Advances in Intelligent Systems and Computing 244

Van-Nam Huynh · Thierry Denœux
Dang Hung Tran · Anh Cuong Le
Son Bao Pham *Editors*

Knowledge and Systems Engineering

Proceedings of the Fifth International Conference KSE 2013, Volume 1

Springer

Advances in Intelligent Systems and Computing

Volume 244

Series Editor

Janusz Kacprzyk, Warsaw, Poland

For further volumes:
http://www.springer.com/series/11156

Van-Nam Huynh · Thierry Denœux
Dang Hung Tran · Anh Cuong Le
Son Bao Pham
Editors

Knowledge and Systems Engineering

Proceedings of the Fifth International Conference KSE 2013, Volume 1

Springer

Editors
Van-Nam Huynh
School of Knowledge Science
Japan Advanced Institute of Science and Technology
Ishikawa
Japan

Thierry Denœux
Universite de Technologie de Compiegne
Compiegne Cedex
France

Dang Hung Tran
Faculty of Information Technology
Hanoi National University of Education
Hanoi
Vietnam

Anh Cuong Le
Faculty of Information Technology
University of Engineering and Technology - VNU Hanoi
Hanoi
Vietnam

Son Bao Pham
Faculty of Information Technology
University of Engineering and Technology - VNU Hanoi
Hanoi
Vietnam

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-02740-1 ISBN 978-3-319-02741-8 (eBook)
DOI 10.1007/978-3-319-02741-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013950936

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains papers presented at the Fifth International Conference on Knowledge and Systems Engineering (KSE 2013), which was held in Hanoi, Vietnam, during 17–19 October, 2013. The conference was jointly organized by Hanoi National University of Education and the University of Engineering and Technology, Vietnam National University. The principal aim of KSE Conference is to bring together researchers, academics, practitioners and students in order to not only share research results and practical applications but also to foster collaboration in research and education in Knowledge and Systems Engineering.

This year we received a total of 124 submissions. Each of which was peer reviewed by at least two members of the Program Committee. Finally, 68 papers were chosen for presentation at KSE 2013 and publication in the proceedings. Besides the main track, the conference featured six special sessions focusing on specific topics of interest as well as included one workshop, two tutorials and three invited speeches. The kind cooperation of Yasuo Kudo, Tetsuya Murai, Yasunori Endo, Sadaaki Miyamoto, Akira Shimazu, Minh L. Nguyen, Tzung-Pei Hong, Bay Vo, Bac H. Le, Benjamin Quost, Sébastien Destercke, Marie-Hélène Abel, Claude Moulin, Marie-Christine Ho Ba Tho, Sabine Bensamoun, Tien-Tuan Dao, Lam Thu Bui and Tran Dinh Khang in organizing these special sessions and workshop is highly appreciated.

As a follow-up of the Conference, two special issues of the Journal of *Data & Knowledge Engineering* and *International Journal of Approximate Reasoning* will be organized to publish a small number of extended papers selected from the Conference as well as other relevant contributions received in response to subsequent calls. These journal submissions will go through a fresh round of reviews in accordance with the journals' guidelines.

We would like to express our appreciation to all the members of the Program Committee for their support and cooperation in this publication. We would also like to thank Janusz Kacprzyk (Series Editor) and Thomas Ditzinger (Senior Editor, Engineering/Applied Sciences) for their support and cooperation in this publication.

Last, but not the least, we wish to thank all the authors and participants for their contributions and fruitful discussions that made this conference a success.

Hanoi, Vietnam
October 2013

Van Nam Huynh
Thierry Denœux
Dang Hung Tran
Anh Cuong Le
Son Bao Pham

Organization

Honorary Chairs

Van Minh Nguyen – Hanoi National University of Education, Vietnam
Ngoc Binh Nguyen – VNU University of Engineering and Technology, Vietnam

General Chairs

Cam Ha Ho – Hanoi National University of Education, Vietnam
Anh Cuong Le – VNU University of Engineering and Technology, Vietnam

Program Chairs

Van-Nam Huynh – Japan Advanced Institute of Science and Technology, Japan
Thierry Denœux – Université de Technologie de Compiègne, France
Dang Hung Tran – Hanoi National University of Education, Vietnam

Program Committee

Akira Shimazu, Japan
Azeddine Beghdadi, France
Son Bao Pham, Vietnam
Benjamin Quost, France
Bernadette Bouchon-Meunier, France
Binh Thanh Huynh, Vietnam
Bay Vo, Vietnam
Cao H, Tru, Vietnam
Churn-Jung Liau, Taiwan
Dinh Dien, Vietnam
Claude Moulin, France
Cuong Nguyen, Vietnam
Dritan Nace, France
Duc Tran, USA
Duc Dung Nguyen, Vietnam
Enrique Herrera-Viedma, Spain
Gabriele Kern-Isberner, Germany
Hiromitsu Hattori, Japan
Hoang Truong, Vietnam
Hung V. Dang, Vietnam
Hung Son Nguyen, Poland
Jean Daniel Zucker, France

Jérôme Lang, France
Jing Liu, China
Jiuyong Li, Australia
Jonathan Lawry, UK
Kenji Satou, Japan
Lam T. Bui, Vietnam
Bac H. Le, Vietnam
Loannis Parissis, France
Marie-Helene Abel, France
Martin Steffen, Norway
Masahiro Inuiguchi, Japan
Michel Riveill, France
Mina Ryoke, Japan
Minh-Dung Phan, Thailand
Mitsuru Ikeda, Japan
Minh L. Nguyen, Japan
Noboru Takagi, Japan
Peter Whigham, New Zealand
Phayung Meesad, Thailand
Quang-Huy Nguyen, France
Quang Uy Nguyen, Ireland
Sabine Bensamoun, France
Sadaaki Miyamoto, Japan
Serge Stinckwich, France
Sébastien Destercke, France
Si Quang Le, UK
Son Doan, USA
Tien-Tuan Dao, France
Tetsuya Murai, Japan
Thanh Binh Nguyen, Vietnam
Thanh Tri Nguyen, Vietnam
Thanh-Thuy Nguyen, Vietnam
The Duy Bui, Vietnam
The Loc Nguyen, Vietnam
Thomas Huynh, USA
Tho Hoan Pham, Vietnam
Thepchai Supnithi, Thailand
The Dung Luong, Vietnam
Tran Dinh Khang, Vietnam
Tsutomu Fujinami, Japan
Tzung-Pei Hong, Taiwan
Vladik Kreinovich, USA
Xiaoshan Li, Macau
Xuan Hoai Nguyen, Vietnam
Xuan-Hieu Phan, Vietnam
Yasuo Kudo, Japan

Contents

Part I
Keynote Addresses

What Ontological Engineering Can Do for Solving Real-World Problems

Riichiro Mizoguchi

Abstract. Ontological engineering works as a theory of content and/or content technology. It provides us with conceptual tools for analyzing problems in a right way by which we mean analysis of underlying background of the problems as well as their essential properties to obtain more general and useful solutions. It also suggests that we should investigate the problems as deeply as possible like philosophers to reveal essential and intrinsic characteristics hidden in the superficial phenomena/appearance. Knowledge is necessarily something about existing entities and their relations, and ontology is an investigation of being, and hence ontology contributes to facilitation of our knowledge about the world in an essential manner.

There exist a lot of problems to be solved in the real world. People tend to solve them immediately after they realize needs to solve them. One of the issues here is that necessary consideration about the nature of those problems is often skipped to get solutions quickly,which sometimes leads to ad-hoc solutions and/or non-optimal solutions. This is why ontological engineering can make a reasonable contribution to improving such situations.

In my talk, after a brief introduction to ontological engineering, I explain technological aspects of ontological engineering referring to my experiences. One of the important conceptual techniques is separation of what and how in procedures/algorithms. Then, I show you a couple of concrete examples of deployment of such conceptual tools in several domains.

Riichiro Mizoguchi
Research Center for Service Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

V.-N. Huynh et al. (eds.), *Knowledge and Systems Engineering, Volume 1*,
Advances in Intelligent Systems and Computing 244,
DOI: 10.1007/978-3-319-02741-8_1,

Argumentation for Practical Reasoning

Phan Minh Dung

Abstract. We first present a short introduction illustrating how argumentation could be viewed as an universal mechanism humans use in their practical reasoning where by practical reasoning we mean both commonsense reasoning and reasoning by experts as well as their integration. We then present logic-based argumentation employing implicit or explicit assumptions. Logic alone is not enough for practical reasoning as it can not deal with quantitative uncertainties. We explain how probabilities could be integrated with argumentation to provide an integrated framework for jury-based (or collective multiagent) dispute resolution.

Phan Minh Dung
Department of Computer Science and Information Management,
Asian Institute of Technology, Thailand

V.-N. Huynh et al. (eds.), *Knowledge and Systems Engineering, Volume 1*,
Advances in Intelligent Systems and Computing 244,
DOI: 10.1007/978-3-319-02741-8_2,

Legal Engineering and Its Natural Language Processing

Akira Shimazu and Minh Le Nguyen

Abstract. Our society is regulated by a lot of laws which are related mutually. When we view a society as a system, laws can be viewed as the specifications for the society. Such a system-oriented aspect of laws have not been studied well so far. In the upcoming e-Society, laws have more important roles in order to achieve a trustworthy society and we expect a methodology which treats a system-oriented aspect of laws. Legal Engineering is the new field that studies the methodology and applies information science, software engineering and artificial intelligence to laws in order to support legislation and to implement laws using computers. So far, as studies on Legal Engineering, Shimazu group of JAIST proposed the logical structure model of law paragraphs, the coreference model of law texts, the editing model of law texts and so on, and implemented their models. Tojo group of JAIST verified whether several related ordinances of Toyama prefecture in Japan contains contradictions or not. Ochimizu group of JAIST studied the model for designing a law-implementation system and proposed the accountability model for the law-implementation system. Futatsugi group of JAIST proposed the formal description and the verification method of legal domains. As laws are written in natural language, natural language processing is essential for Legal Engineering. In this talk, after the aim, the approach and the problems of Legal Engineering are introduced, studies on natural language processing for Legal Engineering are introduced.

Akira Shimazu · Minh Le Nguyen
School of Information Science, Japan Advanced Institute of Science and Technology

V.-N. Huynh et al. (eds.), *Knowledge and Systems Engineering, Volume 1*,
Advances in Intelligent Systems and Computing 244,
DOI: 10.1007/978-3-319-02741-8_3,

Part II
KSE 2013 Main Track

A Hierarchical Approach for High-Quality and Fast Image Completion

Thanh Trung Dang, Azeddine Beghdadi, and Mohamed-Chaker Larabi

Abstract. Image inpainting is not only the art of restoring damaged images but also a powerful technique for image editing e.g. removing undesired objects, recomposing images, etc. Recently, it becomes an active research topic in image processing because of its challenging aspect and extensive use in various real-world applications. In this paper, we propose a novel efficient approach for high-quality and fast image restoration by combining a greedy strategy and a global optimization strategy based on a pyramidal representation of the image. The proposed approach is validated on different state-of-the-art images. Moreover, a comparative validation shows that the proposed approach outperforms the literature in addition to a very low complexity.

1 Introduction

Image inpainting, also known as blind image completion, is not only the art of restoring damaged images; but also a powerful technique in many real-world applications, such as image editing (removing undesired objects, restoring scratches), film reproduction (deleting logos, subtitles, and so on), or even creating artistic effects (reorganizing objects, smart resizing of images, blending images). Recently, it becomes an active research topic in image processing because of its challenging aspect and extensive use in various real-world applications. This topic began by skillful and professional artists in museum to manually restore the old painting.

Digital image inpainting tries to mimic this very precise process in an automatic manner on computers. Because the completion is performed blindly without

Thanh Trung Dang · Azeddine Beghdadi
L2TI, Institut Galilée, Université Paris 13, France
e-mail: {dang.thanhtrung,azeddine.beghdadi}@univ-paris13.fr

Mohamed-Chaker Larabi
XLIM, Dept. SIC, Université de Poitiers, France
e-mail: chaker.larabi@univ-poitiers.fr

V.-N. Huynh et al. (eds.), *Knowledge and Systems Engineering, Volume 1*,
Advances in Intelligent Systems and Computing 244,
DOI: 10.1007/978-3-319-02741-8_4,

reference to original images, the aim of digital image completion is only restoring the damaged image by maintaining its naturalness, i.e undetectable by viewers. However, this task is extremely difficult in the case of high resolution and structured images. On the one hand, the restored parts should not be visible or perceptually annoying to human viewers when filled; on the other hand, the used algorithm needs to be robust, efficient and requiring minimal user interactions and quick feedbacks.

An image inpainting algorithm often works in two stages. First the missing or damaged regions are identified (inpainting regions or target regions). Second, these regions are filled in the most natural manner possible. Up to now, there is no approach for automatically detecting damaged regions to be restored. For the sake of simplicity, they are usually marked manually using image editing softwares. Several approaches have been proposed in the literature and they may be categorized into two main groups [1]: geometry-oriented methods and texture-oriented methods.

The methods of the first group are designed to restore small or thin regions such as scratches or blotches, overlaid text, subtitles, etc. In this group, the image is modeled as a function of smoothness and the restoration is solved by interpolating the geometric information within the adjacent regions into the target region. Approaches falling in this category show good performance in propagating smooth level lines or gradient but they have the tendency to generate synthesis artifacts or blur effects in the case of large missing regions [2, 3, 4].

Whereas, the objective of the methods in the second group is to recover larger areas where the texture is assumed to be spatially stationary. Texture is modeled through probability distribution of the pixel brightness values. The pixel intensity distribution depends on only its neighborhood. This group could be further subdivided into two subgroups named: greedy strategy [5, 6, 7, 8] and global optimization strategy [10, 11, 12]. Greedy strategies have acceptable computation time and take into account human perception features (priority is designed based on the salient structures considered as important for human perception). However, some problems such as local optimization and patch selection may limit the efficiency of these approaches. In contrast, global optimization strategies often provide better results. But, they are computationally expensive. This is mainly due to the fact that time complexity increases linearly both with the number of source pixels and unknown pixels.

In this study, we propose a novel approach for high-quality and fast image completion by combining both greedy and global optimization strategies based on a pyramidal representation of the image [13]. The use of pyramidal representation is twofold: first it allows accounting for the multi-scale characteristics of the HVS; second it offers a good way to accelerate the completion process. It is worth noticing that a perceptual pyramidal representation [16] would be better but at the expense of increased computational complexity.

The proposal is directed by the observation that the human visual system is more sensitive to salient structures being stable and repetitive at different scales. Also, a hierarchical completion is a suitable solution for preserving high frequency components in a visually plausible way, and thus generates high-quality outputs. Namely, a top-down completion is implemented from top level (the lowest resolution) to the

bottom level (the original resolution). A greedy algorithm is applied for the lowest resolution to complete the damaged regions and create a good initialization accounting for the human perception for the next level. At each higher level, a relation map, called shift-map, is interpolated from adjacently lower level and then optimized by a global optimization algorithm, i.e. multi-label graph-cuts [12, 14]. Experimental results highlight a noticeable improvement in both implementation performance and quality of the inpainted image. To affirm the performance of our implementation, the running time is calculated in comparison with some typical inpainting methods. To confirm the quality of our results, the viewer can visually evaluate outputs of inpainting approaches in conjunction with some objective inpainting quality metrics [17, 18].

The rest of the paper is organized as follows. More details of our framework are introduced in section 2. Section 3 is dedicated to experimental results and comparison with the state-of-the-art methods. Finally, this paper ends with some conclusions and future works.

2 Our Proposal

The inpainting problem could be considered as an optimal graph labeling where a shift-map represents the selected label for each unknown pixels and it could be solved by optimizing an energy function using multi-label graph cuts. Because an unknown pixel in the damaged regions could originate from any pixel in the source regions, the global optimization strategies can be computationally infeasible. Moreover, they consider fairly possible label assignments but this does not fit with human perception. In term of inpainting quality, fair assignments may lead to unexpected bias for optimization. In terms of speed, a huge label set requires high computational load.

Our method is designed to overcome these limitations. In order to reduce the memory and computational requirements, a hierarchical approach for optimizing the graph labeling is developed. This hierarchy could provide enough-good results for inpainting problem, even though optimality cannot be guaranteed. In order to take into account human perception, a greedy strategy is applied at the lowest resolution to generate a suitable initialization for the next pyramidal levels. The priority of greedy strategy is designed based on the salient structures considered as one of the most important features for the HVS. An algorithmic description of our framework is given in the Fig. 1.

For details, some notations that are similar to those in paper [7] are adopted. The whole image domain, I, is composed of two disjoint regions: the inpainting region (or target region) Ω, and the source region Φ ($\Phi = I - \Omega$). According to the above idea, a set of images $G_0, G_1, ..., G_N$ with various levels of details is generated using pyramidal operators, where $G_0 = I$ is the input or original image [13]. The inpainting regions are also reduced to the eliminated areas level by level.

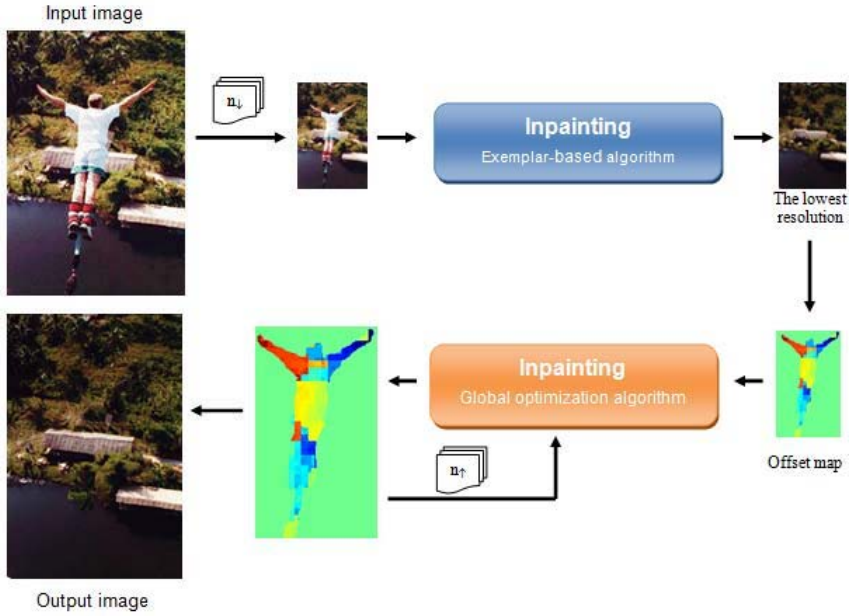


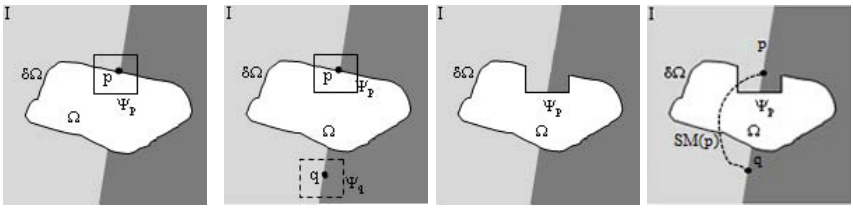
Fig. 1 Our scheme

2.1 Lowest Level Completion

In order to take into account HVS properties, a greedy strategy is applied for the lowest resolution. In our framework, an extension of [7] is developed to complete the reduced inpainting image. The algorithm for a single resolution image repeats the following steps (Fig. 2):

1. *Initialization*: Identify inpaiting boundary, $\delta\Omega$. If there is no pixel on the boundary, the algorithm is terminated.
2. *Priority estimation*: Compute the priority, $P(p)$, for all pixels on boundary, $p \in \delta\Omega$ and select randomly a pixel p with the highest priority.
3. *Patch match*: Find the patch or window Ψ_q that is most similar to Ψ_p thus minimizing mean squared error with existing pixels.
4. *Patch filling*: Fill the missing information in patch Ψ_p by copying the corresponding pixels from patch Ψ_q.
5. *Update*: Update the shift-map, SM_N, defining the relation between filled pixels and their sources and return to the step 1 for next iteration.

In this strategy, a good priority definition is very important because a decision taken based on it could not be changed anymore. Many models for priority have been proposed in the literature [5, 6, 7, 8, 9]. In this work, we used the priority model proposed in [7], namely window-based priority, which is more robust than the



a. Priority estimation b. Patch match c. Inpaint d. Shift definition

Fig. 2 The greedy strategy

others. After inpainting the image at the lowest resolution, a complete shift-map is generated and used as an initialization for the completion of next levels.

2.2 Higher Level Completion

Since the principle of inpainting is to fill in unknown pixels ($p(x_p, y_p) \in \Omega$) using the most plausible source pixels ($q(x_q, y_q) \in \Phi$), a relationship between them needs to be defined. This relation can be characterized by a shift-map determining an offset from known pixel to unknown one for each coordinate in the image (Fig. 3b). The shift-map can be formulated by eq. (1). Then the output pixel $O(p)$ is derived from the input pixel $I(p + SM(p))$.

$$SM(p) = \begin{cases} (\triangle x, \triangle y) & p(x,y) \in \Omega \\ (0,0) & \text{otherwise} \end{cases} \tag{1}$$

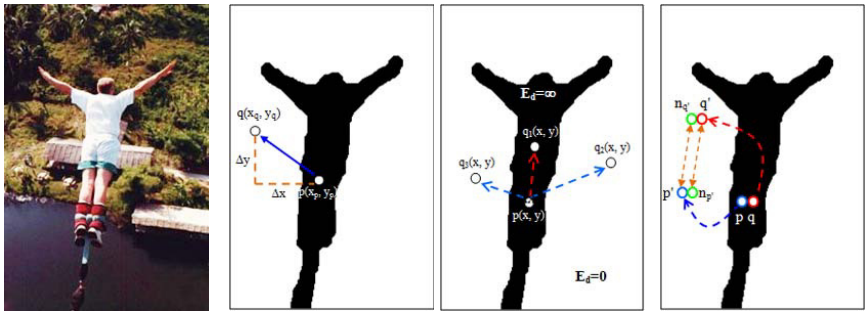
The naturalness of the resulting image is one of the most important issue of inpainting. Therefore, the used shift-map has to comply with such a requirement. In [12], authors proposed a solution to evaluate the shift-map by designing an energy function and optimizing it by a graph-cut algorithm. The energy function is defined as follows:

$$EM = \alpha \sum_{p \in \Omega} E_d(SM(p)) + (1-\alpha) \sum_{(p,q) \in NB} E_s(SM(p), SM(q)) \tag{2}$$

Where E_d is a data term providing external requirements and E_s is a smoothness term defined over a set of neighboring pixels, NB. α is a user defined weight balancing the two terms fixed to $\alpha = 0.5$ in our case. Once the graph and energy function are given, the shift-map labeling is computed using multi-label graph-cuts algorithm [14, 15].

2.2.1 A. Data Term

The data term E_d is used to include external constraints. Because the unknown pixels are filled thanks to the known ones, the data term assumes that no pixels in the hole are used in the output image. The detail of the data term is given by Eq. (3):



(a) An original image (b) An offset (c) Data term (d) Smoothness term

Fig. 3 Algorithm Operators

$$E_d(SM(p)) = \begin{cases} \infty & (x+\triangle x, y+\triangle y) \in \Omega \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In some cases, the specific pixels in the input image can be forced to appear or disappear in the output image by setting $E_d = \infty$. For example, saliency map can be used to weight the data term. Therefore, a pixel with a high saliency value should be kept and a pixel with a low saliency value should be removed (Fig. 3c).

2.2.2 B. Smoothness Term

The smoothness term represents discontinuity between two neighbor pixels $p(x_p, y_p)$ and $q(x_q, y_q)$. In paper [12], the authors proposed an effective formula for smoothness term which takes into account both color differences and gradient differences between corresponding spatial neighbors in the output image and in the input image to create good stitching. This treatment is represented as eq. (4) (Fig. 3d):

$$E_s(SM(p), SM(q)) = \begin{cases} 0 & SM(p) = SM(q) \\ \beta\delta M(SM(p)) + \gamma\delta G(SM(p)) & otherwise \end{cases} \tag{4}$$

where β and γ are weights balancing these two terms, set to $\beta = 1$, $\gamma = 2$ in our experiment. δM and δG denote the differences of magnitude and gradient and they are defined as the follows:

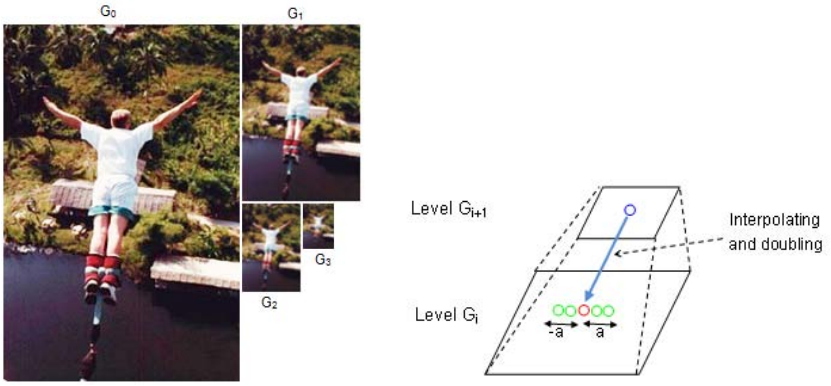
$$\begin{aligned} \delta M(SM(p)) &= ||I(n_{p'}) - I(q')|| + ||I(n_{q'}) - I(p')|| \\ \delta G(SM(p)) &= ||\nabla I(n_{p'}) - \nabla I(q')|| + ||\nabla I(n_{q'}) - \nabla I(p')|| \end{aligned} \tag{5}$$

where, I and ∇I are the magnitude and gradient at these locations. $p' = p + SM(p)$ and $q' = q + SM(q)$ are locations used to fill pixels p and q, respectively. $n_{p'}$ and $n_{q'}$ are two 4-connected neighbors of p' and q', respectively (Fig. 3d).

2.3 *Shift-Map Interpolation*

A full shift-map is first inferred from a completion at the lowest level of pyramid. Then it is interpolated to higher resolutions using a *nearest neighbor interpolation*, and the shift-map values are *doubled* to match the higher image resolution.

At the higher level, only small shifts relative to the initial guess are examined. It means that only some parent neighbors are considered instead of all possible labels. In our implementation, the shift relative for each coordinate varies in range [-a, a], so it takes $(2a+1)^2$ labels for both direction. It is important to note that the data and smoothness terms are always computed with respect to the actual shifts and not to the labels (Fig. 4).



(a) Image gaussian pyramid (b) Interpolation between two adjacent levels

Fig. 4 Interpolation of Shift-Map

3 Experimental Results

This section is dedicated to the study of performance of the proposed algorithm using some typical real images that cover several major challenges for inpainting. In order to try and cover all inpainting methods would be infeasible. For the sake of comparison with literature, three inpainting methods corresponding to algorithms proposed by *A. Criminisi et al* [5] and *T. T. Dang et al* [7] for greedy strategy and *Y. Pritch et al* [12] for global optimization strategy have been implemented. Five images, given on Fig. 6 were chosen for this experiment (including *bungee* (206×308), *angle* (300×252), *silenus* (256×480), *boat* (300×225) and *seaman* (300×218)).

Figure 6 illustrates the results obtained with the proposed approach in comparison to the others. Fig. 6a gives images to be inpainted where damaged areas cover respectively *12.6%*, *5.83%*, *7.74%*, *10.73%* and *14.87%* of the whole image.

To evaluate the quality of inpainting output, some objective inpainted image quality metrics [17, 18] are considered and the metric in [18] is developed because all used images in our experiment are color. The metric values are shown in the table 1 and compared more visually in figure 5.

Table 1 The inpainted image quality metrics

Image	*bungee*	*angle*	*silenus*	*boat*	*seaman*
Size	(206×308)	(300×252)	(256×480)	(300×225)	(300×218)
Damaged Area	12.6%	5.83%	7.74%	10.73%	14.87%
A. Criminisi [5]	0.0685	0.0817	0.0358	0.061	0.0449
T. T. Dang [7]	**0.1157**	**0.0898**	0.0407	0.065	0.0572
Y. Pritch [12]	0.0343	0.0805	0.0289	0.0597	0.0407
Our proposal	0.107	0.087	**0.0407**	**0.069**	**0.0592**

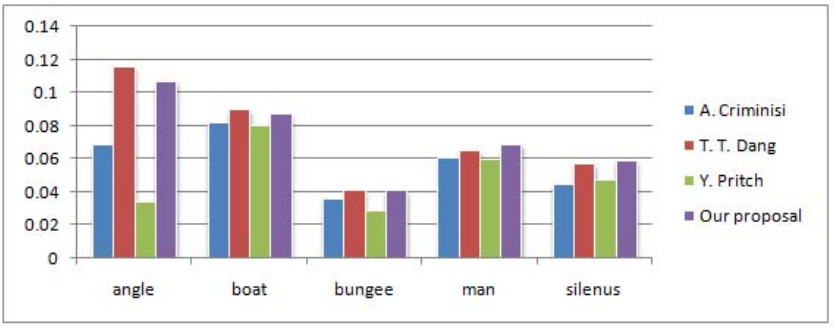


Fig. 5 A chart of quality performance

The performance of the proposed approach is quantitatively evaluated by implementation time in comparison with the other approaches. In order to avoid bias, all approaches are programmed by the same programming language, C/C++ programming language, and implemented on the same PC with the configuration of Intel Core i5, 2.8GHz CPU and 4GB RAM. The running time in seconds of each methods is given in table 2 and shown visually in figure 7. As it can be seen from these results, our method provides an acceptable visual quality, often outperforming the others, with a much faster implementation. Indeed, visual inspection of results shows that the completion performed by our approach looks more natural and more coherent than the other approaches.

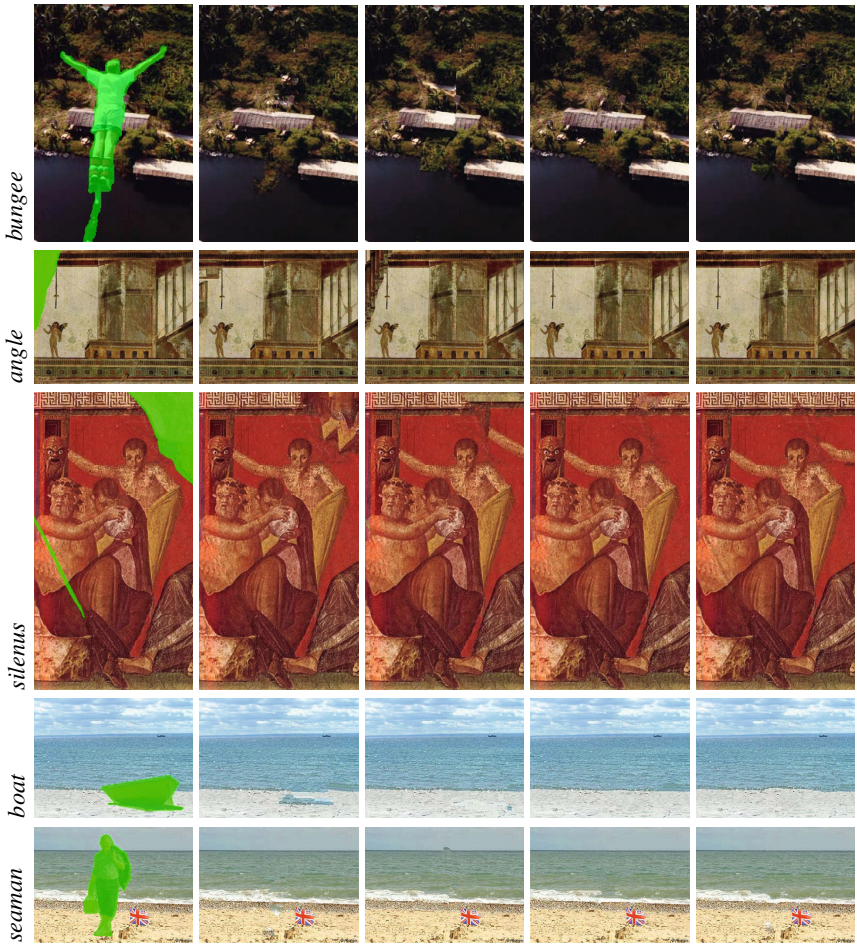


Fig. 6 The experimental results. (a) Image to be inpainted; The outputs when using the methods in (b) [5]; (c) [12]; (d) [7]; (e) our proposal.

Table 2 Computational time (in second) for implemented approaches and the set of used images

Image	*bungee*	*angle*	*silenus*	*boat*	*seaman*
Size	(206×308)	(300×252)	(256×480)	(300×225)	(300×218)
Damaged Area	12.6%	5.83%	7.74%	10.73%	14.87%
A. Criminisi [5]	16.30	8.20	38.29	24.54	27.31
T. T. Dang [7]	15.92	16.36	63.18	50.18	55.16
Y. Pritch [12]	35.39	13.24	57.68	21.18	15.50
Our proposal	**3.32**	**5.81**	**7.53**	**7.25**	**5.97**

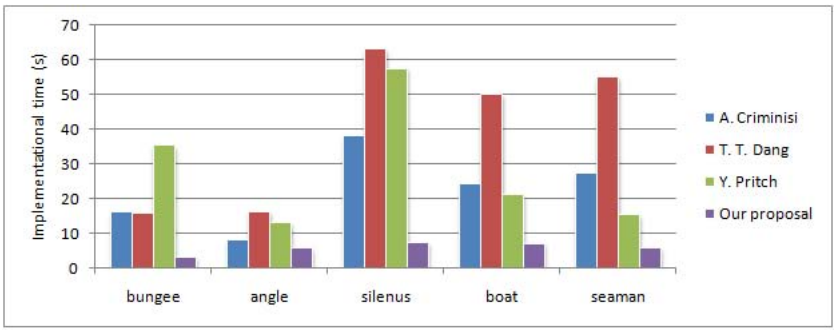


Fig. 7 A chart of implementation performance

4 Conclusions

In this paper, a novel framework of image completion is introduced by combining both greedy and global optimization strategies based on a pyramidal representation of the image. The greedy strategy is applied at the lowest resolution in order to generate a good initialization accounting for human perception. At higher resolutions, the shift map is refined by a global optimization algorithm and multi-label graph-cuts. A comparison with some representative approaches from literature belonging to the second group (i.e. global optimization) is carried out and results show that our approach not only produces better quality of output images but also implements noticeably faster.

The obtained results are very encouraging and a more thorough evaluation procedure, including both objective and subjective evaluation, will be engaged as a future work. Computational complexity issues will be also addressed.

References

1. Arias, P., Facciolo, G., Caselles, V., Sapiro, G.: A Variational Framework for Exemplar-Based Image Inpainting. International Journal of Computer Vision, 1–29 (2011)
2. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 417–424 (2000)
3. Chan, T.F., Shen, J.: Non-texture inpainting by Curvature-Driven Diffusions (CCD). Journal of Visual Communication and Image Representation 4, 436–449 (2001)
4. Tschumperle, D.: Fast anisotropic smoothing of multi-valued images using curvature-preserving pdes. International Journal of Computer Vision 68, 65–82 (2006)
5. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transaction of Image Process 13(9), 1200–1212 (2004)
6. Wu, J., Ruan, Q.: Object removal by cross isophotes exemplar based image inpainting. In: Proceeding of International Conference of Pattern Recognition, pp. 810–813 (2006)

7. Dang, T.T., Larabi, M.C., Beghdadi, A.: Multi-resolution patch and window-based priority for digital image inpainting problem. In: 3rd International Conference on Image Processing Theory, Tools and Applications, pp. 280–284 (2012)
8. Zhang, Q., Lin, J.: Exemplar-based image inpainting using color distribution analysis. Journal of Information Science and Engineering (2011)
9. Cheng, W., Hsieh, C., Lin, S., Wang, C., Wu, J.: Robust algorithm for exemplar-based image inpainting. In: Proceeding of International Conference on Computer Graphics, Imaging and Visualization (2005)
10. Wexler, Y., Shechtman, E., Irani, M.: Space-time video completion. IEEE Transactions Pattern Analysis and Machine Intelligence 29, 463–476 (2007)
11. Komodakis, G.T.N., Tziritas, G.: Image completion using global optimization. In: Proceeding of IEEE Computer Society Conference Computer Vision and Pattern Recognition, pp. 442–452 (2006)
12. Pritch, Y., Kav-Venaki, E., Peleg, S.: Shift-map image editing. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 151–158 (2009)
13. Peter, J.B., Edward, H.A.: The Laplacian pyramid as a compact image code. IEEE Transactions on Communications 31, 532–540 (1983)
14. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(11), 1222–1239 (2001)
15. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive Digital Photomontage. In: Proceedings of SIGGRAPH, pp. 294–302 (2004)
16. Iordache, R., Beghdadi, A., de Lesegno, P.V.: Pyramidal perceptual filtering using Moon and Spencer contrast. In: International Conference on Image Processing, ICIP 2001, pp. 146–149 (2001)
17. Dang, T.T., Beghdadi, A., Larabi, M.C.: Perceptual evaluation of digital image completion quality. In: 21st European Signal Processing Conference, EUSIPCO 2013 (2013)
18. Dang, T.T., Beghdadi, A., Larabi, M.C.: Perceptual quality assessment for color image inpainting. In: IEEE International Conference on Image Processing, ICIP 2013 (2013)

The Un-normalized Graph p-Laplacian Based Semi-supervised Learning Method and Protein Function Prediction Problem

Loc Tran

Abstract. Protein function prediction is a fundamental problem in modern biology. In this paper, we present the un-normalized graph p-Laplacian semi-supervised learning methods. These methods will be applied to the protein network constructed from the gene expression data to predict the functions of all proteins in the network. These methods are based on the assumption that the labels of two adjacent proteins in the network are likely to be the same. The experiments show that that the un-normalized graph p-Laplacian semi-supervised learning methods are at least as good as the current state of the art method (the un-normalized graph Laplacian based semi-supervised learning method) but often lead to better classification accuracy performance measures.

1 Introduction

Protein function prediction is the important problem in modern biology. Identifying the function of proteins by biological experiments is very expensive and hard. Hence a lot of computational methods have been proposed to infer the functions of the proteins by using various types of information such as gene expression data and protein-protein interaction networks [1].

The classical way predicting protein function infers the similarity to function from sequence homologies among proteins in the databases using sequence similarity algorithms such as FASTA [2] and PSI-BLAST [3]. Next, to predict protein function, graph which is the natural model of relationship between proteins or genes can also be employed. This model can be protein-protein interaction network or gene co-expression network. In this model, the nodes represent proteins or genes and the edges represent for the possible interactions between nodes. Then, machine learning methods such as Support Vector Machine [5], Artificial Neural Networks

Loc Tran
University of Minnesota, USA
e-mail: tran0398@umn.edu

V.-N. Huynh et al. (eds.), *Knowledge and Systems Engineering, Volume 1*,
Advances in Intelligent Systems and Computing 244,
DOI: 10.1007/978-3-319-02741-8_5,

[4], un-normalized graph Laplacian based semi-supervised learning method [6], the symmetric normalized and random walk graph Laplacian based semi-supervised learning methods [7], or neighbor counting method [8] can be applied to this graph to infer the functions of un-annotated protein. The neighbor counting method labels the protein with the function that occurs frequently in the protein's adjacent nodes in the protein-protein interaction network and hence does not utilized the full topology of the network. However, the Artificial Neural Networks, Support Vector Machine, un-normalized, symmetric normalized and random walk graph Laplacian based semi-supervised learning method utilizes the full topology of the network. The Artificial Neural Networks and Support Vector Machine are all supervised learning methods. The neighbor counting method, the Artificial Neural Networks, and the three graph Laplacian based semi-supervised learning methods are all based on the assumption that the labels of two adjacent proteins in graph are likely to be the same. However, SVM do not rely on this assumption. Unlike graphs used in neighbor counting method, Artificial Neural Networks, and the three graph Laplacian based semi-supervised learning methods are very sparse, the graph (i.e. kernel) used in SVM is fully-connected.

The Artificial Neural Networks method is applied to the single protein-protein interaction network. However, the SVM method and three graph Laplacian based semi-supervised learning methods try to use weighted combination of multiple networks (i.e. kernels) such as gene co-expression network and protein-protein interaction network to improve the accuracy performance measures. [5] (SVM method) determines the optimal weighted combination of networks by solving the semi-definite problem. [6] (un-normalized graph Laplacian based semi-supervised learning method) uses a dual problem and gradient descent to determine the weighted combination of networks. [7] uses the integrated network combined with equal weights, i.e. without optimization due to the integrated network combined with optimized weights has similar performance to the integrated network combined with equal weights and the high time complexity of optimization methods.

The un-normalized, symmetric normalized, and random walk graph Laplacian based semi-supervised learning methods are developed based on the assumption that the labels of two adjacent proteins or genes in the network are likely to be the same [6]. Hence this assumption can be interpreted as pairs of genes showing a similar pattern of expression and thus sharing edges in a gene co-expression network tend to have similar function. In [9], the single gene expression data is used for protein function prediction problem. However, assuming the pairwise relationship between proteins or genes is not complete, the information a group of genes that show very similar patterns of expression and tend to have similar functions [12] (i.e. the functional modules) is missed. The natural way overcoming the information loss of the above assumption is to represent the gene expression data as the hypergraph [10,11]. A hypergraph is a graph in which an edge (i.e. a hyper-edge) can connect more than two vertices. In [9], the un-normalized, random walk, and symmetric normalized hypergraph Laplacian based semi-supervised learning methods have been developed and successfully outperform the un-normalized, symmetric

normalized, and random walk graph Laplacian based semi-supervised learning methods in protein function prediction problem.

In [13,14], the symmetric normalized graph p-Laplacian based semi-supervised learning method has been developed but has not been applied to any practical applications. To the best of my knowledge, the un-normalized graph p-Laplacian based semi-supervised learning method has not yet been developed and obviously has not been applied to protein function prediction problem. This method is worth investigated because of its difficult nature and its close connection to partial differential equation on graph field. Specifically, in this paper, the un-normalized graph p-Laplacian based semi-supervised learning method will be developed based on the un-normalized graph p-Laplacian operator definition such as the curvature operator of graph (i.e. the un-normalized graph 1-Laplacian operator). Please note that the un-normalized graph p-Laplacian based semi-supervised learning method is developed based on the assumption that the labels of two adjacent proteins or genes in the network are likely to be the same [6].

We will organize the paper as follows: Section 2 will introduce the preliminary notations and definitions used in this paper. Section 3 will introduce the definition of the gradient and divergence operators of graphs. Section 4 will introduce the definition of Laplace operator of graphs and its properties. Section 5 will introduce the definition of the curvature operator of graphs and its properties. Section 6 will introduce the definition of the p-Laplace operator of graphs and its properties. Section 7 will show how to derive the algorithm of the un-normalized graph p-Laplacian based semi-supervised learning method from regularization framework. In section 8, we will compare the accuracy performance measures of the un-normalized graph Laplacian based semi-supervised learning algorithm (i.e. the current state of art method applied to protein function prediction problem) and the un-normalized graph p-Laplacian based semi-supervised learning algorithms. Section 9 will conclude this paper and the future direction of researches of other practical applications in bioinformatics utilizing discrete operator of graph will be discussed.

2 Preliminary Notations and Definitions

Given a graph *G=(V,E,W)* where V is a set of vertices with $|V| = n$, $E \subseteq V * V$ is a set of edges and W is a $n * n$ similarity matrix with elements $w_{ij} > 0$ $(1 \leq i, j \leq n)$.

Also, please note that $w_{ij} = w_{ji}$.

The degree function $d : V \to R^+$ is

$$d_i = \sum_{j \sim i} w_{ij}, \tag{1}$$

where $j \sim i$ is the set of vertices adjacent with i.

Define $D = diag(d_1, d_2, \ldots, d_n)$.

The inner product on the function space R^V is

$$< f, g >_V = \sum_{i \in V} f_i g_i \tag{2}$$

Also define an inner product on the space of functions R^E on the edges

$$< F, G >_E = \sum_{(i,j)\in E} F_{ij} G_{ij} \tag{3}$$

Here let $H(V) = (R^V, <.,.>_V)$ and $H(E) = (R^E, <.,.>_E)$ be the Hilbert space real-valued functions defined on the vertices of the graph G and the Hilbert space of real-valued functions defined in the edges of G respectively.

3 Gradient and Divergence Operators

We define the gradient operator $d : H(V) \to H(E)$ to be

$$(df)_{ij} = \sqrt{w_{ij}}(f_j - f_i), \tag{4}$$

where $f : V \to R$ be a function of $H(V)$.

We define the divergence operator $div : H(E) \to H(V)$ to be

$$< df, F >_{H(E)} = < f, -divF >_{H(V)}, \tag{5}$$

where $f \in H(V), F \in H(E)$

Next, we need to prove that

$$(divF)_j = \sum_{i\sim j} \sqrt{w_{ij}}(F_{ji} - F_{ij})$$

Proof:

$$< df, F >= \sum_{(i,j)\in E} df_{ij} F_{ij}$$

$$= \sum_{(i,j)\in E} \sqrt{w_{ij}}(f_j - f_i) F_{ij}$$

$$= \sum_{(i,j)\in E} \sqrt{w_{ij}} f_j F_{ij} - \sum_{(i,j)\in E} \sqrt{w_{ij}} f_i F_{ij}$$

$$= \sum_{k\in V} \sum_{i\sim k} \sqrt{w_{ik}} f_k F_{ik} - \sum_{k\in V} \sum_{j\sim k} \sqrt{w_{kj}} f_k F_{kj}$$

$$= \sum_{k\in V} f_k (\sum_{i\sim k} \sqrt{w_{ik}} F_{ik} - \sum_{i\sim k} \sqrt{w_{ki}} F_{ki})$$

$$= \sum_{k\in V} f_k \sum_{i\sim k} \sqrt{w_{ik}} (F_{ik} - F_{ki})$$

Thus, we have

$$(divF)_j = \sum_{i\sim j} \sqrt{w_{ij}}(F_{ji} - F_{ij}) \tag{6}$$

4 Laplace Operator

We define the Laplace operator $\triangle : H(V) \to H(V)$ to be

$$\triangle f = -\frac{1}{2} div(df) \tag{7}$$

Next, we compute

$$(\triangle f)_j = \frac{1}{2}\sum_{i\sim j}\sqrt{w_{ij}}((df)_{ij} - (df)_{ji})$$

$$= \frac{1}{2}\sum_{i\sim j}\sqrt{w_{ij}}(\sqrt{w_{ij}}(f_j - f_i) - \sqrt{w_{ij}}(f_i - f_j))$$

$$= \sum_{i\sim j} w_{ij}(f_j - f_i)$$

$$= \sum_{i\sim j} w_{ij} f_j - \sum_{i\sim j} w_{ij} f_i$$

$$= d_j f_j - \sum_{i\sim j} w_{ij} f_i$$

Thus, we have

$$(\triangle f)_j = d_j f_j - \sum_{i\sim j} w_{ij} f_i \tag{8}$$

The graph Laplacian is a linear operator. Furthermore, the graph Laplacian is self-adjoint and positive semi-definite.

Let $S_2(f) = <\triangle f, f>$, we have the following **theorem 1**

$$D_f S_2 = 2\triangle f \tag{9}$$

The proof of the above theorem can be found from [13,14].

5 Curvature Operator

We define the curvature operator $\kappa : H(V) \to H(V)$ to be

$$\kappa f = -\frac{1}{2} div(\frac{df}{||df||}) \tag{10}$$

Next, we compute

$$(\kappa f)_j = \frac{1}{2}\sum_{i\sim j}\sqrt{w_{ij}}((\frac{df}{||df||})_{ij} - (\frac{df}{||df||})_{ji})$$

$$= \frac{1}{2}\sum_{i\sim j}\sqrt{w_{ij}}\left(\frac{1}{\|d_i f\|}\sqrt{w_{ij}}(f_j - f_i) - \frac{1}{\|d_j f\|}\sqrt{w_{ij}}(f_i - f_j)\right)$$

$$= \frac{1}{2}\sum_{i\sim j} w_{ij}\left(\frac{1}{\|d_i f\|} + \frac{1}{\|d_j f\|}\right)(f_j - f_i)$$

Thus, we have

$$(\kappa f)_j = \frac{1}{2}\sum_{i\sim j} w_{ij}\left(\frac{1}{\|d_i f\|} + \frac{1}{\|d_j f\|}\right)(f_j - f_i) \tag{11}$$

From the above formula, we have

$$d_i f = ((df)_{ij} : j \sim i)^T \tag{12}$$

The local variation of f at i is defined to be

$$\|d_i f\| = \sqrt{\sum_{j\sim i}(df)_{ij}^2} = \sqrt{\sum_{j\sim i} w_{ij}(f_j - f_i)^2} \tag{13}$$

To avoid the zero denominators in (11), the local variation of f at i is defined to be

$$\|d_i f\| = \sqrt{\sum_{j\sim i}(df)_{ij}^2 + ?}, \tag{14}$$

where $? = 10^{-10}$.
The graph curvature is a non-linear operator.
Let $S_1(f) = \sum_i \|d_i f\|$, we have the following **theorem 2**

$$D_f S_1 = \kappa f \tag{15}$$

The proof of the above theorem can be found from [13,14].

6 p-Laplace Operator

We define the p-Laplace operator $\triangle_p : H(V) \to H(V)$ to be

$$\triangle_p f = -\frac{1}{2} div(\|df\|^{p-2} df) \tag{16}$$

Clearly, $\triangle_1 = \kappa$ and $\triangle_2 = \triangle$. Next, we compute

$$(\triangle_p f)_j = \frac{1}{2}\sum_{i\sim j}\sqrt{w_{ij}}(\|df\|^{p-2} df_{ij} - \|df\|^{p-2} df_{ji})$$

$$= \frac{1}{2}\sum_{i\sim j}\sqrt{w_{ij}}(\|d_i f\|^{p-2}\sqrt{w_{ij}}(f_j - f_i) - \|d_j f\|^{p-2}\sqrt{w_{ij}}(f_i - f_j))$$

$$= \frac{1}{2}\sum_{i\sim j} w_{ij}(\|d_i f\|^{p-2} + \|d_j f\|^{p-2})(f_j - f_i)$$

Thus, we have

$$(\triangle_p f)_j = \frac{1}{2}\sum_{i\sim j} w_{ij}(\|d_i f\|^{p-2} + \|d_j f\|^{p-2})(f_j - f_i) \tag{17}$$

Let $S_p(f) = \frac{1}{p}\sum_i \|d_i f\|^p$, we have the following **theorem 3**

$$D_f S_p = p\triangle_p f \tag{18}$$

7 Discrete Regularization on Graphs and Protein Function Classification Problems

Given a protein network *G=(V,E)*. *V* is the set of all proteins in the network and *E* is the set of all possible interactions between these proteins. Let *y* denote the initial function in *H(V)*. y_i can be defined as follows

$$y_i = \begin{cases} 1\ if\ protein\ i\ belongs\ to\ the\ functional\ class \\ -1\ if\ protein\ i\ does\ not\ belong\ to\ the\ functional\ class \\ 0\ otherwise \end{cases}$$

Our goal is to look for an estimated function f in *H(V)* such that f is not only smooth on *G* but also close enough to an initial function *y*. Then each protein *i* is classified as $sign(f_i)$. This concept can be formulated as the following optimization problem

$$argmin_{f\in H(V)}\{S_p(f) + \frac{\mu}{2}\|f-y\|^2\} \tag{19}$$

The first term in (19) is the smoothness term. The second term is the fitting term. A positive parameter μ captures the trade-off between these two competing terms.

7.I) **2-smoothness**

When *p=2*, the optimization problem (19) is

$$argmin_{f\in H(V)}\{\frac{1}{2}\sum_i \|d_i f\|^2 + \frac{\mu}{2}\|f-y\|^2\} \tag{20}$$

By theorem 1, we have

Theorem 4: The solution of (20) satisfies

$$\triangle f + \mu(f-y) = 0 \tag{21}$$

Since $\triangle$ is a linear operator, the closed form solution of (21) is

$$f = \mu(\triangle + \mu I)^{-1} y, \tag{22}$$

Where I is the identity operator and $\triangle = D - W$. (22) is the algorithm proposed by [6].

7.II) **1-smoothness**

When $p=1$, the optimization problem (19) is

$$argmin_{f\in H(V)}\{\sum_i \|d_i f\| + \frac{\mu}{2}\|f-y\|^2\}, \tag{23}$$

By theorem 2, we have

Theorem 5: The solution of (23) satisfies

$$\kappa f + \mu\,(f-y) = 0, \tag{24}$$

The curvature κ is a non-linear operator; hence we do not have the closed form solution of equation (24). Thus, we have to construct iterative algorithm to obtain the solution. From (24), we have

$$\frac{1}{2}\sum_{i\sim j} w_{ij}\left(\frac{1}{\|d_i f\|} + \frac{1}{\|d_j f\|}\right)(f_j - f_i) + \mu\,(f_j - y_j) = 0 \tag{25}$$

Define the function $m : E \to R$ by

$$m_{ij} = \frac{1}{2}w_{ij}\left(\frac{1}{\|d_i f\|} + \frac{1}{\|d_j f\|}\right) \tag{26}$$

Then (25)

$$\sum_{i\sim j} m_{ij}\,(f_j - f_i) + \mu\,(f_j - y_j) = 0$$

can be transformed into

$$\left(\sum_{i\sim j} m_{ij} + \mu\right) f_j = \sum_{i\sim j} m_{ij} f_i + \mu y_j \tag{27}$$

Define the function $p : E \to R$ by

$$p_{ij} = \begin{cases} \frac{m_{ij}}{\sum_{i\sim j} m_{ij}+\mu} \; if\; i \neq j \\ \frac{\mu}{\sum_{i\sim j} m_{ij}+\mu} \; if\; i = j \end{cases} \tag{28}$$

Then

$$f_j = \sum_{i\sim j} p_{ij} f_i + p_{jj} y_j \tag{29}$$

Thus we can consider the iteration

$f_j^{(t+1)} = \sum_{i\sim j} p_{ij}^{(t)} f_i^{(t)} + p_{jj}^{(t)} y_j$ for all $j \in V$

to obtain the solution of (23).

7.III) **p-smoothness**

For any number p, the optimization problem (19) is

$$argmin_{f\in H(V)}\{\frac{1}{p}\sum_i \|d_i f\|^p + \frac{\mu}{2}\|f-y\|^2\}, \tag{30}$$

By theorem 3, we have

Theorem 6: The solution of (30) satisfies

$$\triangle_p f + \mu\,(f-y) = 0, \tag{31}$$

The *p-Laplace* operator is a non-linear operator; hence we do not have the closed form solution of equation (31). Thus, we have to construct iterative algorithm to obtain the solution. From (31), we have

$$\frac{1}{2}\sum_{i\sim j} w_{ij}\left(\|d_i f\|^{p-2} + \left\|d_j f\right\|^{p-2}\right)(f_j - f_i) + \mu\,(f_j - y_j) = 0 \tag{32}$$

Define the function $m : E \to R$ by

$$m_{ij} = \frac{1}{2}w_{ij}(\|d_i f\|^{p-2} + \left\|d_j f\right\|^{p-2}) \tag{33}$$

Then equation (32) which is

$$\sum_{i\sim j} m_{ij}\,(f_j - f_i) + \mu\,(f_j - y_j) = 0$$

can be transformed into

$$\left(\sum_{i\sim j} m_{ij} + \mu\right) f_j = \sum_{i\sim j} m_{ij} f_i + \mu y_j \tag{34}$$

Define the function $p : E \to R$ by

$$p_{ij} = \begin{cases} \frac{m_{ij}}{\sum_{i\sim j} m_{ij}+\mu} \; if \; i \neq j \\ \frac{\mu}{\sum_{i\sim j} m_{ij}+\mu} \; if \; i = j \end{cases} \tag{35}$$

Then

$$f_j = \sum_{i\sim j} p_{ij} f_i + p_{jj} y_j \tag{36}$$

Thus we can consider the iteration

$f_j^{(t+1)} = \sum_{i\sim j} p_{ij}^{(t)} f_i^{(t)} + p_{jj}^{(t)} y_j$ for all $j \in V$

to obtain the solution of (30).

8 Experiments and Results

8.1 Datasets

In this paper, we use the dataset available from [9,15] and the references therein. This dataset contains the gene expression data measuring the expression of 4062 S. cerevisiae genes under the set of 215 titration experiments. These proteins are annotated with 138 GO Biological Process functions. In the other words, we are given gene expression data ($R^{4062*215}$) matrix and the annotation (i.e. the label) matrix ($R^{4062*138}$). We filtered the datasets to include only those GO functions that had at least 150 proteins and at most 200 proteins. This resulted in a dataset containing 1152 proteins annotated with seven different GO Biological Process functions. Seven GO Biological Process functions are

1. Alcohol metabolic process
2. Proteolysis
3. Mitochondrion organization
4. Cell wall organization
5. rRNA metabolic process
6. Negative regulation of transcription, DNA-dependent, and
7. Cofactor metabolic process.

We refer to this dataset as **yeast**. There are three ways to construct the similarity graph from the gene expression data:

1. The ε-neighborhood graph: Connect all genes whose pairwise distances are smaller than ε.
2. k-nearest neighbor graph: Gene i is connected with gene j if gene i is among the k-nearest neighbor of gene j or gene j is among the k-nearest neighbor of gene i.
3. The fully connected graph: All genes are connected.

In this paper, the similarity function is the Gaussian similarity function

$$s(G(i,:),G(j,:) = e^{-\frac{d(G(i,:),G(j,:))}{t}}$$

In this paper, t is set to 1.25 and the 3-nearest neighbor graph is used to construct the similarity graph from **yeast**.

8.2 Experiments

In this section, we experiment with the above proposed un-normalized graph p-Laplacian methods with *p=1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9* and the current state of the art method (i.e. the un-normalized graph Laplacian based semi-supervised learning method *p=2*) in terms of classification accuracy performance measure. The accuracy performance measure Q is given as follows

$$Q = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

All experiments were implemented in Matlab 6.5 on virtual machine. The three-fold cross validation is used to compute the average accuracy performance measures of all methods used in this paper. The parameter μ is set to 1.

The accuracy performance measures of the above proposed methods and the current state of the art method is given in the following table 1.

Table 1 The comparison of accuracies of proposed methods with different p-values

Functional classes		1	2	3	4	5	6	7
Accuracy Performance Measures (%)	p=1	**86.20**	**84.64**	**84.72**	**83.94**	**92.71**	**85.16**	**86.72**
	p=1.1	86.11	84.03	**84.72**	83.59	92.53	84.81	86.46
	p=1.2	85.94	84.03	**84.72**	83.68	92.62	84.72	86.46
	p=1.3	85.50	82.12	83.25	82.38	92.27	83.42	85.50
	p=1.4	85.59	83.25	84.11	82.90	92.88	84.64	86.28
	p=1.5	85.50	82.90	83.77	82.73	92.80	84.38	86.11
	p=1.6	85.42	82.64	83.68	82.64	92.88	83.94	85.94
	p=1.7	85.42	82.29	83.33	82.47	92.62	83.85	85.85
	p=1.8	85.42	82.12	83.33	82.55	92.53	83.51	85.59
	p=1.9	85.24	82.12	83.07	82.47	92.27	83.51	85.42
	p=2 (i.e. the current state of the art method)	85.50	82.12	83.25	82.38	92.27	83.42	85.50

From the above table, we easily recognized that the un-normalized graph 1-Laplacian semi-supervised learning method outperform other proposed methods and the current state of art method. The results from the above table shows that the un-normalized graph p-Laplacian semi-supervised learning methods are at least as good as the current state of the art method (p=2) but often lead to better classification accuracy performance measures.

9 Conclusions

We have developed the detailed regularization frameworks for the un-normalized graph p-Laplacian semi-supervised learning methods applying to protein function prediction problem. Experiments show that the un-normalized graph p-Laplacian semi-supervised learning methods are at least as good as the current state of the art method (i.e. $p=2$) but often lead to significant better classification accuracy performance measures.

Moreover, these un-normalized graph p-Laplacian semi-supervised learning methods can not only be used in classification problem but also in ranking problem. In specific, given a set of genes (i.e. the queries) making up a protein complex/-pathways or given a set of genes (i.e. the queries) involved in a specific disease (for e.g. leukemia), these methods can also be used to find more potential members of the complex/pathway or more genes involved in the same disease by ranking genes in gene co-expression network (derived from gene expression data) or the protein-protein interaction network or the integrated network of them. The genes with the highest rank then will be selected and then checked by biologist experts to see if the extended genes in fact belong to the same complex/pathway or are involved in the same disease. These problems are also called complex/pathway membership determination and biomarker discovery in cancer classification.

References

1. Shin, H.H., Lisewski, A.M., Lichtarge, O.: Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics 23, 3217–3224 (2007)
2. Pearson, W.R., Lipman, D.J.: Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America 85, 2444–2448 (1998)
3. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E.L.: Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 14, 1675–1680 (1996)
4. Shi, L., Cho, Y., Zhang, A.: Prediction of Protein Function from Connectivity of Protein Interaction Networks. International Journal of Computational Bioscience 1(1) (2010)
5. Lanckriet, G.R.G., Deng, M., Cristianini, N., Jordan, M.I., Noble, W.S.: Kernel-based data fusion and its application to protein function prediction in yeast. In: Pacific Symposium on Biocomputing, PSB (2004)
6. Tsuda, K., Shin, H.H., Schoelkopf, B.: Fast protein classification with multiple networks. Bioinformatics (ECCB 2005) 21(suppl. 2), ii59–ii65 (2005)
7. Tran, L.: Application of three graph Laplacian based semi-supervised learning methods to protein function prediction problem. CoRR abs/1211.4289 (2012)
8. Schwikowski, B., Uetz, P., Fields, S.: A network of protein–protein interactions in yeast. Nature Biotechnology 18, 1257–1261 (2000)
9. Tran, L.: Hypergraph and protein function prediction with gene expression data. CoRR abs/1212.0388 (2012)

10. Zhou, D., Huang, J., Schoelkopf, B.: Beyond Pairwise Classification and Clustering Using Hypergraphs, Max Planck Institute Technical Report 143, Max Planck Institute for Biological Cybernetics, Tbingen, Germany (2005)
11. Zhou, D., Huang, J., Schoelkopf, B.: Learning with Hypergraphs: Clustering, Classification, and Embedding. In: Schoelkopf, B., Platt, J.C., Hofmann, T. (eds.) Advances in Neural Information Processing System (NIPS), pp. 1601–1608. MIT Press, Cambridge (2007)
12. Pandey, G., Atluri, G., Steinbach, M., Kumar, V.: Association Analysis Techniques for Discovering Functional Modules from Microarray Data. In: Proc. ISMB Special Interest Group Meeting on Automated Function Prediction (2008)
13. Zhou, D., Schölkopf, B.: Regularization on Discrete Spaces. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (eds.) DAGM 2005. LNCS, vol. 3663, pp. 361–368. Springer, Heidelberg (2005)
14. Zhou, D., Schoelkopf, B.: Discrete Regularization. In: Chapelle, O., Schoelkopf, B., Zien, A. (eds.) Semi-Supervised Learning, pp. 221–232. MIT Press, Cambridge (2006)
15. Pandey, G., Myers, L.C., Kumar, V.: Incorporating Functional Inter-relationships into Protein Function Prediction Algorithms. BMC Bioinformatics 10, 142 (2009)

On Horn Knowledge Bases in Regular Description Logic with Inverse

Linh Anh Nguyen, Thi-Bich-Loc Nguyen, and Andrzej Szałas

Abstract. We study a Horn fragment called Horn-$\mathscr{R}eg^I$ of the regular description logic with inverse $\mathscr{R}eg^I$, which extends the description logic $\mathscr{ALC}$ with inverse roles and regular role inclusion axioms characterized by finite automata. In contrast to the well-known Horn fragments $\mathscr{EL}$, DL-Lite, DLP, Horn-$\mathscr{SHI}\mathcal{Q}$ and Horn-$\mathscr{SROI}\mathcal{Q}$ of description logics, Horn-$\mathscr{R}eg^I$ allows a form of the concept constructor "universal restriction" to appear at the left hand side of terminological inclusion axioms, while still has PTIME data complexity. Namely, a universal restriction can be used in such places in conjunction with the corresponding existential restriction. We provide an algorithm with PTIME data complexity for checking satisfiability of Horn-$\mathscr{R}eg^I$ knowledge bases.

1 Introduction

Description logics (DLs) are variants of modal logics suitable for expressing terminological knowledge. They represent the domain of interest in terms of individuals

Linh Anh Nguyen
Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, and
Faculty of Information Technology, VNU University of Engineering and Technology,
144 Xuan Thuy, Hanoi, Vietnam
e-mail: nguyen@mimuw.edu.pl

Thi-Bich-Loc Nguyen
Department of Information Technology, Hue University of Sciences,
77 Nguyen Hue, Hue City, Vietnam
e-mail: ntbichloc@hueuni.edu.vn

Andrzej Szałas
Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, and
Dept. of Computer and Information Science, Linköping University,
SE-581 83 Linköping, Sweden
e-mail: andsz@mimuw.edu.pl

V.-N. Huynh et al. (eds.), *Knowledge and Systems Engineering, Volume 1*,
Advances in Intelligent Systems and Computing 244,
DOI: 10.1007/978-3-319-02741-8_6,

(objects), concepts and roles. A concept stands for a set of individuals, a role stands for a binary relation between individuals. The DL $\mathcal{SROIQ}$ [8] founds the logical base of the Web Ontology Language OWL 2, which was recommended by W3C as a layer for the architecture of the Semantic Web.

As reasoning in $\mathcal{SROIQ}$ has a very high complexity, W3C also recommended the profiles OWL 2 EL, OWL 2 QL and OWL 2 RL, which are based on the families of DLs $\mathcal{EL}$ [1, 2], DL-Lite [4] and DLP [6]. These families of DLs are monotonic rule languages enjoying PTIME data complexity. They are defined by selecting suitable Horn fragments of the corresponding full languages with appropriate restrictions adopted to eliminate nondeterminism. A number of other Horn fragments of DLs with PTIME data complexity have also been investigated (see [12] for references). The fragments Horn-$\mathcal{SHIQ}$ [9] and Horn-$\mathcal{SROIQ}$ [17] are notable, with considerable rich sets of allowed constructors and features.

To eliminate nondeterminism, all $\mathcal{EL}$ [1, 2], DL-Lite [4], DLP [6], Horn-$\mathcal{SHIQ}$ [9] and Horn-$\mathcal{SROIQ}$ [17] disallow (any form of) the universal restriction $\forall R.C$ at the left hand side of $\sqsubseteq$ in terminological axioms. The problem is that the general Horn fragment of the basic DL $\mathcal{ALC}$ allowing $\forall R.C$ at the left hand side of $\sqsubseteq$ has NP-complete data complexity [11]. Also, roles are not required to be serial (i.e., satisfying the condition $\forall x \exists y R(x,y)$), which complicates the construction of (logically) least models. For many application domains, the profiles OWL 2 EL, OWL 2 QL and OWL 2 RL languages and the underlying Horn fragments $\mathcal{EL}$, DL-Lite, DLP seem satisfactory. However, in general, forbidding $\forall R.C$ at the left hand side of $\sqsubseteq$ in terminological axioms is a serious restriction.

In [10] Nguyen introduced the deterministic Horn fragment of $\mathcal{ALC}$, where the constructor $\forall R.C$ is allowed at the left hand side of $\sqsubseteq$ in the combination with $\exists R.C$ (in the form $\forall R.C \sqcap \exists R.C$, denoted by $\forall\exists R.C$ [3]). He proved that such a fragment has PTIME data complexity by providing a bottom-up method for constructing a (logically) least model for a given deterministic positive knowledge base in the restricted language. In [11] Nguyen applied the method of [10] to regular DL $\mathcal{Reg}$, which extends $\mathcal{ALC}$ with regular role inclusion axioms characterized by finite automata. Let us denote the Horn fragment of $\mathcal{Reg}$ that allows the constructor $\forall\exists R.C$ at the left hand side of $\sqsubseteq$ by Horn-$\mathcal{Reg}$. As not every positive Horn-$\mathcal{Reg}$ knowledge base has a (logically) least model, Nguyen [11] proposed to approximate the instance checking problem in Horn-$\mathcal{Reg}$ by using its weakenings with PTIME data complexity.

The works [10, 11] found a starting point for the research concerning the universal restriction $\forall R.C$ at the left hand side of $\sqsubseteq$ in terminological axioms guaranteeing PTIME data complexity. However, a big challenge is faced: the bottom-up approach is used, but not every positive Horn-$\mathcal{Reg}$ knowledge base has a logically least model. As a consequence, the work [11] on Horn-$\mathcal{Reg}$ is already very complicated and the problem whether Horn-$\mathcal{Reg}$ has PTIME data complexity still remained open.

The goal of our research is to develop a Horn fragment of a DL (and therefore a rule language for the Semantic Web) that is substantially richer than all well-known Horn fragments $\mathcal{EL}$, DL-Lite, Horn-$\mathcal{Reg}$, Horn-$\mathcal{SHIQ}$, Horn-$\mathcal{SROIQ}$ as well as Horn-$\mathcal{Reg}$, while still has PTIME data complexity. Recently, we have