Vetenskapsrådet

Application

2015-03-31

2015-05548	Zhang, Yuhang	NT-14

Information about applicant

Name: Yuhang Zhang

Doctorial degree: 2012-10-16

Birthdate: 19820402

Academic title: Doktor

Gender: Male

Employer: No current employer

Administrating organisation: Chalmers tekniska högskola

Project site: 3202 - Signalbehandling och medicinsk teknik

Information about application

Call name: Forskningsbidrag Stora utlysningen 2015 (Naturvetenskap och teknikvetenskap)

Type of grant: Projektbidrag

Focus: Unga forskare

Subject area:

Project title (english): Automatic Image Annotation: Autodidactic Learning from Big Data

Project start: 2016-01-01

Project end: 2019-12-31

Review panel applied for: NT-14

Classification code: 10207. Datorseende och robotik (autonoma system)

Keywords: computer vision, image annotation, machine learning

Funds applied for

Year:	2016	2017	2018	2019
Amount:	968,000	1,002,000	1,034,000	1,070,000

Descriptive data

Project info

Project title (Swedish)*

Automatisk bildannotering: Autodidaktisk inlärning från Big Data

Project title (English)*

Automatic Image Annotation: Autodidactic Learning from Big Data

Abstract (English)*

This project aims to explore the effectiveness of autodidactic learning in automatic image annotation. In an image annotation task, a computer needs to assign semantic labels, such as trees, cars, tables, etc., to images according to their content. Autodidactic learning specifies a learning task, in which the learner has access to very large or infinite resource, but limited budget, time, and external guidance. Our ultimate goal is to allow computers to learn to understand images from uncontrolled data sources, such as the Internet. Automatic image understanding is an urgently needed technique. Due to the wide usage of digital cameras, more and more image repositories become too big to be annotated manually. However, most existing approaches still rely on supervised learning to solve image annotation problem. Labeling images manually is known to be slow, expensive, and error-prone, which is not compatible with the fast speed of information growth. Our approach is to equip computers with the capability of autodidact learning. More specifically, we aim to enable a computer to automatically select useful training data from a big pool, and learn to understand images in an incremental manner from weakly or unlabeled data. In this proposal we show the conceived scenario is already partially realized by active learning, semi-supervised learning, on-line learning, and reinforcement learning. By putting all these pieces together, we can design a new learning framework, take advantage of the availability of big data, and solve a challenging problem. The success of the proposed project will not only push the frontier of image annotation, but also benefit related research fields, such as machine learning and medical image processing.

Popular scientific description (Swedish)*

Det här projektet syftar till att undersöka möjligheterna med autodidaktisk inlärning för automatisk bildannotering. Vid bildannotering, så ska en dator tilldela olika semantiska etiketter, som t ex träd, bil, bord etc., till en bild baserat på dess innehåll. Det är en teknik som det finns stort behov av bland annat på grund av det utbredda användandet av digitala kameror. Det finns fler och fler bildbibliotek som är alltför stora att annotera för hand. Existerande metoder för bildannotering kräver i stor utsträckning s k handledd inlärning vilket även detta kräver mycket manuellt arbete. Sådana angreppssätt skalar inte upp för att klara av de stora bildmängder som redan idag finns tillgängliga. Vårt nya angreppssätt syftar till att ge en dator en autodidaktiskt inlärningsförmåga vilket innebär att datorn själv lär sig att förstå bilder baserat på den rika information som finns på internet. Genom att lösa detta problem hoppas vi även få kunskap om hur våra hjärnor behandlar visuell information vilket är ett öppet problem inom artificiell intelligens.

Project period

Number of project years*

4

Calculated project time*

2016-01-01 - 2019-12-31

Deductible time

Deductible time

Cause	Months

Career age: 29

Career age is a description of the time from your first doctoral degree until the last day of the call. Your career age change if you have deductible time. Your career age is shown in months. For some calls there are restrictions in the career age.

Classifications

Select a minimum of one and a maximum of three SCB-codes in order of priority.

Select the SCB-code in three levels and then click the lower plus-button to save your selection.

SCB-codes*

1. Naturvetenskap > 102. Data- och informationsvetenskap (Datateknik) > 10207. Datorseende och robotik (autonoma system)

Enter a minimum of three, and up to five, short keywords that describe your project.

Keyword 1*

computer vision

Keyword 2*

image annotation

Keyword 3*

machine learning

Keyword 4

Keyword 5

Research plan

Ethical considerations

Specify any ethical issues that the project (or equivalent) raises, and describe how they will be addressed in your research. Also indicate the specific considerations that might be relevant to your application.

Reporting of ethical considerations*

The propose project will be largely based model design, algorithm development, and image-based experiment. The images will either be licensed for scientific research or data publicly available over the Internet. No ethical concern will be involved.

The project includes handling of personal data

No

The project includes animal experiments

No

Account of experiments on humans

No

Research plan

Automatic Image Annotation: Autodidactic Learning from Big Data

Yuhang Zhang, Chalmers University of Technology

1 Purpose and Aims

In the development of our intelligence, who teaches us most, our teachers, our parents, or ourselves?

This project aims to ***explore the effectiveness of autodidactic learning in automatic image annotation.*** In an image annotation task, a computer needs to assign semantic labels, such as trees, cars, tables, *etc.*, to images according to their content. Autodidactic learning specifies a learning task, in which the learner has access to very large or infinite resource, but limited budget, time, and external guidance. Our ultimate goal is to allow computers to learn to understand images from uncontrolled data sources, such as the Internet.

To achieve this objective, we will

1. develop a framework for autodidactic learning;
2. build an autodidactic learning based image annotation system;
3. quantitatively compare the learning efficiency and prediction accuracy of autodidactic learning against existing supervised learning algorithms in image annotation tasks.

The expected outcome are new knowledge in the form of research papers in international conferences and journals, as well as new algorithms in the form of publicly available software.

2 Survey of the Field

2.1 Image Annotation

Over the last decade, significant progress has been made in the research of semantic image annotation. The progress can be easily seen from the continuous emergence of benchmark databases that are increasingly challenging. Back to 2004, one of the most widely used database, Caltech101 [10], contained 9 thousands images and 101 semantic labels. By 2014, the largest database, ImageNet [28], contains 1.4 millions images and 1000 semantic labels. Other notable databases include but are not limited to Corel [9], LabelMe [29], SUN [44], MSRC [34], Standford Background [12], and COCO [23]. Some of these databases contain images that are manually labeled at pixel-level, whereas the others only provide manual labeling at image level with (or without) rough location of each object.

At the same time, algorithms based on diversified models, such as machine translation [9], multinomial distributions [20], bag of features [45], nearest neighbors [13, 14, 41], support vector machines [8, 46], random forests [11], Markov random fields [19, 34, 43], and deep neural networks [17, 35], have been proposed to formulate the mapping between images and semantic labels. Some of these models are still being studied actively today. Nearest neighbor based methods are simple yet possess strong performance guarantee as the amount of the available data grows. Markov random fields are good at modeling the contextual clues in images. Whereas the

other approaches generally require a manual design of image features, deep learning provides an end-to-end solution connecting raw-data and semantic labels.

However, one criticism [39] that has not been properly addressed so far is, algorithms are developed to understand the images in a closed database, rather than in the open world. As a result, not only the models to be developed but also the research efforts for image understanding are "creeping over-fitting" those databases of limited sizes. Such a limitation becomes particularly awkward, when Internet is providing us with easy access to endless images, and Big Data is becoming a household term.

Behind the limitation, there is a realistic dilemma with the existing image annotation algorithms, which generally approach image annotation as a supervised learning problem [6]. On one hand, we cannot annotate all the images in the open world manually. On the other hand, when learning from a tiny training sample, *i.e.*, the manually labeled database relative to the open world, the ignorance of a computer is expected.

It is worth noticing that there is already some preliminary work trying to break through the bottle-neck of training data. Semi-supervised learning [15] uses unlabeled data to reveal the structure information of the feature space. On-line learning [7] allows training data to arrive gradually. In active learning [24, 42], a learner automatically categorizes the training data into simple and hard cases, and ask for manual supervision only for hard cases. Nevertheless, there are still some unrealistic assumptions in these approaches. First, they generally assume the existence of a know-all supervisor, who can answer a learner's questions whenever needed. However, the truth is, humans do not know the answers to all questions. Neither can we work at all time. Second, they assume either a supervisor can select training data for them, or they have sufficient time to go thourgh all the data. However, the truth is, as the numer of semantic labels grows, humans can hardly maintain the quality of the training data [28]. Moreover, there are just ever-growing data available over the Internet, and we cannot wait forever for a learner to go through all of them. As a matter of fact, humans do not learn to read by scaning a dictionary from cover to covre.

Different from previous work, we see the solution in autodidactic learning, which allows a computer to select the learning resource for itself. At the same time, an autodidactic computer will be able to deal with an imperfect supervisor, as well as incomplete supevision.

2.2 Autodidactic Learning

Autodidactic learning plays an important role in the development of humans' intelligence. One typical example is we learn how to use our eyes under limited external supervision. Did we ask our parents 'what is this', when we were young? Yes. Did we ask about everything we see? No. We can recognize millions of objects after asking only a small number of questions, because we only ask valuable questions. We ask only if we are uncertain about our own answer and that an external answer is likely to increase our knowledge. In the other time, we just trust our own answer to minimize the cost of asking questions. Moreover, sometimes we have no one to ask, or do not trust other's answer, or prefer not to bother others. In these situations, we keep the questions in mind and look out for relevant information. We may be able to answer these questions at a later time. That is the ability of autodidatic learning.

Autodidactic learning has not been discussed explicitly in machine learning research, however, its basic components have already been studied under several related topics. **How to accumulate knowledge** is addressed by on-line learning. Typical ways to update the models as more data sequentially arrive include stochastic gradient descent [4], perceptron [3], and Bayesian approach [26]. **How to use unlabeled data** is addressed by semi-supervised and unsupervised learning. Typical approaches include manifold learning [2], graph-based representation [54], and sparse coding [21]. **When to ask for guidance** is addressed by active learning. Typical strate-

gies include uncertainty sampling [22], query-by-Committee [32], expected model change [31], *etc.*. **When to explore new data** is addressed by reinforcement learning with a rich literature discussing the trade-off between exploitation and exploration (see [36] for a review). **How to assemble these solutions into one learning system** is the question to be answered by the proposed project. In particular, we need to figure out a combination of above optional strategies that optimizes the overall performance, reflected by theoretical guarantees and empirical evaluation.

A research field related to autodidactic learning is developmental robotics, which studies machine's life-long and open-ended learning of new skills and new knowledges [1]. In contrast, the autodidactic learning in the proposed project aims to pick up a specific skill, image annotation.

Another notable technique is submodularity functions. The value of submodularity functions in autodidactic learning is, whereas the overall learning problem is in general NP-hard to optimize so approximation is inevitable, submodularity functions can provide a rigorous theoretical approximation guarantee [18] in such a situation.

3 Project Description

In order to quantitatively evaluate the effectiveness of autodidactic learning in image annotation, we will sequentially work on framework construction, algorithm development, and empirical evaluation, three tasks.

3.1 Framework of Autodidactic Learning

Although the research in related fields has already provided rich theoretical basis for autodidactic learning, there are still some new components to be developed.

Task 1.1: Estimate Data Quality A major distinction between the proposed autodidactic learning and existing learning algorithms is the ability to select training data, *i.e.* not only which image to query the supervisor, but also which image to see at the very beginning. Such a capability is necessary, because the data source we aim to learn from, the open world or images over the Internet, is infinitely large.

We assume the learner has a list of labels that are required to learn. We further assume the learner can interact with two or more image retrieval engines, *e.g.*, Google and Flickr, to download images from the Internet. Every time the learner provides a semantic label to an image retrieval engine, which then returns a number of images for free. The retrieval engines may not be perfect, so the learner needs to select the retrieval engines to select the images it sees, in order to maximize the learning performance.

If we can quantify the scope and reliability, of each image retrieval engines, the joint optimization of learning image annotation and engine selection becomes a multi-armed bandit problem. The state-of-the-art solution formulates the problem as Markov decision processes and achieves uniformly maximum convergence rate [5]. There are several ways to quantify the reliability of image retrieval engines, including the class purity as predicted by the learner[1], the model change after learning, or reduction in the expected error after each learning. Note that these criteria were originally used for query selection by active learning [30]. An autodidactic computer will quantify an image retrieval engine's quality with these criteria to optimize data selection.

[1] A pitfall here is the learner may choose to query the engine which only returns simple images.

Task 1.2: Maintain Uncertain Dataset A new challenge faced by the proposed autodidactic learning is the manual labels of some training data may not be given even queried. Since the learner only queries the labels for uncertain data, these data in general cannot be labeled correctly by the learner itself. A simple solution is to abandon these data. However, this is likely to result in big loss of information as the ambiguity of these data suggests they are valuable in estimating the boundary between classes.

Our solution is to maintain an uncertain dataset, which keeps those hard training data whose labels are not yet clear. Data can be removed from this uncertain dataset if their manual labels arrive, or the model can predict their labels confidently, at a later time.

This uncertain set can also guide data selection. Recall in Task 1.1 we need to quantify the quality of each image retrieval engine. We can add the reduction in uncertainty (or equivalently the increase in prediction confidence) into the criteria of data selection.

Task 1.3: Suppress Query Granularity Another issue of particular interest in image annotation is about the granularity of the query. That is, answering a question at image-level, *e.g.*, if it contains a tree, usually costs less than asking a question at pixel-level, *e.g.*, where the tree's boundary is. Previous work [24, 31] on active learning discussed this issue under the circumstance of multiple-instance learning.

In our case, since the data source is weakly controlled, a high-granularity labeling may not even exist for most training data. We see the solution in co-segmentation [27]. Given two images containing similar objects in different backgrounds, co-segmentation can extract the boundary of similar objects in the two images. That is to say, as long as we know the image-level labels of multiple images, we have the chance to automatically recover their pixel-level labels via co-segmentation.

3.2 Autodidactic Learning for Image Annotation

In this task, we apply autodidactic learning to image annotation problem. We implement autodidactic learning with two successful systems for image annotation, nearest neighbors based system and Markov random fields based system.

Task 2.1: Autodidactic Learning for Nearest Neighbor Classifier In our previous work [14], we show image annotation can be approached with a nearest neighbor classifier. Specifically, we segment every image into superpixels [49] and build a k nearest neighbor graph. For each superpixel in a new image, we find its nearest neighbors among labeled superpixels, and transfer the labels from the labeled superpixels to the new superpixel. To efficiently find nearest neighbors among thousands of superpixels, we adopted an approximate algorithm utilizing the nearest neighbor graph. Since nearest neighbor based classification is by nature a data driven approach which can exploit the power of big data, we will primarily test nearest neighbor based methods in the proposed project.

The search and update of the nearest neighbor graph plays a central role in this approach. In our previous work, the nearest neighbor graph are constructed without using the label information. The graph stores all the superpixels in all images, and steadily grows as more images arrive. Such a strategy does not adapt to big data. In the proposed project, we will build the nearest neighbor graph following the heuristic of Hart algorithm [16]. The basic idea is to add a new superpixel into the graph only if it cannot be confidently labeled by the current graph. In this way, the nearest neighbor graph will preserve more superpixels for classes that are difficult to recognize, and exclude redundant superpixels of easy classes. Moreover, the graph can be pruned once in a while to drop those nodes that become easy as the model strenghens incrementally.

In order to distinguish a large number of classes, metric learning will be necessary. We plan to explore large-scale manifold learning [37] in the proposed project. That is, instead of assessing the distance in the raw feature space, we will project data into a low dimensional manifold and use the geodesic distance on the manifold as the similarity measure.

Task 2.2: Autodidactic Learning for Markov Random Fields Markov random fields can explicitly describe the dependency between neighboring pixels and regions in an image, which has been utilized by our previous work [13] as well as many other algorithms [19, 34, 43] for image annotation.

Since the inference of Markov models is by nature more expensive than nearest neighbor searching, in the proposed project, we will use Markov model based method only as a complementary to nearest neighbor based method. In particular, only those data that cannot be predicted confidently by nearest neighbor based method will be passed to Markov model based method.

Another difficulty in Markov model based image annotation is with the training, which is a structured learning problem. Although there are a number of available algorithms [40, 38] for structured learning, how to implement structured learning with incomplete and corrupted data is still an open problem. Our recent work [47] shows a potential solution lies in the pseudo-Boolean formulation. In particular, by representing a structured model with pseudo-Boolean functions, structured learning can be approached by a linear program. Missing labels in this linear program corresponds to a missing constraint, but the linear program is still well-conditioned, and the learning can be approached by maximizing the margin subject to available constraints only.

3.3 Quantitative Evaluation

In this task, we quantitatively evaluate the performance of autodidactic learning in the task of image annotation. Since the learning and testing will be carried out on big data, traditional experiment design is no longer adequate.

Task 3.1: Experiment Design We plan to use the $57,000$ nouns in WordNet [25] as the list of semantic labels we try to learn from the Internet. Most of these words were never included by existing dataset. To provide a starting point for our learning system, we will use crowdsourcing platforms like Amazon's Mechanical Turk to manually label a small number of images for each class. The computer will then work independently by iteratively querying the image retrieval engines. Every once in a while, we will provide some further guidance to the learning system. We upload the uncertain dataset collected by the system onto the crowdsourcing platform to collect supplemental manual labels. The learning will goes on like this.

The trained system will be tested on two types of data. The first type are existing dataset, such as ImageNet [28] and COCO [23]. This test will facilitate the comparison between the proposed approach and previous approaches. The second type are random images on the Internet. A number of random images will be predicted by the trained system first and then manually checked on crowdsourcing platforms. Note that this is different from existing image annotation experiment. In traditional experiments, a computer tries to reproduce humans' result. In the proposed project, human check the results produced by the computer. The reason for such an inversion is to simplify humans' work to allow more images to be manually checked, *e.g.*, after a computer predicts a car in the image a person only needs to tick yes or no. As a result, we will be assessing the accuracy of a computer's prediction, rather than the precision or recall with respect to each class.

Task 3.2: Performance Analysis There are many variables affecting the performance of the proposed image annotation system. In particular, we aims to figure out the correlations between annotation performance and the following variables, through both theoretical analysis and empirical study.

Data selection is a novel capability of the proposed system. We will investigate the difference between learning from selected data and learning from random data. **Manual supervision** is to be significantly reduced in autodidactic learning. We will investigate how critical the remaining supervision is by comparing learning with supervision (humans answer the queries about uncertain data) against learning without supervision (humans do not intervene). **Query granularity** is to be reduced during autodidactic learning. We will investigate if high-granularity supervision is needed at all by comparing system trained with completely low granularity ground truth and system trained with ground truth of mixed granularity.

The propose project will be carried by PI and a new recruited PhD students. The involvement of PI and the PhD student at different stages of the project is shown by the table below with milstones. We believe this project provides good opportunities to the PhD student to pick up frontier techinques in computer vision and machine learning. At the end of this task, we expect to deliver a working system for image annotation.

Table 1: Milestones

Time	Research Activity	Deliverable	Task	People
Month 12	Framework for autodidactic learning	report	1	PI
Month 24	Nearest neighbor based method	software and report	2	PI, PhD
Month 30	Markov random fields based method	software and report	2	PI
Month 36	Experiment design and implementaion	software	3	PI, PhD
Month 42	Performance analysis	report	3	PI, PhD
Month 48	finalize image annotation system	software and report	3	PI, PhD

4 Significance

Automatic image understanding has values in both practical applications and theoretical understanding. Practically, the developed algorithm will allow automatic sorting and searching of images based on their contents. This is an urgently needed technique. Due to the wide usage of digital cameras, more and more image repositories become too big to be annotated manually. Existing approaches for automatic image annotation generally adopts supervised learning. Supervised learning relies on a supervisor to provide informative training data together with ground truth labels. However, manually labeling images is expensive, inefficient, and error-prone. These approaches are not scalable to handle the big amount of images in the open world. Our approach attempts to equip computers with the capability of autodidactic learning. Computers can then learn from the weakly labeled or unlabeled big data over the Internet under limited manual supervision. Theoretically, by solving this problem we will provide also new understanding to a long-lasting question in artificial intelligence, how our brain processes visual information. The discovery in the proposed project will also benefit other related research fields, such as machine learning and medical image computing.

5 Preliminary Results

Results of our previous work foreshadow the success of the proposed project from multiple aspects.

As Table-2-Left shows, the performance of our nearest neighbor based method [14] is comparable to the state of the art. Moreover, by using Markov random fields [13], we can improve

the annotation accuracy on certain images. Table-2-Right shows considering unlabeled image during learning can improve the annotation performance. Note that these experiments are based on databases of limited sizes. We expect considering unlabeled images in the open world will contribute more significantly to the annotation accuracy. Some images annotated by nearest neighbor based method [14] are given by Figure-1-Left.

Table 2: **Left**: image annotation accuracy achieved by two of our previous work and the state of the art. **Right**: improvement in annotation accuracy by learning from both labeled and unlabeled data (semi-super) relative to learning from labeled data only (supervised).

Database	[13]	[14]	S. of the Art
Polo	94.2	91.8	94.2
MSRC	79.0	84.5	87.0
Standford	73.4	79.3	82.9
SIFT FLOW	65.2	78.4	78.6

Database	supervised	semi-super	improve
Polo	91.8	92.5	0.7
MSRC	84.5	86.3	1.8
Standford	79.3	79.6	0.3
SIFT Flow	78.4	78.4	0.0

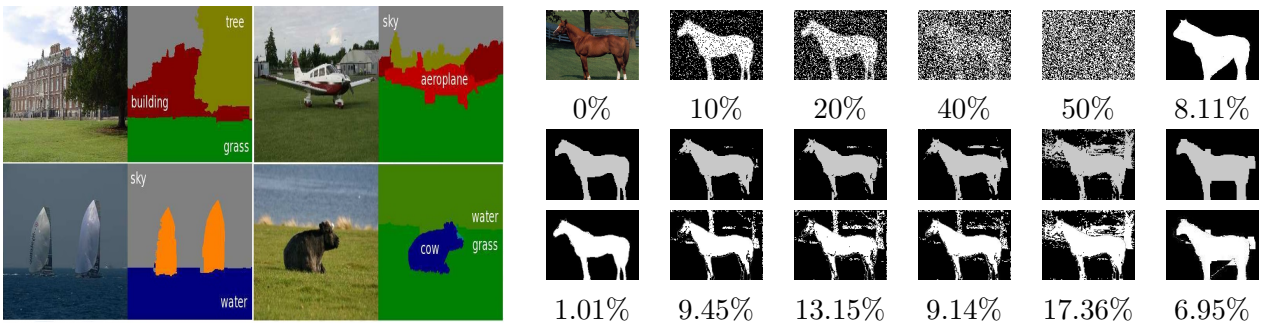

Figure 1: **Left**: images annotated by [14]. Best viewed in color. **Right**: from top to bottom:corrupted ground truth, noise level, segmentation with learned model, restored ground truth, and errors in restored ground truth.

Learning Markov random fields from corrupted data will be researched in the proposed project. Figure-1-Right shows our algorithm's robustness against corrupted data [47]. The purpose is to learn the class of horse from a corrupted ground truth. Impressively, the algorithm can still extract some information from the training image even when 50% of the ground truth are corrupted.

Another problem that will be tackled in the proposed project is to learn high-granularity labeling from incomplete or low-granularity ground truth. Our preliminary study shows co-segmentation is a potential solution. In Figure 2, we only told the computer the two images contain foreground of the same class. The computer then recovered the annotation of the images at pixel level.

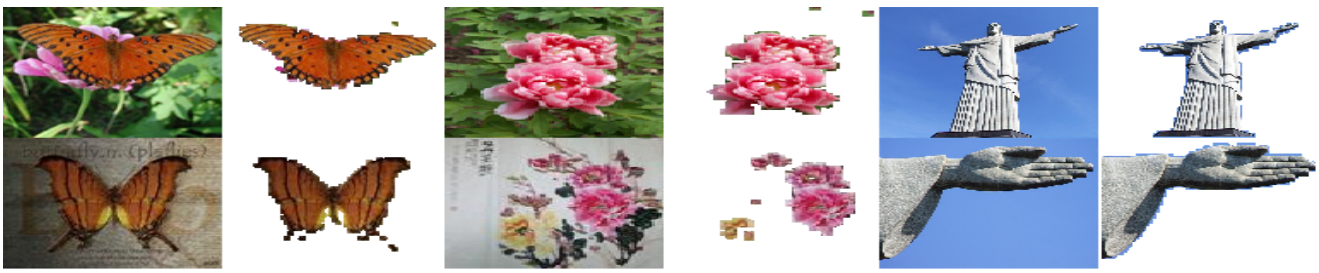

Figure 2: Recover pixel-level image annotation from image-level ground truth.

6 Independent Line of Research

The PI's PhD thesis focuses on developing efficient inference algorithms for Markov random field based image labeling problems, such as image segmentation [49] and stereo matching [48, 50, 51].

During his postdoctoral study in Australia, he worked on semantic image annotation with his supervisor and other collaborators, with focus on nearest neighbor based methods [13, 14, 33]. The PI also have experience in developing learning algorithms for Markov random fields [47]. The PI ever worked on large-scale content-based image retrieval [52, 53], which is strong related to image annotation problems. It is fair to say the proposed project brings together the PI's previous research experience and strength to solve an important problem, image annotation, with a new method, autodidactic learning, in a much larger scale, open world. Carrying out the proposed project will not only consolidate PI's research strength in image labeling problems, but also deliver strong impact to the research field of computer vision and machine learning.

The PI is currently working in the Image Analysis and Computer Vision Group led by Professor Fredrik Kahl of Chalmers University of Technology. The group's research target and strength focuses on global model optimization and medical image analysis. Model inference will be part of the proposed project. At the same time, image annotation algorithms developed by the proposed project can be applied to medical image analysis, such as CT segmentation, too.

7 Form of Employment

The PI will be a postdoctoral fellow at Chalmers. His current contract ends on 28th February of 2016, with extension opportunity based on the availability of research funding. The request budget will cover 30% of his salary from 2016 to 2019.

References

[1] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: A survey. *TAMD*, (1), 2009.

[2] M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. *Mach. Learn.2004*.

[3] L. Bottou. On-line learning in neural networks. chapter On-line Learning and Stochastic Approximations, pages 9–42. Cambridge University Press, New York, NY, USA, 1998.

[4] L. Bottou. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, 2010.

[5] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

[6] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *TPAMI*, 29(3):394–410, March 2007.

[7] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, 11:1109–1135, Mar. 2010.

[8] C. Cusano, G. Ciocca, and R. Schettini. Image annotation using svm. In *Proc. SPIE*, 2003.

[9] P. Duygulu, K. Barnard, J. F. G. d. Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02*.

[10] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, Apr. 2007.

[11] H. Fu, Q. Zhang, and G. Qiu. Random forest for image annotation. In *Proceedings of*, ECCV'12.

[12] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV 2009*.

[13] S. Gould and Y. Zhang. PatchMatchGraph: Building a graph of dense patch correspondences for label transfer. In *ECCV*, 2012.

[14] S. Gould, J. Zhao, X. He, and Y. Zhang. Superpixel graph label transfer with learned distance metric. In *ECCV*, 2014.

[15] M. Guillaumin, J. J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR 2010*, pages 902–909, 2010.

[16] P. Hart. The condensed nearest neighbor rule (corresp.). *IEEE Trans Inf Theory*, 14(3):515–516, May 1968.

[17] R. Kiros and C. Szepesvári. Deep representations and codes for image auto-annotation. In *NIPS*. 2012.

[18] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 9:235–284, June 2008.

[19] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. *ICCV*, 2009.

[20] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In S. Thrun, L. Saul, and B. Schölkopf, editors, *NIPS*, pages 553–560. MIT Press, 2004.

[21] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[22] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of*, SIGIR '94, pages 3–12, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[24] D. Liu, X.-S. Hua, L. Yang, and H.-J. Zhang. Multiple-instance active learning for image categorization. In *Advances in Multimedia Modeling*. 2009.

[25] G. A. Miller. Nouns in wordnet: A lexical inheritance system. *Int. J. Lexicogr.*, Dec. 1990.

[26] M. Opper. On-line learning in neural networks. chapter A Bayesian Approach to On-line Learning, pages 363–378. Cambridge University Press, New York, NY, USA, 1998.

[27] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *Proceedings of*, CVPR '06, 2006.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014.

[29] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, May 2008.

[30] B. Settles. Active learning literature survey. (1648), 2009.

[31] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS*, pages 1289–1296. Curran Associates, Inc., 2008.

[32] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of*, COLT '92.

[33] G. Sheasby, J. Warrell, Y. Zhang, N. Crook, and P. H. Torr. Simultaneous human segmentation, depth and pose estimation via dual decomposition. *workshop of BMVC*, 2012.

[34] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.

[35] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *JMLR*, 2014.

[36] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998.

[37] A. Talwalkar, S. Kumar, M. Mohri, and H. Rowley. Large-scale svd and manifold learning. *JMLR*, 2013.

[38] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: a large margin approach. In *ICML*, pages 896–903, 2005.

[39] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of*, CVPR '11, 2011.

[40] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, Dec. 2005.

[41] Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *ECCV 2012*, pages 836–849, 2012.

[42] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *IJCV*, 108(1-2):97–114, May 2014.

[43] Y. Xiang, X. Zhou, T.-S. Chua, and C.-W. Ngo. A revisit of generative model for automatic image annotation using markov random fields. In *CVPR 2009*, pages 1153–1160, June 2009.

[44] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, June 2010.

[45] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu. Correlative multi-label multi-instance image annotation. In *ICCV 2011*, pages 651–658, Nov 2011.

[46] C. Yang, M. Dong, and J. Hua. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *CVPR 2006*, volume 2, 2006.

[47] Y. Zhang, S. Gould, and F. Kahl. Structured learning with pseudo-boolean functions. *Tech.Report*, 2015.

[48] Y. Zhang, R. Hartley, J. Mashford, and S. Burn. Superpixels, occlusion and stereo. In *DICTA*, 2011.

[49] Y. Zhang, R. Hartley, J. Mashford, and S. Burn. Superpixels via pseudo-boolean optimization. In *ICCV*, pages 1387–1394, Nov 2011.

[50] Y. Zhang, R. Hartley, and L. Wang. Fast multi-labelling for stereo matching. In *ECCV*, volume 6313 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010.

[51] Y. Zhang, R. I. Hartley, J. Mashford, L. Wang, and S. Burn. Pipeline reconstruction from fisheye images. *Journal of WSCG*, 19(2):49–57, 2011.

[52] Y. Zhang, L. Wang, R. Hartley, and H. Li. Handling significant scale difference for object retrieval in a supermarket. In *DICTA*, pages 468–475, Dec 2009.

[53] Y. Zhang, L. Wang, R. I. Hartley, and H. Li. Where's the weet-bix? In Y. Yagi, S. B. Kang, I. Kweon, and H. Zha, editors, *ACCV*, volume 4843 of *Lecture Notes in Computer Science*, pages 800–810. Springer, 2007.

[54] X. Zhu. *Semi-supervised Learning with Graphs*. PhD thesis, Pittsburgh, PA, USA, 2005.

Interdisciplinarity

My application is interdisciplinary

☐

An interdisciplinary research project is defined in this call for proposals as a project that can not be completed without knowledge, methods, terminology, data and researchers from more than one of the Swedish Research Councils subject areas; Medicine and health, Natural and engineering sciences, Humanities and social sciences and Educational sciences. If your research project is interdisciplinary according to this definition, you indicate and explain this here.

Click here for more information

Scientific report

Scientific report/Account for scientific activities of previous project

The proposed project is PI's first project proposal submitted to Swedish Research Council.

Budget and research resources

Project staff

Describe the staff that will be working in the project and the salary that is applied for in the project budget. Enter the full amount, not in thousands SEK.

Participating researchers that accept an invitation to participate in the application will be displayed automatically under Dedicated time for this project. Note that it will take a few minutes before the information is updated, and that it might be necessary for the project leader to close and reopen the form.

Dedicated time for this project*

	Role in the project	Name	Percent of full time
1	Applicant	Yuhang Zhang	30
2	Other personnel without doctoral degree	PhD student	80

Salaries including social fees

	Role in the project	Name	Percent of salary	2016	2017	2018	2019	Total
1	Applicant	Yuhang Zhang	30	204,000	211,000	218,000	226,000	859,000
2	Other personnel without doctoral degree	PhD student	80	435,000	450,000	466,000	482,000	1,833,000
	Total			639,000	661,000	684,000	708,000	2,692,000

Other costs

Describe the other project costs for which you apply from the Swedish Research Council. Enter the full amount, not in thousands SEK.

Premises

	Type of premises	2016	2017	2018	2019	Total
1	office	46,000	48,000	49,000	51,000	194,000
2	IT	14,000	15,000	15,000	16,000	60,000
	Total	60,000	63,000	64,000	67,000	254,000

Running Costs

	Running Cost	Description	2016	2017	2018	2019	Total
1	travel & publication	paper publication and conference attendance fee	35,000	35,000	35,000	35,000	140,000
	Total		35,000	35,000	35,000	35,000	140,000

Depreciation costs

Depreciation cost	Description	2016	2017	2018	2019

Total project cost

Below you can see a summary of the costs in your budget, which are the costs that you apply for from the Swedish Research Council. Indirect costs are entered separately into the table.

Under Other costs you can enter which costs, aside from the ones you apply for from the Swedish Research Council, that the project includes. Add the full amounts, not in thousands of SEK.

The subtotal plus indirect costs are the total per year that you apply for.

Total budget

Specified costs	2016	2017	2018	2019	Total, applied	Other costs	Total cost
Salaries including social fees	639,000	661,000	684,000	708,000	2,692,000		2,692,000
Running costs	35,000	35,000	35,000	35,000	140,000		140,000
Depreciation costs					0		0
Premises	60,000	63,000	64,000	67,000	254,000		254,000
Subtotal	734,000	759,000	783,000	810,000	3,086,000	0	3,086,000
Indirect costs	234,000	243,000	251,000	260,000	988,000		988,000
Total project cost	968,000	1,002,000	1,034,000	1,070,000	4,074,000	0	4,074,000

Explanation of the proposed budget

Briefly justify each proposed cost in the stated budget.

Explanation of the proposed budget*

The total cost can be broken into three categories:

Salary including social fees per year is calculated as

	monthly salary	×	activity rate	×	social fee rate	×	12	×	salary increase	=	total
PI	35	×	0.3	×	1.55	×	12	×	1.05	=	204
PhD Student	28	×	0.8	×	1.55	×	12	×	1.05	=	435

for PI and the new PhD student respectively. The monthly salary is estimated based on the average salary of post-doc and PhD students of Chalmers, respectively. The salaries are estimated to increase by 3.5% every year from 2017.

Running costs cover the expense on travel and publication. It is expected eithter PI or the PhD student will attend one international conference and one European conference per year.

Premises costs consist of office cost and IT cost. Office cost is charged by Chalmers at 7% of the total salary. IT cost includes the expense on buying computers, and paying for crowdsourcing platforms.

Indirect cost is estimated to be 36.7% of the total salary.

The total budget is 4,074,000 SEK.

No other funding applications have been submitted for the proposed project.

Other funding

Describe your other project funding for the project period (applied for or granted) aside from that which you apply for from the Swedish Research Council. Write the whole sum, not thousands of SEK.

Other funding for this project

Funder	Applicant/project leader	Type of grant	Reg no or equiv.	2016	2017	2018	2019

CV

YUHANG ZHANG

yuhang.zhang@live.com

EDUCATION

The Australian National University Sept. 2008 – Oct.2012
Doctor of Philosophy
Thesis: Fast Multi-Labelling in Early Vision
Supervisor: Richard Hartley

The Australian National University Sept. 2006 – Aug. 2008
Master of Philosophy
Thesis: Local Invariant Feature Based Object Retrieval in a Supermarket
Supervisor: Lei Wang

Harbin Institute of Technology Sept. 2001 – Jul. 2005
Bachelor of Engineering
Thesis: Serial Image Based Dynamic Object Tracking
Supervisor: Bingrong Hong

EXPERIENCE

Post-Doc Mar. 2013 - present
· Department of Signals and Systems, Chalmers University of Technology

Research Fellow Nov. 2011 - Feb. 2013
· Research School of Computer Science, The Australian National University

Mathematician Analyst Sept. 2008 - Aug. 2011
· Land and water, CSIRO

Visiting Researcher Sept. 2011 - Nov. 2011
· Mathematical Imaging Group, Lund University

Visiting Researcher Aug. 2011 - Sept. 2011
· Department of Computing, Oxford Brookes University

Visiting Researcher Sept. 2010 - Oct. 2010
· Advanced Laparoscopy and Computer Vision group, Auvergne University

Visiting Researcher Aug. 2010 - Sept. 2010
· LASMEA, Blaise Pascal University

PUBLICATION

Stephen Gould, Jiecheng Zhao, Xuming He, and **Yuhang Zhang**, Superpixel Graph Label Transfer with Learned Distance Metric, European Conference on Computer Vision (ECCV), 2014

Stephen Gould and **Yuhang Zhang**, PatchMatchGraph: Building a Graph of Dense Patch Correspondences for Label Transfer, European Conference on Computer Vision (ECCV), 2012

Yuhang Zhang, Richard Hartley, John Mashford, and Stewart Burn, Superpixels via Pseudo-Boolean Optimization, International Conference on Computer Vision (ICCV), 2011

Yuhang Zhang, Richard Hartley, and Lei Wang, Fast Multi-Labelling for Stereo Matching, European Conference on Computer Vision (ECCV), 2010

Glenn Sheasby, Jonathan Warrell, **Yuhang Zhang**, Nigel Crook, and Philip Torr, Simultaneous Human Segmentation, Depth and Pose Estimation via Dual Decomposition, UK Computer Vision Student Workshop, 2012

Yuhang Zhang, Richard Hartley, John Mashford, and Stewart Burn, Superpixels, Occlusion and Stereo, Digital Image Computing: Techniques and Applications (DICTA), 2011

Yuhang Zhang, Richard Hartley, John Mashford, Lei Wang, and Stewart Burn, Pipeline Reconstruction from Fisheye Images, Journal of WSCG, 2011

Yuhang Zhang, Content-based Image Retrieval: Out of the Clutter in a Big Supermarket, LAP Lambert Academic Publishing, 2010

Yuhang Zhang, Lei Wang, Richard Hartley, and Hongdong Li, Handling Significant Scale Difference for Object Retrieval in a Supermarket, Digital Image Computing: Techniques and Applications (DICTA), 2009

Yuhang Zhang, Lei Wang, Richard Hartley, and Hongdong Li, Where's the Weet-Bix? Asian Conference on Computer Vision (ACCV), 2007

INVITED TALKS

Tips for PhD Students — 4 May, 2012
· The Australian National University

Multi-Labelling in Computer Vision — 12 March, 2012
· Beijing Institute of Technology

Multi-Labelling by Quadratic Pseudo-Boolean Approximation — 21 September, 2011
· Lund University

Superpixels via Multi-label Graph-Cuts — 4 August, 2011
· Oxford University

3D Reconstruction of Buried Pipelines — 26 May, 2009
· CSIRO, Melbourne

COVERED BY PRESS

Image Based Search Helps Find the Weetbix — by A. Hendry
· ComputerWorld, Dec. 2007

Lost and found in the supermarket — by S. Couper
· ANU Reporter, summer edition 2008

GRANT

2011	ANU VC Travel Grant for cross-institute visit
2010	ANU VC Travel Grant for cross-institute visit
2010	ECCV Student Travel Grant
2008 – 2011	CSIRO Flagship Postgraduate Scholarship Package (TOP-UP)
2007	ANU VC Travel Grant for international conference

TEACHING

- ENGN4528 Computer Vision, Lecturer
- ENGN4528 Computer Vision, Tutor
- ENGN8531 Statistical Pattern Recognition and Its applications to Computer Vision, Tutor

REFEREES

Prof.	Richard Hartley	The Australian National University	richard.hartley@anu.edu.au
Dr.	Stephen Gould	The Australian National University	stephen.gould@anu.edu.au
Prof.	Fredrik Kahl	Lund University	fredrik@maths.lth.se

Publication list

Publications of Yuhang Zhang

[1] S. Gould, J. Zhao, X. He, and Y. Zhang, "Superpixel graph label transfer with learned distance metric," in *Computer Vision - **ECCV** 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pp. 632–647, 2014.

[2] S. Gould and Y. Zhang, "Patchmatchgraph: Building a graph of dense patch correspondences for label transfer," in *Computer Vision - **ECCV** 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V*, pp. 439–452, 2012.

[3] Y. Zhang, R. I. Hartley, J. Mashford, L. Wang, and S. Burn, "Pipeline reconstruction from fisheye images," *Journal of WSCG*, vol. 19, no. 2, pp. 49–57, 2011.

[4] Y. Zhang, R. I. Hartley, J. Mashford, and S. Burn, "Superpixels, occlusion and stereo," in *2011 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Noosa, QLD, Australia, December 6-8, 2011*, pp. 84–91, 2011.

[5] Y. Zhang, R. I. Hartley, J. Mashford, and S. Burn, "Superpixels via pseudo-boolean optimization," in *IEEE International Conference on Computer Vision, **ICCV** 2011, Barcelona, Spain, November 6-13, 2011*, pp. 1387–1394, 2011.

[6] Y. Zhang, R. I. Hartley, and L. Wang, "Fast multi-labelling for stereo matching," in *Computer Vision - **ECCV** 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III*, pp. 524–537, 2010.

[7] Y. Zhang, L. Wang, R. I. Hartley, and H. Li, "Handling significant scale difference for object retrieval in a supermarket," in *DICTA 2009, Digital Image Computing: Techniques and Applications, 1-3 December 2009, Melbourne, Australia*, pp. 468–475, 2009.

[8] Y. Zhang, L. Wang, R. I. Hartley, and H. Li, "Where's the weet-bix?," in *Computer Vision - **ACCV** 2007, 8th Asian Conference on Computer Vision, Tokyo, Japan, November 18-22, 2007, Proceedings, Part I*, pp. 800–810, 2007.

[9] Y. Zhang, "Content-based image retrieval: out of the clutter in a big supermarket," *LAP Lambert Academic Publishing*, 2010.

[10] G. Sheasby, J. Warrell, Y. Zhang, N. Crook, and P. H. Torr, "Simultaneous human segmentation, depth and pose estimation via dual decomposition," *Proceedings of the workshop of British Machine Vision Conference (BMVC)*, 2012.

CV

Name: Yuhang Zhang	**Doctorial degree:** 2012-10-16
Birthdate: 19820402	**Academic title:** Doktor
Gender: Male	**Employer:** No current employer

Research education

Dissertation title (swe)
Snabb multi Märkning Early Vision

Dissertation title (en)
Fast Multi-Labelling in Early Vision

Organisation	Unit	Supervisor
Australian National University, Australia Not Sweden - Higher Education institutes	Research School of Information Science and Engineering	Richard Hartley

Subject doctors degree	ISSN/ISBN-number	Date doctoral exam
10207. Datorseende och robotik (autonoma system)		2012-10-16

Publications

Name: Yuhang Zhang	**Doctorial degree:** 2012-10-16
Birthdate: 19820402	**Academic title:** Doktor
Gender: Male	**Employer:** No current employer

Zhang, Yuhang has not added any publications to the application.

Register

Terms and conditions

The application must be signed by the applicant as well as the authorised representative of the administrating organisation. The representative is normally the department head of the institution where the research is to be conducted, but may in some instances be e.g. the vice-chancellor. This is specified in the call for proposals.

The signature*from the applicant*confirms that:

- the information in the application is correct and according to the instructions form the Swedish Research Council
- any additional professional activities or commercial ties have been reported to the administrating organisation, and that no conflicts have arisen that would conflict with good research practice
- that the necessary permits and approvals are in place at the start of the project e.g. regarding ethical review.

The signature *from the administrating organisation* confirms that:

- the research, employment and equipment indicated will be accommodated in the institution during the time, and to the extent, described in the application
- the institution approves the cost-estimate in the application
- the research is conducted according to Swedish legislation.

The above-mentioned points must have been discussed between the parties before the representative of the administrating organisation approves and signs the application.

Project out lines are not signed by the administrating organisation. The administrating organisation only sign the application if the project outline is accepted for step two.

Applications with an organisation as applicant is automatically signed when the application is registered.