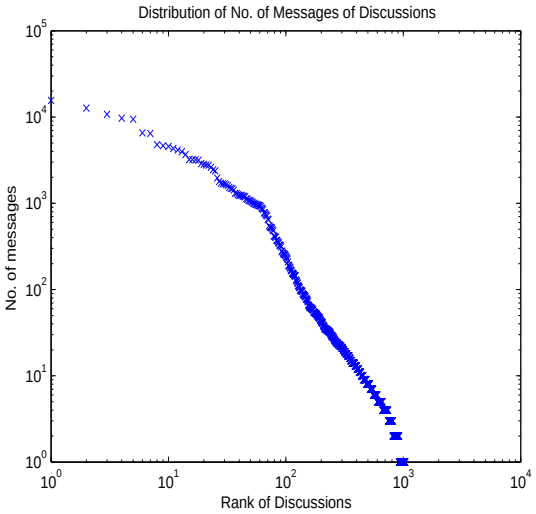


Figure 2. Distribution of No. of messages of discussion vs its rank.

$$P(X = x) = \frac{1}{Z} \prod_k \phi_k(x_{\{k\}}), \tag{1}$$

where $x_{\{k\}}$ is the state of the kth clique (i.e., the state of the variables that appear in the kth clique). Z is known as the partition function and it is given by

$$Z = \sum_{x \in \chi} \prod_k \phi_k(x_{\{k\}}). \tag{2}$$

Usually, Markov networks are represented as log-linear models for the sake of convenient manipulation, especially in cases of evaluating differentials of the networks. In that model, each clique potential is denoted by an exponentiated weighted sum of feature functions (or known as features) of the state. A feature can be any real-valued function of the state and hence the model equation is changed to

$$P(X = x) = \frac{1}{Z} exp(\sum_j w_j f_j(x)) \tag{3}$$

In MLN, $f_j(x)$ is constrained to be binary, $f_j(x) \in \{0, 1\}$. In the translation from the potential function form (1), there is one feature corresponding to each possible state $x_{\{k\}}$ of each clique, with its weight being $log\phi_k(x_{\{k\}})$. This representation is exponential in the size of the cliques. Nevertheless, this translation allows a much smaller number of features (e.g., logical function of the state of the clique) to be specified, and results in a more compact representation than the potential-function form, particularly when large cliques are present.

-

4. Markov Logic Networks

Traditional first-order Knowledge Base (KB) can be regarded as a set of hard constraints on possible worlds. If a world violates even one of them, the word has zero probability to exist. For MLNs, there are also a set of constraints, expressed in formula of first order logic. However, constraints in MLNs are softer: when a world violate any of them, the existence of the world is less probable, but not impossible. Each of the formula is associated with a scalar weight reflecting the importance or hardness of the formula. The probability of existence of a world depends on the weighted sum of the satisfied constraints. The world is more probable if the sum is higher. This is modeled with with (3) where each feature denotes a formula.

Definition 1 *[13] A Markov logic network L is a set of pairs (F_i, w_i), where F_i is a formula in first-order logic and w_i is a real number. Together with a finite set of constants $C = \{c_1, c_2, ..., c_{|C|}\}$, it defines a Markov network $M_{L,C}$ as follows:*

1. *$M_{L,C}$ contains one binary node for each possible grounding of each predicate appearing in L. The value of the node is 1 if the grounded predicate is true, and 0 otherwise.*

2. *$M_{L,C}$ contains one feature for each possible gounding of each formula F_i in L. The value this feature is 1 if the grounded formula is true, and 0 otherwise. The weight of the feature is the w_i associated with F_i in L.*

Thus there is an edge between two nodes of $M_{L,C}$ iff the corresponding grounded predicates appear together in at least one grounding of one formula in L. Under this formulation, an MLN can be viewed as a template for constructing Markov networks. From (3), the probability distribution over worlds x specified by the grounded Markov network $M_{L,C}$ is given by

$$P(X = x) = \frac{1}{Z} exp(\sum_{i=1}^{F} w_i n_i(x)), \quad (4)$$

where F is the number of formulas in the MLN and $n_i(x)$ is the number of true groundings of F_i in x. As formula weight increase, an MLN increasingly resembles a purely logical KB, becoming equivalent to one in the limit of all infinite weights. In this paper, MLN is applied to topic detection of a forum and it is assumed all formulas are function-free clauses and domain closure. This ensures the Markov networks generated are finite and the grounding of predicates and formula are formed simply by replacing the MLN's variable with typed constants in all possible ways. For example, if $C = \{Anna, Bob\}$ and the formula in a MLN is $\forall x$ Smokes(x) $\Rightarrow Cancer(x)$, the features of the Markov network by grounding of the formula is given by $Smoke(Anna) \Rightarrow Cancer(Anna)$ and $Smoke(Bob) \Rightarrow Cancer(Bob)$. See [13] for details.

Inference with MLNs can be done by inferring on the grounded Markov network, in two manners. By using maximum a posteriori (MAP) formulation, the most likely state of a set of query output variables given the state of a set of evidence variables can be founded theoretically, but the process is NP-hard. On the other hand, conditional inference involves computing the distribution of the query variables given the evidence, is #P-complete. For example, if F_1 and F_2 are two formulas, C is a finite set of constants including any constants that appear in F_1 or F_2 (after grounding), and L is an MLN, then

$$\begin{aligned} P(F_1|F_2, L, C) &= P(F_1|F_2, M_{L,C}) \\ &= \frac{\sum_{x \in \chi_{F_1} \cap \chi_{F_2}} P(X = x|M_{L,C})}{\sum_{x \in \chi_{F_2}} P(X = x|M_{L,C})} \quad (5) \end{aligned}$$

where χ_{F_i} is the set of worlds where F_i holds. Computing (5) directly will be intractable because this kind of inference subsumes probabilistic inference, which is #P-complete, and logical inference, which is NP-complete. One of the most widely used approximate algorithms to the evaluation is Markov chain Monte Carlo (MCMC) [6], and in particular Gibbs sampling, which proceeds by sampling each variable in turn given its Markov blanket, and counting the fraction of samples that each variable in each state. In the above example, $P(F_1|F_2, L, C)$ can be approximated using an MCMC algorithm that rejects all moves to states where F_2 does not hold, and counts the number of samples in which F_1 holds.

5. Results

This section describes the discussion clustering of a local popular web forum which provides a discussion cyberspace for people interested in Audio-visual affairs, in particular the high-end or high fidelity (Hi-Fi) equipment. To avoid any advertising effects, the alias **AVForum** is used. In this forum, three distinct discussion boards are available to public users with assigned alias **AvBoard**, **ChatBoard**, and **2ndHandBoard**. In **AvBoard** users are welcome to share their idea on Audio-visual affairs, **ChatBoard** provides a space for unbounded casual chats (except illegal affairs), and **2ndHandBoard** is a platform for people posting advertisement for buying or selling 2nd hand products. In this paper, only the discussions of **AvBoard** are considered and the others are ignored. As it is found that there is a considerable number of discussions having only few people participated, we believe the contents of them should be less important and sometimes some discussions can be regarded as noise or spam. In our experiments, every discussion with less than 8 distinct users are ignored, and users with the total number of messages posted to the forum board less than 50 are also ignored. The processed discussion board finally contains 844 and 703 distinct users and discussions, respectively.

In our study each discussion is represented by a 844 dimension participation frequency vector with individual components denoting the number of messages posted by the corresponding user. Hence, the **AvBoard** can be represented by a 703 x 844 matrix, denoted by D. With this discussion-user matrix, two different formulations of the similarity evaluation between any two discussions are done and compared. In the first setup, discussion-discussion similarity is formed just by inner product of discussion vectors, thus the similarity matrix S is obtained by

$$S = DD^T. \quad (6)$$

The second formulation is done by employing an MLN which evaluates similarity by specifying logic formula to explore relationships among discussions and users. The formula used is depicted in Table 1, which mainly asserts that two discussions are similar if they have similar users. The numeric values before the formula are the weights.

After the similarity matrix for each setup is evaluated, a public domain available clustering tool CLUTO [7] is employed to cluster the discussions into 10 clusters. Discussions of the forum are clustered by maximizing the cost

Table 1. MLN formula for inferring discussion similarity

Predicates Definition:	
user(u)	// u is a user.
discussion(d)	// d is a discussion.
u_join_d(u,d)	// u join the discussion d.
similar_topic(d,d)	//two discussions have a similar topic.
MLN Formula (with weights at the beginning):	
0.1 u_join_d(u1,d1) $\wedge$ u_join_d(u1,d2) $\wedge$ similar_topic(d1,d2)	
0.04 similar_topic(d1,d2) $\wedge$ u_join_d(u1,d1) $\Rightarrow$ u_join_d(u1,d2)	
0.04 similar_topic(d1,d2) $\wedge$ u_join_d(u1,d2) $\Rightarrow$ u_join_d(u1,d1)	

function

$$maxmize \sum_{i=1}^{k} \sqrt{\sum_{u,v \in C_i} sim(u,v)}. \quad (7)$$

where $sim(u, v)$ is the similarity between two discussions u and v, and C_i is a cluster with index i. Two transformed views of the clustering solutions in 3-dimension are shown in Figure 3 and 4, for MLN and inner product similarity formulation respectively. By comparing the figures, it is found that the sizes of clusters formed from inner product formulation do not differ by much. On the other hand, the cluster sizes for the MLN formulation have two extremes. Some clusters are very small in size but there is a very big cluster found, the cluster identified with the number '7'. Since the two clustering solutions are obtained from different similarity formulations, direct comparison with a common measure in clustering quality, such as entropy of individual clusters, is not practical.

A manual study on the discussion contents of the compact clusters obtained from the MLN formulation shows that the discussions within individual clusters are highly related. For the inner product formulation, even some clusters have a dominant discussion topic, there is a considerable number of unrelated and diverse discussions found. As regards the clustering solution for the MLN formulation, the latent topics of the 10 clusters after human evaluation are shown in Table 2. The latent topics for clusters obtained from inner product formulation is also studied, however, it is much difficult to identify the topics as the clusters have very diverse topics. Anyway, some topics are still found a bit dominant in clusters, e.g. Fans club, Vintage Equipment, DIY affairs, etc.

In **AvBoard**, users are provided with a drop-down list to category their discussions. However, it is found that many of them are wrongly placed. Another problem with the list is it is not comprehensive enough and there may be no appropriate categories for discussions. For example, there is no DIY category for users discussing do-it-yourself

Table 2. Latent topics discovered in the AVBoard.

Id	Latent Topic	No. of Discussions
C_0	Fans clubs of a name brand speaker	8
C_1	Fans clubs of two name brands	10
C_2	Vintage Equipment	13
C_3	Hi-Fi Passion sharing	2
C_4	DIY affairs	35
C_5	Equipment recommendation	89
C_6	CD, DVD, LCD TV, CD/DVD Equipment	83
C_7	Miscellaneous	339
C_8	CD players technical affairs, wires	81
C_9	Miscellaneous fans clubs	43

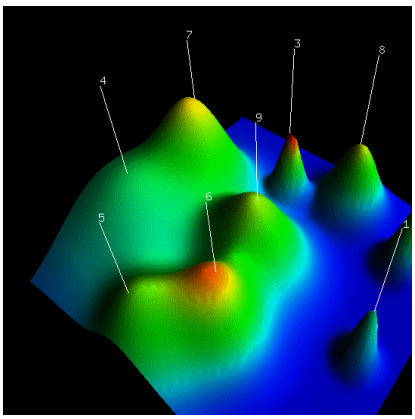


Figure 3. Transformed 3-D view of the clustering solution for the similarity matrix formulated by using MLN.

affairs, a hot hobby. Although there are $general$ and $others$ categories to cater for the situation, it makes the situation worse. Many users just choose one of them regardless of the topics of discussion. Even there is a more suitable category in the list, they still place their discussion in $general$. As a result, we believe that topic detection with MLN is helpful in information retrieval. Finally, it is worth mentioning some interesting discoveries by referring to Figure 3. Cluster C_0, with latent topic "Fans club of a name brand speaker", is quite compact and isolated from other clusters. This is because the contents are specific to a name brand and less related to other Hi-Fi affairs, and only users interested in this name brand speaker will join the discussions. On the other hand, the clusters C_5, C_6 are neighboring with some overlaps. This match the intuitive fact that equipment recommendation and TV/CD/DVD equipment discussions are highly related.

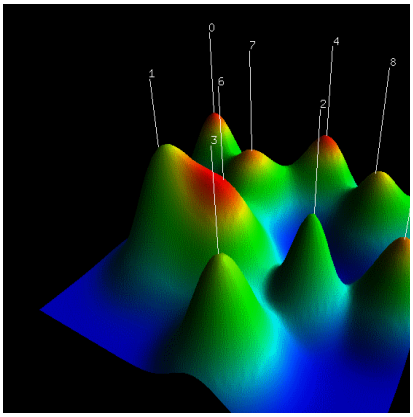


Figure 4. Transformed 3-D view of the clustering solution for the similarity matrix formulated by inner product of discussions.

6 Conclusion

User participation of online forum can be a useful information in topic detection. Comparing with text information, it has the advantages of 'less noisy' and easier to process. In this paper, we use two different formulations, inner product and MLN, to form the discussion-discussion similarity matrix based on user participation. Topic detection is then to be done through clustering and followed by interpreting the solutions. By manual investigation, it is found that similarity measure with MLN formulation is more effective in revealing latent topics. The clusters formed is less noisy and latent topics can be easily identified. It should be noted the degree of effectiveness can be dependent on the nature of the online forum. For example, if an online forum is highly dominated by chatters, it may become difficult to discover latent topics from user participation. Furthermore, this approach is still bounded by a traditional clustering problem, the number of clusters. Too many clusters may split related discussions into separate clusters. On the other hand, too less clusters may merge unrelated discussions together.

References

[1] J. Allan, editor. *Topic Detection and Tracking: Event based Inforamtion Organizaton*. Kluwer Academic Publishers, 2000.

[2] J. Bishop. Increasing participation in online communities: A framework for human-computer interaction. *Comput. Hum. Behav*. 23(4):pages 1881-1893, 2007.

[3] C. Borgs, J. Chayes, M. Mahdian, and A. Saberi. Exploring the community structure of newsgroups. *KDD2004*, pages 783-787.

[4] P.J. Carrington, J. Scott, and S. Wasserman, editors. *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005.

[5] S. Della Pietra, V. Della Pietra, and J. Laffery. Inducing features of random fields. *IEEE Transactions on Pattern analysis and Machine Intelligience*. 19:pages 380-392, 1997.

[6] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. London, UK, Chapman and Hall.

[7] G. Karypis. CLUTO a clustering toolkit. Technical Report 02-017, Dept. of Computer Science, University of Minnesota. Available at http://www.cs.umn.edu/ cluto.

[8] P. Kollock. The economies of online cooperation: Gifts and public goods in cyberspace. In M. Smith and P. Kollock, editors, *Communities in Cyberspace*. Routledge, London, 1999.

[9] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556-562, 2000.

[10] R.D. Nolker and L. zhou. Social computing and weighting to identify member roles in online communities. In *WI'05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 87-93. IEEE Computer Society, 2005.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project. 1998.

[12] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175-186, Chapel Hill, North Carolina, 1994.

[13] M. Richardson, P. Domingos. Markov logic networks. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, Wa.

[14] Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Main_Page.

[15] G.K. Zipf. *The Psychobiology of Language*. Houghton-Mifflin, New York, 1935.

Iterative Feature Selection in Gaussian Mixture Clustering with Automatic Model Selection

Hong Zeng and Yiu-ming Cheung

Abstract

In this paper, we propose a new feature selection method which iterates between a clustering algorithm and two cascade feature selection schemes. In the feature selection phase, not only the most relevant features to the clustering are identified, but also the redundant features among the selected relevant features are removed. Besides, it does not need to particularly assume the explicit parametric form for the probability density function of irrelevant features, thus is more tolerant to the situation when the hypothesis of this form is violated. We experimentally demonstrate the gain in performance our method achieves, using both synthetic and real-world data sets.

1. Introduction

Gaussian mixture (GM) clustering has been widely applied to a variety of fields including data mining, time series forecasting, image processing, and so forth. In general, in order to partition a given data set, GM clustering needs to conduct the model selection, i.e. determine the number of components in a mixture (also called *model order* interchangeably), and estimate the parameters of each component, through the observed data represented as a vector of features (also referred to as attributes, variables or measurements). Unfortunately, among the features simultaneously measured, it is unlikely that all of them will contribute to the grouping task. That is, there may be some irrelevant features in the observations. Under this circumstance, the inclusion of such features could hinder the clustering algorithm to detect the grouping characteristics of data. Furthermore, among the relevant features, some might be redundant as they carry no additional partitioning information beyond that subsumed by the remaining ones. And according to the well-known Occam's Razor principle, given two feature sets of different cardinality that result in the same partition, the bigger of the two is expected to result in a worse predictive performance of the learned model.

In order to obtain an appropriate partition and a subset of the most informative features, the feature selection scheme, which is to identify the features that significantly contribute to the grouping, is often required. It will also brings on other potential benefits, including: reducing the collection and storage requirements, improving the comprehensibility of the resulting partition, etc.. However, in the unsupervised learning scenario, due to the absence of the ground-truth labels that could guide the assessment of the relevance and redundancy for each feature as the reference information, it is a nontrivial task to perform the feature selection. The problem becomes even more challenging when the true number of clusters is unknown *a priori*, what is more, the optimal feature subset and the optimal number of clusters are inter-related: different clustering results might be obtained on different feature subsets.

In the literature, there have been several representative methods that address the issue of the feature selection for the clustering. In the approaches [4, 12], features are typically chosen prior to a clustering algorithm based on the general characteristics of data. Though they significantly reduce the dimensionality, they are successful only to a limited extent, since these selected features might not be necessarily well suited to the mining algorithm [10]. This suggests that the feature selection, should be taken into account jointly with the clustering. Thus, trying to obtain both optima for these two tasks, some approaches, e.g. see [6, 7], wrap the feature selection around the clustering algorithm by first conducting a combinatorial search for candidate subsets in the whole feature space, then evaluating these subsets using the clustering algorithm. Subsequently, the best subset is chosen using a certain criterion during the repeated wrapping around. This kind of approaches may suffer from a heavy computational burden with the time-consuming combinatorial search strategy and the repeated execution of the clustering algorithms. Recently, the approaches in [10, 3] have proposed to tackle these two tasks in a single optimization paradigm. Several preliminary experiments in [10, 3] have shown a promising performance. Nevertheless, such methods suppose that the explicit parametric form of the pdf's for irrelevant features are known *a priori*, which may be impossible from the practical point of

view. Moreover, few of the mentioned methods have considered of selecting the non-redundant relevant features, until recently a few works [17, 18, 8] have addressed this issue yet for supervised learning problems. They have shown that selecting the relevant features may be suboptimal for obtaining a model with considerably good generalization, in case the features are redundant [8].

In this paper, we propose an algorithm geared toward the goal of seeking the smallest feature subset that best represents the partitions of interest. It iterates between clustering and feature selection, in a mutually reinforcing optimization manner. For the clustering algorithm, we adopt an efficient method called the Rival Penalized EM algorithm [2], which is able to determine the number of components automatically and simultaneously with the parameters estimation. In the feature selection phase, not only the feature relevancy analysis is conducted through ranking the features according to a proposed relevance evaluation index, but also the redundant features are removed by applying the Markov Blanket filtering originally utilized in supervised scenarios [9, 17]. Besides, it does not have to assume the explicit parametric form of irrelevant feature distribution, thus is more robust when the assumption is violated.

The remainder of the paper is organized as follows. Section 2 overviews the RPEM algorithm. The proposed feature selection schemes are described in Section 3. Then, Section 4 presents the proposed algorithm in detail, and Section 5 shows the experimental results. Finally, we draw conclusions in Section 6.

2. The clustering algorithm

2.1. Preliminaries

Suppose that the observation data set $\mathbf{X_N} = \{\mathbf{x_1}, \mathbf{x_2} \ldots \mathbf{x_N}\}$ is generated from a mixture of k^* Gaussian components, i.e.,

$$p(\mathbf{x}_t|\Theta^*) = \sum_{j=1}^{k^*} \alpha_j^* p(\mathbf{x}_t|\theta_j^*) \tag{1}$$

with

$$\sum_{j=1}^{k^*} \alpha_j^* = 1 \quad \text{and} \quad \forall 1 \le j \le k^*, \quad \alpha_j^* > 0,$$

where each observation $\mathbf{x_t}(1 \le t \le N)$ is a vector of d-dimensional features: $[x_{1t}, \ldots, x_{dt}]^T$. Furthermore, $p(\mathbf{x}_t|\theta_j^*)$ is the j^{th} Gaussian component with the parameter $\theta_j^* = \{\mu_j^*, \mathbf{\Sigma}_j^*\}$, μ_j^* and $\mathbf{\Sigma}_j^*$ representing the center and covariance of the j^{th} component respectively. α_j^* represents the true mixing coefficient of the j^{th} component in the mixture. The main task of GM clustering analysis is to find an estimate of $\Theta^* = \{\alpha_j^*, \theta_j^*\}_{j=1}^{k^*}$, denoted as $\Theta = \{\alpha_j, \theta_j\}_{j=1}^{k}$, from N observations. A general approach is to search a set of parameters which could reach a maxima of the fitness in terms of *maximum likelihood* (ML) defined below:

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} \{\log p(\mathbf{X_N}|\Theta)\}.$$

When the number of components k is known, the ML estimate of the model parameters could be obtained by the EM algorithm [5, 11, 1]. In most cases, the model order k is often unknown *a priori*, and the maximized likelihood $p(\mathbf{X_N}|\Theta)$ is a nondecreasing function of k. Although some classical model selection criteria, e.g. see [15, 16], have been proposed to penalize this likelihood, they all require to first estimate the parameters for the candidate models, then compare all the candidates in order to determine the optimal model order, which may be somewhat inefficient. Recently, an approach called *Rival Penalized EM* (RPEM for short) [2] has been proposed, by which the order is determined simultaneously with the parameters estimation.

2.2. The rival penalized EM algorithm

RPEM algorithm introduces unequal weights into the conventional likelihood as regularization terms, thus the weighted likelihood is written below:

$$\begin{aligned} Q(\Theta, \mathbf{X}_N) &= \frac{1}{N} \sum_{t=1}^{N} \log p(\mathbf{x}_t|\Theta) \\ &= \frac{1}{N\zeta} \sum_{t=1}^{N} \sum_{j=1}^{k} g(j|\mathbf{x}_t, \Theta) \log p(\mathbf{x}_t|\Theta) \\ &= \frac{1}{N\zeta} \sum_{t=1}^{N} \mathcal{M}(\Theta, \mathbf{x}_t) \end{aligned} \tag{2}$$

$$\begin{aligned} \mathcal{M}(\Theta, \mathbf{x}_t) &= \sum_{j=1}^{k} g(j|\mathbf{x}_t, \Theta) \log[\alpha_j p(\mathbf{x}_t|\theta_j)] \\ &\quad - \sum_{j=1}^{k} g(j|\mathbf{x}_t, \Theta) \log h(j|\mathbf{x}_t, \Theta) \end{aligned} \tag{3}$$

where

$$h(j|\mathbf{x}_t, \Theta) = \frac{\alpha_j p(\mathbf{x}_t|\theta_j)}{p(\mathbf{x}_t|\Theta)}$$

is the posterior probability that $\mathbf{x}_t$ belongs to the j^{th} component in the mixture, and k is greater than or equal to k^*. $g(j|\mathbf{x}_t, \Theta)$'s are designable weight functions, satisfying the constraints below:

$$\sum_{j=1}^{k} g(j|\mathbf{x}_t, \Theta) = \zeta, 1 \le t \le N,$$

and

$$\forall j, g(j|\mathbf{x}_t, \Theta) = 0 \quad \text{if} \quad h(j|\mathbf{x}_t, \Theta) = 0,$$

where ζ is a positive constant. In [2], they are constructed from the following equation (with $\zeta = 1$):

$$g(j|\mathbf{x}_t, \Theta) = (1 + \varepsilon_t) I(j|\mathbf{x}_t, \Theta) - \varepsilon_t h(j|\mathbf{x}_t, \Theta)$$

with

$$I(j|\mathbf{x}, \Theta) = \begin{cases} 1 & \text{if } j = c \equiv \arg\max_{1 \le i \le k} h(i|\mathbf{x}, \Theta); \\ 0 & j = r \neq c. \end{cases} \tag{4}$$

and ε_t is a small positive quantity. This construction of weight functions reflects the pruning scheme: when a sample $\mathbf{x}_t$ comes from a component that indeed exists in the mixture, the value of $h(j|\mathbf{x}_t, \Theta)$ is likely to be the greatest, thus this component will be the winner. Accordingly, a positive weight $g(c|\mathbf{x}_t, \Theta)$ will keep it in the temporary model. In contrast, all other components fail in the competition and are treated as the 'pseudo-components". As a result, the negative weights are assigned to them as a penalty. Over the learning process of Θ, only the genuine clusters will survive, whereas the "pseudo-clusters" will get gradually faded out from the mixture.

The RPEM gives an estimate of Θ^* via maximizing weighted likelihood (MWL) in (2), i.e.,

$$\hat{\Theta}_{MWL} = \arg\max_{\Theta} \{Q(\Theta, \mathbf{X}_N)\}.$$

The more detailed implementation of the RPEM can be found in [2].

3. Unsupervised feature selection

3.1. Selecting the relevant features

A feature should be irrelevant to clustering if the clusters are indistinguishable each other when the observations are projected onto this feature. To illustrate this scenario, we show an example in Figure 1. If we project the two clusters onto the Y axis, it is unable to distinguish these two clusters by the feature Y, because the observations from the two clusters are almost projected onto the same dense region of this dimension. Hence, the feature Y will not be helpful in finding the cluster structure, i.e., it is irrelevant for the clustering. On the contrary, the projections onto the X axis can provide the useful information regarding the cluster structure, thus the feature X is relevant for the clustering.

Based on the above example, we argue that a feature with dense regions in the cluster may not be relevant if its global distribution is also dense. Therefore, we propose the following quantitative index to measure the relevance of each feature:

$$SCORE_l = \frac{1}{k} \sum_{j=1}^{k} Score_{l,j} = \frac{1}{k} \sum_{j=1}^{k} (1 - \frac{\delta_{l,j}^2}{\delta_l^2}),\ l = 1, \ldots, d$$

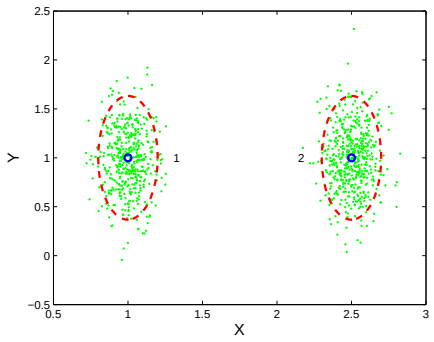


Figure 1. The feature X is relevant to the partitioning, while the feature Y is irrelevant.

where k is the number of clusters, $\delta_{l,j}^2$ is the variance of the j^{th} cluster projected on the l^{th} dimension:

$$\delta_{l,j}^2 = \frac{1}{N_j - 1} \sum_{t=1}^{N_j} (x_{l,t} - \mu_{l,j})^2, \mathbf{x}_t \in j^{\text{th}} cluster,$$

$N_j = \sum_{t=1}^{N} I(j|\mathbf{x}_t, \Theta)$ is the number of data in the j^{th} cluster, and $\sum_{j=1}^{k} N_j = N$. δ_l^2 is the global variance of the whole data on the l^{th} dimension:

$$\delta_l^2 = \frac{1}{N - 1} \sum_{t=1}^{N} (x_{l,t} - \bar{\mu}_l)^2, \bar{\mu}_l = \frac{1}{N} \sum_{t=1}^{N} x_{l,t}.$$

With its form, the $Score_{l,j}$ could be utilized to indicate the relevance of the l^{th} feature with respect to the j^{th} cluster. Thus, the average relevance of the l^{th} feature for the clustering could be represented by the $SCORE_l$. When the $SCORE_l$ receives a value close to the maximum value (i.e. 1), it approximately indicates that all the local variances of the k clusters on this dimension are considerably small in comparison to the global variance of this dimension, which is tantamount to indicating these clusters far away from each other on this dimension. Hence, this feature is very relevant to the partitioning task. Otherwise, the $SCORE_l$ will receive the score close to the minimum value (i.e. 0). To prevent the index from being degenerated in the situation where $\delta_{l,j}^2 > \delta_l^2$, indicating the feature is no more relevant to the j^{th} cluster than to a random sample of data, we should definitely clip the $Score_{l,j}$ at 0.

According to the score of each feature, we could obtain the refined relevant feature subset $R^{'}$ in the following way:

$$R^{'} = F - \{F_l | SCORE_l < \beta, F_l \in F\}$$

where F is the full feature subset, and β is a user-defined threshold value (in general$\beta \in [0, 0.5]$). $R^{'}$ will be input to the succeeding redundancy analysis procedure.

3.2. Selecting the non-redundant features

From the information theory viewpoint, a feature is redundant if it carries the same partition information as that subsumed by the remaining features. Therefore we are able to neglect it without compromising the accuracy of prediction. We now introduce the definition of a feature's Markov Blanket, given by [14], which formulates this idea in supervised scenarios.

Definition 1 (Markov Blanket)**.** *Given a feature F_l, let $M_l \subset F$ ($F_l \notin M_l$), M_l is said to be the Markov Blanket for F_l if:*

$$P(F - M_l - F_l, C|F_l, M_l) = P(F - M_l - F_l, C|M_l),$$

where C is the class label.

Thereby, if a Markov Blanket M_l for F_l can be found in the feature set F, i.e. M_l subsumes the information that F_l has about C, we are able to eliminate the feature F_l from F without affecting the class prediction accuracy.

Since there might not be a full Markov Blanket for a feature, [9] proposed a method which sequentially eliminates such features based on the existence/non-existence of an *approximate* Markov Blanket in the candidate features subset. Broadly, it iteratively constructs a candidate Markov Blanket M_l for F_l, and measures how close M_l is to being a true Markov Blanket for F_l; if M_l is the closest to being a Markov Blanket for F_l, F_l is eliminated, and the algorithm repeats. The closeness of M_l to being a true Markov Blanket for F_l is measured by the expected cross entropy:

$$\Delta(F_l|M_l) = \sum_{f_{M_l}, f_l} P(M_l = f_{M_l}, F_l = f_l) \cdot KL(P(C|M_l = f_{M_l}, F_l = f_l) \parallel P(C|M_l = f_{M_l})) \quad (5)$$

where $KL(. \parallel .)$ denotes the Kullback-Leibler divergence: $KL(P \parallel Q) = \sum_x P(x) \log(P(x)/Q(x))$.

If M_l is a genuine Markov Blanket for F_l, then $\Delta(F_l|M_l) = 0$. An approximate Markov Blanket is formulated by relaxing this requirement with $\Delta(F_l|M_l)$ being very small. The candidate Markov Blanket is constructed by picking up top T features that have the highest Pearson correlation to F_l, where T (the size of Markov Blanket) is often a small integer. The reason of formulating the candidate Markov Blanket in this way is that the features in F_l's Markov Blanket M_l is directly influenced by F_l, while other features are conditionally independent of it given M_l. As the expected cross entropy requires to compute the posterior probability $P(C|.)$, for the computational consideration, it will be convenient to utilize the binary values of original feature values. An applicable discretization [1] method can be found in [17]. The complete Markov Blanket filtering algorithm in [9, 17] is presented in Algorithm 1.

[1] The discretized feature is only used for computing the KL divergence, the Person correlation is still calculated with the original feature values

Algorithm 1: The Markov Blanket filtering algorithm.

Initialize

- $G^{(1)} = F$;

Iterate

- For each feature $F_l \in G^{(m)}$ let M_l be the set of T features $F_i \in G^{(m)} - F_l$ for which the correlation between F_l and F_i are the highest;
- Compute $\Delta(F_l|M_l)$ for each feature l;
- Choose the F_{l_m} that minimizes $\Delta(F_l|M_l)$, and define $G^{(m+1)} = G^{(m)} - F_{l_m}$;

Until $|G^{(m+1)}| = T$.

The order in which the features get sequentially removed by this method, $\{l_1, l_2, \ldots, l_{|F|-T}\}$, corresponds a feature ranking of increasing non-redundancy. That is the feature that appears first in the list (i.e. F_{l_1}, the first one that has been removed from F) is the most redundant among all the features, while the features left after Markov Blanket filtering algorithm has stopped (i.e. $\{F - \{F_{l_1}, F_{l_2}, \ldots, F_{l_{|F|-T}}\}\}$), are the least redundant.

As addressed in the beginning of the subsection 3.2, the Markov Blanket requires class labels, which are not available in clustering problems. To circumvent this, we assume that a set of clusters can be modeled as being a set of different classes. Since we could image that $\min_{F_l \in G^{(m)}} \Delta(F_l|M_l)$, the minimum value of expected cross entropy in the m^{th} iteration, will also increase with m, thereby the most non-redundant features could be simply obtained by:

$$R'' = \{F_{l_m} | \min_{F_l \in G^{(m)}} \Delta(F_l|M_l) > \gamma \cdot \min_{F_l \in G^{(1)}} \Delta(F_l|M_l)\} \cup \{R' - \{F_{l_1}, F_{l_2}, \ldots, F_{l_{|R'|-T}}\}\}$$

where $m = 1, 2, \ldots, |R'| - T$, $F_{l_m} \in R'$, $G^{(1)} = R'$, and γ is a user-defined threshold (e.g., 2).

4. The iterative feature selection and clustering algorithm

So far we have described the two modules for our proposed algorithm: the basic clustering algorithm and the feature selection strategy based on the relevancy and redundancy analysis. Since the optimal number of clusters and the optimal features subset are inter-related, we integrate the feature selection scheme of Section 3 into the RPEM algorithm, and let thw two modules work in an iterative way. Specifically, at the end of each epoch of the RPEM algorithm, the approximately optimal partition can be obtained

on a given feature space. Then the proposed feature selection scheme outputs a refined feature subset in terms of the relevance and non-redundance with respect to this *reference* partition, i.e., the current data partition. Subsequently, a more accurate partition will be performed in the next epoch using the chosen feature set in the current epoch. Algorithm 2 presents the details of the proposed algorithm.

Algorithm 2: Iterative Feature Selection in RPEM clustering algorithm.

input : $\mathbf{X}_N, k_{max}, \eta, epoch_{max}, \beta, \gamma, T$
output: *the most relevant and non-redundant feature subset* $\hat{R}$

$\hat{R} \leftarrow \{F\}$;
$epoch\_count \leftarrow 0$;
while $epoch\_count \leq epoch_{max}$ **do**
  for $t \leftarrow 1$ **to** N **do**
    Step 1: *Calculate* $h(j|\mathbf{x}_t, \hat{\Theta})$*'s to obtain* $g(j|\mathbf{x}_t, \hat{\Theta})$*'s on* $\hat{R}$*;*
    Step 2: *Update parameters* $\hat{\Theta}$ *on* F*;*
    $\hat{\Theta}^{(new)} = \hat{\Theta}^{(old)} + \eta \left.\frac{\partial \mathcal{M}(\mathbf{x}_t;\hat{\Theta})}{\partial \Theta}\right|_{\hat{\Theta}^{(old)}}$;
  end
  $\hat{R} \leftarrow$ `FeatureSelection`(F, β, γ, T);
  $epoch\_count \leftarrow epoch\_count + 1$;
end

In the above algorithm, the weight function $g(j|\mathbf{x}_t, \Theta)$'s are designed as:

$$g(j|\mathbf{x}_t, \Theta) = I(j|\mathbf{x}_t, \Theta) + h(j|\mathbf{x}_t, \Theta), j = 1, \ldots, k_{max}$$

where the $I(j|\mathbf{x}_t, \Theta)$ is defined by (4). It is easy to verify that the above design still satisfies the required constraints on the $g(j|\mathbf{x}_t, \Theta)$. Obviously, such a design gives the winning component only, i.e., the c^{th} component, at each time step an extra award whose value is $I(c|\mathbf{x}_t, \Theta) = 1$. This weight design actually penalizes those rival components in an implicit way. Consequently, it is able to automatically determine an appropriate number of components as well.

Since the RPEM algorithm is able to prune the redundant components, the relevance score calculation in each epoch should be therefore adjusted as:

$$SCORE_l = \frac{1}{k_{nz}} \sum_{j=1}^{k_{nz}} Score_{l,j} = \frac{1}{k_{nz}} \sum_{j=1}^{k_{nz}} (1 - \frac{\delta_{l,j}^2}{\delta_l^2})$$

where k_{nz} is the number of the clusters in the current *reference* partition:

$$k_{nz} = k_{max} - |K|, K = \{j|\alpha_j \equiv 0, j = 1, \ldots, k_{max}\},$$

$|K|$ is the cardinality of the set K, which contains the index variables marking the clusters whose weights have been pruned to zero[2], we should not include such components in the feature relevance score calculation.

[2]In practice, we find the component whose weight is smaller than $\frac{1}{N}$ (N is the number of the observations), it is the case for less than one data in the "cluster", which is of "non-sense".

Procedure `FeatureSelection`(F, β, γ, T)

input : F, β, γ, T
output: $\hat{R}$

`// Selecting the relevant features`
Calculate $SCORE_l, F_l \in F$;
$R' \leftarrow F - \{F_l | SCORE_l < \beta, F_l \in F\}$;
`// Selecting the non-redundant features`
Perform Markov Blanket filtering;
$R'' = \{F_{l_m} | \min_{F_l \in G^{(m)}} \Delta(F_l|M_l) > \gamma \cdot \min_{F_l \in G^{(1)}} \Delta(F_l|M_l)\} \cup \{R' - \{F_{l_1}, F_{l_2}, \ldots, F_{l_{|R'|-T}}\}\}$;
$\hat{R} \leftarrow R''$;

5. Experimental Results

This section shows the experimental results on two synthetic data sets and four real-world benchmark data sets. In all the experiments, the initial number of components k_{max} should be safely large so that the initialization properly covers the data. We therefore set $k_{max} = 10$, and the initial mixing coefficients $\alpha_j = 1/k_{max}$ $(j = 1, \ldots, k_{max})$. The initial centers of each clusters μ_j's were randomly chosen from data points, the initial covariance matrices are fraction (1/5 here) of the mean global diagonal covariance matrix:

$$\mathbf{\Sigma}_j^{(0)} = \frac{1}{5d} trace\left(\frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}_t - \bar{\mu})(\mathbf{x}_t - \bar{\mu})^T\right) I$$

where $\bar{\mu} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t$ is the global mean of the data and I is the identity matrix with proper dimensions. The thresholds β and γ were set to 0.4 and 2, respectively. We utilized a the Markov Blanket of size $T = 2$ to perform the feature redundancy analysis. η is set to 0.01 for synthetic sets and 0.0005 for all the real-world data sets. We found that these parameters and initialization performed reasonably well.

5.1. Synthetic data

Firstly, we investigated the capability of the proposed algorithm to select the relevant and non-redundant features while determine the correct clusters simultaneously. 1000 data points were first generated by the following bivariate Gaussian mixture (plotted in figure 2(a)):

$$0.3*\mathcal{N}\left[\begin{pmatrix}1\\1\end{pmatrix}; \begin{pmatrix}0.1 & 0.0\\0.0 & 0.1\end{pmatrix}\right] + 0.4*\mathcal{N}\left[\begin{pmatrix}1\\5\end{pmatrix}; \begin{pmatrix}0.1 & 0.0\\0.0 & 0.1\end{pmatrix}\right]$$
$$+0.3*\mathcal{N}\left[\begin{pmatrix}5\\5\end{pmatrix}; \begin{pmatrix}0.1 & 0.0\\0.0 & 0.1\end{pmatrix}\right]$$

Obviously, we are unable to discriminate the 3 clusters by a single dimension alone, thus both features are relevant

to the clustering. Then we duplicated the two dimensions and form a 4-dimensional data. Lastly we appended each data with 6 independent variables, sampled from a standard normal distribution, yielding a 10-dimensional data set to be analyzed.

Apparently, either $\{F_1, F_2\}$ or $\{F_3, F_4\}$ can determine the 3 clusters, i.e., one of the two pairs of features are redundant. And the last 6 dimensions are unimodal thus being irrelevant to the clustering. We ran the proposed algorithm 10 times, the 3 components and the non-redundant relevant feature subsets (it has chosen $\{F_3, F_4\}$) were always correctly found in all runs. Figure 2(b) shows the learning curve of the component mixing coefficients in a typical run. Table 1 shows some intermediate execution outputs. In the column of "ranking", the first row of each epoch is the relevance score ($SCORE_l$) whose value is in descending order; the second row is the minimum value of expected cross entropy, with its corresponding feature in the sequential removal order by the Markov Blanket filtering. The two rows under the column of "selected features" list the outcomes of the two feature selection stages.

We then compared our algorithm with the one proposed by Law et, al [10]. The algorithm in [10] makes the *soft* decisions on whether the feature is relevant for the clustering or not, and has to pre-assume the irrelevant features conformed to a Gaussian distribution. Otherwise, its performance would be degraded to a certain degree. To illustrated this, we appended 6 variables uniformly distributed between 0 and 5, to the data with the above four relevant dimensions, while the distribution of the irrelevant features is still assumed to be subjected to Gaussian for the algorithm in [10]. It is found that the algorithm of [10] was unable to give a proper inference about the clusters any more. Instead, it always largely over-fitted the data as illustrated in Figure 3. This implies that the algorithm of [10] is sensitive to the assumed distribution of irrelevant features.

In contrast, the proposed algorithm circumvented this drawback. As shown in Figure 3, it succeeded to infer the true clustering structure in the original feature space. Besides, only two most relevant features (it has chosen $\{F_3, F_4\}$) were used, no redundant features being retained. Figure 3(b) demonstrates its learning curve of the component mixing coefficients. Table 2 shows its intermediate outcomes.

5.2. Real-world data

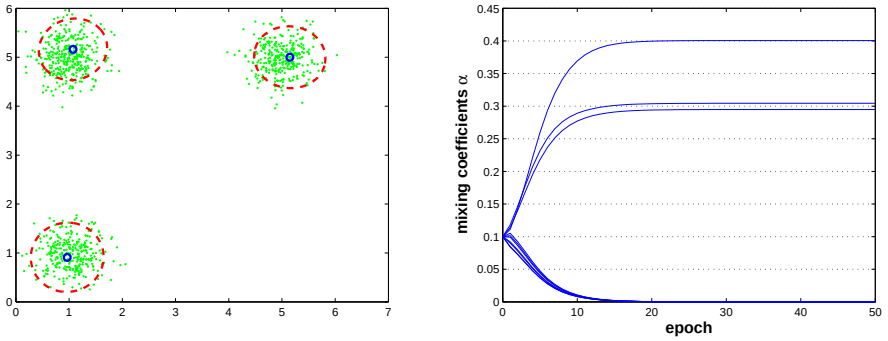
Further, we verified the proposed algorithm (denoted as IRRFS-RPEM) on 4 benchmark real-world data sets [13] (all are normalized in advance). For comparison, we also performed the RPEM algorithm, the algorithm in [10], and a variant of the proposed algorithm (denoted as IRFS-RPEM), in which only the relevancy analysis is carried out in the feature selection phase. We evaluate the clustering accuracy through the *error rate* index. After dividing the original data set into the training set and the testing set of the equal size, we executed the above algorithms on the training set to obtain the parameters of the Gaussian mixture model, then each data point in the testing set was appended a label of the cluster it belonged to, the cluster label was determined by the majority class of the training data assigned to it. The predictive error rate is computed by the mismatch degree between the obtained labels of the testing points and their ground-truth class labels. The mean and the standard deviation of the *error rate*, along with those of the estimated number of clusters in 10-fold runs on the 4 real-world data sets are listed in Table 3.

Table 3. Results of the 10-fold Runs on the Test Sets for Each Algorithm

Data Set	Method	Model Order $mean \pm std$	Error Rate $mean \pm std$
wdbc	RPEM	1.7 ± 0.4	0.2610 ± 0.0781
$d = 30$	algorithm in [10]	5.7 ± 0.3	0.1005 ± 0.0349
$N = 569$	IRFS-RPEM	2.3 ± 0.4	0.1021 ± 0.0546
$k^* = 2$	IRRFS-RPEM	**fixed at 2**	$\mathbf{0.0897 \pm 0.0308}$
sonar	RPEM	2.3 ± 0.8	0.4651 ± 0.0532
$d = 60$	algorithm in [10]	-	-
$N = 1000$	IRFS-RPEM	2.8 ± 0.6	0.3625 ± 0.0394
$k^* = 2$	IRRFS-RPEM	2.7 ± 0.7	$\mathbf{0.3221 \pm 0.0333}$
wine	RPEM	2.5 ± 0.7	0.0843 ± 0.0261
$d = 13$	algorithm in [10]	3.3 ± 1.4	0.0673 ± 0.0286
$N = 178$	IRFS-RPEM	4.7 ± 1.7	0.0492 ± 0.0182
$k^* = 3$	IRRFS-RPEM	3.1 ± 0.5	0.0509 ± 0.0248
ionosphere	RPEM	1.8 ± 0.5	0.4056 ± 0.0121
$d = 32$	algorithm in [10]	3.2 ± 0.6	0.2268 ± 0.0386
$N = 351$	IRFS-RPEM	2.6 ± 0.8	0.2921 ± 0.0453
$k^* = 2$	IRRFS-RPEM	2.5 ± 0.5	$\mathbf{0.2121 \pm 0.0273}$

Each data set has N data points with d features from k^* classes.

It could be observed from Table 3 that both IRFS-RPEM and IRRFS-RPEM have reduced the error rates on all sets compared to the RPEM algorithm. This is because not all features are relevant with respect to the partitioning task. These features with less discriminating power might confuse the RPEM clustering algorithm. Due to the iterative execution of the clustering and the feature selection, the potential optimal cluster-searching space shrank, thus leading to a better performance. Meanwhile, the proportions of the average selected features by IRFS-RPEM and IRRFS-RPEM in the whole feature set for each data sets are reported in Table 4. For the *wdbc* and the *sonar* data sets, IRFS-RPEM and IRRFS-RPEM have selected approximately the same number of features, and have similar predictive performances. A reasonable explanation is that there may be not much redundancy in the selected relevant features for the two data sets. Instead, the *wine* and *ionosphere* sets both seem to



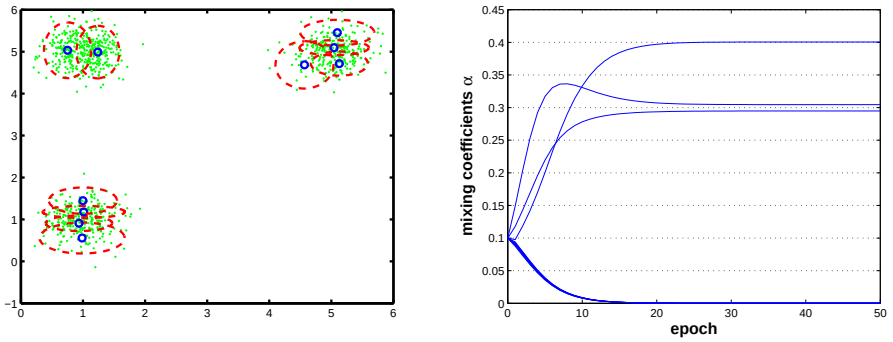
(a) (b)

Figure 2. (a)The bivariate data has 3-cluster structure.(b)The learning curve of ($\{\alpha_j\}_{j=1}^{k_{max}}$) for the proposed algoerithm on the first synthetic data.

Table 1. The intermediate outcomes of the proposed algorithm on the first synthetic data.

epoch	ranking	selected features
1	$0.9734(F_1)\ 0.9728(F_4)\ 0.9718(F_3)\ 0.9716(F_2)\ 0.3590(F_6)\ 0.3311(F_7)\ 0.2202(F_{10})\ 0.1771(F_8)\ 0.1199(F_5)\ 0.1104(F_9)$	$\{F_1, F_2, F_3, F_4\}$
	$0(F_1)\ 0(F_2)$	$\{F_3, F_4\}$
15	$0.8671(F_1)\ 0.8655(F_2)\ 0.8486(F_4)\ 0.8469(F_3)\ 0.2920(F_7)\ 0.2413(F_8)\ 0.2272(F_6)\ 0.2112(F_{10})\ 0.2033(F_5)\ 0.1769(F_9)$	$\{F_1, F_2, F_3, F_4\}$
	$0(F_1)\ 0(F_2)$	$\{F_3, F_4\}$
50	$0.9728(F_2)\ 0.9723(F_1)\ 0.9711(F_4)\ 0.9711(F_3)\ 0.0403(F_7)\ 0.0354(F_5)\ 0.0129(F_8)\ 0.0118(F_9)\ 0.0062(F_{10})\ 0.0051(F_6)$	$\{F_1, F_2, F_3, F_4\}$
	$0(F_1)\ 0(F_2)$	$\{F_3, F_4\}$

In the parentheses, the corresponding feature F_l.



(a) (b)

Figure 3. (a) The clustering results on the second synthetic data set obtained by the algorithm in [10] projected on the first two features: $k = 10$ (over-fitting). (b) The learning curve of ($\{\alpha_j\}_{j=1}^{k_{max}}$) for the proposed algorithm on the second synthetic data: $k = 3$ (correct).

Table 2. The intermediate outcomes of the proposed algorithm on the second synthetic data.

epoch	ranking	selected features
1	$0.9737(F_1)\ 0.9726(F_2)\ 0.9712(F_3)\ 0.9704(F_4)\ 0.3980(F_6)\ 0.3635(F_5)\ 0.2789(F_9)\ 0.2432(F_8)\ 0.1720(F_{10})\ 0.1685(F_7)$	$\{F_1, F_2, F_3, F_4\}$
	$0(F_1)\ 0(F_2)$	$\{F_3, F_4\}$
7	$0.6704(F_1)\ 0.6698(F_4)\ 0.6636(F_2)\ 0.6245(F_3)\ 0.2111(F_8)\ 0.1791(F_6)\ 0.1742(F_{10})\ 0.1127(F_7)\ 0.0691(F_5)\ 0.0385(F_9)$	$\{F_1, F_2, F_3, F_4\}$
	$0(F_1)\ 0(F_2)$	$\{F_3, F_4\}$
50	$0.9737(F_1)\ 0.9726(F_2)\ 0.9712(F_3)\ 0.9704(F_4)\ 0.0403(F_{10})\ 0.0186(F_8)\ 0.0146(F_9)\ 0.0117(F_7)\ 0.0072(F_5)\ 0.0032(F_6)$	$\{F_1, F_2, F_3, F_4\}$
	$0(F_1)\ 0(F_2)$	$\{F_3, F_4\}$

In the parentheses, the corresponding feature F_l.

present the existence of redundancy in the selected relevant features. In particular, for the *ionosphere* data set, the accuracy even improves with further fewer features (nearly 1/3 of its original size) selected by IRRFS-RPEM, compared to the considerably few features already selected by IRFS-RPEM. Although IRFS-RPEM seems to give a better predictive accuracy on the *wine*, it has used more components than the correct order specified.

Table 4. The proportions of the average selected features by IRRFS-RPEM and IRFS-RPEM in the 10-fold runs

Data	IRFS-RPEM	IRRFS-RPEM
synthetic1	40%	20%
synthetic2	40%	20%
wdbc	51.16%	50.33%
sonar	57%	55.83%
wine	83.65%	62.31%
ionosphere	68.13%	34.38%

When comparing the proposed algorithm with the algorithm in [10], although they are comparative in terms of *error rate*, IRRFS-RPEM seems always given a closer estimation of the model order than algorithm in [10], the latter one is more likely to use more components for all the utilized data sets, especially for the *wdbc* dataset. This phenomenon is consistent with the results we have demonstrated on the second synthetic data set.

6. Conclusions

In this paper, we first proposed a new feature relevance measurement index in order to identify the most relevant features, then we introduced the Markov Blanket filter, a technique in supervised learning task, to further reduce the redundancy within the selected relevant feature space but adjusted in an unsupervised scenario. Furthermore, we integrated these two feature selection schemes into the basic RPEM clustering algorithm to derive a new algorithm which iterates between the clustering and the feature selection. Besides, it does not particularly assume the explicit parametric probability distribution function (pdf) for the irrelevant features. From the theoretic analyses and experimental results, we could conclude that the proposed algorithm has at least two virtues: it is effective in eliminating both irrelevant and redundant features, which produces better predictive accuracy in general; it is more tolerant to the degenerating situation when the actual pdf of the irrelevant feature violates the assumed one.

References

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] Y. M. Cheung. Maximum weighted likelihood via rival penalized em for density mixture clustering with automatic model selection. *IEEE Transactions on Knowledge and Data Engneering*, 17(6):750–761, June 2005.

[3] C. Constantinopoulos, M. K. Titsias, and A. Likas. Bayesian feature and model selection for gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):1013–1018, June 2006.

[4] M. C. Dash, K. Scheuermann, and P. H. Liu. Feature selection for clustering-a filter solution. *Proceedings of IEEE International Conference on Data Mining*, pages 115–122, 2002.

[5] A. P. Dempster, N. M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society (B)*, 39(1):1–38, 1977.

[6] J. Dy and C. Brodley. Visualization and interactive feature selection for unsupervised data. *Proceedings of ACM Special Interest Group on Knowledge Discovery in Data*, pages 360–364, 2000.

[7] J. Dy and C. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889, 2004.

[8] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7-8):1157–1182, 2003.

[9] D. Koller and M. Sahami. Toward optimal feature selection. *Proceedings of International Conference on Machine Learning*, pages 284–292, 1996.

[10] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, September 2004.

[11] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[12] P. Mitra, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, March 2002.

[13] D. Newman, S. Hettich, C. Blake, and C. Merz. Uci repository of machine learning databases. 1998.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[15] G. Schwarz. Esting the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[16] C. S. Wallace and P. R. Freeman. Estimation and inference via compact coding. *Journal of the Royal Statistical Society. (B)*, 49(3):241–252, 1987.

[17] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 601–608, 2001.

[18] L. Yu and H. Liu. Efficiently handling feature redundancy in high-dimensional data. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 685–690, 2003.

Reasoning about Decommitment in G-Negotiation Mechanism

Benyun Shi

Abstract

Since computational intensive applications may often require more resources than a single computing machine can provide in one administrative domain, bolstering resource co-allocation is essential for realizing the Grid vision. Given that resource providers and consumers may have different requirements and performance goals, successfully obtaining commitments through concurrent negotiations with multiple resource providers to simultaneously access several resources is a very challenging task for consumers. The novel contribution of this work is devising a concurrent mechanism that (i) coordinates of multiple one-to-many concurrent negotiations between a consumer and multiple resource providers and (ii) manages (de-)commitment for consumer during the one-to-many negotiation in Grid co-allocation. In this work, three classes of commitment strategies for concurrent negotiation are presented. A series of simulations were carried out in two different kinds of settings and favorable results show that these strategies outperformed existing ones.

1. Introduction

A Grid resource management system should bolster co-allocation of computing resources (i.e., allocating to an application multiple resources belonging to possibly different administrative domains) [7, p.2]. Supporting resource co-allocation is essential for realizing the Grid vision because (i) computationally intensive applications may require more resources than a single computing machine can provide in one administrative domain [8, p.1], and (ii) an application may require several types of computing capabilities from resource providers in other administrative domains. Sim [14] argued that software agent, in particular e-negotiation agents, can play an essential role in realizing the Grid vision. While there are existing works on applying bargaining for Grid resources allocation [6] [9-12] [15-16], very few works (except [6][9]) considered negotiation for Grid resource co-allocation. Whereas *SNAP* (Service Negotiation and Acquisition Protocol) [9] searches for the solutions for satisfying simultaneous multiple resources requirements of Grid consumers, it did not specify the strategies for negotiation agents.

Given that resource providers and consumers may have different requirements and performance goals, successfully obtaining commitments through concurrent negotiations with multiple resource providers to simultaneously access several resources is a very challenging task for consumers. Since there may be multiple resource providers providing a specific kind of resource, a consumer may select a required resource by adopting a one-to-many negotiation model. Additionally, for Grid resource co-allocation, resource selection would also involve coordinating multiple one-to-many concurrent negotiations and ensuring that the consumer can successfully obtain all required resources simultaneously. The impetus of this work is devising a concurrent negotiation mechanism that (i) coordinates multiple one-to-many concurrent negotiations between a consumer and multiple resource providers, and (ii) manages (de-)commitment for consumer in each one-to-many negotiation [1] [13] in which both consumers and providers can renege from a deal. This work is part of a research initiative to develop a negotiation mechanism for Grid resource management proposed in [6]. The agendas of this work are to:

1) devise commitment management strategies for the commitment manager in a concurrent G-Negotiation mechanism (section 2) to manage concurrent one-to-many negotiation for each resource (section 3).
2) conduct a series of simulations to compare the commitment management strategies with existing related works [1][13] (section 4).

Whereas section 5 compares the concurrent negotiation mechanism in this work with existing related systems, section 6 concludes this paper by summarizing a list of future works.

2. A Concurrent G-Negotiation Mechanism

This section describes an approach for the grid resource co-allocation problem under a commitment model [1] [13] where reneging from a deal is allowed for both consumer and provider agents. In this work the grid resource co-allocation problem for n kinds of resources is transformed into a problem of n concurrent one-to-many negotiations where each one-to-many negotiation is also a concurrent negotiation for a particular kind of resource R_i, $1 \le i \le n$. Using this mechanism, a consumer in the Grid market negotiates simultaneously with multiple providers that supply possibly different types of resources. Denote $\{O_j^i \mid 1 \le j \le n_i\}$ the set of n_i

resource providers of the resource R_i, $1 \le i \le n$. Each consumer has n resources to acquire and a hard deadline τ_c for acquiring all n resources. Both an agent's preference for a resource and the strategy that it adopts during the negotiation are private information.

By doing this, the negotiation mechanism consists of three components: a *coordinator module* (*CoM*), n commitment managers ($CM_i, 1 \le i \le n$) and each CM_i manages a number of negotiation threads. For each one-to-many negotiation for a particular resource, there exists a commitment manager agent CM_i [1] [13] that manages both commitments and de-commitments. Each CM_i adopts the management strategy in section 3 to decide (i) whether or not to accept a resource provider's proposal or (ii) when to renege from a commitment at each negotiation round. Each negotiation thread (bargaining for a particular resource) follows a *Sequential Alternating Protocol* where at each negotiation round, an agent can (i) accept the proposal from the opponent, (ii) propose its counter-offer, (iii) renege from its commitment or (iv) opt out of the negotiation.

3. Commitment Manager

During each negotiation round, each thread will report its status of negotiation to the commitment manager component. The commitment manager helps the consumer decide whether or not to accept the proposed offers from the resource providers or when to renege from an intermediate commitment. It is not efficient for the consumer agent to easily agree with all acceptable proposals from resource providers and select the best proposal from them because reneging from other deals may be forced to pay many penalty fees to the system in a commitment model. As noted in [1], there exists a trade-off for the consumer agent between the number of agreements it makes and their utility values. In this section, three classes of commitment management strategies are proposed for the consumer agent. These commitment management strategies depend on the time-dependent negotiation strategies in [5][14] and are derived as follows.

Since a resource can be requested by multiple consumers simultaneously, it is possible for a resource provider to renege from the intermediate deal it has already reached with the consumer. In this work, at each negotiation round t, the consumer will first estimate the probability p_{ij}^t of each resource provider O_j^i reneging from the deal based on all proposals it received at the current round if it accepts O_j^i 's proposal. Denote $\mathrm{P}^i(t) = \{\mathrm{P}^i_j(t) | 0 < j \le n_i\}$ as the set of proposals for resource R_i the consumer received at round t, and $Avg(\mathrm{P}^i(t))$ the average value of these proposals. Then, the variance of $\mathrm{P}^i(t)$ can be calculated as

$$D(\mathrm{P}^i(t)) = \frac{1}{n_i}\sum_{k=1}^{n_i}\left[\mathrm{P}^i_k(t) - Avg(\mathrm{P}^i(t))\right]^2$$

If $D(\mathrm{P}^i(t))$ is a large value, it means that $\mathrm{P}^i(t)$ has a sparse value distribution, otherwise, $\mathrm{P}^i(t)$ has a dense value distribution. The consumer's subjective reneging probability p_{ij}^t about resource provider O_j^i reneging from an intermediate deal (if any) at round t can be calculated using the relationship of $Avg(\mathrm{P}^i(t))$, $\mathrm{P}^i(t)$ and $D(\mathrm{P}^i(t))$. For example, when $Avg(\mathrm{P}^i(t)) - \mathrm{P}^i_j(t)$ is much larger than $\sqrt{D(\mathrm{P}^i(t))}$, from the consumer's point of view, the proposal of the resource provider O_j^i is too far away from the average value of other proposals, which means that O_j^i is currently in an advantaged position such that it can be easily chosen by other consumers, the calculated probability should be larger; and vice versa.

Using this calculated subjective reneging probability, the consumer's expected utility $E_{t_c}(U^i(\mathrm{P}^i_j(t_c)))$ from the proposal $\mathrm{P}^i_j(t_c)$ of resource provider O_j^i at current round t_c can be computed as follows

$$E_{t_c}(U^i(\mathrm{P}^i_j(t_c))) = (1 - p_{ij}^{t_c}) \cdot U^i(\mathrm{P}^i_j(t_c))$$

where $U^i(\cdot)$ is the consumer's utility function for resource R_i.

A new proposal $\mathrm{P}^i_j(t_c)$, $(1 \le i \le n)$ from resource provider O_j^i will be accepted as an intermediate deal by the consumer at current round t_c if it satisfies the following conditions.

1) If it already has a commitment with another provider agent O_k^i, at round $t_{ik}(t_{ik} < t_c)$ and this deal has not been broken, the expected utility gained by taking this new proposal must be greater than that of the current deal, which means $E_{t_c}(U^i(\mathrm{P}^i_j(t_c))) > E_{t_c}(U^i(\mathrm{P}^i_k(t_{ik})))$; and the utility gained from the new proposal must be greater than that of current one after having paid the penalty fee, i.e., $U^i(\mathrm{P}^i_j(t_c)) > U^i(\mathrm{P}^i_k(t_{ik})) + \rho_k^i(t_c)$ where the penalty fee

$$\rho_k^i(t_c) = U^i(\mathrm{P}^i_k(t_{ik})) \times (\rho_0^i + \frac{t_c - t_{ik}}{\tau_c - t_{ik}} \cdot (\rho_{\max}^i - \rho_0^i))$$

is calculated following the formula in [1][13].

2) If there is no commitment yet, the consumer will compare all expected utilities of all received proposals with the utility of its next proposal. If there are some proposals whose expected utilities are larger, the consumer will accept the best proposal with the lowest

price among these proposals; otherwise, it makes its proposal and proceeds to next round.

In this paper, a consumer adopts three classes of time-dependent negotiation strategies (*Linear*, *Conciliatory* and *Conservative*) in [5] [14] to generate its proposals. Based on the two conditions above, three classes of commitment management strategies (*Linear-CMS*, *Conciliatory-CMS*, and *Conservative-CMS*) can be specified and they correspond respectively to the *Linear*, *Conciliatory* and *Conservative* time-dependent strategies.

4. Simulations and Experimental Results

To evaluate the effectiveness of the commitment management strategies of the concurrent negotiation mechanism in section 3, a series of experiments were carried out.

1) Objectives and Motivations: The objective is to compare the commitment management strategy in section 3 with that of [1] [13] in a concurrent one-to-many negotiation environment for a particular resource, where de-commitment is allowed for both resource provider and consumer agents.

2) Experimental Settings: Some independent variables are listed in Table I. The other variables are set as follows.

Table I. Independent Variables for First Experiments

Variables	Description	Values
$\rho_0, \rho_{\max}$	Penalty level	[0,1]
$P_{\min}$	Minimum price	1
$P_{\max}$	Maximum price	100
p	Reneging probability for resource providers	[0,1]
n	The number of providers	[1,30]

a) *Initial price and reserve price*: In the experiments, there is one consumer negotiating with multiple resource providers for one resource. Without loss of generality, we assume that there exist intersections between agreement zones (the domain between initial price and reserve price) of the consumer and that of each resource provider. The initial price IP_B of the consumer is first generated uniformly from domain $[P_{\min}, \frac{P_{\max}}{4}]$, and reserve price RP_B uniformly from domain $[\frac{P_{\min}}{2}, \frac{3P_{\max}}{4}]$. Then, the agreement zone of each resource provider is generated as $[IP_{S_j}, RP_{S_j}]$ such that $IP_B \le RP_{S_j} \le RP_B \le IP_{S_j}$.

b) *Deadline*: The deadlines for each resource provider and consumer are uniformly generated from the time region [10,100].

c) *Negotiation strategy*: Resource providers in this experiment make their proposals using time-dependant strategies [5]. Different resource provider has different time preference λ which is chosen from [0.1,10].

d) *Reneging probability*: To compare the commitment management strategy with that of [1] [13], this work simulates two different kinds of reneging probability settings for the resource providers: (i) when the reneging probability of resource provider is fixed to be 0.5 (HALF-PROB), which is also used in [1]; and (ii) when the reneging probability of the resource provider is randomly chosen from region $[0,1]$ (RAND-PROB) (which is more general than (i)).

3) Performance Measure: The experiments compare the commitment management strategy in section 3 with that of [1] [13] in a single one-to-many negotiation for one particular resource rather than multiple concurrent negotiations for multiple resources, only final utility and success rate are used as the performance measure, which are defined as follows:

a) The final utility of the consumer (U_c) is calculated by the following formula in the experiment:

$$U_c = \frac{1}{N}\sum_{i=1}^{N}(U_c^i - \Gamma^i)$$

where $U_c^i = \frac{RP_c^i - P^i}{RP_c^i - MIN_o^i}$ is the utility of the consumer from a deal, (in which MIN_o^i is the minimum initial price of all resource providers, P^i is the price of the deal), Γ^i is the total penalty that the consumer should pay for resource, and $N = 1000$ means that 1000 runs of experiments are carried out.

b) The success rate is defined as the ratio of the success negotiations over the total 1000 runs.

4) Results and Observations: Empirical results are shown in Fig 4.1-4.2.

Fig. 4.1 shows the results of the experiments for the one-to-many concurrent negotiation under different penalty levels. It shows the comparison of the final utilities between commitment management strategy in section 3 and that in [1][13] under two different reneging probability settings. It shows that even using a random time-dependent commitment management strategy, the commitment management strategy in this work results in better final utilities than the Nguyen-Jennings-like commitment management strategy in [1][13] in different penalty levels under two different reneging probability settings. Fig. 4.2 shows the comparison of success rate of commitment management strategies under these two different reneging settings. It shows that the time-dependent commitment management strategy in this work results in higher success rate than the Nguyen-Jennings-

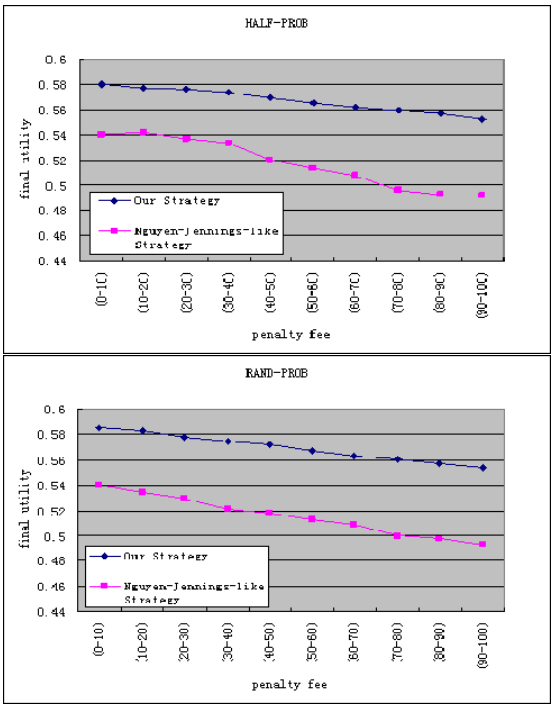


Fig. 4.1 Utility Comparison of Commitment Strategies

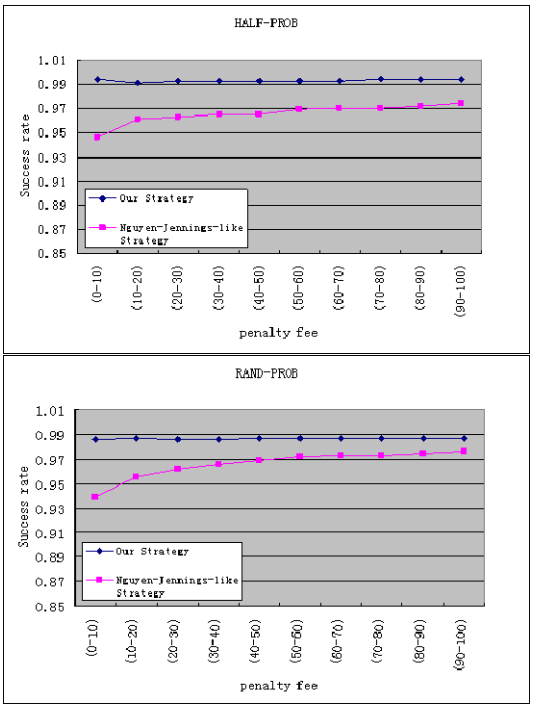


Fig. 4.2 Success Rate Comparison of Commitment Strategies

like commitment management strategy in [1][13].

This is because in [1][13], *the degree of acceptance* is used to determine whether to accept a new proposal when there is no commitment for the consumer. Whereas [1][13] only consider the maximum predicted utility of the next proposal from other resource providers when they calculate the degree of acceptance, which may not necessarily be accurate. In this work, all proposals resource providers proposed at the current round are considered when making decision whether to accept a proposal.

5. Related Work

Since this work explores the issue of applying allocation problem, the areas that relate to this research include: 1) Grid resource co-allocation and 2) concurrent negotiation system.

1) *Grid resource co-allocation*: Most related work in the area of Grid resource negotiation [10-12][15-20] did not consider Grid resource co-allocation. Whereas *SNAP* [9] support Grid resource co-allocation by facilitating the search solutions to satisfying simultaneous multiple resources requirements, [9] did not focusing on specifying the negotiation strategies that agents should adopt.

2) *Concurrent negotiation system*: In [2], Rahwan et al proposed a concurrent negotiation mechanism for coordinating one-to-many negotiations. Additionally, [1][3-4][12-13] extend the mechanism of [2] as a concurrent bi-lateral negotiation system and further introduce a commitment model into their system. In [1-4][12-13], a one-to-many negotiation is modeled as multiple concurrent threads of one-to-one negotiations. This work models concurrent negotiation for Grid resource co-allocation by providing a mechanism for coordinating concurrent multiple one-to-many negotiations, and more effective management strategies for each one-to-many negotiation.

6. Conclusions and Future Work

This paper reports a portion of the work done in a larger project proposed in [6] and more extensive results of this work are reported in a paper co-authored with my supervisor currently under consideration for a conference.

The contribution of this work is devising a concurrent negotiation mechanism (section 2) together with three classes of commitment management strategies (section 3) for managing multiple concurrent negotiations. Favorable empirical results show that the commitment strategies in this work outperformed the commitment strategy in [1][13] in terms of utility and success rate.

Finally, it is noted that this work is still in its infancy and a list of agendas for future works includes: 1) designing coordination strategies to coordinate multiple one-to-many negotiations during the co-allocation, 2) designing algorithms to determine which class of

commitment management strategy should be used in a specific Grid market (for instance, the number of consumer is larger than that of resource providers), and 3) combine commitment managers and the coordinator to evaluate the whole mechanism .

10. References

[1] T.D. Nguyen and N.R. Jennings, "Managing commitments in multiple concurrent negotiations", *International Journal Electronic Commerce Research and Applications, 4 (4)*, 2005, pp. 362-376.

[2] I. Rahwan, R. Kowalczyk, and H. H. Pham, "Intelligent agents for automated one-to-many e-commerce negotiation", *Twenty-Fifth Australian Computer Science Conference, 4*, 2002, pp. 197-204.

[3] T.D. Nguyen and N.R. Jennings, "Concurrent bi-lateral negotiation in agent systems", *In Proc. 4th DEXA Workshop e-Negotiations*, 2003, pp. 839-844.

[4] T.D. Nguyen and N.R. Jennings, "A heuristic model of concurrent bi-lateral negotiations in incomplete information settings", *In Proc. 18th Int. Joint Conf. Artif. Intell.*, 2003, pp. 1467-1469.

[5] K.M. Sim, "Equilibria, Prudent Compromises, and the 'Waiting Game'", *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, Vol. 35, No.4*, Aug. 2005, pp. 712-724.

[6] K.M. Sim, "Relaxed-criteria G-negotiation for Grid Resource Co-allocation", *ACM SIGECOM: E-commerce Exchanges, Vol. 6, No. 2*, Jan. 2007, pp. 37-46.

[7] A. Ali et al, "A Taxonomy and Survey of Grid Resource Planning and Reservation Systems for Grid Enabled Analysis Environment", *Proc. of the 2004 Int. Sym. on Distributed Comp. and Appl. to Business Eng. and Science*, Wuhan, China, 2004, pp. 1-8.

[8] R. Buyya, D. Abramson, and S. Venugopal, "The Grid Economy", *Proceedings of the IEEE, Volume 93, Issue 3*, IEEE Press, New York, USA, March 2005, pp. 698-714.

[9] K. Czajkowski, I. Foster, et al, "*SNAP*: A Protocol for Negotiation of Service Level Agreements and Coordinated Resource Management in Distributed Systems", *Job Scheduling Strategies for Parallel Processing: 8th Int. Workshop*, 2002, pp. 1-10.

[10] K.M. Sim, "From Market-driven Agents to Market-Oriented Grids", *ACM SIGecom: Ecommerce Exchanges, Vol. 5, No. 2*, November 2004, pp. 45–53.

[11] K.M. Sim, "G-Commerce, Market-driven G-Negotiation Agents and Grid Resource Management", *IEEE Transactions on Systems, Man and Cybernetics, Part B, Vol. 36, No. 6*, Dec. 2006, pp. 1381-1394.

[12] T.D. Nguyen and N.R. Jennings, "Coordinating multiple concurrent negotiations", *In Proceeding of the 3rd International Joint Conference on Autonomous Agents and Multi Agent Systems*, New York, USA, 2004, pp. 1064-1071.

[13] T.D. Nguyen and N.R. Jennings, "Reasoning about commitments in multiple concurrent negotiations", *In Proceedings of 6th International Conference on E-Commerce*, Delft, The Netherlands, 2004, pp. 77-84.

[14] K.M. Sim, "From Market-driven Agents to Market-Oriented Grids", *ACM SIGECOM: E-commerce Exchanges, Vol. 5, No. 2*, November 2004, pp. 45–53.

[15] K. Chao et al, "Using Automated Negotiation for Grid Services", *Int. J. of Wireless Information Networks, Vol. 13, No. 2*, Springer, 2006, pp. 141 - 150.

[16] K.M. Sim, "A Survey of Bargaining Models for Grid Resource Allocation", *ACM SIGECOM: E-commerce Exchanges, Vol. 5, No. 5*, January, 2006, pp. 22-32.

Continuous Spatial Queries in Wireless Sensor Networks

Yu Li

ABSTRACT

In this paper, we present our study on processing continuous spatial queries in wireless sensor networks. Wireless sensor networks have the potential to provide a wealth of information about the environments in which they are deployed. The sensor nodes are responsible for capturing environmental data and the base station is responsible for accepting and answering user queries. However, wireless sensor networks surfer from limited energy supply. Thus, energy efficiency is a key consideration in sensor network designs. In this paper, we focus on continuous spatial queries and propose efficient solutions for them to reduce network traffic and extend network lifetime.

1. INTRODUCTION

Wireless sensor networks have the potential to provide a wealth of information about the environments in which they are deployed. A wireless sensor network typically consists of a group of sensor nodes and a base station. The sensor nodes are responsible for capturing environmental data (e.g., temperature), and the base station is responsible for accepting and answering user queries (e.g., "Are there areas in the danger of fire disaster?"). Because the sensed data is stored on individual sensors, the user queries cannot be answered directly by the base station. In order to answer queries, collecting sensor readings is usually necessary. Otherwise, in-network processing is also a possible solution. Both of these approaches require communication between sensor nodes. However, a wireless sensor network is tempered by the limited energy supply of sensor nodes, and radio communication is the main energy consumer [1,2]. Therefore, it is of utmost importance to minimize the communication costs in order to improve energy efficiency of wireless sensor networks.

One of the most important classes of queries is *continuous query*, which requires continuous monitoring of the wireless sensor network and answering the query every time interval. For example, consider a wireless sensor network monitoring the temperature of a fortress. The users may be interest in the areas which are probably in the danger of fire disaster. He/she can issue a spatial query by monitoring the average temperature of the area, e.g., "report the area of sensor nodes with average temperature higher than 50 ℃". When answering the query for the first time, there may be not other choices but only to collect all sensor readings in the sensor network. Afterwards, if we should continuously answer that query, unnecessary radio communications could be saved. For example, the simplest case is that data collection can be totally avoided when there is no any change in sensor readings (so there is no change of the area). And there are complicated cases, e.g., some sensor readings change but the overall shape of the area will not change. This offers opportunities to improve energy efficiency for continuous spatial queries. In this paper, we aim to develop efficient continuous query processing algorithms that maximize the unnecessary radio communications.

There are several types of *continuous queries,* focusing on different aspects of the wireless sensor network. One of them, which have not been extensively discussed yet, is to query the spatial area of the sensor network based on some conditions. These conditions usually indicate some special events, such as being in danger of a fire disaster. The query we presented earlier ("report the area of sensor nodes with average temperature higher than 50 ℃") is an example. In general, we may define it as a spatial query returning an area (in short, spatial query in this paper), which returns a set of sensors indicating the area. As we have noted, these queries should be processed carefully in considering of saving the energy of the whole sensor network. We propose a solution based on the suppression technique in the hope of maximizing the lifetime of the wireless sensor network.

The rest of the paper will be organized as follows. In Section 2, we present a summary of existing solutions of different continuous queries in the literature, including a taxonomy of them. Section 3 formulates the spatial queries we want to study, and present the motivating cases we have investigated. In Section 4, a complete solution based on suppression technique will be proposed, and details will be discussed. Finally Section 5 lists the future works.

2. RELATED WORK

2.1 Background

Processing *continuous queries* in wireless sensor network have attracted many researchers' attention in recent years. Many solutions and systems were published focusing on different kinds of continuous queries [3-14].

The concept of data management in wireless sensor network was firstly introduced in TinyDB project [3]. Besides discussing the general framework of data management in wireless sensor network, acquisitional query processing for aggregation queries was proposed. More important, the general optimization goal, which described as power sensitive dissemination and routing, was also introduced. After that REED [4], which is an extension of TinyDB, discussed how to perform event detection in wireless sensor networks. Bloom filters, partial joins and cache diffusion were introduced as the basic techniques towards to an energy efficiency solution. There are also other works concurrently published on studying the query processing in wireless sensor networks. For example, the work in [5] contributes optimal algorithm for in-network synopsis join processing for wireless sensor networks.

Continuous queries were firstly discussed in model-driven approach as people find that in special systems some communication and sensing cost could be saved by carefully studying the properties of the system. It firstly aggressively

exploits sensor's correlations by constructing a model over all sensor nodes, times, and sensor types [8]. Then try to find correlations by investigate sensor readings. If the sensor reading from some arbitrary point in the network somehow predicts the sensor reading in some other arbitrary point or at the same point in the future, these correlations can be encoded. As the correlations are discovered, energy could be saved by avoiding collect sensing data, so that extend the lifetime of the network. However, model-driven faces the problems of presenting query results, given their inherent uncertainty, to users, especially those without statistics backgrounds [7]. In other words, it is limited by system specify models.

Consequently data-driven approach for processing continuous queries gains more attention. We have seen a number of data-driven approaches exploiting spatial correlation in local ways, typically through clustering [9,10], exploiting temporal correlation on a node-by-node basis [11,12] and trying to do optimize based on the characterization of special kinds of continuously queries [13]. Instead of building system level models, data-driven investigates the properties of the queries and wireless sensor networks to find independent optimizing techniques. Although it may also face the same problem as in model-driven approach, especially due to the influence of message failure, for users (or automated systems) that care about reaching particular certainty thresholds, it is more acceptable.

Suppression is the key technique in data-driven approach for supporting continuous queries without continuous reporting. The network applied data-driven approach, on its own volition, chooses when to push data to the base station. The intuition is if the network and base station can agree on an expected behavior, the network need only report when its readings deviate from that. And the challenge is encoding expected behavior within the network, such that actual behavior can be efficiently evaluated against it. The design space for suppression is enormous. Suppression can be utilized in a multitude of ways. Value based temporal suppression leverages the expectation node values are unlikely to change in a given time step. Spatial suppression is also possible. This approach leverages the expectation that nearby nodes will have similar values, and minimizes the number of messages directed to the base station. The design space extends to more sophisticated schemes that leverage both temporal and spatial correlations. Naturally, the effectiveness of a scheme for a particular deployment depends on how well it captures the correlations existing in the deployment.

2.2 Various Continuous Queries

Before finding each continuous query a solution, classification is necessary in order to study its properties for possible optimization. In following paragraphs, we will review some kinds of continuous queries so far studied in literature.

SELECT * Query: This is the simplest continuous query. For example

```
SELECT * FROM sensor_network S;
```

It simply collects all sensing readings every time interval. In conventional data management, this kind of queries usually has less importance of optimization so that is uninteresting in research. However in wireless sensor networks, it is practical to find it an energy efficient solution. TinyDB proposed semantic routing tree for general optimization. And the CONCH, which is proposed in [6], studied the optimal solution. It argues that the routing tree is usually not optimal when collect sensing readings in wireless sensor network. Its direction is to extend the idea behind semantic routing tree. First it investigated that monitoring changes of sensors can help inferring the data from last time's sensing reading. Second it proposed a cost model of suppression filters installed at different places in the wireless network. At last the construction of optimal semantic routing tree is reduced to a linear programming problem and then solved.

There are some variants of SELECT * query, such as aggregation query. It replaces the returned result with the aggregation of whole wireless sensor network, such as the sum, the average, the maximum and the minimum of sensing readings. Some of them, such as the sum and the average, may apply same techniques in processing SELECT * queries. Others may not be directly solved. However, we only need to build different cost model to take care of them.

Approximate Query: The aggregation queries may have precision requirements. For example

```
SELECT average(*) FROM sensor_network S
       WITH PRECISION 10;
```

Which means it will acceptable even when there is little difference between the result and real data in the wireless sensor network. In conventional data management they are approximate queries. The relaxation of data quality motivates new algorithms in consideration of extending the network's lifetime. Tang's work [13] exploits that tradeoff between data quality and energy consumption. Its basic idea is to calculate a safe bound according to the precision requirement and then allocate the safe bound to each sensor in the wireless sensor network. Only when the bound is broken by sensing reading's change, the data collection is performed. Furthermore, when collecting data, new bound may be calculated and updated for each sensor. Analytical cost model is proposed to achieve an optimal allocation scheme of the safe bound. In general this is also a suppression-like solution, but utilized the unique characteristic of approximate queries.

Spatial Query: The spatial queries query specific spatial events or areas in the sensor networks. For example of the former case, a query trying to find fire events in a wireless sensor network monitoring temperature may require there is a sensor whose temperature is higher than 100 ℃ and its neighbors' temperatures are lower than 50 ℃. Yiu et. al. [14] studied this kind of queries and proposed both acquisitional and distributed solutions. The acquisitional solutions applied the suppression technique as a base line solution. Distributed solutions works in network, and utilize the selectivity to decide which sensor should trigger the searching in sensor network. Selectivity based cost model is built up and optimization techniques is discussed. On the other hand, as far as we know, there is seldom research works on the query spatial areas in sensor network. And our paper will dedicate to that kind of queries.

3. MOTIVATION

3.1 Problem Formulation

In a wireless sensor network monitoring temperature of a fortress, consider following query, which query some areas in the sensor networks that may be in danger of fire disaster.

```
Q1: SELECT area A
    FROM sensor_network S
    WHERE FOR EACH sensor s in A,
          Average(s.temp) > 50 ;
```

The query requires the average temperature of the area should be higher than 50 ℃. The returning result should be an *area*, which is a subset of the wireless sensor network consisting of sensor nodes. We define the area as

Definition *area*: An area in a wireless sensor network can only be (1) one sensor node, or (2) a set of sensor nodes (at least 2) which has at least one neighbor in the area.

In sensor network, if a sensor node can directly communicate (without relay) to another one, it is called a neighbor of the other. So condition (2) ensures that the area is a connected graph.

In particular, as the base station usually has the topology of the whole network, peoples may be more interest in the boundary of an area, which describes the shape of the area.

Definition *boundary of the area*: the boundary of an area A is defined as

$$A_b = \{s_i \mid s_i \in A \wedge \exists s_j \in neighbor(s_i), s_j \notin A\}$$

Which means the boundary is a set of sensors of the area that has at least one neighbor not in the area.

Sometimes the boundary will be the same to the area, such as the area of one sensor node, but more usually there are big areas with small set boundary. Considering the continuously monitoring task, only reporting the changes of boundary of the area to base station is intuitive to save communication cost and extend wireless sensor network's life time. On the other hand, after first time querying the area, with only the boundary we can identify based on the network topology stored in based station. So in later discussion we focus on continuously querying the boundary of the area.

3.2 Motivating Examples

Without loss of generality, given a wireless sensor network shown in Figure 1, sensors will be placed to monitor cells and the neighbors are easy to decide. Suppose that at time *t*, the sensor readings is shown in Figure 1(a) (only part of sensing readings are shown), and by collecting these sensing readings we could identify an area (in shadow) with average temperature higher than 50 ℃. In following time, we consider a filter strategy: assign each sensor a safe bound and only when the change of the sensor's temperature change breaks the safe bound we perform in-network checking and consider possible updates. In other words, changes within the safe bound will be ignored to save energy.

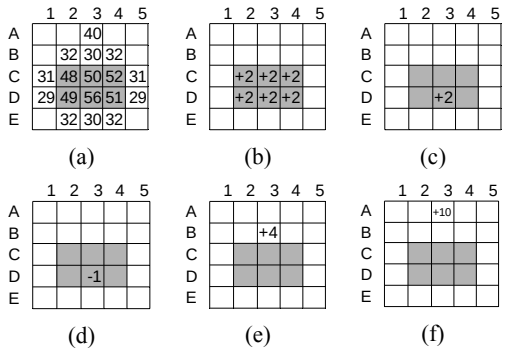


Figure 1 (a) Sensor Network At time t (b)(c)(d)(e)(f) Possible changes of Sensor Network At time t'

Check some examples in Figure 1. Consider temperature increasing first. Inferring from the status of time t (Figure 1(a)), we know that if the total temperature increasing of the area exceeds 12 ℃, there is a neighbor (cell B2, B4, E2 or E4) that could be included into the area. Also it can be that each sensor in the area reads a higher temperature at next time t' (shown in Figure 1(b)). Though whether we can include one of the neighbors could only be clear after we check the changes of the neighbor itself (in case there is a dropping on it), it shows us that 2 ℃ is an increasing bound for possible checking and updating. When the increasing of the area is below 12 ℃, we ignore it and do nothing. And if we uniformly allocate this increasing bound to each sensor node in the area, each one get a bound 2 ℃. Therefore any node breaks its own bound will trigger a new round of checking, such as shown in Figure 1(c). Easy to know, the increasing bound of the area is equal to the minimal amount of temperature need by any neighbors in order to enter the area.

Similar to the increasing case, we can decide decreasing bound for temperature dropping. Also by investigating the initial status in Figure 1(a), we find that the decreasing bound for the area is 6 ℃, and uniformly each sensor gains a bound 1 ℃. So when any sensor reads a temperature dropping exceeds that bound, just as shown in Figure 1(d), a new round of checking should be performed. The dropping bound of the area can be calculated by comparing the sum of temperature of the area and the required average.

Besides sensor nodes in the area, there are other nodes not considered yet, such as neighbors of the area (e.g. cell B3 in Figure 1(a)), and other independent sensor nodes (e.g. cell A3 in Figure 1(a)). Because they are not in the area, temperature decreasing will not change their status. Therefore we focus on studying temperature increasing of them. For independent sensor nodes which are neither the neighbor of the area nor in the area (for short independent sensor nodes in rest of the paper), their increasing bound simply equal to the left amount of temperature to reach the requirement of the query. For example, when reading at least an increasing of 10 ℃ (Figure 1(f)), the sensor in cell A3 can trigger reporting to new area formation. On the other hand, for neighbors of the area, it is not that easy. Because the possible benefit of the extending action of the area, the increasing bound may be less than left amount required by query's condition. For example, sensor in cell B3, if we consider

the area's benefit, an increasing of 4℃ (Figure 1(e)) is enough to possible checking (rather than 20℃).

So far, we describe the motivating examples for us to develop a suppression solution for continuous spatial querying problem. Its key idea is to assign each sensor a set of bounds to suppress small sensor reading changes which do not affect the shape of areas. It divides the problem into 4 sub problems

1. Calculate bounds after the initial executing of query;
2. Allocate the bounds to each sensors;
3. Check and report when bounds is broken;
4. Update bounds after boundary updating.

4. SYSTEM MODEL

Table 1 Notations

k	Required Average
s_i	A sensor node in the sensor network
$neighbor(s_i)$	The set of neighborhood of s_i
$V(s_i,t)$	Sensor s_i's reading at time t
$A(t)$	The area at time t
$A_b(t)$	The area's boundary at time t
$A_n(t)$	The area's neighborhood at time t
$Sum(A(t))$	Sum of sensor readings of $A(t)$
$n(A(t))$	Number of sensors in $A(t)$

4.1 Calculate the Bounds

As mentioned before, there are three kinds of sensor nodes in wireless sensor network after an area is detected, which are (1) the sensor nodes in the area, (2) the neighbors of the area and (3) independent sensor nodes. Each kind has different bounds associated, and we will formally discuss them below.

Sensor nodes in the Area

For each node in the area, there will be two bounds for it – *increasing bound* and *decreasing bound*.

The Increasing Bound

Given a sensor s_j as the neighbor of area A, when the increasing of whole area exceeds

$$k - \frac{Sum(A(t)) + V(s_j,t)}{n(A(t)) + 1}$$

it could be included into the area. Therefore the increasing bound for the area is the minimal cost to include any neighbor of the area, which is

$$E_I = \min_{s_j \in A_n(t)} \{\frac{Sum(A(t)) + V(s_j,t)}{n(A(t)) + 1} - k\}$$

And when allocate it to individual sensor node, below inequality should be guaranteed

$$\sum_{s_i \in A(t)} e_I(i) \le E_I$$

Considering uniform allocation, for each sensor node in the area, the increasing bound is

$$e_I(i) = \frac{E_I}{n(A(t))}$$

The Decreasing Bound

Only when the decreasing will make the sum of sensor readings in area less than the required amount, the area will change its shape. So the decreasing bound is

$$E_D = Sum(A(t)) - k \times n(A(t))$$

And when allocate it to individual sensor node, below inequality should be guaranteed

$$\sum_{s_i \in A(t)} e_D(i) \le E_D$$

Consider uniform allocation, for each node in the area, the decreasing bound is

$$e_D(i) = \frac{E_D}{n(A(t))}$$

The neighbors of the Area

Given a node s_i as a neighbor of the area, as it may receive benefits from the area, with small increasing it could be included into the area. Formally, its increasing bound is

$$\frac{(k+1) \times n(A(t)) - Sum(A(t)) - V(s_i,t)}{k+1}$$

And as discussed there is no decreasing bound for it.

Individual Sensor Nodes

Given a node s_i neither in the area nor as a neighbor of the area, its increasing bound only depends on the sensor reading of itself, which is

$$k - V(s_i,t)$$

And also there is no decreasing bound for it.

4.2 Operations after the Bounds' Broken

Whenever a sensor node breaks its bound, there is still possibility not to violate the bound of the area. In particular, other sensor nodes' change may counteract the bad effect of that sensor node, even when these changes are all in safe bound. For example, when sensor in cell D3 changes as shown in Figure 1(c), if other sensor nodes do not change, or even drop some temperature, the bound of the area will not be broken. Therefore, when a sensor node breaks its bound, we have to travel all nodes in the area in the hope of eliminating the changes. When the increasing bound is broken, as there is possibility of extending the area, the neighbors also should be checked. While traveling the area, a number recording the sum of changes of sensor nodes is maintained, in order to know whether the change is eliminated.

In general, when a sensor node s_i breaks its bound

1. if $s_i \notin A(t) \wedge \forall s_j \in neighbor(s_i), s_j \notin A(t)$ (s_i is a individual sensor node), there is no need to travel and should update to the base station;

2. if $s_i \notin A(t) \wedge \exists s_j \in neighbor(s_i), s_j \in A(t)$ (s_i is a neighbor of the area), traveling the area should be performed;
3. if $s_i \in A(t)$ (s_i is in the area)
    a. if the increasing bound is broken, traveling the area and the neighborhood should be performed;
    b. if the decreasing bound is broken, traveling the area should be performed.

The task of traveling is to compute the sum of changes of sensor nodes. So for each sensor node, it may be visited many times but the changes can only be counted in once. The strategy for traveling should ensure that the visiting times are minimized while all nodes are visited. Naïve strategies such as randomly walking may waste unnecessary radio communications so it is not preferred. In following paragraphs, we introduce a travel strategy based on layers.

	1	2	3	4	5	6	7	8
A	0	0	0	0	0	0	0	0
B	0	1	1	1	1	1	1	0
C	0	1	2	2	2	2	1	0
D	0	1	2	3	3	2	1	0
E	0	1	2	3	3	2	1	0
F	0	1	2	2	2	2	1	0
G	0	1	1	1	1	1	1	0
H	0	0	0	0	0	0	0	0

Figure 2 Example of Layers of the Area

As shown in Figure 2, we organize sensors in the area and its neighborhood into different layers. For each layer, we assign a number to it according to its minimal distance to the boundary of the area. The boundary will be layer 1, and the neighborhood of the area will be layer 0. With the help of layers we design following strategy for travel:

Start from any sensor s_i of layer *k* of the area (max layer is *m*)

1. visit the layer k in specific direction, until there is no node to visit;
2. if k > 0, jump to k-1 layer, perform same travel to step 1. Loop until there is no sensor node to visit in layer 0;
3. if s_i 's layer < m, then jump to k+1 layer, perform same travel to step 1. Loop until there is no sensor node to visit in layer m.

The jump action between layers is easy to perform as the distance between sensor nodes of layers is fixed. And this strategy ensures that we only whole area once with addition of several jumps between layers, whose cost is equal to communicate between two neighbor sensor nodes. And some details may change in different situations. For example, in step 2 we do not need to visit layer 0 when travel after the decreasing bound is broken, or when travel is triggered by the increasing of neighbor sensors.

4.3 Update the Boundary and Bounds

After travel all nodes in the area, we get the sum of changes. If the sum does not exceed the bound of the area, we ignore it and do not send any message to base station. On the other side, as the shape of the area changes, we have to update the boundary to base station, and then update bounds for sensors.

Update the Boundary

Before update the boundary to base station, we have to consider which sensor nodes will be added into or removed from the area. In different situations, the processing is different.

Create new Area triggered by Individual Sensor Node

When the increasing bound of an individual sensor is broken, a new area is created. We could also try to extend the new area by including the sensor's neighbor, which is fairly easy. Finally report all sensor nodes in new area to base station.

Add new sensor node triggered by Sensor Node in the Area

The sensor triggered this update is added to the area. However, it may also extend to its neighbors, when the increasing amount is big enough. In detail, keep extending when

$$\frac{Sum(A(t')) + V(s_i, t)}{n(A(t')) + 1} \geq k$$

$Sum(A(t'))$ is the new sum of sensor reading, whose initial value can be easily got before travel to layer 0 according to

$$Sum(A(t')) = Sum(A(t)) + \Delta$$

Δ is the sum of change of the area. After add a new sensor node to the area, the sum of the area is updated as

$$Sum(A(t')) = Sum(A(t)) + s_j$$

s_j is the newly added sensor node. Finally report all newly added sensor nodes to base station.

Add new sensor node triggered by Neighbors

In this case we travel along layer 0, check whether the node could be added according to same equation in last situation. After adding all nodes in layer 0, each newly added node's neighbor should also be checked according to the same way. Finally report all newly added sensor nodes to base station.

Remove sensor node

In general, if there is need to remove some node, remove the one breaking the bound will work. After removing it, all other nodes do not break their bound, and therefore fulfill the query. However, sometime a node can not be simply removed, because the left area may not fulfill the query (we call this kind of sensor node is the *critical sensor node*). Consider that situation, suggest s_j is a *critical sensor node*, then we have

$$Sum(A(t)) - V(s_i, t) - (n(A(t)) - 1) \times k < 0$$

Which is equal to

$$0 < Sum(A(t)) - n(A(t)) \times k < V(s_i, t) - k$$

We can use the inequality to identify it.

And similarly consider an area A not fulfilling the query, which means

$$Sum(A(t)) - n(A(t)) \times k < 0$$

Suggest that after removing node s_k, the area will fulfill the query, then

$$Sum(A(t)) - V(s_k, t) - (n(A(t)) - 1) \times k > 0$$

Therefore we have

$$0 > Sum(A(t)) - n(A(t)) \times k > V(s_k, t) - k$$

According to above inequality, we can identify which sensor nodes should be removed when the area does not qualify the query condition. And extending to finding multiple sensor nodes to remove is fairly easy.

Processing at the Base Station

The change of boundary can be reported as the newly included or removed sensor nodes of the area. At the base station,

1. if the node is reported as newly included into the area, add it as the boundary, and remove nodes which are the neighbor of the reported one and on last time's boundary from the boundary set.
2. if the node is reported as newly removed from the area, besides removing it from the boundary set, also add nodes which are the neighbor of the reported one and in last time's area to the boundary set.

Update the Bounds

After update boundary changes to base station, the bounds of the area should be correspondingly updated. As the base station may not have the newest sensor reading of all sensor nodes, we may try to infer it from the changes, or some times even perform the recalculation process.

Recalculation

When there is new area created, we should recalculate the bounds for it. New area could be created when (1) an individual sensor node breaks its increasing bounds, or (2) after removing some nodes from the existing area. New area's bound can only be calculated from the scratch because there is no history data to utilize. So the base station should trigger a traveling for each area to collect the sum of sensor reading of each area, as well as neighbor's sensor readings, in order to calculate new bounds.

Updating Bounds

When there is no new area created, but only adding or removing sensor nodes to the area, the bounds can be updated based on history data. As mentioned before, before travel to layer 0, we could easily get the new sum of sensor readings of the area ($Sum(A(t'))$). We could also report this value back to the base station, and utilize it as follows.

If some node s_j is added into/removed from the area, the new sum of sensor readings of new area ($Sum(A'(t'))$) can be calculated as

$$Sum(A'(t')) = Sum(A(t')) \pm V(s_j, t')$$

Then we can compute new increasing bound and decreasing bound for new area as described in section 4.1. Noted that in order to compute the new increasing bound, we may also order base station to collect the sensor reading of neighbors of new area.

After the calculation of new bounds, we can allocate it to each sensor node and spread them along with routing tree of the wireless sensor network.

5. FUTURE WORK

In former sections of the paper, we presented a suppression solution for spatial queries for areas. Its basic idea is to suppress unnecessary checking and updating when the changes are small or not enough to modify the shape of the area. Currently it is a preliminary solution for the problem, and there are directions to further optimize it.

Optimize the Allocation of Safe Bounds

As mentioned in section 3.2, we need an algorithm to allocate the overall safe bound of the area to its sensors. We currently uniformly allocate the safe bound in the solution presented in section 4. However, there are facts indicating that uniform allocation may not be optimal [13]. So we should build an analytical cost model of the energy cost and safe bound, and consider optimization strategies based on it. In particular, mathematical solution may not be able to apply due to the efficiency or complexity, so that we may also need to consider possible approximate algorithm with acceptable computation cost.

Optimize the Checking Process

When a sensor's safe bound is broken, we need to invoke a checking process to valid whether it also breaks the overall safe bound of the area. In practical, several sensors may break their bounds together at the same time, along with many changes eliminating each other. Therefore a method guiding the traveling and checking process in consideration of that will help to avoid many unnecessary communications. This could also be another possible direction to further optimize the checking process.

And on the other hand, implementing a simulation system to study the performance of current solution is also helpful in further optimize the solution. We plan to build our simulation program in NS2 simulation system in future.

6. REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," IEEE Communications Magazine, vol. 40, no. 8, pp. 102–114, 2002.

[2] R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin, "Habitat monitoring with sensor networks," Communications of the ACM, vol. 47, no. 6, pp. 34–40, June 2004.

[3] S.R. Maden, M.J. Franklin, J.M. Hellerstein, and W. Hong, "TinyDB: An Acquisitional Query Processing System for Sensor Networks," ACM TODS, vol. 30, no. 1, pp. 122-173, 2005.

[4] D.J. Abadi, S. Madden, and W. Lindner, "REED: Robust, Efficient Filtering and Event Detection in Sensor Networks," Proc. of VLDB'05, pp. 769-780, 2005.

[5] U. Srivastava, K. Munagala and J. Widom, "Operator Placement for In-Network Stream Query Processing," Proc. of PODS'05, pp. 250-258, 2005.

[6] A. Silberstein, R. Braynard, and J. Yang, "Constraint Chaining: On Energy-Efficient Continuous Monitoring in Sensor Networks", Proc. of SIGMOD'06, pp. 157-168, 2006.

[7] T. S. Rappaport, "Wireless Communications: Principles and Practice," Prentice Hall, 1996.

[8] C. Intanagonwiwat, R. Govindan, and D. Estrin, “Directed diffusion: A scalable and robust communication paradigm for sensor networks,” Proc. ACM MobiCom’00, pp. 56–67, Aug. 2000.

[9] H. Gupta, V. Navda, S. Das, and V. Chowdhary, "Energy-Effecient Gathering of Correlated Data in Sensor Networks," Proc. of ACM Intl. Symp. on Mobile Ad Hoc Networking and Computing Computing, pp. 402-413, 2005.

[10] Y. Kotidis. "Snapshot Queries: Towards Data-Centric Sensor Networks". In Proc. of ICDE'05, pp. 131-142, 2005.

[11] A. Jain, E. Chang, and Y. Wang, "Adaptive Stream Resource Management Using Kalman Filters," Proc. of SIGMOD'04, pp. 11-22, 2004.

[12] A. Deligiannakis, Y. Kotidis, and N. Roussopoulos, "Compressing Historical Information in Sensor Networks," Proc. of SIGMOD'04, pp. 527-538, June 2004.

[13] X. Tang and J. Xu, "Extending Network Lifetime for Precision-Constrained Data Aggregation in Wireless Sensor Networks," Proc. of Infocom'06, pp. 1-12, 2006.

[14] M.L. Yiu, N. Mamoulis and S. Bakiras, "Evaluation of Spatial Pattern Queries in Sensor NetWorks," HKU CS Tech Report TR-2007-02, 2007.

Precise Modeling of Saturation Throughtput of IEEE 802.11 Point-to-Point Link

Yong Yan, Xiaowen Chu
Department of Computer Science, Hong Kong Baptist University
Email: {yyan, chxw}@comp.hkbu.edu.hk

***Abstract* - Wireless mesh networks have attracted extensive research interests in recent years. With the maturity and pervasive existence of IEEE 802.11a/b/g technology, 802.11 protocol is considered as a promising protocol used for constructing the backbone of wireless mesh networks. In a multi-channel multi-interface wireless mesh network, point-to-point 802.11 wireless link can provide the highest throughput, and it is therefore critical to understand the 802.11 throughput performance in a point-to-point configuration. This paper presents a simple yet precise Markov model for the analysis of point-to-point 802.11 link performance in terms of saturation throughput. Different from previously proposed analytical models, our model is "precise" in the sense that we do not make any approximation. Our analytical model is validated by computer simulation results, for both 802.11b and 802.11g configurations. Our analytical results also show that the current default setting of $CW_{\min}$ cannot achieve the best performance. By configuring $CW_{\min}$ to a suitable value, the system throughput can be improved by 2.5% and 5.5% for 802.11b and 802.11g respectively.**

I. Introduction

Wireless mesh networks (WMNs) are gaining significant progress in both academia research and commercial deployment in recent years [6, 7]. It has been shown that the aggregated system throughput can be significantly improved by exploiting multi-channel and multi-interface technique [8, 9, 10]. A typical wireless mesh network is shown in Fig. 1, whose backbone consists of a set of wireless mesh routers. Wireless stations (i.e., end users) can access the Internet by associating with a nearby wireless mesh router. With a suitable channel assignment scheme and the help of directional antenna, it is possible that two nearby wireless mesh routers are connected by dedicated point-to-point wireless channel. This can greatly increase the capacity of wireless mesh backbone because each channel is shared by only two wireless mesh routers.

IEEE 802.11 is almost pervasively used by wireless LANs. The supported data rate of 802.11 has also been increased from 11Mbps (802.11b) and 54Mbps (802.11a/g) to 300Mbps or even 600Mbps (802.11n). It is therefore very promising to use IEEE 802.11 protocol in wireless mesh networks. Due to the limited number of wireless channels, it is very critical to improve the utilization of the scarce wireless spectrum. Our paper aims to analyze the saturation throughput of IEEE 802.11 point-to-point link, and also to find a suitable system parameter which can lead to the best throughput.

The performance of IEEE 802.11 has been actively studied in the last years. Most of them focus on the scalability of 802.11 DCF, i.e., how to handle a large number of wireless stations. Cali et al. [4] derives an analytical model for 802.11 DCF using a *p*-persistent backoff scheme to approximate the original binary exponential backoff scheme. Bianchi [5] proposes a Markov chain model to derive the saturation throughput by assuming a constant and independent collision probability of a packet transmitted by each station. This assumption is accurate only when the number of stations in the wireless LAN is fairly large, however.

Our paper aims to propose a precise analytical model to calculate the saturation throughput of an 802.11 point-to-point link. The main difference between our model and existing models is that, our model is precise without making the assumption that the collision probability is constant and independent. The limitations of our model are: (1) we can only model a point-to-point link, i.e., there are only two senders in the WLAN; (2) our model assumes $CW_{\max} = \mathbf{2}CW_{\min}$. This assumption is reasonable because in a point-to-point link, the saturation throughput is almost a constant value for all values of $m \geq \mathbf{1}$ where $CW_{\max} = \mathbf{2}^m CW_{\min}$.

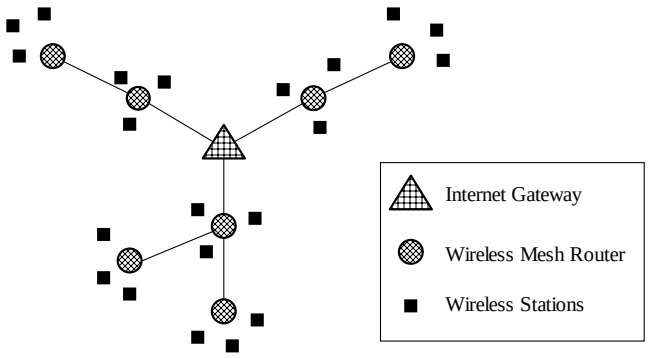


Figure 1: A multi-channel multi-interface wireless mesh network

The rest of the paper is organized as follows. Section II briefly introduces the 802.11 Distributed Coordination Function (DCF) protocol. Section III presents a Markov model used to calculate the saturation throughput for a point-to-point 802.11 DCF link. In Section IV, we validate our model by comparing with simulation results, for both 802.11b and 802.11g configurations. Section V discusses the impact of

CW_{min} on the performance of 802.11 point-to-point link. Finally, Section VI concludes the paper.

II. Background

The IEEE 802.11 standard is working on both the physical (PHY) and medium access control (MAC) layers of the network. Other than considering about the physical details, we will concentrate on the MAC layer protocol itself.

The basic access method in the 802.11 MAC protocol is DCF (Distributed Coordination Function) known as carrier sense multiple access with collision avoidance (CSMA/CA) [1]. DCF employs a distributed CSMA/CA algorithm and an optional virtual carrier sense using RTS and CTS control frames. When using the DCF, before initiating a transmission, a station senses the channel to determine whether another station is transmitting [4]. If the medium is found to be idle for an interval that exceeds the Distributed InterFrame Space (DIFS), the station proceeds with its transmission. However if the medium is busy, the transmission is deferred until the ongoing transmission terminates. A random interval, henceforth referred to as the backoff interval, is then selected; and used to initialize the backoff timer. The backoff timer is decreased as long as the channel is sensed idle, stopped when a transmission is detected on the channel, and reactivated when the channel is sensed idle again for more than a DIFS. The station transmits when the backoff timer reaches zero. CSMA/CA is a strategy that intends to avoid collisions, but it can not eliminate collisions. When more than one node are counting down their backoff timers, there's a probability that some of them have their timers reach zero at the same time slot, and start transmitting at the beginning of next time slot simultaneously, which causes a collision. The collision probability increases with the number of active senders in the network. Fig. 2 illustrates the mechanism of DCF. [4]

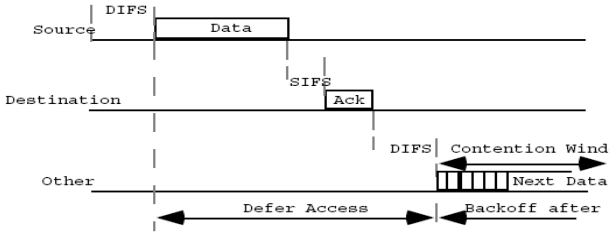


Figure 2: DCF: the basic access mechanism of 802.11 networks. [1]

DCF requires each sender to wait for a random backoff period after the channel is idle for DIFS, it adopts a slotted binary exponential backoff scheme. The backoff time is calculated as below,

$$BackoffTime = Random() \times aSlotTime$$

where *Random()* indicates a uniformly distributed random integer between $[0, CW-1]$ where CW represents the value of contention window, which starts from CW_{min}, doubled each time a retransmission occurs, until reaching the maximum value CW_{max}. Fig. 3 shows the basic mechanism of Contention Window, and Fig. 4 shows an example of the increase of CW value.

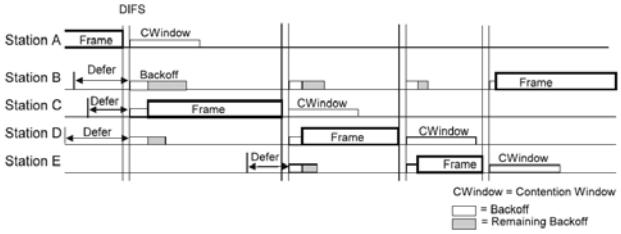


Figure 3: The basic mechanism of Contention Window

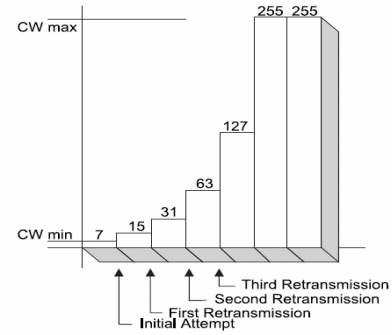


Figure 4: An example of exponential increase of *CW*

III. The Markov Model

The simplest Ad Hoc network consists of two nodes, that is, a Point-to-Point network. The pre-assumption of our model is that only two nodes are in the network, each of them always having packets to send to the other one directly, and of more importance is that, the probability density function of both nodes' backoff timers are uniformly distributed in the same range, which means, both nodes are competing for the channel equally. In this model, we set the CW_{max} as twice of CW_{min}, which indicates a 2-level exponential backoff scheme. Although in the 802.11g standard it is a multi-level exponential backoff scheme with the CW_{max} value of 1023, we will show that there is negligible difference between 2-level and multi-level schemes.

A. The Proposed Model

We use a Markov chain to model the network of two senders. This model focuses on the difference between two backoff timers. Let each state in the markov chain represent the current absolute difference, in time slots, between the two backoff timers of these two senders. For instance, state *i* represents that currently, one node has a backoff timer, longer than the other's *i* time slots. Apparently, collision will happen at state 0, because the two senders have the same backoff timers, and every non-zero state implies that a successful transmission is just ahead. Fig 5 shows the Markov chain, we divide the states into two levels, the states from 0 to $CW_{min}-1$

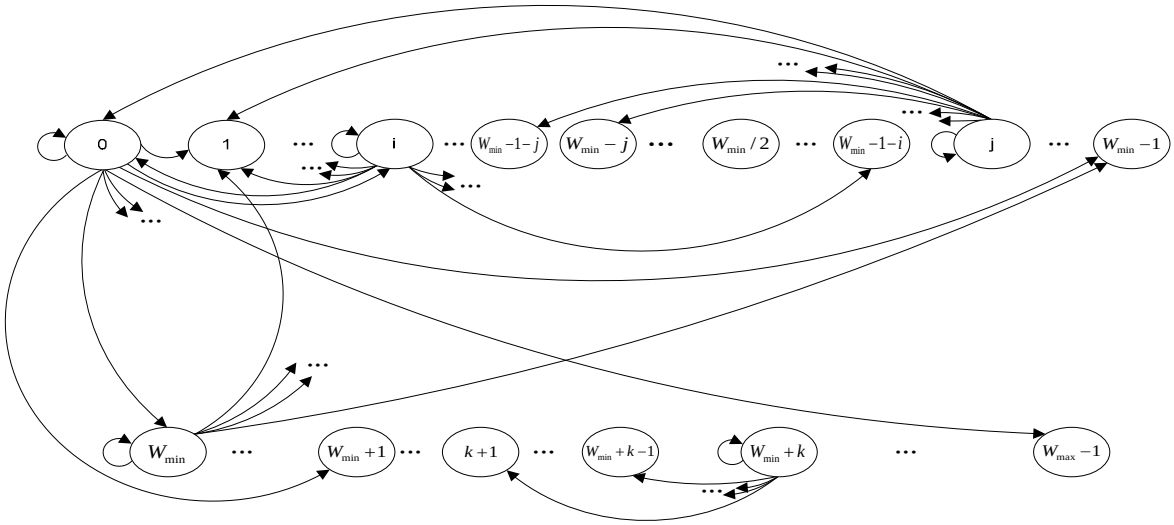


Figure 5: The markov model ($CW_{\min}$ and $CW_{\max}$ are expressed as $W_{\min}$ and $W_{\max}$)

belong to the Low-Level, and the High-Level includes states from $CW_{\min}$ to $CW_{\max}-1$.

Let $P\{j \mid i\}$ denote the transition probability from state *i* to state *j*, the transition probabilities are:

$$P\{j \mid i\} = 1/CW_{\max} \quad i = 0,\ j = 0 \tag{1}$$

$$P\{j \mid i\} = 2(CW_{\max} - j)/CW_{\max}{}^2 \quad i = 0 \quad j \in [1, W_{\max} - 1] \tag{2}$$

$$P\{j \mid i\} = 1/CW_{\min} \quad i \in [1, CW_{\min} - 1] \quad j = 0 \tag{3}$$

$$P\{j \mid i\} = 2/CW_{\min} \quad i \in [j, CW_{\min} - 1 - j],\ j \in [1, (CW_{\min}/2) - 1] \tag{4}$$

$$P\{j \mid i\} = 1/CW_{\min} \quad i \in [1, (CW_{\min}/2) - 1],\ j \in [i+1, CW_{\min} - 1 - i] \tag{5}$$

$$P\{j \mid i\} = 1/CW_{\min} \quad i \in [CW_{\min}/2, CW_{\min} - 1],\ j \in [CW_{\min} - i, i] \tag{6}$$

$$P\{j \mid i\} = 1/CW_{\min} \quad i \in [CW_{\min}, CW_{\max} - 1],\ j \in [i - CW_{\min} + 1, i] \tag{7}$$

At the Low-Level, since state 0 represents collisions, equation (1) and (2) account for the process following a collision, in particular, $P\{0 \mid 0\}$ is the probability that system encounters two consecutive collisions. Equation (3) represents the process that a successful transmission is followed by a collision. Equation (4) accounts for the process of backward transition, which means a new random backoff timer makes a new difference that is smaller than the previous difference. In contrast, Equation (5) and (6) are the forward transition.

Once the state is at $CW_{\min}$ or higher, the system is at the High-Level. A non-zero state always implies a successful transmission ahead, which will be followed by a selection of random number between 0 and $CW_{\min} - 1$, therefore, all states in the High-Level can only have backward transitions to their previous $CW_{\min}$ states, which is represented by Equation (7).

```
      i:
j:      0,1,2,3,4,5,6,7,8,9,0,1,2,3,4,5,6,7,8,9,0,1,2,3,4,5,6,7,8,9,0,1
 0  {32*Z,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0},
 1  {62*Z,X,X,X,X,X,X,X,X,X,X,X,X,X,X,I,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0},
 2  {60*Z,I,X,X,X,X,X,X,X,X,X,X,X,X,X,I,I,I,1,0,0,0,0,0,0,0,0,0,0,0,0,0},
 3  {58*Z,I,I,X,X,X,X,X,X,X,X,X,X,I,I,I,I,I,1,0,0,0,0,0,0,0,0,0,0,0,0,0},
 4  {56*Z,I,I,I,X,X,X,X,X,X,X,X,X,I,I,I,I,I,I,I,1,0,0,0,0,0,0,0,0,0,0,0},
 5  {54*Z,I,I,I,I,X,X,X,X,X,X,I,I,I,I,I,I,I,I,I,1,0,0,0,0,0,0,0,0,0,0,0},
 6  {52*Z,I,I,I,I,I,X,X,X,X,I,I,I,I,I,I,I,I,I,I,1,0,0,0,0,0,0,0,0,0,0,0},
 7  {50*Z,I,I,I,I,I,I,X,X,I,I,I,I,I,I,I,I,I,I,I,1,0,0,0,0,0,0,0,0,0,0,0},
 8  {48*Z,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,1,0,0,0,0,0,0,0,0,0,0},
 9  {46*Z,I,I,I,I,I,I,I,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,1,0,0,0,0,0,0,0,0},
10  {44*Z,I,I,I,I,I,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,I,1,0,0,0,0,0,0,0},
11  {42*Z,I,I,I,I,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,1,0,0,0,0,0},
12  {40*Z,I,I,I,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,1,0,0,0},
13  {38*Z,I,I,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,1,0,0},
14  {36*Z,I,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,1,0},
15  {34*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I,0},
16  {32*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,I,I},
17  {30*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I,I},
18  {28*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I,I},
19  {26*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I,I},
20  {24*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I,I},
21  {22*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I,I},
22  {20*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I,I},
23  {18*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I,I},
24  {16*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I,I},
25  {14*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I,I},
26  {12*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I,I},
27  {10*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I,I},
28  { 8*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I,I},
29  { 6*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I,I},
30  { 4*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I},
31  { 2*Z,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,I},
```

Figure 6: Transition Probability matrix

To be exact, the transition probability matrix of this markov model is shown in Fig. 6. Note that the value in this matrix is set to represent a network whose $CW_{\min}$ is 16 and $CW_{\max}$ is 32. In this matrix, I represents the value $1/CW_{\min}$, X represents $2/CW_{\min}$, and Z represents $1/CW_{\max}{}^2$. There are $CW_{\max}$ equations and $CW_{\max}$ unknown parameters in this matrix. If we denote each entry in this matrix as $m[i, j]$, and the solution to this matrix is $q[i]$, these equations can be written by:

$$q[j] = \sum_{i=0}^{CW_{\max}-1} q[i]m[i, j], \quad j \in [0, CW_{\max} - 1] \tag{8}$$

According to the property of Markov chain, this matrix can be easily solved by using the recursive manner. The solution *q[i]* means the probability that the difference between two backoff timers is *i* time slots.

B. System Idle time Distribution

We denote *X* and *Y* as the random variable of both random backoff timers. At the Low-Level, both nodes have their uniform distribution of random backoff timers: $f_x(X) = f_y(Y) = 1/CW_{\min}$, where X and Y range from 0 to $CW_{\min} - 1$. By solving the markov matrix, we can get the probability distribution of the difference of backoff timers in two levels; we can denote the solution $q[i]$ as $f_i(|X'-Y'|)$.

With the above solution, now we are ready to calculate system throughput. In terms of throughput analysis, what really matters is the system idle time cost by the random backoff period, and this idle time always equals the shorter backoff timer at each transmission round. In other words, we are about to use the solution $f_i(|X'-Y'|)$ to find the probability $p_j(Min|X'-Y'|)$, p_j is the probability that system is idle for *j* time slots in one transmission round, it can be calculated as following:

$$p_0 = q[0]/CW_{\min} + (1-q[0])/CW_{\min} = 1/CW_{\min} \tag{9}$$

$$p_j = \sum_{i=0}^{j-1} q[i]/CW_{\min} + q[j](CW_{\min} - 1 - j)/CW_{\min} + q[0](1/CW_{\max}{}^2 + 2(CW_{\max} - 1 - j)/CW_{\max}{}^2)$$

$$j \in [1, CW_{\min} - 1] \tag{10}$$

$$p_j = 2(CW_{\max} - 1 - j)/(CW_{\min}CW_{\max}{}^2) + (CW_{\max} - 1 - j)/(CW_{\min}CW_{\max}{}^3) \quad j \in [CW_{\min}, CW_{\max} - 1] \tag{11}$$

Here, with $p_j(Min(X'-Y'))$ we got the probability distribution of system idle time cost by random backoff.

C. Throughput Analysis

In this paper, throughput is defined as the percentage of the time cost by transmitting successful payload $T_{SucPayload}$, in the overall time T_{total}, which includs T_{suc}, the time cost by successful packets, and T_{col}, the time cost by collision packets, i.e.,

$$Throughput = \frac{T_{SucPayload}}{T_{total}} = \frac{N_{SucPacket} \times Payload / LinkRate}{T_{suc} + T_{col}} \tag{12}$$

where $N_{SucPacket}$ $Payload$, T_{suc} and T_{col} can be calculated by:

$$N_{SucPacket} = TotalRound \times (1 - q_0) \tag{13}$$

$$Payload = LLCHdr + IPHdr + UDPHdr + UDPpayload \tag{14}$$

$$T_{col} = TotalRound \times q_0 \times \sum_{i=0}^{CW_{\max}-1} p_i(i \times SlotTime + TC) \tag{15}$$

$$T_{suc} = TotalRound \times (1 - q_0) \times \sum_{i=0}^{CW_{\max}-1} p_i(i \times SlotTime + TS) \tag{16}$$

where TC and TS are the basic time cost in collision and successful transmissions, excluding the backoff time.

For 802.11b, TC and TS can be expressed as:

$$TC = T_{PLCP} + MAC\_PSDU / LinkRate + DIFS + PROP \tag{17}$$

$$TS = TC + SIFS + T_{PLCP} + ACK\_PSDU / AckRate + PROP \tag{18}$$

$$MAC\_PSDU = Header(MAC + LLC + IP + UDP) + UDPpayload + FCS \tag{19}$$

$$T_{PLCP} = PLCPpreamble / PreambleRate + PLCPheader / HeaderRate \tag{20}$$

For 802.11g, TC and TS can be expressed as:

$$TC = PLCP / PlcpRate + N_{symbol} \times SymbolRate + DIFS + PROP \tag{21}$$

$$TS = TC + SIFS + PLCP / PlcpRate + ACK\_PSDU / AckRate + PROP \tag{22}$$

$$MAC\_PSDU = Header(MAC + LLC + IP + UDP) + UDPpayload + FCS + Service + Tail \tag{23}$$

$$N_{symbol} = \lceil MAC\_PSDU / SymbolSize \rceil \tag{24}$$

where $SymbolRate$ is 4μs/symbol and $SymbolSize$ is 216 bits in the 802.11g standard, and N_{Symbol} accounts for the number of symbols needed to encode a MAC layer data unit, that is, a MAC_PSDU. 802.11g encodes every 216 bits into one symbol when the Link rate is 54Mbps.

We now can get the system throughput by using (13)-(24) to replace (12).

IV. Model Validation

A. Throughput Analysis of 802.11b

In this paper, our parameters in simulation and model are all set according to the 802.11 standard. Table 1 lists some of the PHY and MAC parameters used in 802.11b standard.

Fig. 7 presents the throughput analysis results for 802.11b point-to-point network. We can observe that our model result is very close to simulation result. We also compare our model to Bianchi's model [5] in terms of throughput, under the same set of PHY and MAC parameters as in the IEEE latest standard. It is shown that our model exactly approaches the simulation result, while Bianchi's model has a little deviation, because Bianchi's model assumes a constant and independent collision probability.

Meanwhile, in Fig. 7 we also present the throughput when CW_{max} is set to 1024, as same in the IEEE standard. Although very close, the throughput of 1024-CW_{max}, is a little lower than the throughput of 64-CW_{max}. This result means in a Point-to-Point link, using lager CW_{max} will reduce the system throughput.

Table 1 IEEE 802.11b PHY Characteristics [1]

SlotTime	20 μs
SIFSTime	10 μs
DIFS	28μs (2×SlotTime+SIFS)
aCWmin	32
aCWmax	1024
PLCP Preamble	72 bits
PreambleRate	1 Mbps
PLCP Header	48 bits
HeaderRate	2 Mbps
MAC_Header	192 bits
Supported Rates	1,2,5.5,11 Mbps
LinkRate	11 Mbps
AckRate	2 Mbps

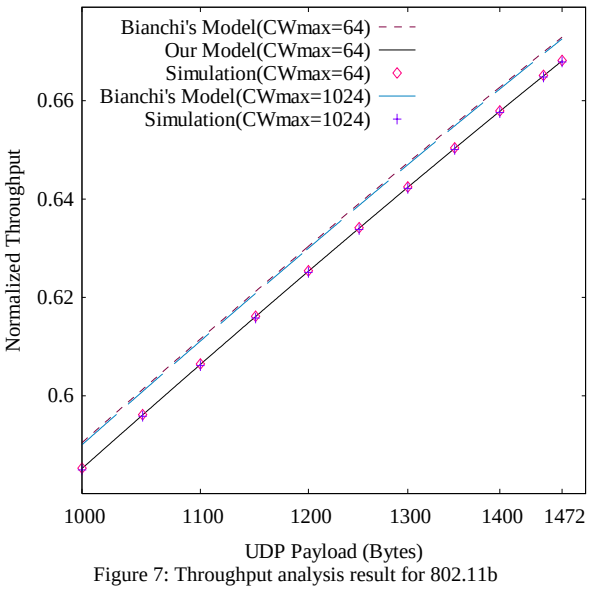


Figure 7: Throughput analysis result for 802.11b

B. *Throughput Analysis of 802.11g*

The IEEE 802.11g standard brings changes and additions to IEEE Std 802.11, 1999 Edition. Table 2 lists part of the IEEE 802.11g PHY and MAC parameters.

Fig. 8 shows the throughput results of 802.11g. Similar to 802.11b, our model is precise, while Bianchi's model is not very close to the simulation result; the throughput of 1024-CW_{max}, is a little lower than the throughput of 32-CW_{max}. A special phenomenon in this figure is that the curve is sawtooth, this is due to the feature of symbol encoding in 802.11g. When sending a packet, 802.11g encodes every 216 bits into one symbol; when the payload is not enough for a multiple of 216 bits, 802.11g adds padding bits. As long as the payload increases, every time when it is needed to use one more symbol for encoding, the new symbol will only carry one byte payload and 208 padding bits. Since throughput calculation should not include the padding bits, thus, sawtooth will appear periodically.

Table 2 IEEE 802.11g PHY Characteristics [3]

SlotTime	20μs (9μs for short)
SIFSTime	10 μs
DIFS	28μs (2×SlotTime+SIFS)
aCWmin	16
aCWmax	1024
PLCP	20 bits
Service+Tail	16 bits + 6 bits
MAC_Header	192 bits
Supported Rates	1,2,4,5.5,6,9,11,12,18,24,36,48, and 54 Mbp/s
LinkRate	54 Mbps
AckRate	24 Mbps
SymbolRate	4 μs/symbol
SymbolSize	216 bits

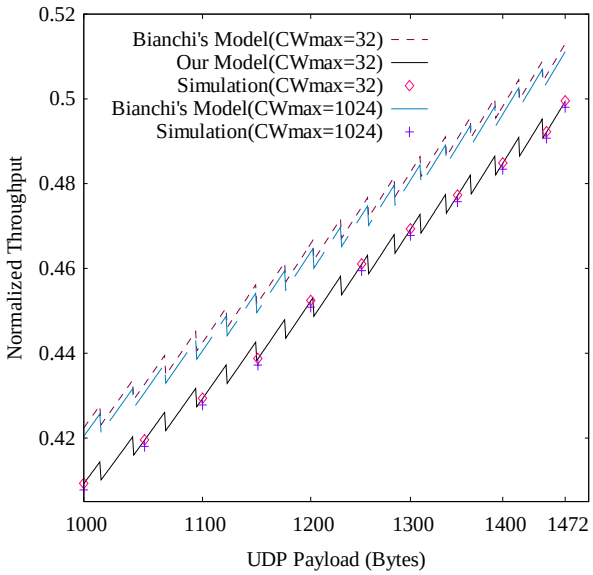


Figure 8: Throughput analysis result for 802.11g

C. Random Backoff Interval Results

Fig. 9 compares our model to the simulation and result of system idle time cost by the random backoff. It is seen that our model is almost identical to the simulation result.

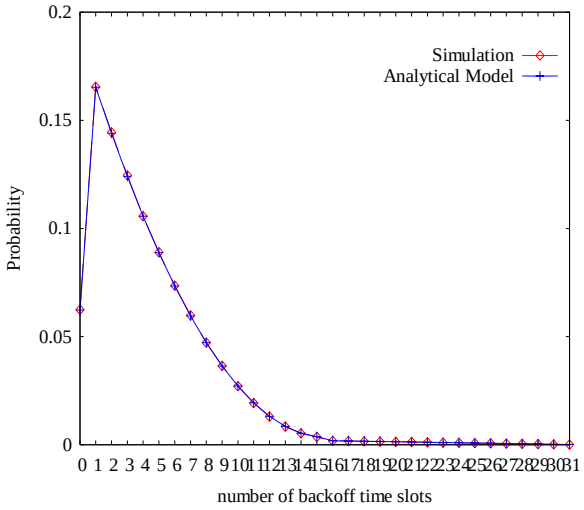


Figure 9: Comparison of system idle time distribution

V. Optimal Contention Window Size

Random backoff period is a critical factor to the system throughput, and the backoff period itself is largely dependent on the Contention Window size, i.e., the value of $CW_{\min}$ and $CW_{\max}$. Smaller contention window brings a higher probability of collision, while larger contention window causes longer delay. To achieve best network performance, care should be taken to set the $CW_{\min}$ and $CW_{\max}$. In the current IEEE standard, $CW_{\max}$ is set to 1024, and $CW_{\min}$ is set to 32 for 802.11b, and 16 for 802.11g.

In order to analyze the link performance of point-to-point Ad hoc network, we also tested the system throughput under different CW values in 802.11b and 802.11g, the simulation results are presented in Fig. 10 and 11.

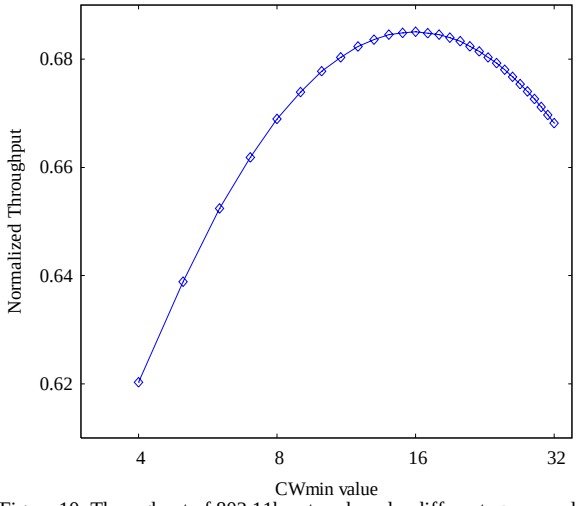


Figure 10: Throughput of 802.11b network under different $CW_{\min}$ values

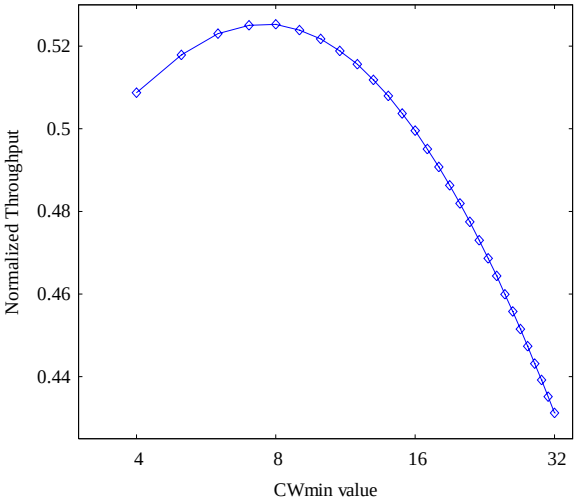


Figure 11: Throughput of 802.11g network under different $CW_{\min}$ values

We can observe that the value $CW_{\min}$ strongly influences the system throughput. For an 802.11b point-to-point network, the best throughput can be achieved when $CW_{\min}$ is set to 16, which produces a 2.5% improvement to the IEEE 802.11b standard $CW_{\min}$ value. For 802.11g, the best $CW_{\min}$ value is 8, with an improvement of 5.5% to the IEEE 802.11g standard $CW_{\min}$.

VI. Conclusions

In this paper, we proposed a precise markov model to analyze link performance in IEEE 802.11 Point-to-Point networks. This model focuses on the system idle time cost by the random backoff period. A Two-level exponential backoff scheme is modeled where $CW_{\max}$ is set to twice of $CW_{\min}$, and we demonstrate that this model works well even though the actual $CW_{\max}$ is several times of $CW_{\min}$. Simulation results prove that our model matches both of them. Meanwhile, by

this model we can also get the theoretical maximum system throughput. At last, we argue that the $CW_{\min}$ value could be set to achieve best performance in point-to-point 802.11 networks.

REFERENCES

[1] ANSI/IEEE Standard 802.11, 1999 Edition, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.

[2] IEEE Standard 802.3 - 2005 edition.

[3] IEEE Standard 802.11g - 2003 edition.

[4] F.Calì, M. Conti, E. Gregori, "IEEE 802.11 wireless LAN: capacity analysis and protocol enhancement", in Proc. of IEEE Infocom'98, pages 142-149, March 29 - April 2, 1998.

[5] Giuseppe Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function", *IEEE Journal on Selected Areas in Communications*, Vol. 18, NO. 3, pages 535–547, March 2000.

[6] R. Karrer, A. Sabharwal, and E. Knightly, "Enabling large-scale wireless broadband: the case for TAPs," in Proc. of HotNets, Cambridge, MA, 2003.

[7] I. F. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," Computer Networks Journal (Elsevier), vol. 47, no. 4, pp. 445-487, March 2005.

[8] A. Raniwala and T. cker Chiueh, "Architecture and algorithms for an IEEE 802.11-based multi-channel wireless mesh network," in Proc. of IEEE Infocom'05, 2005.

[9] P. Kyasanur and N. Vaidya, "Capacity of multi-channel wireless networks: Impact of number of channels and interfaces," in Proc. of Eleventh ACM MobiCom'05, 2005.

[10] P. Kyasanur, J. So, C. Chereddi, and N. Vaidya, "Multi-Channel mesh networks: challenges and protocols", in IEEE Wireless Communications, April 2006.

[11] J. Shi, T. Salonidis, and E. Knightly, "Starvation mitigation through MultiChannel coordination in CSMA multihop wireless networks," in Proc. of ACM MobiHoc'06, 2006.

Object Tracking Using Information Content

Chang Liu, Pong C Yuen, Guoping Qiu

Abstract

The conceptual model of visual saliency in human vision system has been employed in extracting salient feature from images and multimedia data processing in the last decade. This paper proposes to employ the visual saliency for object detection and tracking. The crucial factor is to compute a saliency map(s) such that visual attention can be performed. This paper proposes a new method for saliency map construction based on information theory and spatial-temporal model, called information saliency map (ISM). The ISM provides rich information content of the video. Moreover, both spatial and temporal information are used to compute the ISM. Object detection and tracking are then performed based on the ISM. Two popular and publicly available visual surveillance databases from CAVIAR and PETS are selected for evaluation. Experimental results show that the proposed method is robust for object detection and tracking in complex background and illumination changes. The average detection rate is 90.35% while the false alarm rate is 2.46% in CAVIAR (INRIA entrance hall) dataset with ground truth data. Comparison between the proposed method with two current state of the art methods, namely mean-shift and Gaussian mixture model, is also reported.

1. Introduction

Object detection and tracking is the first and important step for a computer vision system. It is also one of the most active research areas in computer vision because of the wide range of applications such as visual surveillance, human identification, human behavior recognition, event recognition and traffic congestion control. It is easy to note that object detection and tracking is the core component in all these applications. Therefore, an efficient and robust tracking algorithm under different situations such as illumination change, occlusion, complex background, is required. A number of good object detection and tracking methods have been proposed in the last decade[3, 7, 10, 12, 14, 20, 22]. However, these methods do suffer from a limitation(s) such as manually initialization and sensitive to illumination changes.

The goal of object detection and tracking process is to find out the object(s) of interest and to keep track its motion for further analysis. Moreover, in surveillance applications, we are interested in all types of objects including human, cars, animals and even luggage. From this perspective, the system should detect and track all potentially interested objects. In other words, the surveillance system should be able to narrow down and rank the regions of interest for further processing. This is similar to our human visual system (HVS) which is able to reduce the amount of incoming visual data to a small but relevant amount of information, called visual saliency, for higher level cognitive processing.

The conceptual model of visual saliency has been employed in (image) scene analysis[8, 11]. The crucial factor is how to compute the saliency map. The methods[19, 1, 9, 4] determine the saliency map based on the spatial and temporal information separately. The advantage is computational simple while the saliency map may be sensitive to illumination changes and may not be suitable for tracking purpose. Along this line, this paper presents a new method to construct the saliency map based on information theory and spatial-temporal model. Shannon information theory shows that uniqueness or rarely happened events contain high information while common or frequently happened events imply low information. This is in line with our HVS as well as the visual saliency model. In order to overcome the illumination sensitivity, both spatial and temporal information will be considered as a 3D volume to compute saliency map. This will average out the illumination effect. In turn, the saliency map will be insensitive to illumination changes. In this way, an information saliency map (ISM) is then generated. Object detection and tracking is then performed using the ISM.

The use of ISM for object detection and tracking offers the following advantages:

- Object detection and object tracking are performed simultaneously.
- ISM is insensitive to the illumination changes

- ISM provides levels of saliency, which is an additional information for further (high level) processing

The rest of this paper is organized as follows. Section 2 will give a brief review on the existing methods on object detection and tracking. The proposed method will be discussed in Section 3. Experimental results and conclusion are then presented in Sections 4 and 5 respectively.

2. Previous Works

Many algorithms/methods have been proposed in object detection and tracking. Here, we would loosely categorize into two approaches, namely background substraction[15, 13] and object-based[22, 14].

Background subtraction extracts object position by computing the difference between the current frame and the reference frame which is considered as the background image. Background subtraction approach is widely used because it can detect both the moving objects and static objects. The crucial factor in this approach is how to update the background. The Gaussian Mixture Model(GMM) proposed by Stauffer and Grimson[18] is one of the most popular background updating models. This method models each background pixel by a mixture of K Gaussian distributions. When a new pixel value comes in, the mixture model provides additional information to update the model. Since the original GMM is computational expensive, different efficient updating algorithms [12, 23, 6] have also been proposed.

The object-based is another popular approach. This approach assumed that the object of interest has been segmented out and the appearance information, such as color, histogram of the object can be obtained. In order to track the object, this is equivalent to finding a motion vector indicating where the object has moved. Meanshift [3, 22, 2] is one of the most popular methods in this approach. Particle filter[7, 14, 20, 5] is another popular method to estimate motion vectors. This method samples the posterior distribution estimated in the previous frame and propagating these samples to form the posterior for the current frame. Particle filter can cope well with multimodal and non-Gaussian probability density function of motion parameters. However, it requires a number of samples to find a fair maximum likelihood estimate of the current state.

3. Proposed Method Using Information Content

Based on the conceptual model of visual saliency, this paper proposes to employ the information theory and spatial-temporal model to compute the information saliency map (ISM) which reflects the information content on each pixel. Object detection and tracking are then performed based on the ISM. Unlike existing approaches which separate detection process and tracking process, the proposed method performs object detection and object tracking simultaneously.

3.1. Information theory

Consider a discrete random variable $X \in 1,...,K$, suppose the event $X = k$ is observed, the Shannon's self-information content [17] of this event $I(k)$ is defined as follows,

$$I(k) = \log_2 1/p(X = k) = -\log_2 p(X = k) \quad (1)$$

It means that the information of event k is inversely proportional to the probability of the observation of event k, a rarely happen event contains high information while an event which happen frequently contains low information.

3.2. Information saliency map

In order to compute the information saliency map, we need to calculate the information on each pixel based on Eq (1). In turn, we need to estimate the probability of each pixel. Qiu et al. [16] has proposed a spatial-temporal model to calculate that probability. They divide a frame into a number of blocks with smaller size. For each block, they construct a spatial-temporal 3D volume which considers the current frame and previous frames. Their model is theoretically sound, but computing each block probability in the spatial-temporal 3D volume in DCT domain is computational expensive when the number of previous frames is large. Inspired by the Qiu et al.'s model, we consider each frame in the 3D volume independently and then re-combine 3D volume to compute the ISM.

The structure of ISM can be represented as the following:

$$ISM(t) = \begin{pmatrix} Info(1,1,t) & \dots & Info(1,w,t) \\ \vdots & \ddots & \vdots \\ Info(h,1,t) & \cdots & Info(h,w,t) \end{pmatrix} \quad (2)$$

Where the ISM in time t can be divided into several blocks: $Info(r,s,t), r = \{1,2,...,h\}, s = \{1,2,...,w\}$.

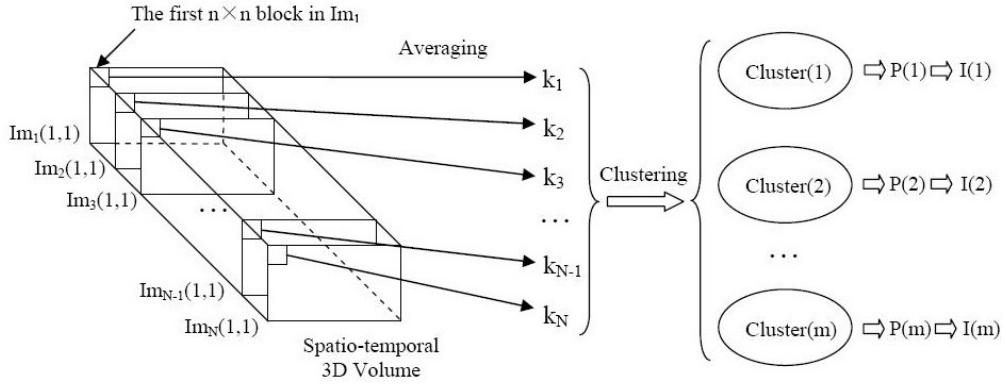


Figure 1. Computing the information saliency map of current frame

The block diagram of the proposed method in computing the ISM is shown in Figure 1. Consider a $N-frame$ (N=20 in this paper) spatial-temporal volume which consists of the current frame and the previous $N-1$ frames. Each frame Im is then divided into a number of blocks with smaller size: $\{Im(1,1), Im(1,2), ..., Im(h,w)\}$. Information saliency map of each block will be computed individually and the information content of block $Im(r,s)$ in time t will be represented by $Info(r,s,t)$, where

$$Info(r,s,t) = -\log[P(B(r,s,t)|V(r,s,t))] \qquad (3)$$

$B(r,s,t)$ represents block $Im(r,s)$ at time t, and $V(r,s,t)$ represents the spatio-temporal volume containing $B(r,s,t)$ as the current image block. The current block probability density function is determined by considering the DC coefficient of each block $\{k_1, k_2, ..., k_N\}$ which can be calculated by the block mean value.

To compute the information I(k) from (1), the probability of variable k needs to be computed from the DC coefficients $\{k_1, k_2, ..., k_n\}$. The straightforward method is to make use of histogram, but it requires a pre-defined bin (histogram) width. Since the probability is calculated on-line and the number of data (N) is 20, the fixed width histogram may introduce large error. Instead, we propose an adaptive method to construct the histogram based on the clustering technique. By clustering the DC coefficients into different clusters, each cluster can be consider as a bin with adaptive bin wide. To do so, k-mean is one of the possible choices. However, k-mean is an iterative algorithm. It would be computational expensive because we need to calculate around 200 blocks for each frame. Therefore, we propose a simple but effective method as shown in Algorithm 1. Suppose K=$\{k_1, k_2, ..., k_N\}$ is a sorted DC coefficients. Two consecutive coefficients will belong to the same cluster if $(|k_i| - |k_{i+1}|)| \,/\, (|k_i| + |k_{i+1}|) < \alpha$, ($\alpha$ = 0.05 in this paper) where $i = 1, 2...N-1$. The probability of each cluster is then computed by dividing the number of coefficients in each cluster by N (total number of coefficients). Then the information content of each block is calculated using (1). The information saliency map is then generated for the current frame. Figure 2(a) shows the frame 70^{th} of the video Browse1 from CAVIAR dataset. The video Browse1 consists of two people browsing around the entrance hall. Figure 2(b) shows the corresponding ISM. It can be seen that the locations of the two people are clearly indicated in the ISM.

Given sorted K=$\{k_1, k_2, ..., k_N\}$,
j=1, cluster(j)=1;
for i=1 to N-1
  if dist$(k_i, k_{i+1}) < \alpha\,(|k_i| + |k_{i+1}|)$
    cluster(j) = cluster(j)+1;
  else
    P(j) = Number of coefficients in cluster(j)/N;
  j = j+1, cluster(j) = 1;

Algorithm 1. PDF computation

3.3. Object tracking using ISM

The larger the value in the ISM, the larger the information at the corresponding position. Basically, object detection and tracking based on the ISM can be performed by defining a threshold value. A typical result is shown in Figure 3 where the upper row shows the original video Browse1 at frame 20, 50 and 70 while the lower row shows the corresponding ISM and tracking results.

(a) Original image (b) ISM

Figure 2. Information saliency map for the 70^{th} frame in video Browse1 in CAVIAR database

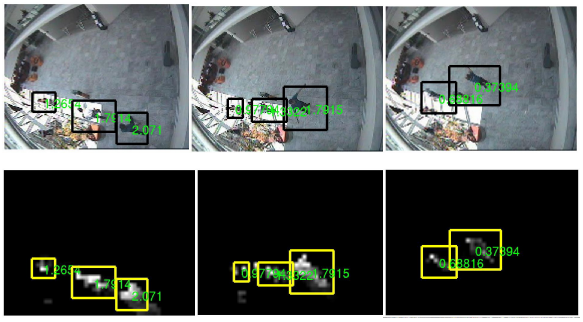


Figure 3. From left to right column:frame 20, 50, 70 from video "*Browse*1" CAVIAR database

However, this straightforward detection method suffers from two limitations. The first limitation is that there may have false detections because of the present of noise. In order to avoid the noise effects in the block information saliency image, we apply an averaging filter to the ISM. The rationale is that if a block within a moving object has a relatively high information saliency value, its neighbor blocks should also have high probability to contain high information. Otherwise, it must be noise. This simple process can remove noise effectively. However, the drawback is that an object with similar size may also be treated as noise.

The second limitation is that the human/object will be lost tracked if he/she changes from moving to stationary. In such a case, the information content will decrease to a smaller value. This can be illustrated using the example in which an object is changed from motion to stationary and starting moving again. The object motion and the corresponding ISM value are recorded and shown in Figure 4. It can be seen that the video can be divided into three video segments, namely motion, stationary and motion. In the stationary part, the saliency value is below threshold and is closed to zero. If we only based on the saliency value, the object will be lost tracked. This drawback can be solved by monitoring the falling edge of the saliency curve. Whenever, there is a falling edge, the object of interest becomes stationary and the area is also classified as a region of interest under tracking. Whenever, we find that there is a rising edge of the saliency curve, that object is in motion again. The video sequence from this curve is show in Figure 5.

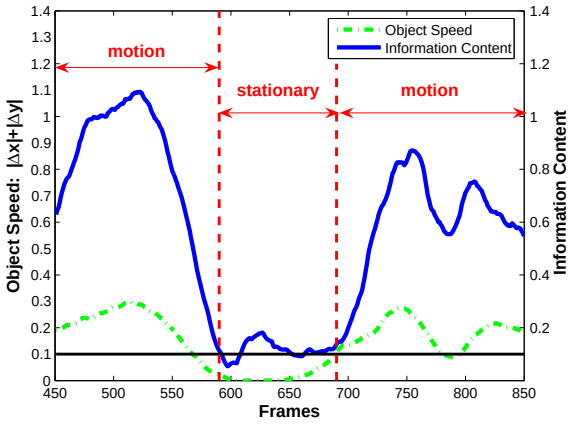


Figure 4. Relation of object speed and object information content, the green dashdotted line represents object moving speed, the blue solid line represents the object information content, the horizontal black solid line represents a predefined information content threshold to identify if an object is static, it is set to be 0.1 here

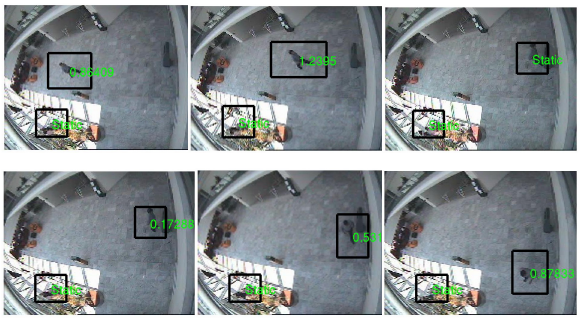


Figure 5. information content tracking, the six figures are respectively from Frame 500, 550, 600, 700, 750, 800 in CAVIAR "*Browse*1" database, the value in the rectangle correspond to the object information content or object status

4. Experimental Results

The experimental results in this section is divided into two parts. First, two popular and publicly available surveillance datasets, namely CAVIAR[25] and PETS[24],are used to evaluate the performance of the proposed method. Second, the proposed method is compared with existing methods. Two recently developed methods, namely improved meanshift[22] and

Gaussian mixture model (GMM)[21], are selected for comparison.

4.1. Evaluation of the proposed method

CAVIAR database[25] consists of 3 datasets, namely INRIA entrance hall, shopping mall frontal and shopping mall side. Only INRIA entrance hall dataset is selected to evaluate our method because only this dataset has the ground truth data. The INRIA entrance hall dataset has six types of events, namely "*Browsing*", "*Fighting*", "*Groups_meeting*", "*Leaving_bags*", "*Rest*" and "*Walking*", totally 28 video sequences. These video sequences are captured from inclined look-down camera with a wide angle. People appear in front of the camera have different sizes and body figures, which make it difficult to detect object. Moreover, the bottom left part region of the video sequences is under severe illumination condition.

The detection rate and the false detection rate of each video sequence are recorded and tabulated in Table 1. The average detection rate is 90.35% while the false detection rate is 2.46%. The results are encouraging. In particular, we would like to point out that our proposed method is able to detect and track objects under both illumination changes and small motion which can be demonstrated using the video "*Fight_OneManDown*". The experimental results are shown in Figure 6 where 3 key frames are selected to show the tracking process. It can be seen that the woman standing at the left side of the image is not a visual salient region until she moves and its information content value exceed a certain value at the 270 th frame. She is kept tracked under our attention even she becomes stationary again. Figure 7 is another example to illustrate the results of our proposed method under illumination changes.

Figure 6. information content tracking, the three figures are respectively from Frame 220, 270, 350 in CAVIAR "*Fight_OneManDown*" database, the value in the rectangle correspond to the object information content

The PETS2001[24] database consists of 20 video sequences. All video were captured at outdoor environment, various objects including humans, bicycles and vehicles. Moreover, there are global illumination changes during a short period of time. Since we would like to test the video sequence with large illumination changes, 4 out of 20 video sequences are selected to evaluate our proposed method. Six frames of one of the serve change of illumination video are shown in Figure 8 (filename:*Dataset*3_*Testing*_1). In this video, the background is under a 20-second global illumination changing process (due to sunlight occluded by cloud). The result shows that our method is robust to global illumination changes. Since no ground truth data is available in PETS2001 database, no statistic on detection rate nor false detection rate is reported.

Figure 7. information content tracking, the three figures are respectively from Frame 570, 630, 690 in CAVIAR "*Meet_Split_3rdGuy*" database, the value in the rectangle correspond to the object information content

Database	TP	FP	TG	FAR	TRDR
Browse	6722	204	7298	2.9%	92.1%
Fight	5060	165	5625	3.2%	90.0%
Meet	7327	165	7815	2.2%	93.8%
LeftBag	7220	140	8702	1.9%	83.0%
Rest	4768	113	5322	2.3%	89.6%
Walk	5277	103	5512	1.9%	95.7%
Average	**1299**	**32**	**1438**	**2.46%**	**90.35%**

Table 1. Tracking results on CAVIAR database, 28 video sequences totally, TP:True Positive, FP:False Positive, TG:Total Ground truth, FAR:False Alarm Rate, FAR=FP/(TP+FP), TRDR:Tracker Detection Rate, TRDR=TP/TG

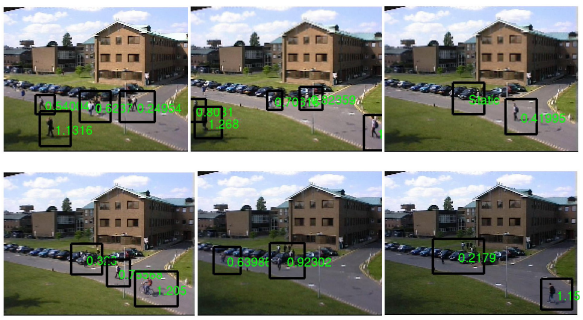


Figure 8. information content tracking, the six figures are respectively from Frame 2720, 2800, 2880, 2930, 3080, 3150 in PETS2001 "*Dataset*3_*Testing*_1" database, the value in the rectangle correspond to the object information content

4.2. Comparing the proposed method with existing methods

The objective of this section is to compare the proposed method with two existing methods, namely improved mean-shift [22] and Gaussian mixture Model [21] using the INRIA entrance hall dataset in CAVIAR.

The improved mean-shift method requires a manually selection of object for tracking. Experimental results show that when the illumination changes is mild, the tracking results are very good. However, when there is a serve change of illumination, the result may not be satisfactory. This can be illustrated using the video Browse4. In the video, a person walks close to the bottom left part where has a strong lighting. The first row of Figure 10 shows the result. It can be seen that at the beginning (before frame 980), the tracking results are good. When the person moves closer (frame 995) to the strong lighting region, the tracking result turns bad. The tracking results are unsatisfactory afterwards (frames 1005, 1025 and 1050). Please refer to the supplementary video for the full tracking result. Since the mean-shift method requires manually selection of object region, the statistics on detection and false detection rates may not be fully reflected the mean-shift algorithm's performance, but also depends on the initialization. Therefore, no statistics on mean-shift method is reported.

The improved adaptive Gaussian mixture model constructs a probability density function for each pixel independently and pixel- level background subtraction is performed to find the region of interest. Experimental results show that GMM is able to model the background very well, even for serve illumination changes. However, GMM suffers from a drawback. When an object is moving slowly or with relatively small motion, GMM may mis-classify that region(s) as background and update the background accordingly. As a result, the object will then be missed. This situation can be illustrated using the example in Figure 10. A human at the bottom left region has a slow motion and stationary at certain period of time in the video. The detection results using GMM are shown in the third row in Figure 10. It can be seen that GMM is not able to locate that region of interest and considers as illumination noise. As a comparison, the results of the ISM generated using our method are shown in the last row in Figure 10. Our method is able to locate the relatively small motion object most of the time. In order to give a quantitative comparison, the ROC curves for the GMM method and the proposed method are plotted in Figure 9. It can be seen that the proposed method outperforms the GMM.

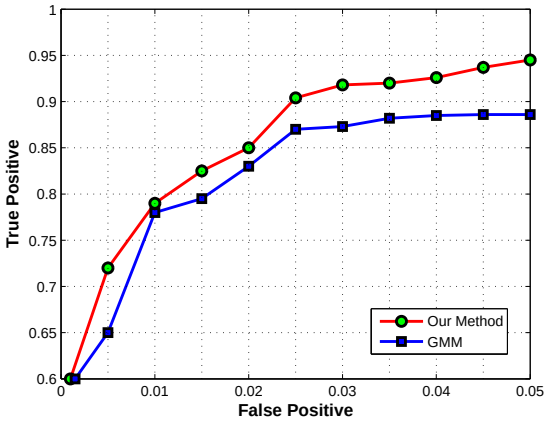


Figure 9. ROC curves for GMM[21] and our proposed method in CAVIAR INRIA entrance hall dataset

5. Conclusions and Future Works

This paper has proposed a new idea for object detection and tracking based on visual saliency. A novel object detection and tracking method based on information theory and spatial-temporal model has been developed and reported. The proposed method determines an information saliency map (ISM) which shows the information content of each pixel. The ISM not only provides the saliency of each pixel for object detection and tracking, but also gives additional higher level object information such as the object motion speed. Two publicly available databases have been selected to evaluate the proposed method and the results are encouraging. The detection rate and false detection rate on CAVIAR INRIA entrance hall dataset are 90.35% and 2.46% respectively. Comparison with two popular approaches, namely mean-shift and Gaussian mixture model, are also reported. Experimental results show that the proposed method is robust to illumination changes and manually object initialization is not required.

Although this paper has successfully demonstrated the feasibility of using visual saliency for object detection and tracking, the computational time (on P4 3GHz personal computer using Matlab implementation) is around two second/frame.

Recently we have introduced a new method called adaptive principle components selection algorithm to realize dimension reduction, and modeled the vector variables in the feature space with Multivariate Gaussian Distribution, the new algorithm shows more accurate experimental results and can be computed much faster than our previous algorithm.

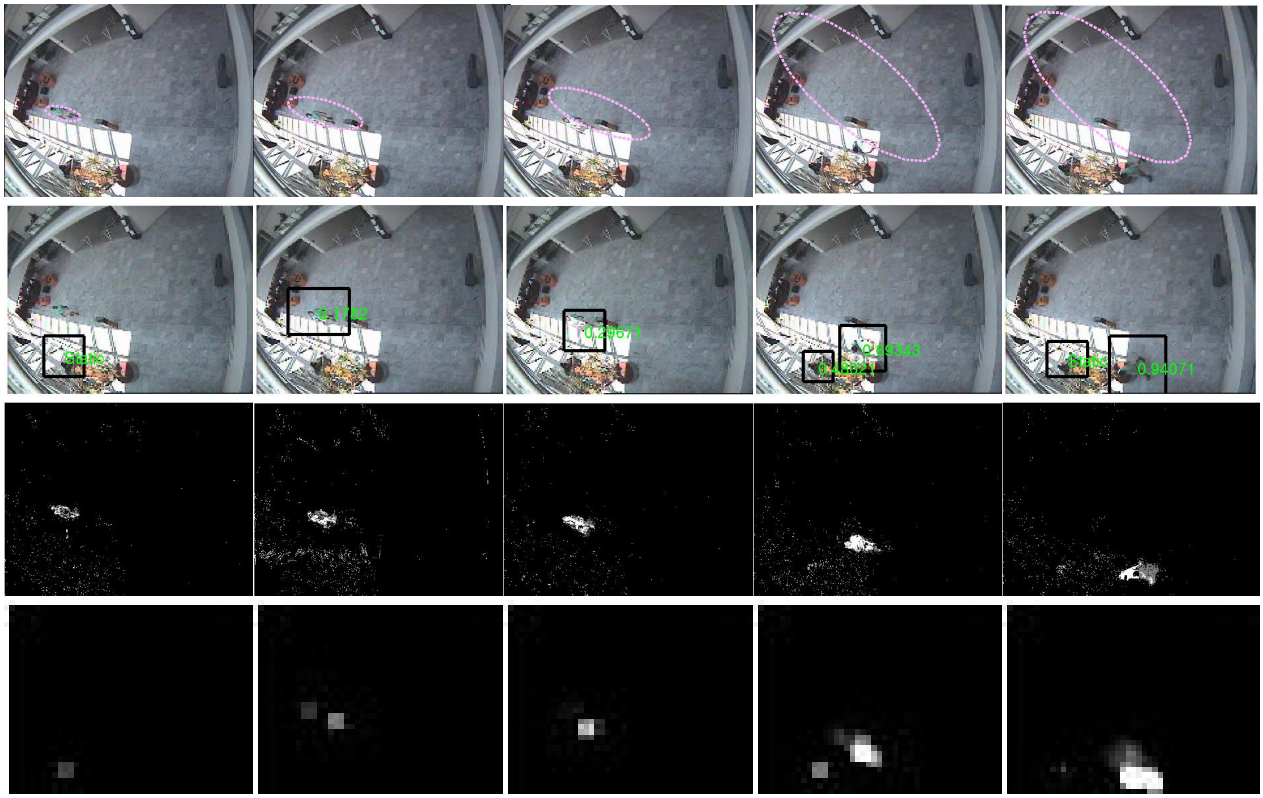


Figure 10. Results on improved meanshift[22] (1^{st} row), the proposed method (2^{nd} row), GMM[21](3^{rd} row) and the ISM (4^{th} row); the five columns are from Frame 980, 995, 1005, 1025, 1050 respectively in CAVIAR "*Browse*4" database. The evaluation programs for mean-shift and GMM methods were downloaded at http://staff.science.uva.nl/ zivkovic/PUBLICATIONS.html.

References

[1] N. D. Bruce. Features that draw visual attention: an information theoretic perspective. *Neurocomputing*, 65-66:125–133, 2005.

[2] M. A. Carreira-Perpinan. Acceleration strategies for gaussian mean-shift image segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1160–1167, 2006.

[3] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 142–149, 2000.

[4] N. V. Dashan Gao. Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, pages 282–287, 2005.

[5] G. Fan, V. Venkataraman, L. Tang, and J. Havlicek. A comparative study of boosted and adaptive particle filters for affine-invariant target detection and tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 131–138, 2006.

[6] J. Goldberger and H. Greenspan. Context-based segmentation of image sequences. *IEEE Transactions on PAMI*, 28(3):463–468, 2006.

[7] M. Isard and A. Blake. Condensation-conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29(1):5–28, 1998.

[8] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on PAMI*, 20(11):1254–1259, 1998.

[9] V. S. James W. Davis. Fusion-based background-subtraction using contour saliency. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2005.

[10] D.-S. Jang and H.-I. Choi. Active models for tracking moving objects. *Pattern Recognition*, 33(7):1135–1146, 2000.

[11] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal on Computer Vision*, 45(2):83–105, 2001.

[12] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *The 2nd European Workshop on Advanced Video-based Surveillance Systems*, pages 149–158, 2001.

[13] D.-S. Lee. Effective gaussian mixture learning for video background subtraction. *IEEE Transactions on PAMI*, 27(5):827–832, 2005.

[14] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, pages 28–39, 2004.

[15] T. Parag, A. Elgammal, and A. Mittal. A framework for

feature selection for background subtraction. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1916–1923, 2006.

[16] G. Qiu, X. Gu, Z. Chen, Q. Chen, and C. Wang. An information theoretic model of spatiotemporal visual saliency, to appear, international conference on multimedia and expo. 2007.

[17] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[18] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[19] J. van de Weijer, T. Gevers, and A. D. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on PAMI*, 28(1):150–156, 2006.

[20] C. Yang, R. Duraiswami, and L. Davis. Fast multiple object tracking via a hierarchical particle filter. In *International Conference on Computer Vision*, pages 212–219, 2005.

[21] Z. Zivkovic. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, 2006.

[22] Z. Zivkovic and B. Krose. An em-like algorithm for color-histogram-based object tracking. In *proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 798–803, 2004.

[23] Z. Zivkovic and F. van der Heijden. Recursive unsupervised learning of finite mixture models. *IEEE Transactions on PAMI*, 26(5):651–656, 2004.

[24] http://ftp.pets.rdg.ac.uk/pets2001/.

[25] http://homepages.inf.ed.ac.uk/rbf/caviar/.

Behavior of Virtual Human in equipment maintenance

HE Yuesheng

Abstract

The paper presents an original integration approach for the behavior of virtual human in the virtual environment to maintain the digital design of equipments. The virtual human is able to be controlled by users automatically. The approach has been proven being practical in the designed work of machine. The main purpose of the approach is to represent virtual human's behavior in macro orders. The next step of the work is to make virtual human have ability of perception and reaction. Then we will have more autonomous virtual human.

Keywords: virtual human, virtual environment, equipment maintenance, autonomous behavior, perception

1. Introduction

Virtual human, the representation of the geometric and behavioral characters of human in the virtual environment, is one of the new research areas of computer science. Virtual human includes some fields such as Computer Graphics, Robotics and Machine Learning. As the developing of the VR technology and its application, the research on virtual human has attracted many researchers. It has wide application areas including industry, medicine, military and entertainment. In the industry area, virtual human can effectively support the ergonomics and training of operation. It can be used in each periods of the digital product management including design, producing, maintaining and training. In the analysis process of the maintainability and maintenance process, using virtual human, the human factors of the virtual prototype can be analyzed. It has such important meaning for the human-centered product design, so research on the frame of virtual human software is important for the application of virtual human in the fields of the maintainability during the design of industry products.

To achieve the above goal, we should face the challenge from five aspects as Norman Badler described[1] :

- Create an interactive computer graphics human model.
- Endow it with reasonable biomechanical properties.
- Provide it with "human-like" behaviors.
- Use this simulated figure as an agent to effect changes in its world.
- Describe and guide its tasks through natural language instructions.

To describe human's action by using "language" is a convenient way to put the virtual human's behavior and other elements in the virtual environment together. Therefore, the behaviors constitute a powerful vocabulary for postural control. The manipulation commands provide the stimuli; the behaviors determine the response. The rationale for using behavioral animation is its economy of description: a simple input from the user can generate very complex and realistic motion.

By designing a simple set of rules for how objects behave, the user can control the objects through a much more intuitive and efficient language because much of the motion is generated automatically. Several systems have used the notion of behaviors to describe and generate motion. The most prominent of this work is by Prof. Magnenat Thalmann[7], who used the notion of behavior models to generate animations of Marilyn Monroe. By using the virtual human and describe her behavior, Prof. Thalmann rebuilt the classical scene of Miss Monroe's movie. In fact behaviors have been applied to articulated figures describe a computational environment for simulating virtual actors, principally designed to simulate actors and actresses for animation purposes is a very interesting region. Most of the action of the actors is in walking, and a gait controller generates the walking motion. Prof. Norman Badler used the same theory to construct virtual human software which is now used successfully by the U.S. military forces and NASA in the battle simulating and equipment maintenance[1][3].

To analysis the human factor as early as possible, we designed a virtual human system which has the ability to interact with the digital model of the equipment or the virtual environment.

We used a language to describe the virtual human's behavior in the task of equipment maintenance. So the designer will be convenient to make the animation to simulate the whole maintenance process. Moreover, the process will be easy for people to analysis.

In the 2nd part, it will present the method to describe the element of virtual human's action. 3rd part will be about how to describe the action in the task of maintenance. The 4th part will give the experimental result. Then the 5th will discuss the future work.

2. Element of virtual human's action

To describe a virtual human's body model in the virtual environment being constructed by the computer, it should be easy to be controlled. First of all, it should be convenient to simulate real human's action in the real environment.

We define the virtual human's model structure as the figure below[4],

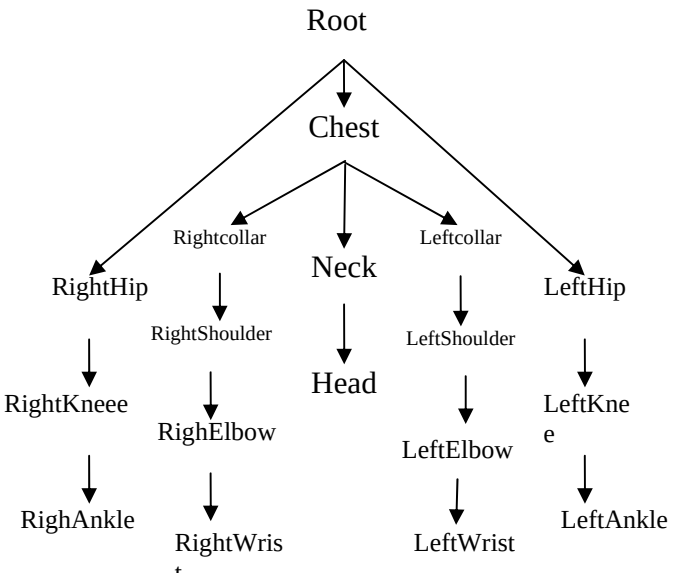


Figure 1 Structure of virtual human's body model

The root site for the figure is the one at the top of the tree, and its global location is taken as given, that is, not dependent on any other element of the environment.

Follow this tree structure, the virtual human has been defined as some segments which have been connected by joints.

Each joint corresponds to a certain set of coordinate transforms. The computation of the coordinate transforms for each segment and site in the downward traversal of the tree require inverting the site locations that connect the segment to other segments lower in the tree. It may also require inverting joint displacements if the joint is orient upwards in the tree. In fact, this is not expensive computationally because the inverse of a homogeneous transform is easy to compute, through a transpose and the dot product.

As the figure shows, the body of virtual human has been linked by some "chains " of the joints and segments. Every chain of joints is a set of joints which are controlled as one entity. Everyone is treated as a group. Internal joint angles are not visible outside of the group: they are driven by the group driver. The driver is nothing but a mapping from a number of parameters constituent joints. Those independent parameters will be called group angles and be used to accomplish the coordinate transforms.

The group angles of the joint groups are subject to linear constraints of the form[2] :

$$\sum_{i=1}^{n} a_i \theta_i \le b_i$$

where θ s are group angles and n is the number of θ s, or number of DOFs(degrees of freedom) of the joint group. "a" is the weight of every angle and "b" is the limit. There may be many such constraints for each group.

By this rule, every virtual human's postal will be represented as a set of angels of the joint groups (θ s). Thus, the moving motion of virtual human is represented as the coordinate transform of the root point of the body model.

As the movement and posture have been defined, the next step is to solve the problem of how to describe the elemental action of a virtual human. So, it is described a finite automata as below[10].

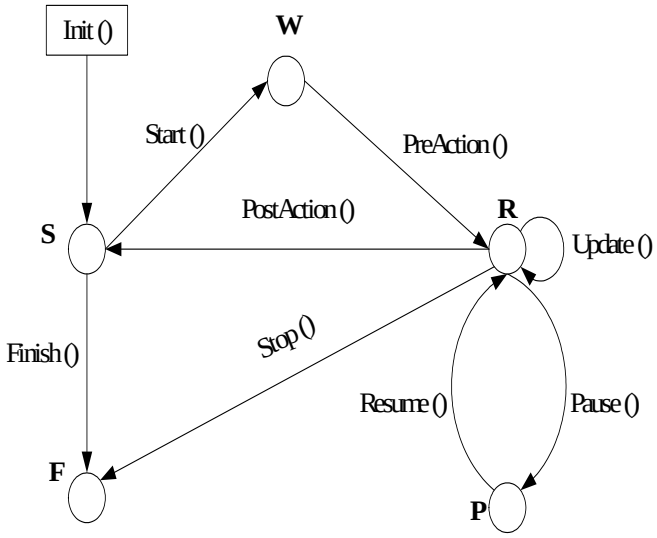


Figure 2 States of action finite automata

The 5 different States are:

1. S -- Stopped : accomplished the initialization
2. W--Waiting: ready to start the simulating loop
3. R --Running: running the simulating loop
4. P --Paused: simulating loop is paused
5. F --Finished: the action is finished

Therefore the action of virtual human in the virtual environment which has been constructed by the computer is a 5-states finite automata. In other words, the interaction between the virtual human and virtual environment is described by a finite automata. With this description, the virtual human will simulate real human's action in the real environment. We will easy to make the animation of the virtual human's behaviors in the virtual environment.

3. Action of Task in the maintenance

To use the virtual human to simulate the process of maintenance, the way of divide the whole task into small piece of actions must be designed.

As Prof. Ranko Vujosevic's research (University of Iowa), every maintain work can be divided into 4 different levels.

Based on the research, we design the virtual human's behavior in the maintenance as below[9],

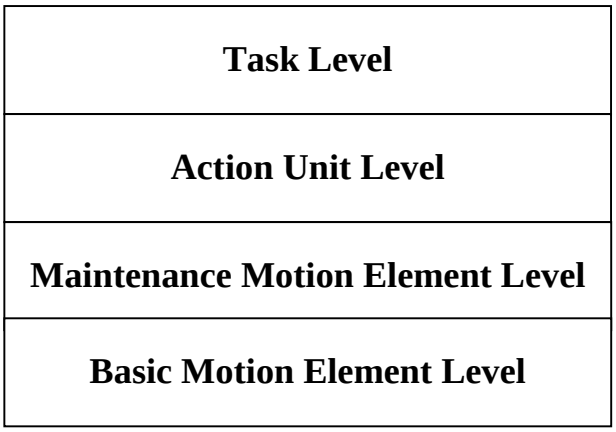


Figure 3 Levels of equipment maintenance task

As the figure shows, the whole task is described as 4 different levels,

1. Task level – represent a whole certain task, for instance to change the broken accessory;
2. Action Unit Level – sequence of virtual human's action, for instance move to a certain place and hold a certain tool;
3. Maintenance Motion Element Level – kind of actions related with the maintenance, for instance "hold", "move", "screw" etc. ;
4. Basic Motion Element Level – element of basic virtual human's action, for instance the coordinate transform of joint angles been driven by the finite automata.

The next step is to describe the procedure of every maintenance task. It is still be treated as a sequence of virtual human's behaviors as the figure below[10],

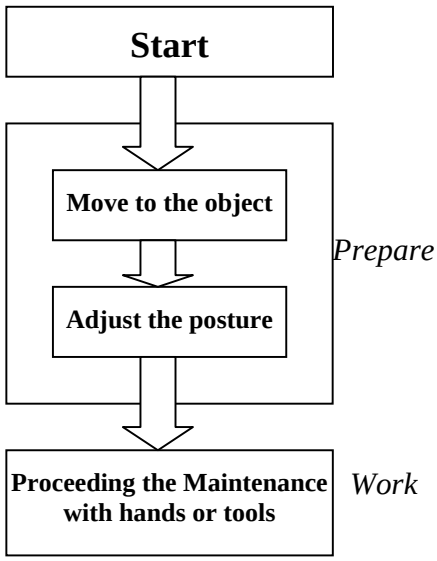


Figure 4 Procedure of a maintenance task

In the figure, the task is a procedure that has been consisted by 3 stages,

1. Start – init the whole environment, locate the virtual human, machine and tools;
2. Prepare – the virtual human move to the machine, then get the proper tool and act the proper posture;
3. Work – the virtual human maintain the machine, the main work of this stage is about virtual human's hands.

Under the rules of levels and stages, virtual human's action will be separated and constructed in different classes, the figure below shows the classes of arm's motion of virtual human[9],

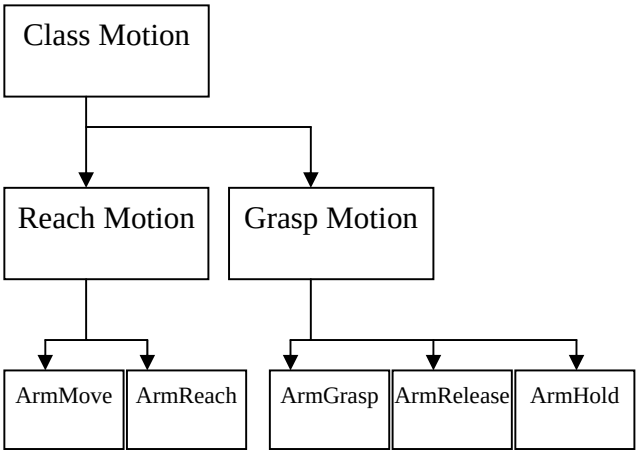


Figure 5 Class library structure to describe the behavior of arms

The topmost class Motion is in the "Basic Motion Element Level". Second level classes are in the "Maintenance Motion Element Level". Thus, the third level classes are in the "Action Unit Level".

By this method, the virtual human's behavior of equipment maintenance has been described in the OOP (Object Orient Program) language.

4. Experimental Result of the virtual human system

Based on the software ~ "Jack", we used the python language to construct a virtual human system to accomplish the task of equipment maintenance. The figure below shows the virtual human walked towards an object.

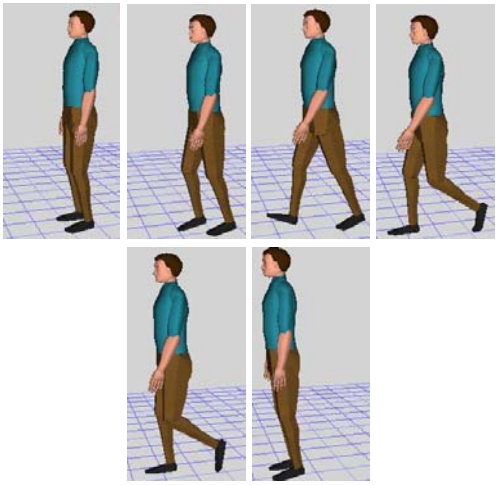


Figure 6 Walk of Virtual Human

Then the below figure shows the motion of virtual human's arm.

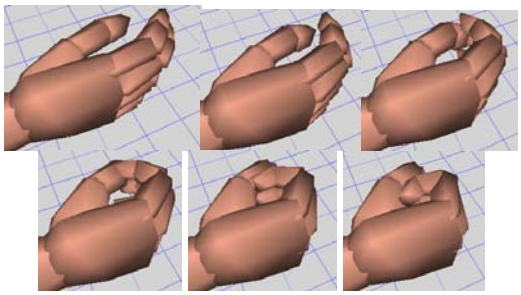


Figure 7 Hand behavior of virtual human

The result shows that the elemental actions of virtual human have been described by the language well.

Moreover, the system simulates the behavior of virtual human in equipment maintenance very well.

Figure 8 Virtual human to accomplish the maintenance task

Through the figure, it shows how the virtual human system simulates the behavior of dissembling a battery from a space station, then using a tool to fix it.

Because the virtual human's actions are represented by the language, the whole procedure has been made up in a short time and the animation's effect is satisfied.

5. Future work and discussion

By describing the element of virtual human's action as a language and analyzing the stages of maintenance, we design a system to simulate the behavior of virtual human in equipment maintenance.

It provides a good way to analyze human's working behavior and basic element of original design of equipment in the stage of digital one.

However, we used the "macro" order to control virtual human, the judgment should be made by designer. So the next step is to make the virtual human more automated, or, gives "intelligence" to virtual human.

One of the important research areas is to research the human-like behavior in an unknown-environment. How the virtual human will learn to adapt an unknown-environment by himself, moreover, how the virtual human learn to fix the digital equipment by himself are very interesting. Other important area is to research the multi-human's behavior in the virtual environment. Thus, our future work in this area is to research the "automatic" behavior of virtual human.

Most of the aspects for the research are about the "learning" ability of virtual human.

The reinforcement learning technique has proved to be fast, simple and robust. It has also succeeded in automatically learning behavioral models for difficult tasks. For instance, it is used successfully in the robot control. Thus, we believe it will be as useful to the computer graphics community as to the real word of robot by using a technique based on the reinforcement learning (for instance the Q-learning) approach[5][6].

On the other hand, because the virtual human should recognize and work in a word of computer graphics and images, and should consider many complex situations such as time and other virtual human's actions, we believe the approach such as the SVM will be very helpful.

References

[1] Norman I Badler. Virtual humans for animation, ergonomics, and simulation[J]. Nonrigid and Articulated Motion Workshop, 1997. Proceedings. IEEE Published: 1997, Page(s): 28 -36.

[2] Norman I. Badler, C. B. Phillips, B. L. Webber. Simulating Humans: Computer Graphics, Animation, and Control. Oxford University Press, 1993.

[3] Norman I. Badler, Jan Allbeck, Liwei zhao, Meeran Byun. Representing and Parameterizing Agent Behaviors. In Proceedings of Computer Animation'02, 133-143, 2002.

[4] Li Yan, Wang Wei, Lu Xiaojun, An Anthropometry-based Method for Virtual Human Modeling. Journal of System Simulation, 15(supplementary issue): 210-212, 2003.

[5] T. Conde, D. Thalmann, Learnable Behavioural Model for Autonomous Virtual Agents : Low-Level Learning, In Proceedings of Fith International Conference on Autonomous Agents and Multiagent Systems 2006 (AAMAS-06), Hakodate, Japan, May 2006, pp. 89-96

[6] T. Conde, D. Thalmann, An Integrated Perception for Autonomous Virtual Agents: Active and Predictive Perception, Computer Animation and Virtual Worlds, Volume 17, Issue 3-4, John Wiley, 2006

[7] Moccozet L, Thalmann N. M., Dirichlet Free-Form Deformation and their Application to Hand Simulation[J], Proceedings Computer Animation'97, IEEE Computer Society, 1997, pp.93-102.

[8] Molet T, Boulic R, Thalmann D. A real-time anatomical converter for human motion capture[J]. In Euro graphics Workshop on Computer Animation and Simulation, 1996, 79-94.

[9] He Yuesheng, Li Yan, Lu Xiaojun, A Software Architecture of a Virtual Human Factors Analysis System for Maintenance Engineering, Computer Simulation, Issue4,2006

[10] Xiaojun Lu, Yan Li, Hangen He, Yunxiang Ling: Research and Implement of Virtual Human's Walking Model in Maintenance Simulation. Edutainment 2006: 1027-1036

[11] N. Magnenat-Thalmann, A. Foni, G. Papagiannakis, N. Cadi-Yazli. Real Time Animation and Illumination in Ancient Roman Sites. The International Journal of Virtual Reality, IPI Press, Vol.6, No.1, pp. 11-24. March 2007.

Image segmentation based on the method of the maximal variance and improved genetic algorithm

Jianjia Pan

Abstract

Aiming for the problem of falling into local optimum when searching for the optimal threshold of the image using normal genetic algorithm, this paper presents a new method based on the maximal variance and improved genetic algorithm to segment the face image. This new method uses the maximal variance of the face gray image as the fitness and changes the problem of image segmentation into a problem of optimization. Adopting genetic algorithm which is characteristic of robustness and adaptability can increase efficiency. As a result, this new method can obtain the optimal segmentation result when applied to different face images. Experiments show that using this method to search for the global threshold can converge the optimal value and decrease the searching time.

Keywords: genetic algorithm, improved genetic algorithm, the maximal variance, image segmentation.

1. Introduction

Image segmentation is the fundamental and typical technology of image processing and computer vision, its aim is to segment the image to some areas with different characters and then select the interested target, it can provide reference for subsequent classification, recognition and searching. This process is a little difficult section, accurate segmentation will affect and decide the accurate degree of analysis and comprehension.

The method of image segmentation usually includes thresholding segmentation, edge detection etc. Thresholding segmentation is an available method. It has presented many methods to confirm the threshold, such as using the edge gray for the threshold, the minimum of the histogram as the threshold to separate the image and using statistical method to affirm the threshold.

The maximal variance criterion to affirm the optimal threshold is the statistical method. This method does not set the parameters artificially, it is a method of choosing the threshold automatically. This method is not only suitable for a single thresholding choice of two regions but also for multiple thresholding choice of multi-regions. The process of searching the threshold by the method of the maximal variance is to search the optimum, so we can adopt the genetic algorithm which is characteristic of fast optimizing to optimize and raise efficiency. However, adopting the normal genetic algorithm to search the global threshold can not converge the optimal value and converges very slowly and affect the application of the genetic algorithm.

This paper presents a new method based on the maximal variance and improved genetic algorithm to segment the image, which can separate the object from the background and get the satisfied results. This algorithm is characteristic of robustness, adaptivity and parallelism etc

2. The principle of genetic algorithm and its improvement

2.1 The principle of genetic algorithm

Genetic algorithm is an adaptive random searching algorithm based on natural selection and genetic mutation. In this algorithm, the chromosome is a binary encoding. There are many chromosomes, each encoding individuals is a candidate. The chromosome is the evolutional subject. The breeding operators includes reproduce, crossover and mutation which are dubbed genetic operators. At each generation, the population is evaluated on basis of fitness. The elites which have higher fitness will have more opportunity to reproduce compared to the rest of the population. GA recombines the individuals in this generation to produce new individuals of the next generation using two genetic

operators named crossover and mutation. Encoding, breeding operators and fitness measures insure the ability of GA to find good solutions efficiently. The filial generation includes much information of the parental generation and exceeds the parental generation as a whole, which can ensure the individuals develop forward and obtain the optimal solution

2.2 Improved genetic algorithm

The principal problems which affect the application of genetic algorithm are the overall searching capability and the rate of convergence. It has been proved that the normal genetic algorithm can not converge the optimal solution by modeling and assaying GA using Markov chain[1].The method of reserving the optimal individual[2] which are proposed by Grefenstettc can converge the globally optimal solution, but the rate of convergence is slow which affects the application of the genetic algorithm. In the evolutionary process, it should insures diversity of the population in order to insures the globally astringency of the algorithm.

On the other hand, it should enable the individual to develop optimally in order to quicken the rate of convergence. However, this method will decrease the diversity of the individual and trap in the local optimum easily. Thus many people recently have research on the improvement of the genetic algorithm.

This paper adopts an improved genetic algorithm[3] to segment the image, this method is not limited to some section but to design global, which adds some new operations such as mass selection, competition, filtering, locally optimum and replenishing the new generation dynamically. In this way both global searching and converging rate can be looked after.

The practical procedure of improved genetic algorithm is depicted as follows:

(1) Determining the GA parameters: the size population is M, generation gap G is 95 percent of M, the length of chromosome is l, the crossover rate is Pc, the mutation rate is Pm, the convergence parameter is ε ;
(2) Initiating the population: selecting M binary random serials which length is l as initial population;
(3) Judging the condition of convergence is satisfied or not (see Eqs.(2)). Feeding out the result if the condition is satisfied, otherwise you should continue;
(4) Sorting the fitness values of the parental generations, retaining the optimal M(1-G) individuals of the parental populations;
(5) Breeding operators(including crossover and mutation) is on basis of mass selection and competition: Using the method of competition to select 2 parental individuals to cross and mutate 3 or 5 times and generate 6 to 10 individuals. Selecting one individual which have the maximal fitness value to get into the next generation. Operating over and over until the number of the next generation come to certain MG;
(6) Sorting the fitness of the new generation;
(7) Greedy operation which searches the local optimum: Optimal operation on the first 3~5 individuals, namely searching among the neighborhood of the individual some times randomly(usually 10~20 times).Shifting the original value if the preferable value is found; otherwise the original value is unchanged;
(8) Filtering operation: Filtering the worse one of two individuals which fitness value and Hamming distance are both lower than the threshold until Y individuals are left;

Judging Y is equal to M or not, if Yes, get down to the step (3),otherwise replenishing the new generation dynamically: mutating the first aM individuals of the last generation 3~5 times randomly and then generate M-Y new individuals to get into the next generation, afterwards get down to the step (3).

3. Image segmentation based on the maximal variance and improved genetic algorithm

Adopting the criterion of the maximal variance to confirm the optimum to segment the image, since the process of searching the threshold by the method of the maximal variance is to search the optimum, so we can adopt the genetic algorithm which is characteristic of fast optimizing to optimize and raise efficiency. However adopting the normal genetic algorithm to search the global

threshold can not converge the optimal value and converges very slowly and affect the application of the genetic algorithm. So this paper presents a new method based on the maximal variance and improved genetic algorithm to segment the image, which can separate the object from the background and get the satisfied results.

3.1 Confirming the optimal threshold by the criterion of the maximal variance

The method of the maximal variance[4] which is on the basis of the method of least squares is proposed by Ostu. This algorithm is relatively simple and wide-ranging, thus attracts the most attention.

Supposing the gray-level value of the image is L, the threshold T can separate the pixels into two groups. The gray-level values of the group 1 are higher than T, the ones of the group 2 are lower than T. Let the pixel numbers of the group 1 be n_1,the average gray-level of these pixels be m_1,the variance is σ_1; the pixel numbers of the group 2 is n_2,the average gray-level of these pixels is m_2,the variance is σ_2;the average gray-level value of the image is m; then the variance within two groups is:

$$\sigma_I^2 = n_1\sigma_1^2 + n_2\sigma_2^2$$

the variance between two groups is:

$$\sigma_B^2 = n_1(m-m_1)^2 + n_2(m-m_2)^2$$

$$= n_1 n_2 (m_1 - m_2)^2$$

the optimal threshold value T is:

$$T = \max[\sigma_B^2(t)/\sigma_I^2(t)] \text{（t=1,2, … ,L）} \quad (1)$$

3.2 Confirming the optimal threshold based on improved genetic algorithm

Encoding: The image to be processed has 256 gray-levels, so the segmenting threshold is between 0 and 255. We adopt binary encoding techniques, 8 bits in one threshold segmenting. Each chromosome represents a segmenting threshold. The initial population with which the genetic process begins can be chosen randomly, and its fitness ranges from low to high.

Decoding: Decoding binary genome, the value is between 0 and 255.

Fitness: Adopting equation (1) as fitness.

The condition of convergence:

$$\left|\frac{Mf(n) - Mf(n-1)}{Mf(n)}\right| < \varepsilon \quad (2)$$

Where *Mf*(n) is the maximal fitness in the *n*th iteration, 0< ε <1.It may be exist *Mf*(n)=*Mf*(n-1) in the evolutionary process, so we stipulate a minimum iteration number N*min* in the course of practical evolvement. When the iteration number is larger than N*min* and equation (2) is satisfied, the evolutionary process is terminated.

4. Experimental results

This paper adopts two images with gray-level of 256 as the experimental subjects; both of them are face images. The GA parameters are: M=100, l=8, Pc=0.65, Pm=0.08, ε =0.0001, N*min*=20.The experimental results show that adopting the method of the maximal variance based on improved genetic algorithm to search for the threshold can segment the image effectively. As a result, the object can be separated from the background.

Comparing fig.(b) and fig.(c) between Fig.1 and Fig.2, we find that the segmenting result adopting the traditional algorithm is not ideal. However, adopting the improved genetic algorithm which can approach to the global optimal threshold provides satisfactory experimental results. And more information is taken from the original image for the further process. Besides, this improved algorithm can decrease the convergence time compared to the traditional one.


(a) The original image


(b) Segmentation result by traditional algorithm


(c)Segmentation result by improved algorithm

Fig.1 Segmentation result


(a)The original image


(b) Segmentation result by traditional algorithm


(c) Segmentation result by improved algorithm

Fig.2 Segmentation result

5. Conclusion

Aiming for the problem of falling into local optimum when searching for the optimal threshold of the image using normal genetic algorithm, this paper presents a new method based on the maximal variance and improved genetic algorithm to segment the face image. In this way the disadvantages of normal genetic algorithm on the aspects of global searching and converging rate can both be solved. Experiments show that using this method to search the global threshold can converge the optimal value and decrease the searching time.

References

1.LIU-Feng, LIU-Guizhong. Global convergence and convergence rate for genetic algorithm [J]. Journal of Systems Engineering, 1998,13(4):79-85.

2. HE-Lin, WANG-Kejun. Elitist preserved genetic algorithm and its convergence analysis [J].Control and Decision,2001,15(1):63-66

3. XU-Lu. Improved genetic algorithm and its application in image processing [D].2000, 3:18-21

4. Ostu N. A Threshold Selection Method From Gray Level Histogram. IEEE Transactions Sys Man and Cybernetics, 1979.8:62-66

Face Recognition Using Wavelet Packet Decomposition and Support Vector Machines

Limin Cui

Abstract

Automated face detection and recognition is one of the most attentional branches of biometrics and it is also the one of the most active and challenging tasks for computer vision and pattern recognition. This paper present a method for face recognition based on wavelet packet decomposition and SVMs. The experiments show that the method has better performance than the traditional ones, and reduce the greatly computation.

Keywords: face recognition; wavelet packet decomposition; SVMs; multi-class classification

1. Introduction

The face recognition is one of the most active research areas in biometric, pattern recognition, and computer vision because of its many important applications in fields such as security, human-computer interaction, and surveillance. Various approaches for face recognition have been proposed, and survey in this area can be found in [1-4]. Two issues are central to face recognition as follows:

(i) Feature Extraction: what features can represent a face. The objective is to find techniques that can reduce dimensionality and increase their discriminative power.

(ii) Classification: how to classify a new face image based on the chosen feature representation.

Many representation approaches have been introduced in the recent years. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are the most popular approaches. PCA chooses the subspaces in the function of data distribution [5]. LDA selects the subspaces that yield maximal inter-class distance and keeping the intra-class distance small [6]. Independent Component Analysis (ICA) can be seen as an extension to PCA except that the distribution of the components is designed to be non-Gaussian [7]. ICA provides a more powerful data representation than PCA. Literature [8] introduced a Gabor-Fisher Classifier method, which couples Gabor Wavelets, PCA and Enhanced Fisher Discriminant Model together. However, the methods mentioned above are linear, and not effective for nonlinear distribution of face patterns. Some researchers proposed kernel subspace approaches such as Kernel Principal Component Analysis (KPCA) [9] and Kernel Fisher Discriminant Analysis (KFDA) [10, 11].

While regarding classification methods, nearest neighbor [5], convolutional neural networks [12], Bayesian Classification [13], AdaBoost methods [14], and Support Vector Machines (SVMs) [15-17] have been widely used. Among various classifiers, SVMs Classifier has attracted much attention and achieves success in face recognition. As a new universal learning machine, SVMs are proposed by Vapnik et al. in the statistical theory and structural risk minimization [18], and have proven to be a powerful approach in many areas. The applications of SVMs to face recognition have been proposed by Osuna et al [15] recently. SVMs are kernel based approaches that use a kernel function to map the input data into a high-dimensional feature space, and then construct an optimal separating hyperplane in that space. The kernel functions can be linear, Gaussian, polynomial and RBF kernels and satisfy Mercer's condition.

Over the past few years, some researchers study on wavelet and various classifiers to face detection and recognition [19-22]. The literature [19] proposed a multi-stage approach to build a face detection system, and [20] used PCA and wavlet subband to deal with face recognition. The method of local discriminant wavelet packet is proposed in [21]. Wavelet Packet and SVMs are used for face recognition in [22].

In this paper, we propose a new method for face recognition based on wavelet packet decomposition of face images and SVMs classifier. Firstly, we use wavelet packet to decompose the images, and each face image is described by a subset of band filtered images containing wavelet coefficients. From these wavelet coefficients we extract feature vectors. Different from literature [22] that used norm to extract features, moments combine feature vecoters. Then SVMs are used to classify these feature vectors. Section 2 reviews the basic wavelet packet theory. In the section 3, we give a general introduction to SVMs for classification. A face recognition system based on the proposed method is proposed in section 4. Experimental results are presented in section 5 and finally, section 6 gives the conclusions.

2. Wavelet Packet

Wavelet transform has been widely used in many fields, such as numerical analysis, signal analysis, image processing, pattern recognition, and so on, because of its advantages, such as good time and frequency localizations. The wavelet packet decomposition is generalization of the classical wavelet decomposition that allows more flexibility in image decomposition and dimensionality reduction. In this section, we first introduce notations for wavelet packet analysis and then extract feature vectors.

2.1. Wavelet Packet Decomposition

Let φ and ψ be the scaling function and the mother wavelet function. Suppose φ and ψ satisfy the following two-scale relations respectively [23]

$$\varphi(x) = \sum_k h_k \varphi(2x-k), \quad \psi(x) = \sum_k g_k \varphi(2x-k)$$

where $\{h_k\}$ are the two-scale coefficients with $g_k = (-1)^k h_{1-k}$. Let μ_0 denote the scaling function and μ_1 denote the mother wavelet function, then wavelet packets can be described by the following collection of basis functions:

$$\mu_{2n}(x-l) = \sum_k h_{k-2l} \sqrt{2} \mu_n(2x-k) \tag{1}$$

$$\mu_{2n+1}(x-l) = \sum_k g_{k-2l} \sqrt{2} \mu_n(2x-k) \tag{2}$$

where p is a scale index, and l is a translation index. The relationship between wavelet packets of different scales can be specified as follows:

$$\sqrt{2}\mu_n(2x-k) = \sum_l [\overline{h}_{k-2l} \mu_{2n}(x-l) + \overline{g}_{k-2l} \mu_{2n+1}(x-l)]$$

Let $U_j^n = span\{2^{j/2} \mu_n(2^j x-k)\}$, from (1) and (2) we have

$$U_1^n = U_0^{2n} \oplus U_0^{2n+1}$$

Generally, we can obtain

$$U_{j+1}^n = U_j^{2n} \oplus U_j^{2n+1} \tag{3}$$

From (3), we know as following:

Unlike the classic wavelet transform that recursively decomposes only low-pass sub-band, the wavelet packet transform decomposes both sub-bands at each level. The result is a wavelet decomposition tree. A facial image is decomposed by wavelet transform and wavelet packet transform as shown in Figure 1 and Figure 2.

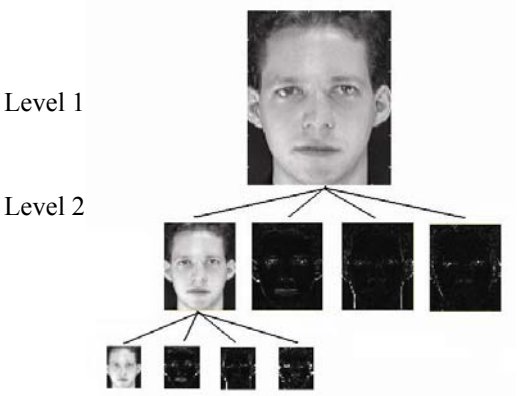


Figure 1. 2-level wavelet decomposition of facial image.

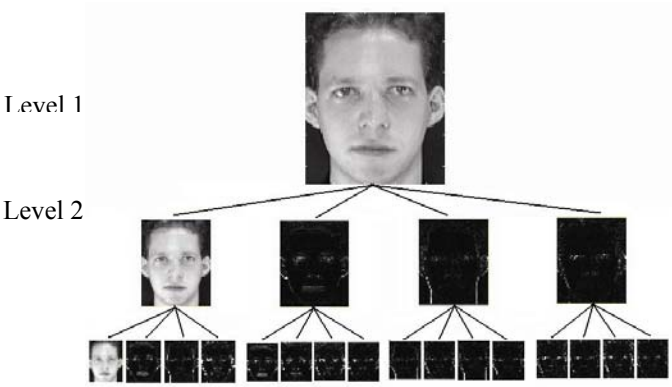


Figure 2. 3-level wavelet packet decomposition of facial image.

2.2. Feature Vectors Extraction

As mentioned above, wavelet packet decomposition with depth 2 is performed. There is no need to use a deeper decomposition because the image size is small after two levels decomposition. Then 6 images, including one image of approximation and 15 images of details are obtained as shown in Figure 2. These images can be described by 16 wavelet coefficient matrices in level 2. These 16 matrices are huge to difficult for computation. For considering computation and complexity, we need to redect dimension of 16 wavlet coefficient matrics.

In this paper we adopt mean value μ_i and variance σ_i^2 of each matrice to represent face feature vector v as follows:

$$v = \bigcup_{i=1}^{16} \{\mu_i, \sigma_i^2\}$$

where $i = 1 \cdots 16$.

3. Support Vector Machines Classification

In this section we will introduce the basics of SVMs for binary and multi-class classification problems.

3.1 Binary Classification

In the binary classification case, the objective of SVMs is to find the optimal separating hyperplane with a maximum margin between two classes. The SVMs classifier is given by (4) as follows:

$$y = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b) \quad (4)$$

where $K(x, x_i)$ is the kernel function, x_i is the vector of the i^{th} training samples, $y_i \in \{-1, +1\}$ is a class label, and N is the number of training samples. $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_N\}$ is learned by solving the following quadratic programming problem:

$$\begin{cases} \max \quad W(a) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ sbuject \quad to \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C \end{cases}$$

C is predefined parameter that is a trade-off between a wide margin and a small number of margin failures.

3.2 Multi-class Classification

There are two basic strategies for multi-class problem: One-vs-All [24] and One-vs-Another [25] Strategy.

(i) There are m SVMs trained in the One-vs-All approach. Each of the SVMs separates a single class from all left classes.

(ii) There are m^2 SVMs trained in the One-vs-Another approach. Each SVM separates a pair of class. The pairwise classifiers are composed in trees, where each tree node represents an SVM.

In this paper, we use these two methods for multi-class classification.

4. The Proposed System

A face recognition system is obtained by the proposed method. The system consists of two stages, namely training and recognition stage as shown in Figure 3.

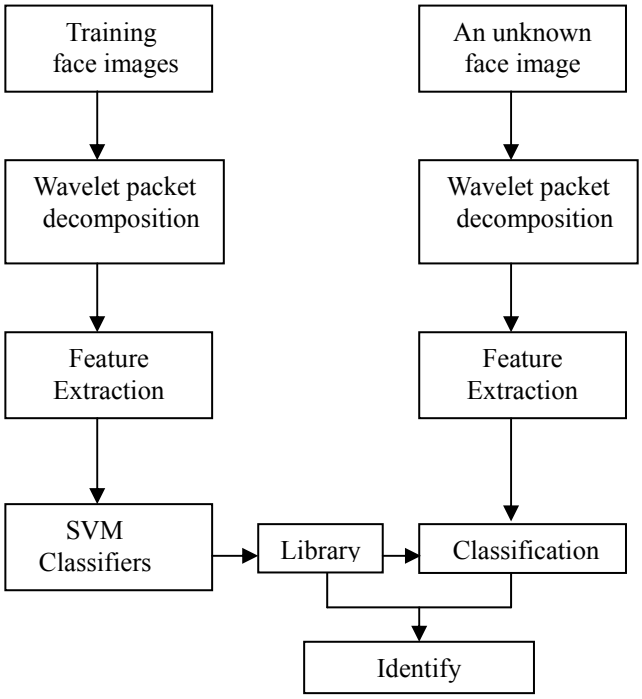


Figure 3. The frame of face recognition system based on wavelet packet decomposition and SVMs

5. Experiments

Daubechies wavelet is used as the mother wavelet because we not only need smooth, compactly-supported orthonormal wavelets, but also as few non-zero expansion coefficients as possible.

5.1. Part I: ORL Face Database

The experiment is performed on the Cambridge ORL face database, which contains 40 distinct persons as shown in Figure 4. Each person has ten different images. There are variations in facial expressions such as open or closed eyes, smiling or nonsmiling, and glasses or no glasses. All the images were taken against a dark homogeneous background with the subjects in an up-right, frontal position, with tolerance for some side movements. There is also some variations in scale. We show four individuals as shown in Figure 5.

In our face recognition experiments on the ORL database, we select 200 samples (5 for each individual) randomly as the training set. The remaining 200 samples are used as the test set. Using our method, we can obtain high recognition rate as shown in Table 1.

5.2. Part II: Yale Face Database

This experiment is performed on the Yale database for comparing the ability of proposed system in recognizing frontal face images with different facial expressions, illuminations, and occlusion by glasses. In Yale database, there are 15 persons as shown in Figure 6. We show four individuals as shown in Figure 7.

There are 165 face images in Yale database. In our face recognition experiments on the Yale database, we select 75 samples (5 for each individual) randomly as the training set. The remaining 90 samples are used as the test set. Using our method, we can obtain high recognition rate as shown in Table 2.

Figure 4. Forty distinct persons in ORL face database.

Figure 5. Four individuals in the ORL face database. There are 10 images for each person.

Table 1. Experiment results on ORL face database

Methods	The Proposed Method (One-vs-All)	The Proposed Method (One-vs-Another)	PCA	LDA
Accurate Rate	97.5	98.2	85.4	96.1

Figure 6. Fifteen distinct persons in Yale face database.

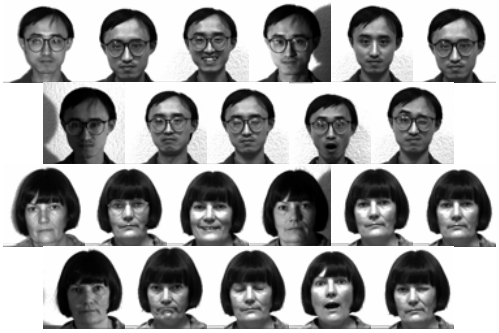


Figure 7. Two individuals in the Yale face database. There are 11 images for each person.

Table 2. Experiment results on Yale face database

Methods	The Proposed Method (One-vs-All)	The Proposed Method (One-vs-Another)	PCA	LDA
Accurate Rate	94.4	96.3	78.6	93.2

6. Conclusions

This paper present a face recognition system based on wavelet packet decomposition and SVMs. Different from traditional methods, we took full advantage of the time-frequency localization properties of wavelet packet transform to reduce the computational load and gain good recognition rate.

References

[1] Samal A. and Iyengar P. A., “Automatic recognition and anlysis of human face and facial expressions: A survey”, Pattern Recognition, Vol. 25, pp. 65-77, 1992.
[2] Chellappa R. Wilson C. L., and Sirohey S., “Human and machine recognition of faces: A survey”, Proceedings of IEEE, Vol. 83, No. 5, pp. 705-740, May 1995.
[3] Shigeru Akamatsu, “Computer recognition of human face - A survey”, Systems and Computers in Japan, Vol. 30, No. 10, pp. 76-89, Aug. 1999.
[4] Zhao W., Chellappa R., Phillips P.J., and Rosenfeld A., “Face recognition: A literature survey”, ACM Computing Survey, Vol. 35, No. 4, December 2003, pp. 699-458.
[5] Turk M. A. and Pentland A. P., “Eigenfaces for recognition”, Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, Mar. 1991.
[6] Belhumeur P. N., Hespanha J. P., and Kriegman D. J., “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 711-720, July 1997.
[7] Bartlett M. S., Lades H. M., and Sejnowski T. J., “Independent component representations for face recognition”, Proceedings of the SPIE, Conference on Human Vision and Electronic Imaging III, Vol. 3299, pp. 528-539, 1998
[8] Liu C. and Wechsler H., “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition”, IEEE Transactions on Image Processing, Vol. 11, No. 7, pp. 467-476, 2002.
[9] Schölkopf B., Smola A. J., and Müller K. -R., “Nonlinear component analysis as a kernel eigenvalue problem”, Neural Computation, Vol. 10, No. 5, pp. 1299-1319, 1998.
[10] Mika S., Rätsch G., Weston J., Schölkopf B., and Müller K. -R., “Fisher discriminant analysis with kernels”, Neural Networks for Signal Processing IX, New York: IEEE Press, pp. 41-48, 1999.
[11] Mika S., Rätsch G., Weston J., Schölkopf B., Smola A. J., and Müller K. -R., “Invariant feature extraction and classification in kernel spaces”, Advances in Neural Information Processing Systems, MIT Press, Vol. 12, pp: 526-532., 2000.
[12] Lawrence S., Giles C. L., Tsoi A., and Back A.. “Face recognition: A convolutional neural network approach”, IEEE Transactions on Neural Networks, Vol. 8, No. 1, pp. 98–113, 1997.
[13] Moghaddam B., “Principal manifolds and bayesian subspaces for visual recognition”, In International Conference on Computer Vision (ICCV’99), pp. 1131–1136, Sep. 1999.
[14] Freund Y. and Schapire R., “A decision-theoretic generalization of on-line learning and an application to boosting”, .Journal of Computer and System Sciences, Vol. 55, No.1, pp. 119–139, Aug. 1997.
[15] Osuna E., Freund R., and Girosi F., “Training support vector machines: An application to face detection”, Proceedings of IEEE on Computer Vision and Pattern Recognition (CVPR), pp.130–136, June 1997.
[16] Romdhani S., Torr P.H.S., Schölkopf B., and Blake A., “Computationally efficient face detection”, Proceedings of International Conference on Computer Vision (ICCV), pp. 695–700, July 2001.
[17] Heisele B., Serre T., Prentice S., and Poggio T., “Hierarchical classification and feature reduction for fast face detection with support vector machines,” Pattern Recognition., Vol.36, No.9, pp. 2007–2017, Sept. 2003.
[18] Vapnik V., “The nature of statistical learning theory”, New York: Springer-Verlag, 1995.
[19] LE D.D., and Satoh S., “A multi-stage approach to fast face detection”, IEICE Transactions on Information and Systems, Vol. E89-D, No. 7, July 2006.
[20] Feng G. C., Yuen P. C., and Dai D. Q., “Human face recognition using PCA on wavelet subband”, Journal of Electronic Imaging, Vol. 9, No. 2, pp. 226-233, 2000.
[21] Liu C. C., Dai D Q., and Yan H., “Local discriminant wavelet packet coordinates for face recognition”, Journal of Machine Learning Research, 2007.
[22] Yang J., Gang F. L., and Jiang F. J., “Human face recognition based on wavelet package and SVM”, Computer Simulation, Vol.21, No. 9, Sep. 2004.
[23] Chui C. K., “An introduction to wavelets”, Academic Press, Boston, 1992.
[24] Vapnik, V., “Statistical Learning Theory”, Wiley and Sons, Inc., New York, 1998.
[25] Kreβel, U., “Pairwise classi_cation and support vector machines”, In Advances in Kernel Methods Support Vector Learning, Cambridge, MA: MIT Press, pp. 255-268, 1999.