

Omar Arif

Robust Target Localization and Segmentation

Application of Kernel-based statistical methods to
computer vision

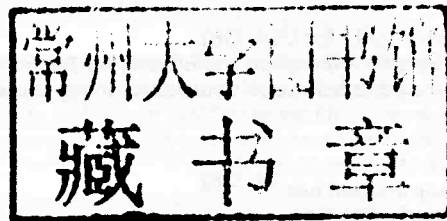


LAMBERT
Academic Publishing

Omar Arif

Robust Target Localization and Segmentation

Application of Kernel-based statistical
methods to computer vision



LAP LAMBERT Academic Publishing

Impressum/Imprint (nur für Deutschland/ only for Germany)

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Alle in diesem Buch genannten Marken und Produktnamen unterliegen warenzeichen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Coverbild: www.ingimage.com

Verlag: LAP LAMBERT Academic Publishing AG & Co. KG
Dudweiler Landstr. 99, 66123 Saarbrücken, Deutschland
Telefon +49 681 3720-310, Telefax +49 681 3720-3109
Email: info@lap-publishing.com

Herstellung in Deutschland:

Schaltungsdienst Lange o.H.G., Berlin

Books on Demand GmbH, Norderstedt

Reha GmbH, Saarbrücken

Amazon Distribution GmbH, Leipzig

ISBN: 978-3-8433-5038-9

Imprint (only for USA, GB)

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders.

The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this works is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Cover image: www.ingimage.com

Publisher: LAP LAMBERT Academic Publishing AG & Co. KG
Dudweiler Landstr. 99, 66123 Saarbrücken, Germany
Phone +49 681 3720-310, Fax +49 681 3720-3109
Email: info@lap-publishing.com

Printed in the U.S.A.

Printed in the U.K. by (see last page)

ISBN: 978-3-8433-5038-9

Copyright © 2010 by the author and LAP LAMBERT Academic Publishing AG & Co. KG and licensors

All rights reserved. Saarbrücken 2010

Omar Arif

Robust Target Localization and Segmentation

SUMMARY

This thesis aims to contribute to the area of visual tracking, which is the process of identifying an object of interest through a sequence of successive images. Some of the challenges associated with these tasks are image noise, occlusions, background clutter, complex object shapes, etc.

The work contained in this thesis explores kernel-based statistical methods. These methods map the data to a higher dimensional space where the tasks of classification and clustering are easily carried out. There are two problems related to the mapping: The out-of-sample and the pre-image problem. A pre-image framework for some of the manifold learning and dimensional reduction methods is developed.

Two algorithms are developed for visual tracking that are robust to noise and occlusions. In the first algorithm (Chapter 3), a KPCA-based eigenspace representation is used. The de-noising and clustering capabilities of the KPCA procedure lead to a robust algorithm. This framework is further extended in Chapter 6 to incorporate the background information in an energy based formulation, which is minimized using graph cut. Chapter 7 extends this framework to track multiple objects using a single learned model.

In the second method, a robust density comparison framework is developed (Chapter 5) that is applied to visual tracking (Chapter 8), where an object is tracked by minimizing the distance between a model distribution and given candidate distributions.

The superior performance of kernel-based algorithms comes at a price of increased storage and computational requirements. A novel method is proposed in Chapter 4, that takes advantage of the universal approximation capabilities of generalized radial basis function neural networks to reduce the computational and storage requirements for kernel-based methods. The ideas developed are general and are applicable to other kernel-based methods, such as KPCA and support vector machines.

TABLE OF CONTENTS

LIST OF TABLES	iv
LIST OF FIGURES	v
SUMMARY	viii
I INTRODUCTION AND BACKGROUND	1
1.1 Target Segmentation	3
1.1.1 Active Contour based	4
1.1.2 Graph cut	5
1.2 Target Localization	6
1.3 Organization and Contributions of the thesis	7
II KERNEL-BASED STATISTICAL METHODS	9
2.1 Mercer kernels	9
2.2 Kernel Principal Component Analysis	10
2.3 Contributions	13
2.4 Multi-dimensional Scaling	13
2.5 Isomap	14
2.6 Spectral Clustering and Diffusion Maps	14
2.7 Locally Linear Embedding	15
2.8 Kernelized Locally Linear Embedding	16
2.9 Experiments	17
III KPCA-BASED EIGENSPACE REPRESENTATION FOR TRACKING	21
3.1 Introduction	21
3.2 Related Work	21
3.3 KPCA-based Eigenspace Representation of the Target	24
3.3.1 Extracting Target Feature Vectors	24
3.3.2 Target Eigenspace Representation	24
3.3.3 Properties of the Eigenspace Representation	25
3.4 Similarity Measure in KPCA Space	26
3.4.1 Feature Vector Similarity Function	27

3.4.2	Region similarity measure	29
3.4.3	Evaluation of Region Similarity Function on a 1D Synthetic Signal Detection	29
3.5	Object Tracking	32
3.5.1	Variational Target Localization	32
3.5.2	Segmentation by Thresholding	33
3.6	Experiments	34
3.6.1	Target Localization and Segmentation	34
3.6.2	Target Localization	38
3.6.3	Target Localization on IR Sequences	41
3.6.4	Orientation Tracking	41
3.7	Conclusion	45
IV	KERNEL MAP COMPRESSION	46
4.1	Introduction	46
4.2	Contribution	47
4.3	Existing Efforts	47
4.4	Kernel Map Compression	49
4.4.1	Setup	50
4.4.2	Procedure	51
4.4.3	Choosing Initial center locations.	52
4.4.4	Approximating Several Functions at Once	53
4.4.5	Achieving Further Compression	54
4.4.6	Computational Cost	54
4.4.7	Pre-image Computations	55
4.5	Application	55
4.5.1	Synthetic Datasets	55
4.5.2	Speeding up Support Vector Machines	58
4.5.3	Efficient KPCA-based Gesture and Face Recognition	61
4.6	Conclusion	64
V	ROBUST DENSITY COMPARISON	65
5.1	Introduction	65

5.2	Contribution	65
5.3	Maximum Mean Discrepancy	66
5.4	Robust Maximum Mean Discrepancy	66
5.4.1	Robust Density Function	67
5.4.2	Robust maximum mean discrepancy	68
5.4.3	Summary	68
5.4.4	Toy Example	69
5.5	Conclusion	70
VI	KPCA-BASED ENERGY FOR GRAPH CUT	71
6.1	Introduction	71
6.2	Proposed Algorithm	72
6.2.1	Joint Appearance-Spatial Target and Background Model	72
6.2.2	Energy Formulation	73
6.3	Summary of the Procedure	74
6.4	Results	74
VII	TRACKING MULTIPLE OBJECTS	78
7.1	Parameterized Appearance Function	78
7.2	Overall Training and Tracking Procedure	79
7.3	Experimental Results of Tracking Multiple Personnel	80
VIII	LOCALIZATION THROUGH DENSITY COMPARISON	83
8.1	Variational Target Localization	83
8.2	Results	84
IX	CONCLUSION	88
	APPENDIX A MEAN SHIFT	90
	REFERENCES	91

LIST OF TABLES

1	Average PSNR(db) over ten digits	18
2	Tracking Results: ¹ Number of reinitialization required to complete the tracking. \times indicates more than 5. ² Ratio of correct number of estimation to the total number of frames. \times indicates less than 70%.	40
3	Performance comparison of approximation methods for approximating the sinc curve. method - p means that the corresponding method used p support vectors or reduced data points to approximate the sinc curve. The proposed method (KMC) is able to reach the performance of standard SVM by using just 7 data points (compression ratio of 24) with degradation of less than 5 %	57
4	Character recognition database. Top: number of SVs for the original SVM and the number of test errors for each classifier. Bottom: number of test errors for each reduced classifier. Method - $p\%$ means that for each classifier, the space was reduced to $p\%$ of the original space. Second to last column shows error rate across all 26 classifiers. The last column shows the % degradation.	60
5	USPS handwritten digit. Top: number of SVs for the original SV and the number of test errors for each classifier. Bottom: number of test errors for each reduced classifier. KMC- p and Burges- p mean that for each classifier, the space was reduced to p points. The last three columns show error rate across all classifiers, compression ratio which is the ratio of the full space to the reduced space and % degradation.	62
6	Sign language recognition: ¹ success rate of identifying all the 2040 images; ² success rate for test cases. KMC- p and Burges- p mean that for each eigenvector the space was reduced to p points using the corresponding method. Second to last column shows compression ratio, while the last column shows the percentage degradation when testing all images.	63
7	Face recognition: ¹ success rate of identifying all the 400 images. ² success rate for 100 test cases. KMC- p mean that for each eigenvector the space was reduced to p GRBFs. Second to last column shows compression ratio, while the last column shows the percentage degradation when testing all images.	64
8	Results: Error is estimating the location of the target. X indicate the tracker lost track within 100 frames. Striked out numbers indicate the tracker lost track after 100 frames.	75
9	Tracking sequence	85

LIST OF FIGURES

1	Segmentation: without shape prior, with shape prior, shape prior and occlusion, shape prior and occlusion with manual localization (from left to right).	2
2	Tracking with and without Kalman filtering	3
3	Outline of the thesis	7
4	Toy example: Dot product in the mapped space can be computed using the kernel in the input space.	10
5	KPCA eigenspace representation. All points vectors in the input space are mapped nonlinearly to a Hilbert space where eigenvalue decomposition results in an m -dimensional KPCA space.	11
6	Pre-image computation for different methods	17
7	Image de-noising, Gaussian noise level $\sigma = .5$	19
8	Image de-noising, Gaussian noise level $\sigma = 1$	19
9	Pre-image computations. The pre-images without the red bounding box are the ones whose embedding was approximated using interpolation in the embedding space.	20
10	KPCA vs PCA: KPCA has a larger theoretical number of retainable eigenvectors versus PCA, therefore KPCA can capture more clusters/features.	26
11	Noise/outlier rejection: Top row shows that the projections onto the leading eigenvectors remain the same as in Figure 10. The reconstructions in the bottom row show that the eighth and beyond eigenvectors capture noise.	27
12	The target object color values lie on the surface of the ellipse in KPCA space, while other color values lie interior to the ellipse. The square distance from the origin can be used as a similarity function.	28
13	Region similarity measure for the first three eigenvectors. For each eigenvector, the region similarity peaks at the location of the target object. A target can be robustly located even under occlusions, as the eigenvectors corresponding to visible portions will score higher and thus have more influence.	30
14	Template signal (red) overlayed on a sample corrupted signal (blue).	31
15	ROC curves for Gaussian and Log-Normal noise.	31
16	ROC curve for various occlusion levels.	31
17	The similarity measure represents an unnormalized density estimate. Mean-shift can be used to find the mode of the density function.	33
18	Segmentation by thresholding	34

19	Sequence 1: Video with artificial occlusions. Eigenvectors corresponding to the visible parts are used to track through the occlusion.	36
20	Sequence 2. Eigenvectors corresponding to the visible parts are used to track through the occlusion.	36
21	Sequence 3. Tracking and segmenting a small object in cluttered environment with illumination changes.	37
22	Sequence 6 and 7. Tracking construction workers in cluttered environment with occlusions.	37
23	Tracked objects from test sequences.	39
24	Matched object regions for a target in the sequence Multiperson A using the proposed, ensemble, covariance and mean-shift trackers (top to bottom). The track point is more stable in the proposed methods.	40
25	In the Caviar data set, the target object is occluded by people coming from the opposite direction. In PETS 2009 data set, the target object is successfully tracked in a crowded scene.	41
26	Tracking of visible-infrared sequence. The feature vectors per pixel are formed using the color, infrared, and spatial values, i.e., $u = [I(x), IR(x), x]$	42
27	Beach ball orientation tracking. White line with circle indicates the orientation.	42
28	Tracking of a fish sequence.	43
29	Ground truth comparison for fish sequence. Red: Ground truth, Blue: Proposed tracker.	43
30	Tracking of a worm under microscope.	44
31	Ground truth comparison for worm sequence. Red: Ground truth, Blue: Proposed tracker.	44
32	Kernel map compression (KMC). The red arrow shows the proposed approach to approximate the relationship between the input space and the feature subspace directly, circumventing the feature space.	50
33	Sinc curve synthetic experiment for SVM regression.	56
34	Performance comparison for the synthetic 2D example for KPCA.	58
35	Performance comparison for the synthetic 2D example for KPCA.	58
36	Samples for testing reduced SVM.	59
37	Character recognition: % degradation vs. compression curves	61
38	Character recognition: Performance comparison for the case of reducing the space to 20% of the original space	61
39	Performance comparison for the USPS data set.	62
40	Samples for testing reduce KPCA	63

41	Non-parametric density estimation of multi-modal, noisy Gaussian distribution.	68
42	MMD vs robust MMD.	69
43	Illustration of the effect of noise on the difference between the the two distributions. The samples from the two distributions are shown in red and blue.	69
44	Inadequacy of thresholding for segmentation.	71
45	Sequence 1: Pink line indicates the trajectory created by the proposed tracker and the blue diamonds indicate the location of snapshots in (b).	76
46	Sequence 2 and 3. The proposed tracker is not confused by the same distribution of the pants color of the two people being tracked.	76
47	Sequence 4. The background color distribution is similar to the color distribution of the shirt of the target.	77
48	Sequence 5.	77
49	Three spatially similar targets.	79
50	Parameterized appearance function to generate feature vectors. The person in Figure 49(a) is modeled by providing few templates (a), from which the appearance information is automatically extracted in (b) using statistical analysis. From the statistical appearance analysis, a spatial segmentation model is generated (c). The spatial model is used to identify the appearance function parameters.	80
51	Sequence 1: Tracking of three people. Segmentation results are not shown in the middle frame for better visualization of target objects	81
52	Sequence 2: Tracking of two people	82
53	Sequence 3: Tracking of three people	82
54	Sequence 4: Tracking of three people	82
55	Construction Sequence. Trajectories of the track points are shown. Red: No noise added, Green: $\sigma = .1$, Blue: $\sigma = .2$, Black: $\sigma = .3$	85
56	Face sequence. Montages of extracted results from 90 consecutive frames for different noise levels.	86
57	Fish Sequence.	87
58	Jogging sequence.	87

CHAPTER I

INTRODUCTION AND BACKGROUND

Computer vision aims to artificially replicate the human visual perceptions. It deals with the science and technology of processes related to the acquisition of images, analysis of images and sequence of images to extract useful information, and to the development of artificial cognitive systems that “see.”

Image segmentation and visual tracking are two important components of computer vision. The former deals with an image, while the latter is concerned with the sequence of images. Segmentation aims to partition an image into smaller more meaningful parts, which are related with respect to some common aspect. Visual tracking, on the other hand, is the process of locating an object of interest in a sequence of successive images. For a deformable object, the motion of the object, over a sequence of images, is described by an overall global motion (pose), and the local deformation of the object [94]. In this respect, at each frame of the sequence, segmentation is required to delineate the target from the background. Tracking a deformable object is therefore, the process of estimating the pose parameters (*target localization*) and the local deformation (*target segmentation*) of the object. In this sense, visual tracking encompasses segmentation. Algorithms differ in whether they perform localization only or both i.e. localization and segmentation. The former can be called *transformation based* and the latter *contour based*. Objects can be represented by their appearances such as color, texture, edges etc, which provide characteristic information about the target object. This characteristic information is encoded into a cost or similarity functional. Tracking algorithms then find the target object that is optimum with respect to a pre-determined similarity functional.

Visual tracking is a challenging task. In some cases, the image information (appearance) may not be sufficient enough to identify the target object. This may happen due to a number of reasons, such as noise, occlusions, complex object shapes and presence of target

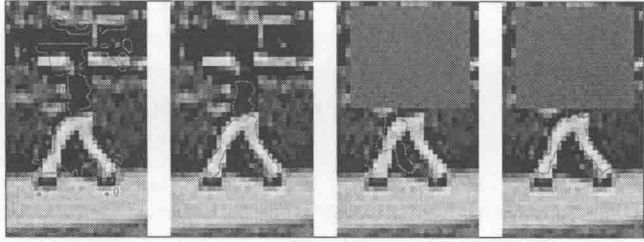


Figure 1: Segmentation: without shape prior, with shape prior, shape prior and occlusion, shape prior and occlusion with manual localization (from left to right).

features in the background etc. In such cases, visual tracking can be simplified by imposing a shape/spatial constraint on the tracking process. These constraints can come in the form of smoothness priors on the contour that segments the object from the background, or in the form of incorporating prior known information on the shape of the object to be segmented. For example, in the first image in Figure 1, the target object is segmented without any shape prior. The segmentation is noisy, due to the presence of the target features (color) in the background. The second image produces a correct segmentation by incorporating shape priors. However, the addition of shape information requires extra effort on the part of tracker. The prior shapes need to be aligned with the pose/location of the object, which is being tracked, for correct segmentation. This is not trivial since the pose is not known and has to be estimated along with the segmentation. If the pose can not be adequately estimated, then the tracking results are meaningless. The effect of incorrect target localization on target segmentation is shown in Figure 1-third image. In the last image, the target is localized manually, which results in correct segmentation. So correct target localization is critical to shape based trackers.

Apart from imposing constraints on the shape of the object, tracking can also be simplified by imposing constraints on object motion. The object motion can be constrained to be constant velocity or constant acceleration based on prior information [110]. Figure 2 shows an example of successful tracking using a constant velocity motion constraint. These methodologies form the basis of well known tracking algorithms like Kalman filter

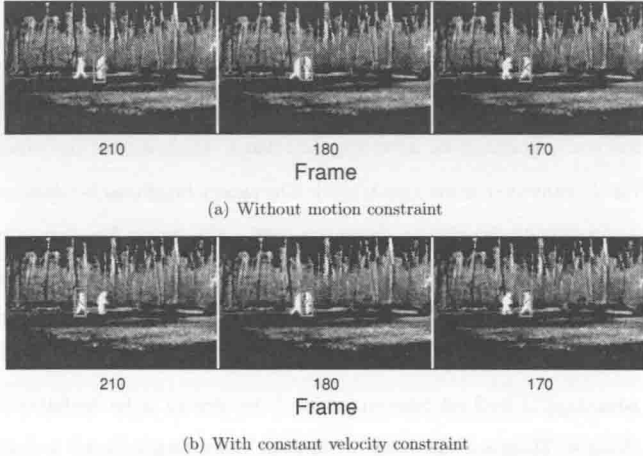


Figure 2: Tracking with and without Kalman filtering

[50] and particle filter [44]. These techniques have been used in tracking algorithms such as [18, 28, 49, 101]. Filtering techniques will not be pursued in this research.

The computation complexity of visual tracking algorithms is an important issue. Tracking algorithms should be able to work in real time.

This thesis aims to contribute to the above mentioned challenges and requirements for visual tracking. It does this by developing novel algorithms for target localization and segmentation, which are robust to noise and massive occlusions. The algorithms are based on kernel-based statistical methods. Also, a novel method is provided to reduce the computation complexity of the algorithms.

Below, target segmentation and localization, as carried out by different methods, is explained briefly. We will be mostly concerned with how shape information is incorporated and used to perform localization and segmentation.

1.1 Target Segmentation

The aim of target segmentation is to delineate the object from its background. Several algorithms and techniques have been developed for segmentation. The more recent ones are discussed below.

1.1.1 Active Contour based

The basic idea in active contour based segmentation algorithms is to evolve a closed contour, starting from an initial estimate, such that it encircles the object region. Evolution of the contour is governed by an energy functional, which defines the fitness of the contour to the hypothesized object region [110]. The energy functional is minimized when the contour delineates the object from the background. The energy functional can be based on local information such as the image gradient [52, 54], global features such as color [41, 109, 114], or mean image intensity inside and outside of the evolving contour [25]. Level set methods [89] have been successfully used for implicit representation of the contour. The most important advantage of level set representation of the contour is its flexibility in allowing topology changes. There is a vast body of literature concerning level sets and active contour, see for example [68, 69, 70, 78].

More recently, shape information is incorporated into the active contour framework to make the segmentation robust to noise and occlusion [33, 34, 35, 40, 60, 79, 98]. Leventon et al. [60] define the shape term E_{shape} using a probabilistic approach based on principal component analysis (PCA). The set of training shapes are represented by their signed distance functions [89], and PCA is applied to obtain a reduced representation. A probability density function is defined over the parameters of the reduced representation to obtain the shape energy. The level set function is evolved using both the image and the shape terms, which draw the level set function towards the most probable shape according to the learned distribution. Tsai et al. [98] incorporate the shape model, derived also by performing PCA on a collection of signed distance maps of the training shapes, into region-based active contour ([25]). The problem is reformulated to directly optimize the parameters associated with the first few eigenvectors. Cremers et al. [33] use kernel density estimator to define the probability density on the space of signed distance function representing the prior shapes. They show that this approach captures nonlinear shape variability. Freedman et al. [40] track by combining density matching and shape priors. For density matching, Bhattacharyya measure is used to define the distance between a model intensity distribution and the intensity distribution of an estimated image region. The tracker is expressed as a PDE-based curve

evolution, which is implemented using level sets. Dambreville et al. [35] combine intensity based segmentation with prior shape knowledge learned using Kernel PCA (KPCA). A binary representation of shapes is used. KPCA is shown to outperform linear PCA, by allowing only shapes that are close enough to the training data. Dynamical shape priors have also been used to improve the tracking of deformable objects in the presence of noise and occlusions. [30, 77].

1.1.2 Graph cut

Image segmentation can also be formulated as a graph partitioning problem [23]. Let \mathcal{R} be the set of all pixels in the image, and let $\mathcal{L} = \{0, 1\}$ be a label assignments on \mathcal{R} . The label 1 means the pixel belongs to the target, while the label 0 means it belongs to the background. The segmentation problem is cast as that of finding a labeling $l : \mathcal{R} \rightarrow \mathcal{L}$, minimizing an energy $E(l)$, modeled by:

$$E(l) = E_d(\mathcal{I}, l) + E_s(l). \quad (1)$$

E_d is the data term which measures how well the labeled pixels fit the image model. The standard data term is

$$E_d(\mathcal{I}, l) = \sum_{u \in \mathcal{R}} F_u(l_u), \quad (2)$$

where F_u measures how well label l_u fits pixel u . To realize the energy on the graph, each pixel is considered a node of the graph with two additional terminal nodes, the target and the background terminal nodes. To encode the data term on the graph, each pixel is connected to the target and background terminal nodes with edge weights $F_u(1)$ and $F_u(0)$, representing the cost of assigning a pixel to target and background respectively. E_s is the regularization or boundary term. Let \mathcal{N} be a neighborhood system on \mathcal{R} , then E_s is realized by connecting each pair of pixels $(u, v) \in \mathcal{N}$ with a non-negative edge weight measuring the penalty for assigning two neighboring pixels to different regions. The mincut of the graph represents the segmentation that best separates the target from its background and minimizes the energy $E(l)$.

Prior shape information can be encoded in the graph cut by either imposing it on the edges between pixels and the terminals nodes [63, 105], or by defining the edge weights