

Learning Structured Features for Computer Vision

Sebastian Nowozin, Gökhan Bakır¹, Koji Tsuda



For many computer vision applications low-level image features are used to process and analyze image data. High-level vision tasks such as object class recognition and action recognition are often approached by learning statistics on such low-level features. In our work, we instead *learn* structured features directly, in order to exploit discriminative structure usually not available in simple statistics. The methods developed should be broadly applicable, employ computationally tractable training procedures, have interpretable classification rules and obtain high accuracy on benchmark data sets.

In [1] we propose an unsupervised and supervised model for learning discriminative sets of low-level “visual word” features. Thus, feature selection is performed in the power set of all visual words. The set approach is further generalized to handle pairwise geometry constraints between visual words by selecting discriminative subgraph features from the set of all labeled connected subgraphs defined on the image. In [2] we extend the approach to handle temporal structure for action recognition in videos. Human actions such as “walking” and “boxing” are seen as temporal sequence of “video word” features and optimal discriminative subsequence features are selected such that temporal ordering constraints are preserved.

We evaluate our proposed methods on standard benchmark data sets for object class recognition such as the PASCAL VOC 2006 challenge. The approach proposed in [1] achieves state-of-the-art performance. The action recognition method is benchmarked on the KTH action recognition data set, showing excellent performance [2].

The developed implementations of the proposed algorithms are made available as open-source software on the author’s home-page.

References

1. Nowozin, S., K. Tsuda, T. Uno, T. Kudo, G. Bakır: Weighted Substructure Mining for Image Analysis. *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, 1–8, IEEE Computer Society, Los Alamitos, CA, USA (2007).
2. Nowozin, S., G. Bakır, K. Tsuda: Discriminative Subsequence Mining for Action Classification. *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV 2007)*, 1919–1923. IEEE Computer Society, Los Alamitos, CA, USA (2007).



Example of a the feature space used: fully connected graph defined on Harris-Laplace feature points. Discriminative subgraphs are selected from the set of all connected subgraphs [1].

¹ Now at: Google, Zurich, Switzerland

Discriminative Object Localization

Christoph H. Lampert, Matthew B. Blaschko

Object localization, also called object detection, is an important task for image understanding, e.g. in order to separate an object from the background, or to analyze the spatial relations of different objects. Despite many decades of computer vision research, the question how arbitrary object categories can be detected and recognized automatically in natural images is still far from solved. As a field of active research it borders machine learning, visual perception and image processing.

Many popular techniques to detect objects in images rely on localized classification. They evaluate a quality function for many different candidate boxes in the image, ideally for all possible ones. Consequently, two of the most important questions in the area of object localization are:

- How does one train a quality function for the task of object localization?
- How does one evaluate the quality function efficiently over all candidate regions?

Our hypothesis was that these two aspects are inherently linked, and that better localization results can be achieved by developing a framework that solves both problems jointly.

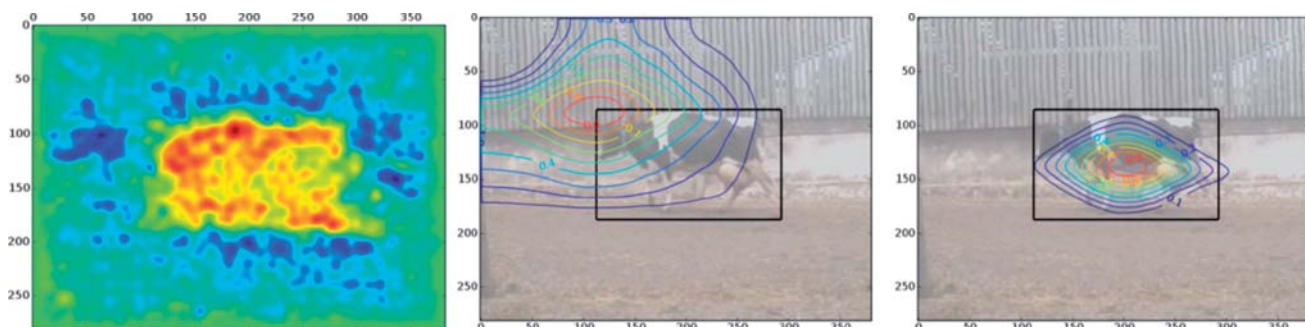
Contrary to previous approaches we see object localization not as a sequence of classification tasks, but as a problem of regression in a structured output space that consists of all possible bounding boxes in an image. By this approach, we make the test phase of the algorithm, the prediction of a bounding box for a new image, consistent with the training material, images and their bounding boxes.

To solve the structured regression problem, we introduce the concept of a *restriction kernel* as a joint-kernel function and apply the structured support vector machine framework. We obtain training and test procedures that allow to learn a system for object localization from examples in a globally optimal yet efficient way. With only a single free parameter required for training and none for testing, the resulting system is much easier to use than previous approaches and applicable to a wide range of localization problems. At the same time, it performs object localization with high speed and high accuracy, as is proved e.g. by winning 5 out of 20 first places in 2007's international PASCAL challenge on visual object categorization.



References

1. **Lampert, C. H., M. B. Blaschko**, T. Hofmann: Beyond Sliding Windows: Object Localization by Efficient Subwindow Search. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 1–8, IEEE Computer Society, Los Alamitos, CA, USA (2008).
2. **Lampert, C. H., M. B. Blaschko**: A Multiple Kernel Learning Approach to Joint Multi-Class Object Detection. In *Proceedings of the 30th DAGM Symposium*, 31–40. (Eds.) Rigoll, G. Springer, Berlin, Germany (2008).
3. **Blaschko, M. B., C. H. Lampert**: Learning to Localize Objects with Structured Output Regression. In *Computer Vision, ECCV 2008*, 2–15. (Eds.) Forsyth, D., P. Torr, A. Zisserman, Springer, Berlin, Germany (2008).



Object Localization. Left image: a per-pixel weight distribution that is learned by structured regression. Middle/right image: contour lines of the resulting quality function for predicting the top-left and center point of the object's bounding box.

Detection of Steganography in Images Using Statistical Models

Valentin Schwamberger, Matthias O. Franz¹, Bernhard Schölkopf



Steganography is the art of information hiding. Unobtrusive carriers are used that do not raise suspicion when being transmitted or stored. The actual data is embedded in these carriers. Security agencies assume that criminals use such technologies for undiscovered communication or for saving critical data. Typical carriers are digital images (Figure 1).

Counter technologies and algorithms that attack steganography are called *steganalysis*. Most steganalytic methods are not capable of detecting general steganographic manipulations in images, since they are tuned to specific steganographic algorithms. The few currently available universal steganalytic algorithms are relatively insensitive towards small embeddings [3]. This is due to the problem of detecting a tiny manipulation (the embedded data) in a large amplitude signal (the carrier image). The goal of this project is to develop a state-of-the-art universal steganalytic algorithm building on a previously developed predictive image model [2]. A higher sensitivity should be achieved. The image model predicts coefficient values from neighborhoods in a Laplacian pyramid or wavelet representation of the image based on the coefficient statistics of the image [1]. To this end, a Bayesian predictor is utilized. Since an embedded message cannot be predicted from the neighborhood statistics of the image, it must be part of the prediction error of the model. Thus, by analyzing the prediction error of the model instead of the whole image, we effectively remove most of the carrier signal. The residuum is much more affected by the embedding manipulation than the full image, which results in a better detectability. The algorithm extracts a set of statistics from the logarithmic prediction error. Finally, a support vector machine is trained on this data for classification (Figure 2).

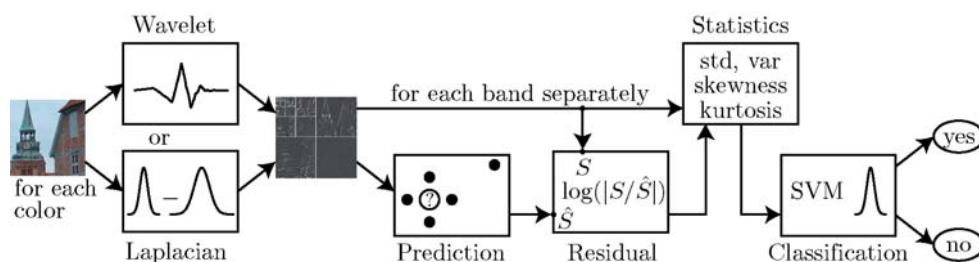
Using this image model, high prediction performance for color images has been achieved so far. Since there exists a strong correlation between color channels of an image, the prediction is very accurate for this case. Thus, the error is small and the extracted statistics result in a good discriminability when plugged into the non-linear support vector machine. For black and white images, the predictability is lower, but higher embedding rates can still be detected. The prediction with the Bayesian image model turned out to be more reliable than the standard least-squares approach. Thus, the noise estimation works well. Besides that, extensions of the model – for example, using a larger number of neighbors – should further improve the detectability of embedded messages.

References

1. Buccigrossi, R. W., E. P. Simoncelli: Image Compression via Joint Statistical Characterization in the Wavelet Domain. *IEEE Transactions on Image Processing* 8(12), 1688–1701 (1999).
2. Franz, M. O., B. Schölkopf: Implicit Wiener Series for Higher-order Image Analysis. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, 465-472. (Eds.) Weiss, Y., L. K. Saul, L. Bottou, MIT Press, Cambridge, MA, USA (2005).
3. Lyu, S., H. Farid: Steganalysis Using Higher-order Image Statistics. *IEEE Transactions on Information Forensics and Security* 1(1), 111–119 (2006).



1. The image on the right hand side has been embedded in the original image on the left hand side (using simple least significant bit embedding). The resulting image is in the middle.



2. At a glance: A flow diagram of the developed algorithm for steganalysis.

¹ Now at: University of Applied Sciences, Konstanz, Germany

Non-monotonic Likelihood Maximization for Image Reconstruction in Tomography

Suvrit Sra, Dongmin Kim¹, Bernhard Schölkopf

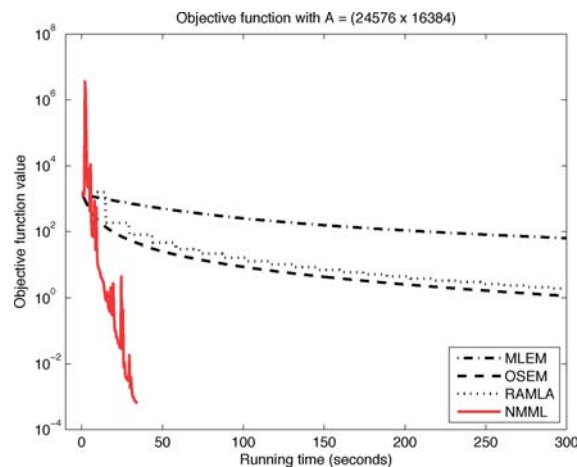
Maximizing some form of Poisson likelihood (either with or without penalization) is central to image reconstruction algorithms in emission as well as transmission Tomography. We introduce NMML, a non-monotonic algorithm for maximum likelihood image reconstruction in emission and transmission Tomography [1]. A vast number of image reconstruction algorithms have been developed in tomography, and new ones continue to be designed. However, methods based on expectation maximization (EM) and the ordered-subsets (OS) framework seem to have enjoyed the greatest popularity. Our new method, NMML, differs fundamentally from methods based on EM: 1) it does not depend on the concept of *optimization transfer* (or surrogate functions), and 2) it is a rapidly converging *non-monotonic* descent procedure. The most important property of NMML is however its simplicity that is supplemented by efficiency and scalability. We also provide a theoretical convergence analysis that is supported by extensive empirical evaluation. NMML may be viewed as a specialized gradient-projection

algorithm, but without the corresponding expensive line-search steps. We compare our method with well-established EM based methods that are common in medical tomography [2]. Our initial results are promising. NMML significantly outperforms the EM based methods (Figure 1) for emission reconstruction. We are currently investigating results for the transmission case, which is numerically a more difficult optimization problem. Figure 2 shows initial results of image reconstruction using penalized likelihoods. NMML yields a smoother image than OSEM given the same amount of running time.

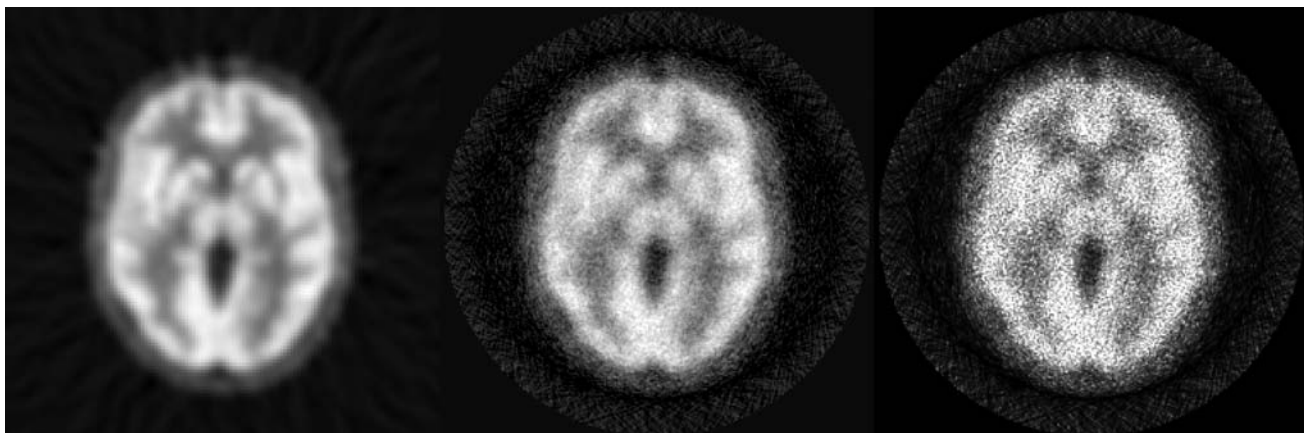


References

1. Sra, S., D. Kim, B. Schölkopf: Non-monotonic Poisson Likelihood Maximization. *MPI-Technical Report*, 170, (2008).
2. Fessler, J.: Image Reconstruction: Algorithms and Analysis. Book under preparation, (2008).



1. Running times of NMML in comparison to other methods.



2. Regularized reconstruction comparison. Left: original; center: NMML; right: OSEM.

¹ University of Texas at Austin, USA

Automatic 3D Face Reconstruction from Images or Video

Kwang In Kim¹, Pia Breuer², Wolf Kienzle³, Volker Blanz², Bernhard Schölkopf



Reconstruction of 3D faces from images or video can afford numerous interesting applications including generation of facial animation and face recognition that is tolerant to variations in pose and illumination. While there are various approaches for this problem, algorithms that rely only on *single* still images are of particular interest due to their little amount of restrictions on the application environment, i.e., they can be applied to any face image without requiring the existence of multiple simultaneous views or consecutive monocular views. One successful method in this category is to fit 3D face models to single images.

However, existing methods for fitting 3D face models require user interaction. For example, several systems require users to provide the locations of critical feature points such as eye and mouth corners.

In this project, we investigate an algorithm for fully automated reconstruction of 3D faces that

- can be applied either to single still images or to raw monocular video streams,
- operates across a wide range of poses and illuminations,
- produces close-to-photorealistic 3D reconstructions.

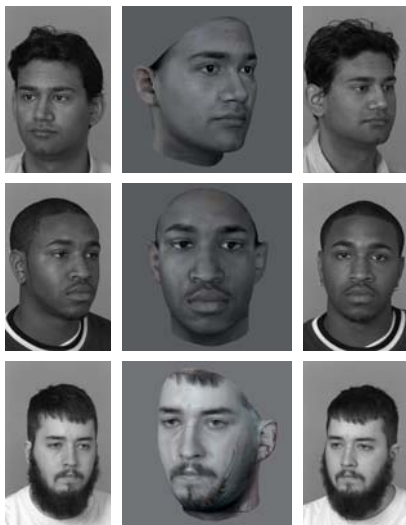
The algorithm is composed of three stages: 1. detect a single face in a given input image or video; 2. detect facial features in the

face image; and 3. fit a 3D face model. A Support Vector Machine (SVM)-based detector is adopted for face detection. The detections confidences provided by the SVM are used to single out the best face from all possible detections. Then, from this face image, a set of candidates for the locations of feature points are generated using a regression-based estimation followed by a classification-based refinement. From these candidate feature locations, a single best configuration is chosen by exploiting a prior on the configuration of feature points plus the scores

obtained by fitting a simplified 3D morphable model. This is then fed in to the full fitting model to guide the fitting process. Below, we show examples of 3D face reconstruction from images (Figure 1) and video (Figure 2) [1].

References

1. Breuer, P., K. I. Kim, W. Kienzle, B. Schölkopf, V. Blanz: Automatic 3D Face Reconstruction from Single Images or Video. In Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008), 1–8, IEEE Computer Society, Los Alamitos, CA, USA (2008).



1. Examples of face reconstruction from still images. From left to right: original image, novel view of the automated reconstruction, and original second view of the person (not used for the reconstruction).



2. An example of reconstruction from a webcam video. A single frame is chosen based on face detection scores. Left: original frame and right: automatically reconstructed head. The bottom row shows 6 sample frames.

Bayesian Color Constancy

Peter V. Gehler, Carsten Rother¹, Andrew Blake¹, Tom Minka¹, Toby Sharp¹

Color Constancy is the tendency to perceive surface color consistently, despite variations in ambient illumination. In computational color constancy one aims to estimate the true reflectances of visible surfaces in an image. This estimation in turn can be used to remove the color cast of the scene illuminant, a process also known as Auto-White Balance. In our work we built up on a Bayesian formulation of the image formation and show that the illuminant estimation greatly benefits from improved priors for illuminant and reflectances. Those priors are obtained from a new dataset which we collected for the purpose of benchmarking color constancy algorithms. This work appeared in [1].

Most generally, illumination variations occur both within scenes, and from scene to scene, and theories such as the “Retinex” have been devised to explain color constancy under such conditions. In our work we follow a line of research that assumes a uniform illumination of the scene. Even this problem is severely under-constrained in principle but we can resolve this by exploiting assumptions about the variability of scene reflectance. The Bayesian approach models this variability of reflectance and illuminant as random variables and then estimates the scene illuminant from the posterior distribution conditioned on image intensity data.

We collected a new dataset using a high quality digital SLR camera in RAW format, free of any color correction. A copy of the MacBeth ColorCheckerChart was placed in each scene which allows us to accurately estimate the dominant scene illuminant. A total of 568 images are available both indoor (246) and outdoor (322).

The new dataset is used to benchmark different color constancy algorithms. A main result is that the use of accurate priors for both reflectance and illumination of the scene devised from this data gives an improvement of the Bayesian color constancy algorithm. Another finding is that on outdoor images alone there is no improvement to be gained over the auto white balance setting of the digital camera. For indoor images however big gains in performance can be achieved.



References

1. Gehler, P. V., C. Rother, A. Blake, T. Minka, T. Sharp: Bayesian Color Constancy Revisited. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 1–8 (2008).



1. An example indoor image from the new image database we collected. The image was converted from the RAW format using the Auto White Balance setting of the Canon camera. In the lower left corner of the image one sees the copy of the MacBeth ColorCheckerChart which is used to estimate the color cast of the scene.



2. The same image as in Figure 1 after correction with our Bayesian Color Constancy algorithm. The yellow cast of the image was successfully removed. The dominant illuminant of this scene was a neon light which could apparently not be corrected by the auto-white balance of the digital camera.

¹ Microsoft Research Cambridge, UK

Automatic Image Colorization via Multimodal Predictions

Guillaume Charpiat¹, Matthias Hofmann², Ilya Bezrukov, Yasemin Altun, Bernhard Schölkopf



We consider the task of predicting how an image would look if it had been produced by a different tool, e.g., in medical imaging, how to transform a Magnetic Resonance scan into a Computed Tomography scan [p. 54], or, in photography, how to transform a grayscale image into a color one. In all cases, the issue is to model automatically, from a training set of image pairs (A_i, B_i) a function which transforms the input image A_i into the output image B_i .

On the one hand, because of the usually very high dimension of the space of images (e.g. $2 \cdot 10^5$ for the space of all 400×500 px images), the training set cannot be dense and consequently one cannot use standard learning tools involving directly distances between images. On the other hand, there is most often no one-to-one correspondence between input and output pixel values, so one cannot directly predict pixel values. The question is then to know what to learn.

In a first approximation, one can assume that the probability of an output value at a pixel depends mostly on its neighborhood in the input image, and that this local conditional probability does not depend on the pixel location (but only on the values in its neighborhood). The problem becomes more tractable and consists in learning these conditional probabilities.

There is consequently a need for extracting information from local image descriptors, for predicting several assignment possibilities at the pixel level, and for computing the whole output image based on local possibilities and spatial coherence. In [1] we propose a general approach for image prediction, applied to the colorization task: how to assign color to a grayscale image, given an example set of color images. We compute texture features to describe local grayscale neighborhoods, and choose the psychophysical L-a-b color space as a metric space for colors. We learn, with standard machine learning tools, the probability of all possible colors given a grayscale

neighborhood, instead of predicting only the most probable color as in most approaches. We also learn the probability of color variation given a grayscale neighborhood, in order to predict how homogeneous or heterogeneous colors are supposed to be at this location. This defines a spatial coherency criterion. The problem is then mathematically stated so that a graph cut technique can be applied, which returns the solution very fast.

In spite of the limited information provided by the texture descriptors, this new approach matches the state of the art in automatic colorization. Real-time user intervention is also possible if color landmarks are needed.

We have also applied methods of structured prediction learning [2] to obtain a prediction function for image colorization that works discriminatively instead of generatively. This includes the solution of several *maximum-a-posteriori* inference problems, for which we use loopy belief propagation and tree-reweighted message passing. Our initial experiments suggest that structured SVMs can be applied successfully to the task of intermodality image prediction. We plan to improve the speed of training and extend the method for the application on three-dimensional MR data.

References

1. Charpiat, G., M. Hofmann, B. Schölkopf: Image Colorization via Multimodal Predictions. In *Proceedings of the 10th European Conference on Computer Vision (ECCV)*, 126–139, (Eds.) Forsyth, D., P. Torr, A. Zisserman, Springer, Berlin, Germany (2008).
2. Tsochantaridis, I., T. Joachims, T. Hofmann, Y. Altun: Large Margin Methods for Structured and Interdependent Output Variables. In *Journal of Machine Learning Research (JMLR)* 6, 1453–1484 (2006).



Example of colorization. Top row : (left) one painting by Da Vinci as a training set, (middle) a second one as a test image, (right) the predicted image. The border is not colored because of the window size needed for texture descriptors. Bottom row : (left) color variation predicted (white stands for homogeneity and black for color edge), (middle) most probable color at the local level (for information purpose), (right) color chosen by the graph cuts, taking into account all color probabilities and color variation predicted. Note that blue is the most probable color at the pixel level in several skin regions (because sky and skin textures are very similar), but that on the global scale, with skin surroundings which cannot be confused with sky, this color is not probable for spatial coherency reasons.

¹ Now at: INRIA Sophia-Antipolis, France; ² Also at: Universitätsklinikum Tübingen, Germany and University of Oxford, UK

Image Denoising Using Optimized Dark Frames

Manuel Gomez Rodríguez¹, Jens Kober, Bernhard Schölkopf

Long exposure photographs contain substantial amounts of noise, drastically reducing the Signal-To-Noise Ratio (SNR). Camera-dependent noise reduction applied directly to raw photographs before any further processing can confer substantial benefits. In commercial cameras, this is usually done by subtracting a “dark frame” taken immediately after the actual photograph using identical settings but closed shutter. The goal of this work is to find an efficient method to denoise long exposure photographs taken with a given camera under unknown temperature conditions.

We have developed a novel optimization approach for this problem. It is based on a library of dark frames previously taken under varying conditions of temperature, ISO setting and exposure time, and a quality measure or prior for the class of images to denoise. Our method automatically computes a synthetic dark frame that, when subtracted from an image, optimizes the quality measure. For specific choices of the quality measure (e.g., squared image gradient), the denoising problem reduces to a quadratic programming (QP) problem resembling that of a support vector machine that can be solved efficiently. We show experimentally that it is sufficient to consider a limited subsample of pixels when evaluating the quality measure in the optimization, in which case the complexity of the procedure does not depend on the size of the images but only on the number of

dark frames. We provide quantitative experimental results showing that our method automatically computes dark frames that are competitive with those taken under idealized conditions (controlled temperature, ISO setting, exposure time, and averaging of multiple exposures), and superior to those taken under realistic conditions (varying temperature). We provide application examples in astronomical image denoising. The method is validated on two CMOS cameras as well as on a CCD camera.

By building on dark frames recorded with the same camera as the photograph to be denoised, our approach is specific to the camera used to take the pictures, rather than being a generic image processing method. It can be combined with classical image processing methods such as wavelet denoising or bilateral filtering, which further improves the results.

In future work, we plan to investigate better image quality measures and other ways of estimating synthetic dark frames.



Reference

1. **Gomez Rodriguez, M., J. Kober, B. Schölkopf:** Denoising Photographs Using dark Frames Optimized by Quadratic Programming. *Proceedings of the IEEE International Conference on Computational Photography ICCP*, (2009).



Original image



Denoised Image

Denoising results for a single image recorded using a CMOS digital SLR; exposure time 60 seconds at ISO 1600.

¹ Stanford University, USA

Online Blind Image Deconvolution for Astronomy

Stefan Harmeling, Michael Hirsch, Suvrit Sra, Bernhard Schölkopf



When we observe an astronomical object from the surface of the earth its emitted or reflected light invariably passes through the atmosphere, which yields a blurry observed image. Deblurring such images is therefore a fundamental problem in astronomy. The particular challenge is that the blur is not only unknown, but is also continually changing in time as well as spatially due to refraction-index fluctuations caused by atmospheric turbulence. It is well known that exposure times on a time scale where the turbulences are stationary (i.e. shorter than a tenth of a second) yield images that contain high-frequency information of the celestial object. This fact is exploited in Speckle Imaging, which is a collection of various techniques for recovering high frequency information encapsulated in short-time exposure images. Due to the stochastic nature of atmospheric turbulence, Speckle Imaging techniques have to take a large number of images into account to actually regain diffraction-limited information.

In our work we introduce an efficient online scheme that models the multi-frame deconvolution problem as an online learning problem. We assume that the images are obtained in a streaming fashion, and that they are potentially too numerous to save or process efficiently. While this may appear artificial at first glance, the rate of data acquisition in astronomy has been increasing tremendously over the years, to the point where data recorded in current sky surveys will take years to be analyzed.

Our approach has the following advantages:

- It has low resource requirements because the images are tackled in a streaming fashion; this obviates the need to store the images, and results in a more efficient algorithm (in terms of run-time and memory usage).
- Typical online methods usually require the setting of a learning rate. Our method is based on multiplicative updates to the image estimate that effectively bypasses the need to set a learning rate, which simplifies the algorithm and its implementation.
- Our method is very simple, and it could be implemented in the imaging hardware itself to provide dynamic deblurring, if desired.

First results described in [1] are depicted in the Figure which shows from left to right the evolution of our algorithm over its first 20 iterations. The top row shows the recorded images (the input to our algorithm), the middle row shows the current reconstruction (the output) and the bottom row the averaged input images (for comparison). Our reconstruction after 20 iterations has reduced the blur existing in all recorded images and shows two stars. In future work we will improve and refine our method by incorporating prior knowledge and by generalizing the image transformation model.

References

1. **Harmeling, S., M. Hirsch, S. Sra, B. Schölkopf:** Online Blind Image Deconvolution for Astronomy. *Proceedings of the IEEE International Conference on Computational Photography (ICCP 2009)* (accepted).



Temporal evolution of our algorithm applied to an image sequence of the double star system Epsilon Lyrae 1. From left to right the first 20 iterations are shown. In each column from top to bottom: original frame, estimated image, averaged image.