DISS. ETH NO. XXXX

Towards Cost-Effective and Performance-Aware Vision Algorithms

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Science (Dr. sc. ETH Zürich)

presented by

Dengxin Dai
Master of Science
born November 15, 1986
citizen of China

accepted on the recommendation of

Prof. Dr. Luc Van Gool, examiner
Prof. Dr. Bernt Schiele, co-examiner

2016

DEDICATION GOES HERE...

Abstract

Computer vision has leaped forward during the last decade, and now is able to recognize objects of thousands of categories and reconstruct 3D scenes at city- or world-scale. However, the field still has to find means to keep up with the exploration of the massive amounts of data being captured on a daily basis. This is mainly due to the lack of sufficient training annotations and the lack of computational resources. The thesis is dedicated to mitigate the problem.

Firstly, we elaborate two strategies to reduce the annotation costs in order to train vision algorithms: (1) developing smart annotation approaches for efficient, large-scale annotations; and (2) learning better feature representations using unlabeled data; We develop algorithms for the strategies and show - in the context of recognition tasks - that they are able to considerably reduce the annotation costs for the training of recognition algorithms.

Secondly, in addition to reducing annotation cost, we also examine how to reduce the computational cost associated with the training and testing of vision algorithms. This research has lead to two contributions: (1) two efficient solvers for linear and kernel SVM+, significant speeding up the training process of SVM+ to explore privileged information; and (2) a method to allow computationally cheap features to imitate alternative features that perform better but are computationally more expensive. The imitation significantly improves the performance of the cheap features while retaining their efficiency.

Thirdly, as images keep growing in size, vision algorithms need to be more intelligent and self-aware of their performance. To this aim, we have developed two approaches: (1) Performance Forecasting to cheaply predict the success of vision algorithms for particular samples, which can be used for better resource allocation; and (2) Scale-Aware Image Segmentation to re-organize image segmentation hierarchies to better couple hierarchy depth and segmentation scale. The two methods also show potentials in reducing computational time of consecutive vision algorithms.

Zusammenfassung

Zusammenfassung geht hier...

Acknowledgements

Acknowledgement goes here.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Contributions 3

1.3 Organization 4

2 Conclusion **6**

2.1 Contributions 6

2.2 Perspectives 7

1

Introduction

A picture is worth a thousand of words, and often comes with a story that the photographer wanted to deliver. We probably still remember the moments when our grandparents were telling the stories behind the old pictures at our home, be it about birthday parties, be it about holidays. The purpose of taking pictures has been extended dramatically in the last decades, especially since the invention of digital cameras and the Internet. The inventions make it effortless ever to take pictures, store them, and share them with friends and families. While still used to remember such special moments of our daily life, most images and videos are taken now to support many other practical applications — cameras have been widely used for public security, for product control in factories, for product description in E-commerce, for real-world object reconstruction, for robot control, to name a few.

1.1 Motivation

While the quantity and the quality of images taken everyday are steadily increasing, it is hard to take its full advantage if the data is unorganized. In computer vision, algorithms are developed to tackle the fundamental problems of processing and understanding the visual data, in order to better suits the application context. The most extensively researched problems include image classification, object detection, action recognition, human pose estimation, object tracking in videos, 3D object reconstruction. With algorithms from these sub-fields, images and videos can then be converted into semantic information so that the data can be used to its full potential.

The last decade has witnessed a great progress in computer vision, which comes as an outcome of three factors: 1) introduction of large-scale human annotations, such as ImageNet [Deng et al., 2009] for image classification, PASCAL VOC [Everingham et al., 2010] and MSCOCO [Lin et al., 2014] for object detection, KITTI [Geiger et al., 2013] for road scene understanding; 2) development of sophisticated, statistical learning approaches including Support Vector Machines, Adaboost, Random Forests, and Deep Neural Networks; and 3) accessibility to powerful computing infrastructures such as the GPUs and many software frameworks, such as CuNN, Caffe, Torch, and Theano. These advantages have made the field achieved multiple milestones: computer vision is able to recognize objects of thousands of categories in high accuracy [Krizhevsky et al., 2012a; Simonyan & Zisserman, 2015], transcribe images/videos directly into natural languages [Vinyals et al., 2015], and generate realistic images [Gregor et al., 2015].

Despite the great achievement, the field is still left behind by the exploration of visual data. Visual content is exploding – millions of videos are generated and consumed every second, from the footage of surveillance cameras to the over two billion images and videos uploaded daily to social platforms. However, intelligent and insightful understanding of these data is still far from being reached. We believe there are at least three reasons behind this: 1) visual annotation is expensive to obtain; 2) algorithms in computer vision are often computationally heavy; and 3) developed methods often are not performance-aware.

The first reason is evidenced by the fact that most of the benchmarks still only cover a small portion of the visual data that is available in the wild, and are collected under quite controlled scenarios. The speed of visual annotation cannot match that of visual data acquisition — taking pictures, storing and sharing them are effortless nowadays, while annotating visual data in the form that machine can learn on is tedious and very time-intensive. For example, YouTube has 100 hours of videos (100 million frames) updated every minute, while it required 19 man-years to label 1.2 million internet images [Deng et al., 2009]. The annotations are done in the form of bounding boxes, rather than in more dedicated forms such as full segmentation masks.

The second reason follows the fact that visual data is very high-dimensional, rendering the algorithms computationally heavy in both training and testing. For instance, training a state-of-the-art image classifier [Simonyan & Zisserman, 2015; He et al., 2015] can take days or weeks even with modern GPUs, computing the optical flow of a standard video dataset [Karpathy et al., 2014] can simply take several days, and computing the similarity among images of a large dataset can be very expensive as well [Gong et al., 2013]. This heavy computation hinders the community from quickly exploring new models, easily deploying the trained models to power-limited devices, and readily scaling the developed models.

The last one is due to the trend that the community is mostly focused on developing the next best method for task X, and rarely on improving the self-awareness of the methods, be it a measure of model uncertainty [Kendall et al., 2015; Kondermann et al., 2008], be it the usefulness to down-streamed vision tasks [Yao et al., 2011; Jain et al., 2015]. This performance blindness leads to ineffective solutions in many situations, such as the same amount of resource is allocated to every image regardless of its complexity (difficulty), and the advance made in all the sub-fields cannot be synergized effectively.

1.2 Contributions

The aforementioned problems deliver a strong need to reduce the annotation cost and computational cost of current computer vision methods, and to learn to predict their performance. This thesis is dedicated to provide a collection of methods to attack these problems from multiple perspectives.

First, we elaborate two strategies to reduce the annotation cost:

- Developing efficient visual annotation approaches. A natural and efficient annotation method is developed for object recognition by letting annotators speak. Since drawing scribbles and speaking are very natural to human, our method unleashes the expressive ability of annotators and solves the *what* and *where* problems of object annotation both at the same time, leading to an approach which draws a good trade-off between recognition accuracy and annotation cost.

- Learning feature representation with unlabeled data. A new method is developed to learn a new feature representation on top of standard feature representations. The leaning takes advantage of discriminative learning and ensemble learning, and is able to generate new features specifically tailored to the data at hand.

Secondly, we examine how to reduce the computational cost associated with the training and testing of vision algorithms:

- Developing efficient training algorithms to the standard approach SVM+. SVM+ has shown excellent performance in visual recognition tasks for exploiting privileged information in the training data. We propose two efficient algorithms for solving the linear and kernel SVM+. Experiments show that our proposed algorithms achieve significant speed-ups to the state-of-the-art solvers for SVM+.

- Proposing Metric Imitation (MI), a method to allow computationally cheap features to imitate alternative features which perform better but are computationally more expensive. The leaned transformation significantly boost the performance of cheap features while retaining their efficiency.

Lastly, we investigate performance-aware vision algorithms by making the following contributions:

- Predicting how likely vision algorithms succeed on particular samples. It is true for every vision task that not all images are equally difficulty. We examine how to learn this on the task of example-based texture synthesis (ETS). ETS has been widely used to generate high quality textures of desired sizes from a small example.

However, not all textures are equally well reproducible that way. We predict how synthesizable a particular texture is by ETS and find that texture synthesizability can be learned and predicted efficiently.

- Learning scale-aware image segmentation. Hierarchical image segmentation provides segmentation at different scales in a single tree-like structure. However, they are not aware of the scale information of the regions in them. As such, one might need to work on many different levels of the hierarchy to find the objects in the scene. This work predicts the scales of the regions to modify their depth in the tree to better couple tree depth and region scale. The output of our method is an improved hierarchy, which improves the quality of the hierarchical segmentation representations.

- Evaluating the usefulness of image super-resolution methods to other computer vision tasks. Although it might be believed that image super-resolution is helpful for other vision tasks, this work has formalized the conception and conducted quantitative evaluation. The work can serve as an inspiration for the community to evaluate image super-resolution with respect to the helpfulness to other vision tasks, and to apply it as a pre-processing component if the input images are of low-resolution.

1.3 Organization

Since extensive research has been done in similar spirit, the thesis begins with Chapter **??** examining related work in a broad context. Our developed approaches are presented in Chapters **??** – **??**. They are written to be generally self-contained and can be read independently. Finally, Chapter 2 concludes this thesis. A detailed overview of the remaining chapters follows:

In Chapter **??**, *Related Work*, we provide a short literature overview of previous art in the direction of reducing the annotation cost and computational cost of vision algorithms and towards self-aware algorithms.

In Chapter **??**, *Efficient Visual Annotation with Speech Recognition*, we present our efficient visual annotation approach Draw&Tell and show its efficiency in the context of semantic image segmentation. In order to solve the *what* and *where* problems in visual annotation both at the same time, we let annotators speak the name of the object while they draw strokes on it. The speech is recognized by a speech recognition engine specifically trained for the purpose, and an extension to the fully convolutional neural network is made to learn from the stroke-based 'coarse' annotation. The approach draws a good trade-off between recognition accuracy and annotation cost. The work in this chapter was originally presented in [Dai et al., 2016a].

In Chapter **??**, *Representation Learning with Unlabeled Data*, we present our motivation and method of learning a new feature representation with unlabeled data. We then evaluate the method in the context of semi-supervised image classification and image clustering. This work was originally presented in [Dai et al., 2012] and in [Dai & Van Gool, 2013].

In Chapter **??**, *Fast Training Algorithms for SVM+*, we present two efficient algorithms for training the linear and kernel SVM+. New problems with fewer constraints are formulated in the dual domain, making the problem solvable efficiently by the SMO algorithm of one-class SVM. Experiments show that our proposed algorithms are significantly faster than the the state-of-the-art solvers for SVM+. This work was originally presented in [Li et al., 2016].

In Chapter **??**, *Efficient Metric Computation via Imitation*, we present a method called Metric Imitation (MI) to efficiently compute the distance among images. MI learns a transformation to cheap features so that the distance with the transformed features can approximate the distance with better-performing but computationally-expensive features. The method was originally presented in [Dai et al., 2015].

In Chapter **??**, *Performance Prediction: Succeed or Fail?*, we present our approach of predicting the success of example-based texture synthesis. To this aim, we collected a dataset with $21,302$ annotated textures and annotated them according to the synthesizability — the quality of synthesized results by texture synthesis methods. A set of relevant features are then defined to regress the value of synthesizability. Extensive experiments show that texture synthesizability is learnable. This work was originally presented in [Dai et al., 2014].

In Chapter **??**, *Performance Prediction: Under-, Properly-, or Over-Processed?*, we present a method to predict the scale of image segments relative to the corresponding objects, and to then apply the prediction to re-align the results from general hierarchical image segmentation so that the depth in the tree structure and the scale of the regions is better coupled. The output of our method is an improved hierarchy, which improves the quality of the hierarchical segmentation representations. The work was originally presented in [Chen et al., 2016].

In Chapter **??**, *Performance Evaluation: Helpful for Other Tasks?*, we evaluate the usefulness of image super-resolution methods to other computer vision tasks. Sixes state-of-the-art computer vision algorithms are evaluated on four popular computer vision tasks: boundary detection, semantic image segmentation, digit recognition, and scene recognition. The work confirms that image super-resolution is helpful for other vision tasks when the state-of-the-art approaches are used. The work was originally presented in [Dai et al., 2016b].

2

Conclusion

This thesis presented methods towards cost-effective and performance-aware vision algorithms. The primary focus was on effective algorithms in terms of annotation cost, for which we presented two different approaches, one for efficient visual annotation and one for learning with unlabeled data. In addition, we examined how to reduce the computational cost associated with the training and testing of vision algorithms in the context of learning with privileged information and metric learning. Finally, we investigated how performance-ware algorithms can be learned, and then be used to facilitate downstreamed applications. A more detailed look at the specific contributions are summarized below.

2.1 Contributions

In Chapter **??**, *Efficient Visual Annotation with Speech Recognition*, we proposed an efficient annotation method for semantic image segmentation by leveraging the power of speech recognition. In this method, we allow annotations to speak objects's names while they draw strokes on them. Object names are recognized by a combination of speech recognition and webly-supervised object recognition. The drawn strokes are then converted to semantic heatmaps for the corresponding classes by ensemble interactive segmentation. Finally, an extension to the standard fully convolutional networks [Long et al., 2015] is made to accommodate the 'weak' annotation. The method yields comparable results to the same CNNs architecture trained with standard annotations of full segmentation masks, while being 10x faster.

In Chapter **??**, *Representation Learning with Unlabeled Data*, we developed a method to learn new feature representations by exploring the data distribution patterns of unlabeled data. The leaning takes advantages of discriminative learning and ensemble learning to effectively exploit manifold-smoothness assumption: surrogate classes are sampled from the manifold of data samples, on which discriminative learning is performed to extract the high-level knowledge; the noise of the sampled training training set is mitigated by ensemble learning. The learned discriminative classifiers are used to generate the new high-level feature representations. Experiments on eight datasets show that the learned representations are superior to the standard feature representations.

In Chapter **??**, *Fast Training Algorithms for SVM+*, we developed efficient training algorithms to the standard approach SVM+. New fomulations with fewer constraints are formulated in the dual domain, making it solvable efficiently by the SMO algorithm of one-class SVM. Experiments show that our proposed algorithms achieve significant speed-ups to the state-of-the-art solvers for SVM+.

In Chapter **??**, *Efficient Metric Computation via Imitation*, we developed a method to allow computationally cheap features to imitate better-performing but computationally more expensive features. We treat the problem as a transfer learning, where the neighborhood property expensive features are quantified into manifold structures. The manifold structures are view-independent, and can then be transferred to the space of cheap features. Finally, a linear transformation of the cheap features is learned so that the manifold can be approximated as well as possible. The leaned transformation significantly boosts the performance of cheap features while retaining their efficiency. Experiments on multiple experiments validate the efficacy of the method.

In Chapter **??**, *Performance Prediction: Success or Fail?*, We predicted how likely texture synthesis succeed on particular texture samples. To the aim, we collected a dataset of $21,302$ textures and performed 4 standard synthesis methods on it. The synthesized results were annotated based on their visual quality in terms of three levels. A set of relevant features were defined in order to learn a regressor to predict the performance of the methods. Extensive experiments show that the performance of texture synthesis methods can be predicted accurately.

In Chapter **??**, *Performance Prediction: Under-, Properly-, or Over-Processed?*, we presented a method to predict whether a segment from image segmentation methods is under-segmented, properly-segmented, or over-segmented. The prediction is then used to re-align the tree structure of hierarchical image segmentation, so that the depth of a regions is better coupled with its scale. The improved hierarchy improves the quality of the hierarchical segmentation representations.

In Chapter **??**, *Performance Evaluation: Useful for Other Vision Tasks?*, we extensively evaluated the usefulness of image super-resolution for four standard vision tasks. This work has formalized the conception that image super-resolution is helpful for other vision tasks when the input images are of low-resolution.

2.2 Perspectives

Deeper integration of vision and speech. In Chapter **??**, vision and speech is integrated to recognize the name of the object. In this work, we only fuse the class probability from the speech recognition engine and the webly-supervised object recognizer. A deep integration of the two streams of information probably will help for better accuracy, especially given the great success of multi-stream convolutional neural networks in integrating different sources of information, such as RGB color and optical flow [Simonyan & Zisserman, 2014]. Although the current method yields good recognition accuracy, the accuracy drops as the number of object increases. This implies that a better (deeper) integration of the two sources of information is necessary if the number of object classes of interest is very large.

In a broader context, a better integration of vision and speech can lead to voice-user interface to pictures. Voice-user interface (VUI) has become more commonplace, and people are increasingly taking advantage of its good values ? it is hands-free, eyes-free, and far more mobile. As we people in 21th century communicate in pictures, there is a strong need to have a VUI to talk to pictures in order to perform scene-relevant tasks, such as image tagging, story logging, and visual question answering. Also, a VUI to pictures can offer assistance to blind people, and dictate robotics to react to real visual scenes (situations) in a hand-free manner such as in clinical robotic surgery. We hope this work can inspire more research effort in integrating vision and speech.

Other forms of annotation for the place of objects. The location of object is indicated by scribbles in Chapter **??**. While scribbles are very natural to draw and they are very suitable for stuff classes such as sky and road, other forms of annotation might be more informative for well-shaped objects such as cars and pedestrians. Examples of other annotation forms include coarse bounding boxes and ellipses. Bounding-box is especially helpful, given the fact that numerous learning approaches have been developed for bounding-box based annotations. Furthermore, investigating when to use what forms of the annotation is interesting, and a solution to it will boost the annotation efficiency even further.

Combing with other cost-effective annotation methods. As stated in Related Work **??**, there exist many techniques for cost-effective annotations, most of which are perpendicular to our method, such as active learning, human-in-the-loop, crowd-sourcing, and gaming. Also, our method is 'naturally' suitable for mobile devices, in which the microphone is already integrated and the hand-input devices are not as easily accessible as that for work stations. Annotating a large dataset is always a practically valuable contribution to the community. Crowd-sourcing the method or making the method available as an app for mobile devices seems to be a good step towards constructing a large dataset with a large number of objectsj.

Integrating ensemble projection into deep neural network. The representation learning by ensemble projection (Chapter **??**) is performed on an ensemble of training set with sampled surrogate class labels. This scenario shares high similarity with that of the feature learning work developed in [Dosovitskiy et al., 2014]. Inspired by [Dosovitskiy et al., 2014], we find the possibility of formulating our ensemble projection as an 'unsupervised' neural network training problem. A straightforward solution is to add the non-linearity transformation of ensemble projection to the top of the standard neural network [Chatfield et al., 2014] and use the classification scores of the surrogate classes to fine-tune the network.

Distilling knowledge of neural networks via metric imitation. The cheap features used in our Metric Imitation (Chapter **??**) are all shallow features, such as Gist and LBP, and the transformation learned is a linear transformation. Extending it to network imitation is a great merit to have: to fine-tune (corresponding to a non-linear transformation) a shallower network by imitating the distance metric computed by a deep neural network. With the advance of deep learning, there are numerous of deep neural networks developed [Krizhevsky et al., 2012b; Simonyan & Zisserman, 2015; Szegedy et al., 2015; He et al., 2015]. The networks are of very different complexity: some are deeper than others, some are wider than others. Often, the top-performing ones are the deepest ones. Knowledge distillation [Hinton et al., 2014] is a technique developed train lighter (shallower) networks by imitating deeper networks – to produce similar class scores for a large pool of images. Investigating the potential of using metric imitation for knowledge distillation is intriguing, and constitutes our future work.

Performance prediction for more tasks. xxx

Bibliography

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*.

Chen, Y., Dai, D., Pont-Tuset, J., & Van Gool, L. (2016). Scale-aware alignment of hierarchical image segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

Dai, D., Kroeger, T., Li, W., & Van Gool, L. (2016a). Draw&tell: Efficient annotation for semantic image segmentation by drawing and speaking. In *in submission to ECCV*.

Dai, D., Kroeger, T., Timofte, R., & Van Gool, L. (2015). Metric imitation by manifold transfer for efficient vision applications. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3527–3536).

Dai, D., Prasad, M., Leistner, C., & Gool, L. V. (2012). Ensemble partitioning for unsupervised image categorization. In *ECCV*.

Dai, D., Riemenschneider, H., & Van Gool, L. (2014). The synthesizability of texture examples. In *CVPR*.

Dai, D., & Van Gool, L. (2013). Ensemble projection for semi-supervised image classification. In *International Conference on Computer Vision*, (pp. 2072–2079).

Dai, D., Wang, Y., Chen, Y., & Van Gool, L. (2016b). Is image super-resolution helpful for other vision tasks? In *WACV*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dosovitskiy, A., T. Springenberg, J., Riedmiller, M., & Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, (pp. 766–774).

Everingham, M., Van Gool, L., Williams, C., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *IJCV*, *88*(2), 303–338.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*.

Gong, Y., Kumar, S., Rowley, H. A., & Lazebnik, S. (2013). Learning binary codes for high-dimensional data using bilinear projections. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gregor, K., Danihelka, I., Graves, A., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. In *ICML*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Hinton, G., Vinyals, O., & Dean, J. (2014). Distilling the knowledge in a neural network. In *Deep Learning and Representation Learning Workshop, NIPS*.

Jain, M., van Gemert, J. C., & Snoek, C. G. (2015). What do 15,000 object categories tell us about classifying and localizing actions? In *Computer Vision and Pattern Recognition (CVPR)*, (pp. 46–55).

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

Kendall, A., Badrinarayanan, V., & Cipolla, R. (2015). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*.

Kondermann, C., Mester, R., & Garbe, C. (2008). A statistical confidence measure for optical flows. In *ECCV*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012a). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, (pp. 1097–1105).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. In *NIPS*.

Li, W., Dai, D., Tan, M., Xu, D., & Van Gool, L. (2016). Fast algorithms for linear and kernel svm+. In *Computer Vision and Pattern Recognition (CVPR)*.

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

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1–9).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *CVPR*.

Yao, A., Gall, J., Fanelli, G., & Gool, L. V. (2011). Does human action recognition benefit from pose estimation? In *British Machine Vision Conference (BMVC)*.

List of Publications

Journal Publications

1. A. Yao, J. Gall, L. Van Gool. Coupled Action Recognition and Pose Estimation from Multiple Views. *International Journal of Computer Vision (IJCV)*, 2012. 100(1), 16-37.

2. J. Gall, A. Yao, N. Razavi, L. Van Gool and V. Lempitsky. Hough Forests for Object Detection, Tracking and Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2011. 33(11), 2188-2202.

3. A.Y.J. Yao and W. Einhaeuser. Colour aids late but not early stages of rapid natural scene recognition. *Journal of Vision*, 2008. 8(16):12, 1-13.

Refereed Conference Proceedings

1. A. Yao, J. Gall, C. Leistner and L. Van Gool. Interactive Object Detection. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

2. A. Yao, J. Gall, L. Van Gool, and R. Urtasun. Learning Probabilistic Non-Linear Latent Variable Models for Tracking Complex Activities. *Neural Information Processing Systems (NIPS)*, 2011.

3. A. Yao, J. Gall, G. Fanelli and L. Van Gool. Does Human Action Recognition Benefit from Pose Estimation? In *Proceedings British Machine Vision Conference (BMVC)*, 2011.

4. A. Yao, D. Uebersax, J. Gall and L. Van Gool. Tracking People in Broadcast Sports. In *Proceedings German Association for Pattern Recognition (DAGM)*, 2010.

5. J. Gall, A. Yao and L. Van Gool. 2D Action Recognition Serves 3D Human Pose Estimation. In *Proceedings European Conference on Computer Vision (ECCV)*, 2010.

6. A. Yao, J. Gall and L. Van Gool. a Hough Transform-Based Voting Framework for Action Recognition. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Others

1. D. Waltisberg, A. Yao, J. Gall and L. Van Gool. Variations of a Hough-Voting Action Recognition System. *Proceedings International Conference on Pattern Recognition (ICPR) Contests*, 2010.

2. G. Fanelli, A. Yao, P.L. Noel, J. Gall and L. Van Gool. Hough Forest-Based Facial Expression Recognition from Video Sequences. *International Workshop on Sign, Gesture and Activity (SGA)*, 2010.

Curriculum Vitae

Personal Data

Name	Stefan Saur
Date of birth	1^{st} December 1979
Place of birth	Buchen (Odenwald), Germany
Citizenship	German

Education

2005 – 2009	*ETH Zurich, Computer Vision Laboratory, Switzerland* Doctoral studies
2004	*National University of Singapore, Singapore* Semester abroad
2000 – 2005	*University of Karlsruhe, Germany* Studies of Electrical Engineering and Information Technology Graduation with the degree Dipl.-Ing.
1990 – 1999	*Ganztagsgymnasium Osterburken, Germany*

Work Experience

2005 – 2008	*ETH Zurich, Computer Vision Laboratory, Switzerland* Teaching and research assistant
2000 – 2005	*Siemens AG, Germany* Several internships, semester project, and master thesis
2001 – 2009	*Webdesign, self-employed*

Awards

2004	*Baden-Württemberg Stipendium*, Landesstiftung Baden-Württemberg
2000	*IPP award*, University of Karlsruhe, Germany