Péter Pázmány Catholic University
Faculty of Information Technology

University of Bordeaux 1
Laboratoire Bordelais de Recherche en Informatique

Video Event Detection and Visual Data Processing for Multimedia Applications

Theses of the Ph.D. Dissertation

Dániel Szolgay

Supervisors:
Tamás Szirányi, D.Sc.
Jenny Benois-Pineau, D.Sc.

Scientific adviser:
Tamás Roska, D.Sc.
ordinary member of the
Hungarian Academy of Sciences

Budapest, 2011

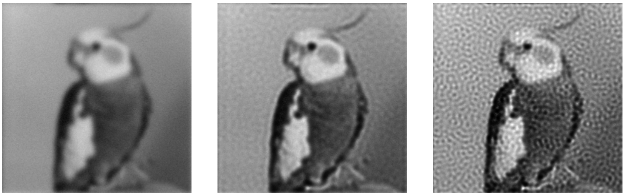
1 Introduction and aim

In the last 30 years image processing has become a mature engineering discipline and it has become an indispensable tool for many fields like medical visualization, law enforcement, human computer interaction, industrial inspection and security or medical surveillance.

The evolution of technology in the last decade opens up new possibilities, and the new possibilities set up new challenges. In the early days of digital image processing there were only digital images to process in a relatively low number. Around the turn of the millennium videos appeared and in parallel the constantly growing size of the image databases exceeded the manually manageable limit. New methods were required to handle the new challenges: content based image retrieval, video coding, event detection in videos have become part of digital image processing. Nowadays everyone can easily access digital cameras and make digital video recordings, hence the amount of video data is rapidly increasing. At the same time the type of video content has become more challenging, since generally neither the "cameraman" nor the device is professional. Blurry, noisy recordings with practically random camera motion need to be analyzed. Obviously to detect events in these kinds of recordings the whole process from low- to high-level has to be adapted to the task. This work is concerned with low- and mid-level image processing problems, that need to be solved to handle these new kinds of videos efficiently. The first two parts of the dissertation address basic image enhancement problems such as optimizing deconvolution for image deblurring, and extraction of the geometrical structure of the image by decomposing it into texture and geometrical components, while in the third part, higher level video understanding will be examined, where the task is the detection of moving objects and their separation from a cluttered background in videos recorded with a moving camera.

Image restoration is practically as old as image processing itself, constantly waiting for newer and better solutions. Deconvolution of blurred images, like the ones taken with strongly moving wearable cameras, gives a new motivation to solve an old challenge. Beside motion, there could be many other reasons of image blur like defocusing, atmospheric perturbations, optical aberrations. For these reasons, which are common in aerial, satellite or medical imaging, the acquired images are corrupted and restoration is needed. The distortion of the image is generally modeled as convolution: the original unknown image is convolved with a Point Spread Function (PSF) that describes the distortion. The goal is obvious: restore the original image as well as possible based on the blurry measurement and, in some cases, the PSF. The problem is ill-posed, since there is more than one image that would seem as a good solution. Hence it is a common drawback of non-regularized iterative deconvolution methods that after some iterations they start to amplify noise (see Fig. 1). Our goal was to automatically find an optimal stopping condition for these algorithms where the reconstructed image is as close to the original (unknown) image as possible.

The decomposition of an image into geometrical (cartoon) and noise like (texture) components is a fundamental task for both videos and still images. It can help image compression,

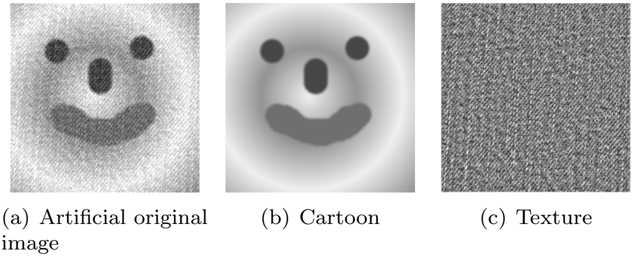


(a) Blurred Image (b) Deconvolved image after 14 iterations (c) Deconvolved image after 60 iterations

Figure 1: An example how non-regularized deconvolution methods amplify noise if not stopped at the optimal iteration.

denoising, image feature selection, or it can be a preprocessing step for video event detection: the same way as shadow, reflection, smoke/fog removal, the elimination of texture from the video frames aids the higher level understanding of the video. Theoretically the two parts are independent of each other: the cartoon image contains only geometrical information while the texture image, as complementary of the cartoon, is free of geometrical information (see an example on Fig. 2).

Separating foreground objects from the background is a fundamental module for many video applications, as it is commonly used to bootstrap higher-level analysis algorithms, such as object-of-interest detection, tracking, or content based video indexing, which could be applied for security or medical surveillance. The task is challenging for still camera recordings, but if wearable cameras are used, then strong motion and parallax,



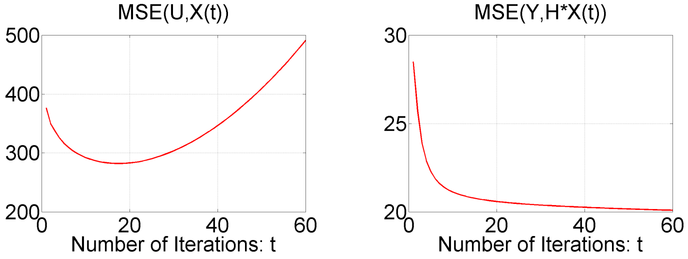
(a) Artificial original image (b) Cartoon (c) Texture

Figure 2: An example of cartoon/texture decomposition.

low quality of signal (reduced by motion blur) makes the problem even more complex. Generally low level algorithms (such as deblurring, denoising, morphological enhancement) are used as pre- or post-processing to achieve better result.

2 Methods used in the experiments

In the last decades a lot of methods were developed in order to restore the original image from the blurred, noisy measurement. We were working with an iterative non-regularized method called Richardson-Lucy algorithm [7, 8]. In case of deconvolution we do not know the original image U, only the blurry measured one Y can be used to guide us toward U. If $X(t)$ is the output of the method after t iteration, then a goal function of the method is usually based on minimizing $|Y - H * X(t)|$. Obviously the goal is to find the minimum of $|U - X(t)|$ or the similar $MSE(U, X(t))$ function, and stop the deconvolution there. However, these measures are not directly computable since U is unknown. We can only access $MSE(Y, H * X(t))$, which is not appropriate for the estimation of the ideal stopping point as it is clearly visible on Fig.3.



(a) The unknown $MSE(U, X(t))$ (b) The measurable $MSE(Y, H * X(t))$

Figure 3: The measureable Mean Square Error: $MSE(Y, H * X(t))$ function do not follow the unknown $MSE(U, X(t))$ function. Where U is the original image, $X(t)$ is the reconstructed image at iteration t and H is the PSF. Consequently, the ideal stopping point cannot be estimated based on $MSE(Y, H*X(t))$.

We have calculated the independence of $X(t)$ and the difference of two consecutive iteration $X(t) - X(t-1)$ using ADE [9] as a measure of independence. To prove the efficiency of the proposed method a test environment was built, where the U is known and the quality of the result is measurable.

We were also working on a special part of denoising where the aim is to separate the image into cartoon and texture components. A new method was developed, which uses the BLMV filter introduced in [10] to initialize an Anisotropic Diffusion (AD) filter [11]. The iterative AD is stopped automatically based on the orthogonality criterion of the cartoon and texture component [12], utilizing Angle Deviation Error (ADE) [9] measure. The algorithm was compared to the state-of-the-art methods using artificial images where the ground truth cartoon and texture components are available, which makes numerical evaluation possible, and also on non-artificial images for visual comparison, which is the most widely used evaluation method for cartoon/texture decomposition.

These algorithms and the evaluation environment was implemented in MATLAB® [13].

To detect camera independent motion I have built a framework with three main parts (see Fig. 4): (1) Motion Compensated Frame Differencing, (2) Estimation of Foreground Filter Model, (3) Detection of Moving Objects.

The compensation of camera motion is a must in step (1). It is done by Hierartchical Block Matching (HBM) [14] and a Global Motion Estimator (GME) [15].

After the compensation, the two frames have the same coordinate system and an error image can be calculated as a difference of compensated frames. This error image should contain only the foreground regions. To eliminate false positives we use Probability Density Function (PDF) estimation in the second step. The samples for this estimation are coming from the Modified Error Image (MEI), built in step (1). The MEI contains the information of the motion compensated difference of

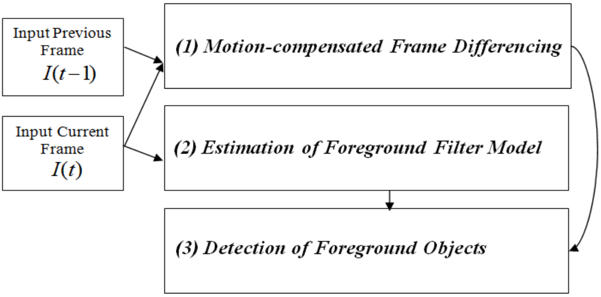

Input Previous Frame
$I(t-1)$
Input Current Frame
$I(t)$
(1) Motion-compensated Frame Differencing
(2) Estimation of Foreground Filter Model
(3) Detection of Foreground Objects

Figure 4: Diagram of the foreground object extraction method.

the actual and the previous frames and the color information of the actual frame.

For background color density estimation we used sample-point estimator, with Gaussian kernel.

The 3rd part is the detection of moving objects, which contains a thresholding of the previously built PDFs and the DB-SCAN [16] clustering algorithm, which works on the detected foreground pixels, using a 7 dimensional feature vector to describe a pixel.

The framework was implemented in C++ using the OpenCV library [17]. The algorithm was implemented for one thread, nevertheless a short research was done on the possibilities of parallelization on GPU.

3 New scientific results

1. Thesis: *The stopping condition is a common problem for the non-regularized iterative deconvolution methods. a novel method has been introduced for automatically estimating the stopping condition based on the orthogonality of the change of the estimated signal between two consecutive iterations and the signal itself. An effective lower bound estimate has been provided to the conventional ad-hoc methods and proved experimentally the efficiency of the proposed method for different noise models and a wide range of noise levels.*

The publications of the author connected to the thesis: [2,3].

Finding the optimal stopping point for iterative deconvolution methods is an ill-posed problem. In a real life problem scenario only the acquired image and the Point Spread Function (PSF) is available. In general, non-optimal ad-hoc methods are used to stop the iteration.

We have introduced a novel method for calculating the ideal stopping point for iterative non-regularized deconvolution processes, using the Angle Deviation Error (ADE) [9] measure instead of the commonly used Mean Square Error (MSE) measure.

The proposed method is capable of estimating the optimal stopping point of iterations based on the independence of an actual estimated signal and its gradient, indicating when an aimless section of the iterations is just starting, when the image is not enhanced anymore and only noise is added to it.

The proposed measure, $ADE(X_e(t), X(t))$ contains only measurable images and provides a reasonable solution for the stopping problem: at the minimum of $ADE(X_e(t), X(t))$ the change between two consecutive iterations $X_e(t)$ has the highest possible independence of the actual reconstructed image, hence we can assume that at this point $X_e(t)$ contains mostly independent noise and not structural information of the image, and further iteration will not enhance the image quality.

The method was tested with the well known Richardson-Lucy [7, 8] deconvolution algorithm with different noise models (Gaussian, Poisson) and wide range of noise levels. It does not require any input parameter or manual calibration. The correlation between the result of the theoretically best solution $(MSE(U, X(t)))$ and the result of the proposed method $(ADE(X_e, X_re))$ is 0.6726. If we regard the correlation not in iteration number but in image quality, the value is even higher: 0.9556. We can conclude that the proposed method outperforms the generally used ad-hoc methods.

2. Thesis: *A novel axiomatic method has been proposed for the automatic separation of geometrical and textural components of the image. The heart of the algorithm is the Anisotropic Diffusion (AD), whose iteration is stopped adaptively to the image content, based on ADE orthogonality measure. It has been proved experimentally that the proposed method separates cartoon and texture components of the image with better quality than the recently published algorithms.*

The publication of the author connected to the thesis: [4]

The aim of the Anisotropic Diffusion [11] is to blur and filter the image from noise while it keeps the strong edges. For this it uses a weight function, which hinders the diffusion in the directions orthogonal to edges and allows it along the edges or in edge-free territories.

AD, as proposed in [11], is not suitable for cartoon/texture decomposition, since texture may contain high magnitude edges, which should be blurred and cartoon may contain weaker edges, which should be preserved.

The proposed algorithm utilizes cartoon image of the BLMV non-linear filter [10] to initialize the weights for AD. In this image the textured regions are already blurred somewhat, hence the AD does not keeps them, while the main edges are preserved therefore the weighted inhibition of the AD will keep them untouched.

The iterative AD is stopped automatically based on the orthogonality of the two components using ADE measure. The proposed algorithm offers theoretically clear solutions for the main issues of the decomposition into cartoon and textured partitions:

- Adaptive scale definition by using locally optimal BLMV filter tuned by ADE measure;
- Anisotropic Diffusion, initialized by the new adaptive BLMV to better separate texture from cartoon;
- Orthogonality criterion for the quality measure of the decomposition (stopping condition to AD).

Our method was compared to the state-of-the-art methods of the field (TVL1 [18], ROF [19], DPCA [20], DOSV [21], AD [11] , BLMV [10]) using artificial images for numerical evaluation and real life images for visual comparison. The visual evaluation on real images is the most widely used method in spite of its subjectivity. Both evaluation approaches shows the proposed method superiority and contrary to the other algorithms it does not require precise manual tuning of the parameters, only a range of parameter values should be given as a starting condition.

3. Thesis: *Based on kernel density function estimation a novel method has been developed for moving foreground object extraction in sequences taken by a wearable camera (25fps, 320x240 frame size), with strong and unpredictable motion.*

The publications of the author connected to the thesis: [1,5]

Foreground extraction on wearable camera recordings is a challenging task since the camera motion is unpredictable and strong, and motion blur and intensive noise corrupt the quality of images.

Working with moving cameras the estimation and compensation of the camera motion is the first step towards moving foreground detection. We have performed a Hierarchical Block-Matching (HBM) [14] and affine Global Motion Estimation (GME) [15] to compensate camera motion.

After this step two consecutive frames of the video can be represented in the same coordinate system and the error image can be calculated as the absolute difference of the two frames.

This error image should contain high values only on the pixels corresponding to moving foreground objects, while the static background points should have low values after the difference calculation. Due to changes of the perspective, quantization errors and errors of the motion compensation the error image cannot be used as foreground model, because of the large number of false positives. A Modified Error Image (MEI) has been formed, which contains the color information of the original frame and the motion information of the error image.

To separate pixels of moving objects from pixels in static contours present in MEI due to the noise, Probability Density Function (PDF) estimation of the background and probabilistic decision rule is used.

The estimation of the PDF was done based on samples from a spatial-temporal patch with kernel density estimation [22], using Gaussian kernel. It is called spatiotemporal according to the choice of sample points: spatial neighborhood and temporal history of a pixel are both used.

For the bandwidth calculation we propose to use the distance from all the k nearest neighbors, instead of the distance from the kth alone, since the latter may give us false result when the number of sample points is strongly limited. In the given circumstances (low number of sample point, strong noise) the sample point selection technique has key importance.

A common approach for selecting the sample points in case of still cameras for a given (x, y) coordinate is to use the n previous measurements taken at the same (x, y) position [23]. When the camera is moving the case is different. Even after motion compensation the real background scene position that corresponds to the (x, y) pixel in one frame, might move a little, due to errors of camera motion compensation, or quantization. Assuming that this spatial error is random, the values selected in a small patch centered on the pixel (x, y) are used. Based on the values of the M measurement matrix, which contains the last n motion compensated frames, a joint PDF is built for the color channels of each non zero pixel of the current MEI.

Once the PDF has been built for each pixel in the current frame, we can proceed to the detection of foreground moving objects. Here the pixels will be first classified as foreground or background based on an adaptive threshold that considers the PDFs characteristics. Then the detected pixels will be grouped into clusters (moving objects) with DBSCAN algorithm [16] on the basis of their motion, color and spatial coordinates in the image plane.

It has been proved in an experimental way that the proposed framework gives better detection results than the widely used Stauffer-Grimmson method [24]. The calculations are done in offline mode at the moment, since the computational cost is too high for real-time processing.

4 Application of the results

Wearable video monitoring has a lot of potentials in the fields of health care, security and social life. It can be an important tool for diagnosing aged dementia, where the traditional methods may fail, since the patients cannot or voluntarily will not help the physicians to diagnose their disease. Using video logs about the life of the patients can help the doctors in their work. For security surveillance it can be an effective tool using together with static cameras or in cases when the use of static cameras is not an option (e.g. police patrols).

Blogging and life logging is becoming more and more popular. The author writes down his or her life like in a diary, but using the possibilities of the electronic world, uploading pictures, videos and music. A research project of Microsoft, called SenseCam [25] is helping the users to build a diary with photos using a special wearable camera that documents the users whole day with pictures. (This is a way of modern entertainment, but also it could be used in health care curing patients with memory disorders.) With wearable cameras and the necessary handling algorithms video diaries would also be available for the blogging society.

Our work in separation of the foreground and the background is just the first step toward content based search of videos, which is one of the most intensively researched areas of multimedia and computer vision.

The decomposition of an image into cartoon and texture components could be a starting point for many algorithms. It could be useful for image compression where compressing the cartoon and the texture components separately can provide better results [26]. Such a coding proved to be efficient in the past [27, 28]. It is applicable for image denoising [19] since zero mean oscillatory noise can be regarded as a fine texture, image feature selection [18] and main edge detection as illustrated in [10] etc. In motion estimation it could be used to eliminate the effect of noise, which may reduce the precision of the estimation.

Deconvolutional methods are widely used in image processing where defocusing is an issue: from microscopy to astronomy. It could be a preprocessing step for videos taken by wearable cameras, where motion blur corrupts the frames. Although nowadays regularization is the main trend, non-regularized methods are also capable producing results comparable to the state of the art [29]. For non-regularized methods the stopping problem is a key issue. The method we were working on offers a logically sound and effective solution to this problem.

Acknowledgements

First of all I would like to thank my supervisors Professor Tamás Szirányi and Professor Jenny Benois-Pineau, for their consistent help and support in many ways and their guidance during my studies.

The advices, help, and encouragement of Prof. Tamás Roska are kindly acknowledged.

I thank all my colleagues whose ideas and advices assisted me during my work.

The support of the Péter Pázmány Catholic University and the University of Bordeaux 1, where I spent my Ph.D. years, is gratefully acknowledged. My studies in Bordeaux were financed by the French Government through "Bourses Eiffel" and "Bourses pour doctorat en cotutelle".

I am very grateful to my mother and father and to my whole family who always encouraged me during the long years and

supported me in all possible ways.

Author's Publications

[1] D. Szolgay, J. Benois-Pineau, R. Megret, Y. Gaestel, and J.-F. Dartigues, "Detection of moving foreground objects in videos with strong camera motion," *Pattern Analysis and Applications.* accepted in 04.04.2011.

[2] D. Szolgay and T. Szirányi, "Orthogonality based stopping condition for iterative image deconvolution methods," in *Computer Vision ACCV 2010*, vol. 6495 of *Lecture Notes in Computer Science*, pp. 321–332, Springer Berlin / Heidelberg, 2011.

[3] D. Szolgay and T. Sziranyi, "Optimal stopping condition for iterative image deconvolution by new orthogonality criterion," *Electronics Letters*, vol. 47, no. 7, pp. 442–444, 2011.

[4] D. Szolgay and T. Sziranyi, "Adaptive image decomposition into cartoon and texture parts optimized by the orthogonality criterion," *IEEE Transactions on Image Processing.* Submitted in May 2011.

[5] R. Megret, D. Szolgay, J. Benois-Pineau, P. Joly, J. Pinquier, J.-F. Dartigues, and C. Helmer, "Wearable video monitoring of people with age dementia : Video indexing at the service of health care," in *International Workshop on Content-Based Multimedia Indexing, 2008.*, pp. 101 – 108, june 2008.

[6] D. Szolgay, C. Benedek, and T. Sziranyi, "Fast template matching for measuring visit frequencies of dynamic web advertisements," *Proceedings of VISAPP 2008, Third International Conference Computer on Vision Theory and Applications*, pp. 228–233, 2008.

References

[7] W. Richardson, "Bayesian-based iterative method of image restoration," *JOSA*, vol. 62, pp. 55–59, 1972.

[8] L. Lucy, "An iterative technique for rectification of observed distributions," *The Astronomical Journal*, vol. 79, pp. 745–765, 1974.

[9] L. Kovacs and T. Sziranyi, "Focus area extraction by blind deconvolution for defining regions of interest," *IEEE Tr. Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1080–1085, 2007.

[10] A. Buades, T. Le, J.-M. Morel, and L. Vese, "Fast cartoon + texture image filters," *IEEE Transactions on Image Processing*, vol. 19, no. 8, pp. 1978 –1986, 2010.

[11] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 629–639, 1990.

[12] J.-F. Aujol and G. Gilboa, "Constrained and snr-based solutions for tv-hilbert space image denoising," *J. Math. Imaging Vis.*, vol. 26, pp. 217–237, November 2006.

[13] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[14] M. Bierling, "Displacement estimation by hierarchical block matching," pp. 942–951, 1988.

[15] M. Durik and J. Benois-Pineau, "Robust motion characterisation for video indexing based on mpeg2 opticalflow," *In Proc. of the International Workshop on Content-Based Multimedia Indexing*, pp. 57–64, 2001.

[16] M. Ester, H. peter Kriegel, J. S, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," pp. 226–231, AAAI Press, 1996.

[17] G. Bradski and V. Pisarevsky, "Intel's computer vision library: applications in calibration, stereo segmentation, tracking, gesture, face and object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, 2000. Proceedings*, vol. 2, pp. 796 –797, 2000.

[18] W. Yin, D. Goldfarb, and S. Osher, "Image cartoon-texture decomposition and feature selection using the total variation regularized L1 functional," in *Variational, Geometric, and Level Set Methods in Computer Vision*, pp. 73–84, Springer, 2005.

[19] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, November 1992.

[20] F. Zhang, X. Ye, and W. Liu, "Image decomposition and texture segmentation via sparse representation," *Signal Processing Letters, IEEE*, vol. 15, pp. 641 –644, 2008.

[21] R. Shahidi and C. Moloney, "Decorrelating the structure and texture components of a variational decomposition model," *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 299 –309, 2009.

[22] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[23] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," vol. 2, pp. 302 –309, june-july 2004.

[24] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 747 –757, aug 2000.

[25] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood, "Sensecam: a retrospective memory aid," *International Conference on Ubiquitous Computing, LNCS 4206*, pp. 177–193, 2006.

[26] N. Sprljan, M. Mrak, and E. Izquierdo, “Image compression using a cartoon-texture decomposition technique,” *Proc. Int. Work. on Image Analysis for Multimedia Interactive Services (WIAMIS)*, p. 91, 2004.

[27] M. Kunt, A. Ikonomopoulos, and M. Kocher, “Second-generation image-coding techniques,” *Proceedings of the IEEE*, vol. 73, no. 4, pp. 549 – 574, 1985.

[28] D. Barba and J.-F. Bertrand, “Optimization of a monochrome picture coding scheme based on a two-component model,” in *9th International Conference on Pattern Recognition, 1988.*, pp. 618 –622 vol.1, nov 1988.

[29] S. C. L. Zou, H. Zhou and C. He, “Dual range deringing for non-blind image deconvolution,” *International Conference on Image Processing*, pp. 1701–1704, 2010.