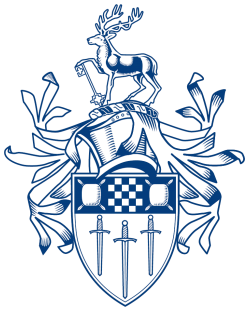
Pose-Invariant 2D Face Recognition by Matching Using Graphical Models

Shervin Rahimzadeh Arashloo

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey



Center for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

September 2010

Summary

The thesis presents a 2D face recognition system using Markov random field matching methodology for establishing dense correspondences between a pair of images in the presence of pose changes and self-occlusion. The proposed method, which exploits both shape and texture differences between images, achieves very competitive performance compared to the current approaches. The algorithm bypasses the need for geometric pre-processing of face images. By virtue of the matching methodology embedded in the algorithm, the proposed approach can cope with moderate translation, in and out of plane rotation, scaling and perspective effects. Also by employing a graphical model based approach, the proposed system circumvents the need for non-frontal images being available for training a pose-invariant face recognition system. In contrast to the state-of-the-art approaches based on 3D models, the approach operates on 2D images and bypasses the need for 3D face training data and avoids the vagaries of 3D face model to 2D face image fitting.

From the point of view of object recognition based on graphical models, the matching energy in graph based approaches is shown to exhibit certain drawbacks and should not be used as a similarity criterion for the hypothesis selection directly. The main shortcomings of the energy functional (using at most pairwise potentials) are identified and a plausible energy normalization scheme is proposed.

In order to reduce the computational burden of the inference in the model, two multi-scale processing approaches are proposed. One employs the super-coupling transform in order to solve the matching problem in a multiresolution fashion. The other is heuristic but surprisingly leads to good results.

Last but not least, a sparse graphical model for facial feature localization is proposed. The method takes advantage of the sparsity of facial image features in order to speed-up the matching process. The conditional dependencies between different groups of image primitives are included as higher order interactions based on point distribution models and linearity-based priors. The sparse model has been successfully applied to the task of facial feature localization and also as an initialization step to speed-up inference in a more costly matching approach.

Key words: Face Recognition, Pose-invariance, Markov Random Fields, Image Matching, Object Recognition.

Email: S.Rahimzadeharashloo@surrey.ac.uk

WWW: http://www.ee.surrey.ac.uk/CVSSP/

Acknowledgements

I would like to thank my supervisor Prof. Josef Kittler for his invaluble guidance and constructive comments at any time during the course of this work. Special thanks are also due to Dr. William Christmas for his support and help in using the RAVL C++ library.

I would also like to thank my family for all their support throughout the 3 years of duration of this work and also all my friends in Guildford.

Acronyms

Acronym	Meaning
2D	Two-Dimensional
3D	Three-Dimensional
AAMs	Active Appearance Models
BP	Belief propagation
DoG	Difference of Guassian
EBGM	Elastic Bunch Graph Matching
EER	Equal Error Rate
FA	False Acceptance
FR	False Rejection
GPU	Graphical Processing Unit
HTER	Half Total Error Rate
ICM	Iterated Conditional Mode
LB	Lower Bound
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LLDA	Locally Linear Discriminant Analysis
LLR	Locally Linear Regression
MAP	Maximum A Posteriori
ML	Maximum Likelihood
MRF	Markov Random Field
PDM	Point Distribution Model
PCA	Principal Component Analysis
RANSAC	RAndom SAmple Consensus
RGT	Renormalization Group Transform
RL	Relaxation Labeling
ROC	Receiver Operating Characteristic
SIFT	Sacle Invariant Feature Transform
TRW-S	Sequential Tree ReWeighted
TER	Total Error Rate
ULBP	Uniform Local Binary Pattern

Mathematical Notation

Symbol	Meaning
x_E	State of the edge/hyperedge E
x	Joint state of all variables in a graphical model
$\mathbb{R}$	The set of real numbers
$\mathscr{V}$	The node-set of a graphical model
$\mathscr{E}$	The edge-set of a graphical model
$P(x)$	Probability of the joint state x in a graphical model
Ψ_E	Factor E of a graphical model
β	inverse of the temperature in a Gibbs distribution
$\mathscr{G}$	A probabilistic graphial model
∂A	Markov Blanket of the set A of variables
$\|E\|$	Cardinality of the set E
exp	exponential
A^T	Transpose of the vector/matrix A
En	Energy of a probability
ϕ	Sufficient statisttics
Ω	Cumulant function
$\mathscr{I}$	Index set of sufficient statistics
$\mathbb{X}_u$	Domain of variable u
$[\![.]\!]$	Iverson brackets
θ	Parameters of an exponential probability distribution
$\mathbb{E}_P[.]$	Expectation with respect to the probability distribution P
$\mathbb{M}(\mathscr{G})$	Marginal polytope of the graphical model $\mathscr{G}$
$\mathbb{L}(\mathscr{G})$	Pseudo-marginal polytope of the graphical model $\mathscr{G}$
μ	A vector included in the marginal polytope
τ	A vector included in the pseudo-marginal polytope
arg max	Maximizers of a function
w	A ponit in the PCA subspace
ρ	Probability distribution defining a convex combination
$\equiv$	Reparameterization via message passing
Φ_E	Min-marginal associated with the hyperedge E
Tr	Spatial transformation
σ	standard deviation
tanh	Hyperbolic tangent
χ^2	Chi-squared distance
α	Weight parameter for shape and texture scores fusion

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An ever-increasing demand for face recognition technology in various domains has stimulated lots of intensive research in this field in the past couple of decades. A primary area of interest is security and surveillance to which face recognition has contributed a lot for its superior adaptability to the requirements of real world applications. Among the appealing characteristics of face recognition technology are its capability to operate at a distance and without prior knowledge or permission of the subject. This technology has also stepped into everyday life by providing commercial applications for online image search and tagging, analysis of personal photos *etc*. Considering the span of application domains, more research is required to enhance current methods or design new ones to deal with emerging new and more demanding scenarios.

1.1 Face Recognition Systems

Two different scenarios are considered for face recognition: verification and identification. In a verification scenario, the system confirms a person's identity by examining the similarities of the captured image with the templates corresponding to the claimed identities, which are stored in the system. As a result, a one to one comparison is required in order to decide whether the person presenting herself/himself to the system is the person she/he claims to be. In contrast, in an identification scenario, the system identifies a person by checking the whole database for the closest match resulting in a one to many search, producing a similarity score for each comparison being made. In order for the system not to get confused between genuine claims (clients)

and incorrect claims (imposters) the similarity score should exceed a certain threshold. The system thus will either find a match and identify the individual or it will flag the biometric test as ”unknown subject”. More specifically, in a verification task, the subject claims an identity by different means such as passwords, personal identification numbers (pin codes), signatures and documents and after collecting his image the system compares the person with the claimed identity, whereas in an identification scenario, the subject does not claim any identity and the system compares the characteristics of the subject's image with the entire database to give or deny permission for a specific task.

1.2 Challenges in Face Recognition

In spite of the success of face recognition systems in controlled conditions, the performance of these systems is not satisfactory in realistic situations and under degraded conditions such as viewpoint, illumination or expression changes or occlusion, disruption due to accessories, ageing and *etc.* The reason for this is that the between-class separability is eroded by the within-class variance attributed to changes in image capture conditions. Although the foregoing problems almost affect most object recognition systems, the situation is particularly challenging for faces as nonrigid objects with a limited number of samples and high dimensionality of the face image data. This means that if a system has already seen the clients under all possible conditions (pose and illumination changes, occlusion, *etc.*) then recognition under all conditions would be much easier. However, clearly this is not feasible almost in any scenario.

Among the various factors which adversely affect the performance, the two most challenging problems for a face recognition system are pose and illumination changes [147]. These variations are unavoidable when face images are captured in uncontrolled and non-cooperative situations such as surveillance videos or everyday personal photos.

In the case of pose changes, the following factors are responsible for the degradation of the performance: first, the misalignment remaining even after aligning images using eye coordinates. This is in contrast to the frontal pose where aligning the images using only eye coordinates is sufficient to achieve very high performance in practice. Second, the non-planarity of the face causes two side-effects: partial self-occlusion which basically makes parts of the useful

information unaccessible and second, non-linear and complicated changes in the appearance of visible parts. When the effects of uneven illumination conditions further aggravate the situation, algorithms designed for frontal case fail.

The challenging problem of recognition of faces in the presence of pose variation has attracted lots of research and various methods have been proposed for this problem. However, not many realistic and real-world-applicable methods exist. One appealing characteristic of a recognition system is the minimum requirement for training data. This property is advantageous both in terms of training time and generalization capacity of the algorithm. This characteristic is not very common to many available automatic face recognition algorithms as most of these approaches use large training data corresponding to different poses. A problem that may occur in case of using limited training samples and complex models is that of over-fitting. Most of the time, the number of training images used to construct a model is small compared to the dimensionality of facial images. As a result, employing complex classifiers for recognition might be at the risk of over-fitting to the training set which subsequently may result in decreased performance in cross database validations.

A closely related problem in some of these methods is caused by adopting a generative approach. This means inferring new data in desired conditions based on a set of observations. More specifically, in the case of pose-invariant face recognition one may use training data to infer the frontal pose of a non-frontal test image to be subsequently used in a frontal face recognition system. Inference based on training data imposes a limit on the reconstruction accuracy. As a matter of fact, atypical features in a test image cannot be recovered in an analysis-through-synthesis generative framework. Subsequently, a degradation in recognition performance results. Reconstructing new images in desired poses can be avoided by using only the visible parts of the face image, which can be facilitated by establishing dense correspondences between a gallery and visible parts of a test image. In such an approach, for example, in the case of pan movement of head, by exploiting the face symmetry, only half of the face closer to the camera (which is believed to be visible) may be used for recognition.

A further issue is related to the geometric alignment of images. In recognition in frontal pose, faces are treated as 2D objects. As a result, any geometric misalignment can be corrected using only 2 or 3 points. The detection of these points imposes another restriction on auto-

matic face recognition. The limitation, stemming from the need for at least two fiducial points corresponding to eye coordinates in frontal pose, becomes more severe in non-frontal poses, requiring more landmark points to align images. This is a result of non-planarity of the face which makes it hard to be characterized accurately in any framework. In spite of the advances in automatic detection and localization of facial features, the problem is not completely solved. The challenge is obvious even for face recognition in frontal pose (a problem not considered very difficult in the field) where manual annotation of images results in superior performance compared to the case where eye coordinates are detected automatically. Consider a case in more challenging conditions: subject blinks, pose of the head changes, lighting varies unevenly and *etc.* These problems have forced many algorithms to consider the recognition problem at a higher level *i.e.* assume landmark points are detected, and then simply focus on finding suitable representations/features for face images for the final classification stage. As a result, the availability of annotated data is taken for granted and manual annotation information is widely employed in such methods. Currently, there are methods for pose-invariant recognition of faces which assume 100 landmark points are available prior to recognition. Clearly, these assumptions are far from reality. However, as having correspondence information between images can lead to improvements in performance, in extreme cases, one may wish pixel-wise correspondences between images being available for recognition. The problem is that in real-world scenarios such information will not be available to the system and it will be unavoidable to obtain it *automatically*.

A suitable tool to address the aforementioned limitations and problems to some extent is the undirected graphical models for object recognition also known as Markov random fields (MRF). Such methods handle variations of object appearance by virtue of part-based representation and by considering inter-relationships between parts. They also require a minimum number of gallery images and can be used with only one template per class. Furthermore, in order to find landmark points automatically, one can take advantage of such methods. As a result, they possess very appealing properties and are ideally suited as a basis for the design of face recognition algorithms to operate in a pose-invariant recognition scenario.

The recognition of objects in this framework is very often formulated as minimization of a cost/energy functional encoding the deviations from a particular model of interest. The deviations for example may correspond to textural content, shape of objects *etc.* Then the goal is to

find the solution corresponding to the minimum energy. A common drawback of these methods is the computational complexity of the energy minimization task. However, as denoted earlier and shown in the following chapters, one can take advantage of multi-resolution analysis or exploit the sparseness of facial features to reduce the computational complexity. Closely related to the complexity problems in this framework is the incorporation of higher order priors of an object into the energy. By prior we mean a source of information that is not necessarily available in a single test image under analysis but is *assumed* to be a characteristic of a class of objects. The parameters of such characterization may be estimated by statistical analysis of a group of objects belonging to the same category. The incorporation of such sources of information in a graphical model-based representation has been challenged by curse of dimensionality. That is, minimizing the cost function becomes very inefficient as soon as more and more complicated relationships between different parts of an object are included as deviation factors into the cost function. Recent advances in the field of MRF optimization have addressed this issue to some extent and provided more efficient ways for the minimization of energy in the presence of some specific higher order prior information.

1.3 Contributions

Appealing to the success of part-based approaches to object recognition which exploit contextual relationships between object primitives in recognition problems, we investigate the applicability of the idea to face recognition. The contributions of the thesis can be summarized as follows. From a high level point of view, the thesis proposes a face recognition method with the following characteristics:

- A face recognition system using MRF matching methodology for establishing dense correspondences between facial images in the presence of pose changes and self-occlusion. The method takes into account both shape and texture differences between images and achieves very competitive performance to the current approaches.
- The method circumvents the need for geometric pre-processing of face images. As a result, the proposed approach can cope with moderate translation, in and out of plane rotation, scaling and perspective effects.

- By employing an MRF based approach, the proposed system circumvents the need for non-frontal images for training a pose-invariant face recognition system.
- The method alleviates the self-occlusion problem by exploiting facial symmetry and using half-face information for recognition in the case of pan movement of head.
- In contrast to the state-of-the-art approaches based on 3D models, the approach operates on 2D images and bypasses the need for 3D face training data and the vagaries of 3D face model to 2D face image fitting.

In the context of MRF-based matching and recognition, our contributions can be summarized as below:

- From the point of view of object recognition based on MRF models, the matching energy in graph-based approaches is shown to exhibit certain drawbacks and should not be used as a similarity criterion for the hypothesis selection directly. The main shortcomings of the energy functional (using at most pairwise potentials) are identified and a plausible energy normalization scheme is proposed.
- Since one of the bottlenecks of MRF-based approaches is the computational complexity of the optimization algorithms, two multi-scale processing approaches are proposed. One employs the super-coupling transform in order to solve the matching problem in a multi-resolution fashion. The other is heuristic but surprisingly leads to good results.
- Last but not least, a sparse graphical model for facial feature localization is proposed. The method takes advantage of the sparsity of facial image features in order to speed-up the matching process. The conditional dependencies between different groups of image primitives are included as higher order interactions based on point distribution models. The sparse model has been successfully applied to the task of facial feature localization and also as an initialization step to speed-up inference in a more costly matching approach.

1.4 Organization

Chapter 2: Related Work

We begin in Chapter 2 by reviewing the literature on pose-invariant face recognition. The study categorizes these approaches into different groups as: multi-view systems, generative approaches, discriminative approaches and graph-based methods. We discuss their differences, weaknesses and strengths while listing some of the well known approaches in each category. The review identifies the graph-based methods as the most promising approach to pursue in the thesis.

Chapter 3: Background

In this chapter we introduce graphical models in more detail. The chapter will expand on the motivation for the development and application of such methods, basic definitions and problem formulation in this framework. The exposition then follows by the discussion of probabilistic inference applicable to graphical distributions and provides some insight into a family of algorithms for maximum a posteriori (MAP) inference.

Chapter 4: Image Matching

Chapter 4 introduces an MRF-based image matching method we adopt as a basis for recognition. The role of unary and binary measurements and the idea of node decomposition and the optimization method used for inference are explained. Examples of application of the method will be presented and discussed.

Chapter 5: Face Recognition Based on Image Matching

In Chapter 5 we develop a face recognition method based on the matching method introduced in Chapter 4. Different measures such as dynamic block adaptation and correlation modeling as well as label pruning are introduced to increase the effectiveness of the approach. We experimentally evaluate the method on the XM2VTS and the CMU-PIE databases in verification and identification scenarios and compare the results to those obtained by the state-of-the-art approaches. We also provide the results of an experimental evaluation of the methodology in object recognition on the SOIL database to illustrate the applicability of the approach to general object recognition problems.

Chapter 6: Multi-scale Matching

In Chapter 6 we consider two multi-scale image matching methods in order to speed-up inference and establish pixel-wise correspondences. One is heuristic and the other is based on

super-coupling transform to maintain consistency between the probability distributions at different scales. We provide experimental evaluation of both approaches on the CMU-PIE and FERET databases and compare the results to other methods.

Chapter 7: Face Representation Using a Sparse Model

Chapter 7 presents a new sparse MRF-based method for facial feature localization. In the proposed approach, different measures for modeling texture and shape variation of faces including higher order shape models based on point distribution models are considered. After describing the optimization approach, the method is evaluated on a problem of facial feature localization in frontal and rotated images of the XM2VTS database and also in real images collected from Google. The method is then used as an initialization step for the method of Chapter 6 and speed-up gains are achieved.

Chapter 8: Conclusions and Future Work

In closing, we summarize our research and propose possible directions for future research suggested by this work.

Chapter 2

Related Work

2.1 Introduction

In this chapter we briefly review some of the most prominent approaches taken to tackle the face recognition problem under varying pose. The synopsis will be organized into four groups based on the general concepts of classification motivating each approach. We will list the most well known methods in each class and also indicate where our approach sits in the organizational framework. We will also include some approaches in the discussion which are not explicitly used for pose-invariant recognition but have some common characteristics with other methods within each category. The databases used in the thesis are introduced next. The chapter is mainly focused on the approaches which make use of 2D images rather than those which use captured 3D shape information for recognition. For a review on the latter the reader is referred to [30].

2.2 Categorization

The various approaches to pose-invariant face recognition can be roughly divided into four major groups. These are multi-view systems, generative methods, discriminative approaches and graph-based methods.

2.2.1 Multi-view Systems

The earliest attempt of generalizing across different poses is represented by the multi-view methods. These methods are direct extensions of the systems operating on the frontal pose, storing multiple templates in different poses for each class. Generally, after estimating the pose of a test image, it is aligned with the selected gallery images and then a similarity criterion is evaluated which is used for decision making in the last stage.

One of the earliest multi-view algorithms is described in [25]. Having estimated the pose of the test image, the algorithm geometrically aligns the probe images to candidate poses of the gallery subjects using the automatically determined locations of three feature points. This alignment is then refined using optical flow between each pair of probe and gallery images. Finally, normalized correlation score is used for recognition. Good recognition results are reported on a limited database of 62 subjects imaged in a number of moderate pose changes ranging from -30° to $+30^\circ$ (yaw) and from -20° to $+20^\circ$ (pitch).

The popular eigenface approach of Turk and Pentland [127] is extended in [101] to handle multiple views. The authors compare the performance of two systems based on eigen-analysis. One is a parametric eigenspace constructed using all views from all subjects while the other is a view-based eigenspace approach which considers a separate eigenspace for each view. The methods are tested on a database of 21 people recorded in nine evenly spaced views from -90° to $+90^\circ$. It was found that the view-based eigenspace approach outperforms the parametric eigenspace by a small margin.

Other work in [128] employs multi-linear algebra as an expansion to the formal eigenface method to decompose the effects of viewpoint, illumination *etc.* The so-called *tensor face* method maps a probe image into a set of candidate subspaces where the nearest neighbor among all different subspaces corresponding to different imaging conditions is employed for decision making.

As a more recent attempt utilizing multiple gallery images in different views is the work in [117]. There, the authors make use of frontal and semi-profile images to construct composite face images. The approach consolidates the information represented by multiple images through the application of a registration and blending procedure. The goal is to avoid storing

multiple templates of each subject while obviating the complexity of generating 3D structures of images. For recognition a texture based approach using Gabor features is employed.

A common drawback of the multi-view systems is that they need multiple images of subjects at different poses in the gallery. This kind of information is not available in certain scenarios. The requirement for a large memory for storage is another drawback. In addition, efficient application of view-based methods requires good pose estimation of the probe image. Even in the presence of multiple gallery images in different poses, in practice there could be misalignment between the test image and those of the nearest gallery images which will eventually either decrease the overall performance or necessitate refining the alignment using *e.g.* optical flow. More recent works in the area of pose-invariant face recognition have moved away from using multiple gallery images towards using gallery images in a single pose (very often frontal).

2.2.2 Generative Methods

The second group of methods take a different approach. These methods do not store multiple gallery images. Instead, they reconstruct a novel image in a desired pose. The reconstructed image is either directly or indirectly used for classification in the next stage. The methods falling in this category operate either in 2D or 3D. In 2D versions of this class, although multiple gallery images of different poses or illuminations are not stored as templates, they are used to construct the model. In the 3D counterparts, 3D information is used to build the face model.

As an example of the 2D learning-based methods for virtual view synthesis is the work in [52]. There, the authors make use of the idea that a set of images of an object under all possible illumination conditions in a specific pose form a convex cone in the image space. Then, using a number of training images of each face taken under different lighting conditions, the shape and albedo of the face are reconstructed which in turn, provides a generative model to render or synthesize images of the face under novel poses and illumination conditions. For recognition, a nearest neighbor classifier is employed to assign the test image to the closest approximated illumination cone.

A different approach is to assume that a 3D shape can be modeled by a linear combination of prototypical views of an object in 2D. Under this assumption a rotated view of an object can be represented as a linear combination of the rotated views of the prototype objects. Using

this idea, in [24] the authors extended the previous work in [25] by using one example view of each subject and synthesizing virtual views using prior knowledge of shape and texture of faces in 2D. For recognition, an input image is repeatedly matched against all example views of all subjects in the gallery. For matching against a gallery view, the authors first geometrically register the test image to gallery images using an affine transformation followed by optical flow and then compute correlation score for classification.

In another work in [56], the authors propose a different way to deal with pose variations in 2D. Using a training set of sparse face meshes, a point distribution model (PDM) is built and the parameters responsible for controlling the appearance changes in shape due to pose variations are identified. Based on this investigation, two different approaches are proposed for pose correction. The first one is a method in which the pose parameters of both faces are set to those of frontal face images. In the second approach, one face adopts the pose parameters of the other. After obtaining pose corrected faces exploiting facial symmetry, using thin plate spline warping, virtual views are synthesized. For recognition, the virtual faces are fed into a system that makes use of Gabor filtering. It was also shown that if only pose parameters of faces are modified, client specific information is not lost in the warped image and better discrimination between subjects can be achieved.

A well known example of the 2D generative approaches is the active appearance models (AAMs) [37]. These are statistical models of shape and texture variations constructed for a class of images (faces). For training, a set of annotated images are used. Applying the eigen-analysis, main modes of texture and shape variations along with the correlations between them are learned. Fitting the model to a new image entails an analysis-through-synthesis process, that is, fitting is refined in each iteration by measuring the difference between the image reconstructed using the model and the original image. The fitting is made more efficient by learning the relationships between perturbations in model parameters and the induced errors. Once the model is fitted, the model parameters can be mapped into those that affect mainly pose and those that affect other variations. Ignoring the parameters responsible for pose changes, the only parameters that affect identity can be used for recognition [44]. Another approach is to render the face image in a new pose and use conventional face recognition algorithms [61].

One of the problems associated with the active appearance model framework is that as soon as

in-depth rotation of face becomes large and some features become occluded, the assumptions behind the model break down and the model collapses. In order to deal with larger rotation angles, researchers have examined different approaches. For example the authors in [40] use a number of models trained for different view points. If the pose is known in advance the most suitable model in terms of its pose is selected. In cases the pose is not known a priori, all different models are fitted and the one that matches the model best is selected.

Another example of 2D methods is the work in [34] which is a learning based regression method proposed to generate virtual frontal views of rotated images. The authors first validate the assumption that the mapping between a non-frontal face image and its frontal version can be approximated by a linear function. Formulating the estimation of the linear mapping as a prediction problem, the authors present a regression-based solution. In order to improve the estimation accuracy, the locally linear regression (LLR) was further proposed. In LLR, first a dense sampling of the non-frontal face image is performed to obtain a number of overlapped local patches. In the next step, the linear regression technique is applied to each small patch for the prediction of its virtual frontal patch. A cylindrical model for the head is employed to approximately predict correspondence between frontal and non-frontal local patches. Through the combination of all these patches, the virtual frontal view is generated.

In another work [146] a method aspiring to use mugshot databases effectively is proposed to generate virtual views. The approach uses frontal and side-face images to handle pose variations in face recognition. Virtual views in arbitrary poses are generated in two different shape modeling and texture synthesis stages. For modeling the shape, a multilevel minimization method is used to produce 3D face shapes. For texture synthesis, after analyzing face surface properties, desired views in different poses are rendered. In order to construct the model the authors use 80 manually labeled landmark points. The identity is determined by appearance-based face recognition.

In [106], a bayesian generative model is proposed to transform features from the identity space into the observation space using a deterministic function. Several local models describing the change of each local feature are involved. The parameters of the transformation functions are learned using the EM algorithm, assuming that the pose is known a priori.

In [59], instead of gray pixel values, eigen-light-fields are used to tackle the pose variation

problem in an appearance based system. Light field of an object is defined as the set of radiances of light along all rays in the scene radiated from an object. It is shown that the light-fields provide considerable information about the shape of an object which can be used to distinguish between them. Performing eigen-analysis on the set of light-fields of an objects, single gallery and probe images were assumed to be drawn from a larger data pool containing images of different poses and illumination conditions. Recognition is then formulated as a missing data problem. Using the prior knowledge of the distribution of the complete gallery set, the missing information is inferred and a few eigenvectors of the light-field are used for recognition.

A commonly followed method in the context of pose-invariant face recognition is to reconstruct virtual views using 3D models, the most well known one in this category being the 3D morphable model [26]. These methods do not need multiple images from different view points to construct the model. Instead, 3D laser scanned data is employed for training. In these techniques, shape and texture are learned offline and statistical models are constructed for each of them using eigen-analysis. For a new test image, the morphable model is fitted to the probe image and shape and texture parameters of the unknown probe image are recovered. From a recognition point of view, two approaches are plausible. One approach is to directly use the inferred parameters of the model for recognition, as *e.g.* the works in [89] and [108]. The second approach is to apply a geometric normalization step and render the probe image in a desired pose (usually frontal) and then use 2D-based methods for recognition [122]. Although the state-of-the-art methods are represented by this category, they still suffer from unresolved problems. The most important one is that in 3D geometric normalization based approaches, the recovered shape and texture are completely determined by the model fitted to the query 2D face image which has the capacity to reconstruct only the information captured during statistical learning. As a result, these approaches cannot recover atypical features that have not been observed in the training set. Moreover, the high computational complexity of the 3D methods in this category makes them unsuitable for real-time applications. Another drawback is the necessity to label landmarks to initialize the fitting which is usually carried out manually.

2.2.3 Discriminative Methods

In discriminative approaches, novel faces are not generated. In fact, features used are projected into a pose-invariant space in which recognition is performed based on similarity in this space. Similar to the generative methods, during test, multiple gallery images of each subject are not always needed, but for training, images corresponding to different poses or lighting conditions are employed to construct the model.

As an example in this class, in [76] the authors take advantage of a locally linear discriminant analysis (LLDA) approach to model nonlinearities of the data caused by pose changes in order to extract class-discriminative features. The idea is that any global nonlinear data manifold can be locally aligned based on the local linearity assumption. Input vectors are mapped to each local feature space to yield locally linearly transformed classes which maximize the inter-class covariance while minimizing the intra-class variance. Unlike view-based approaches, the local discriminant functions are estimated simultaneously to encode the relationships of different pose groups. The proposed approach for multiclass nonlinear discrimination is claimed to be computationally more efficient compared to similar approaches.

In [67], the problem of tuning multiple kernel parameters for the kernel-based linear discriminant analysis (LDA) method is addressed. The kernel approach is then employed for face recognition by mapping the input image to a high-dimensional feature space. The experimental evaluation showed that the proposed approach was a feasible one to tackle the pose problem. An algorithm was proposed to automatically tune the parameters of the Gaussian radial basis function kernel for the kernel subspace LDA. The proposed approach to parameter tuning improved the generalization capability of the kernel LDA method by maximizing the margin criterion while maintaining the eigenvalue stability of the kernel-based LDA method.

A further example of this category, is the work in [94] in which the extracted Gabor jet features at several locations on the face image are transformed in order to predict their appearance when viewed from the frontal pose. In order to estimate the transformation between different parts of a test image and gallery images a maximization step over the recognition performance is performed.

In [74], the authors propose a probabilistic approach to handle changes in the viewing angle. Using three landmarks, the images are first normalized and then divided into different subre-

gions from which features are extracted. The similarity of an image pair is computed as the sum of similarities between different subregions. In order to take into account the change in appearance of different regions due to pose variations, the method uses images of different poses to learn the distribution of feature similarities across different poses. Two methods for recognition over different poses are proposed for two different scenarios: when the pose is known apriori and when it is not. The results for the case when the pose is known is better compared to the other one which marginalizes the probabilities over all poses. The authors further investigate the discriminatory capacities of different regions of the face. The result is intuitive as it predicts that partly occluded half of the face due to pan movement is less discriminative compared to the visible half closer to the camera.

In a somewhat similar approach, in [13], taking into account the fact that local regions of the face change differently in appearance as the viewpoint varies, the authors divide images into non-overlapping patches and try to learn how they deform and appear as a result of pose changes. A learning-based method called stack-flow is proposed to estimate the optimal alignment between the images rotated in-depth and estimate patch deformations, using face data captured under different poses. The similarity of patches is measured based on their gray levels, and the overall similarity of two faces is determined by a probabilistic reasoning over local patch similarities without any attempt to use shape information explicitly. Like many other methods, the approach needs prior registration of the images using eye coordinates.

A relatively different approach to tackle the pose variation problem is based on the idea proposed in [97]. The idea is that 2D projections into an eigen-space of a 3D object when viewed from different angles, form a 2D manifold. In [57], the authors, based on this idea and using densely sampled image sequences of rotated faces, construct eigen-signatures for gallery images. Radial basis functions networks are then employed to generate eigen-signatures from one single probe image. For recognition, the distances between the projection of a test image to the eigen-space and the eigen-signatures created from gallery images, are measured. Good generalization was achieved with half profile training views. However, the recognition rates for tests across wide pose variations (*e.g.* frontal gallery and profile test) were low.

As an example of the works in the literature making use of dense correspondences between images for recognition is the work in [33] which uses stereo matching for recognition. The

method suffers from a number of problems. The error induced by the assumption that the pixels have only 1D disparities is problematic when the change of viewing angle results in 2D disparities even after aligning the two images using eye coordinates. Also, the similarity measure used in [33] is based upon the gray scale content of images only, ignoring the geometric distortion which is useful for recognition.

In [142], a component-based discriminative approach is proposed to tackle face retrieval under image variations including pose changes. After detecting a face, the method locates five facial components using a neural-network based component detector. Local descriptors are extracted by steerable filter outputs and aggregated to form a face descriptor. Given a large database a number of first K candidates are extracted based on local features. The top K identities are then sorted according to a low dimensional global face descriptor. In the final stage, multiple gallery images of each subject are used for comparison with a query image.

In another work [141] an implicit image matching is proposed to handle variations in image pose. Instead of directly solving the matching and correspondences analysis problem between a pair of images, the authors augment the feature vectors with the spatial location of the extracted local features. The method which works by detecting the face first and subsequently the eyes, normalizes the image both geometrically and photometrically. After augmenting the features which are based on difference of Gaussian (DoG) filters [138], a quantization step is performed to convert the augmented feature vectors into histograms which are then used for classification.

Although discriminative approaches seem less restrictive in terms of their ability for image interpretation, a common drawback of these approaches is that in order to construct the model, a large number of images either in different poses or under different illumination conditions are required making them less attractive in cases where only one gallery image is available per subject.

2.2.4 Graph-based Methods

Graph-based methods constitute the last category in our classification of pose-invariant methods for face recognition. In these approaches, images are assumed to be built up from different parts interacting together. The success of these methods in the context of general object matching/recognition tasks relies on different facts. First, this methodology allows different parts

of an object to be considered independently of other non-neighboring parts which is useful in dealing with occlusion and cluttered background. This property is also desirable when different parts of an object undergo different geometric distortions *e.g.* in the case of pose changes. On the other hand, unlike part-based approaches which discard information regarding the configurational arrangements of object primitives [42], by exploiting the structure of the objects, the matching process becomes less ambiguous in these methods which subsequently improves performance. Furthermore, these approaches require a minimum number of training images and good performance can be achieved even by using one gallery image per class.

Although in some of the works we review in the following no attempt is made to recognize faces across poses changes, we briefly review them because of their similarities to the approach we take in the thesis. Our approach which will be described in more detail in the following chapters exploits the match probability in a graphical framework for decision making in face recognition across different poses.

A well known example of graph-based approaches is the elastic bunch graph matching (EBGM) approach presented in [140]. A number of manually selected fiducial points are considered as graph nodes. Local measurements for the nodes are based on complex Gabor wavelets and the similarities between them are computed using the magnitude ignoring phase information. In order to handle variations in appearance of different fiducial points (*e.g.* differently shaped eyes, mouth, nose) the authors proposed a stack-like structure called *bunch graph* in which the local characteristics of each node covered a range of variations for the node. For graph matching, the similarity is defined as the sum of node similarities plus geometric distortion of the assumed structure. A heuristic method is employed for matching. For recognition, the authors only consider local node similarities.

In [132], the authors formulate face recognition as a bayesian decision making problem based on graphical models. Local features are obtained using Gabor filter outputs projected in the LDA subspace and the contextual relationships between the object parts are modeled by a linear penalty function. For inference on the graph, a heuristic search method is employed in conjunction with the similarity measure between two faces. The method is evaluated on two databases and a performance improvement is reported over some other non-graphical approaches.

In [99], a face is represented by a graphical model. Straight line segments are extracted and

used as face features, *i.e.* nodes of the graph. Adding proper binary relations between nodes, a partial matching is used to match different pairs of graphs and to select the best match for recognition. By virtue of partial graph matching the method achieves robustness against partial occlusion to some extent.

In [68], the authors employ a graph-based method in which individual nodes are local patches of images. For recognition, each patch is assigned to one possible class of identities. Inference is made using belief propagation algorithm [100] and for classification a majority vote rule is used. The method implicitly assumes that images are registered. A direct way of generalizing the method for handling patch deformations is not proposed although it appears necessary for matching images rotated in depth. The binary relations of a pair of nodes are adopted as a penalty for being assigned to different identities only, and as a result local patches are not penalized for being assigned to the same spatial location in one image. In addition, the inference based on belief propagation used in the method is not the best choice as better inference methods are available.

In [77] authors consider face graph matching using SIFT features [90] as node attributes. Different heuristic matching strategies are considered for graph matching. The similarity measure is considered as the sum of unary similarities between nodes, ignoring shape distortion in the similarity criterion.

2.3 Databases and Benchmarking

Databases play an important role in the advancement of scientific research. Consider the case where every method is tested on a different collection of images. Taking into account the fact that it is almost impossible to maintain exactly the same imaging conditions for each image collection, it would be impossible to compare the performance of different methods objectively. As a result, different databases have been collected and made available to the scientific community in order to provide a fair evaluation and comparison of different methodologies. In the following we describe three face databases along with a general object database which we will use in our experimental evaluations.

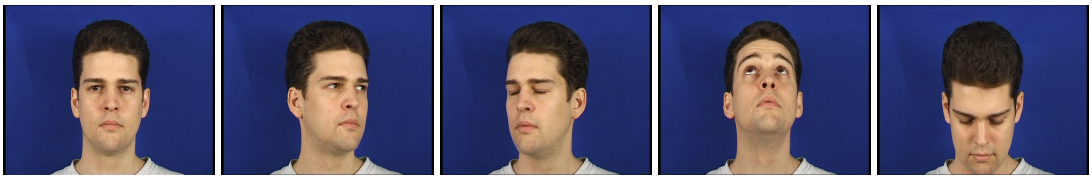


Figure 2.1: Illustration of pose variation in the XM2VTS database.

2.3.1 XM2VTS

The Extended M2VTS (XM2VTS) [95] database is a large multi-modal dataset designed for evaluating multi-modal verification methods. The database includes still color images, audio data, video sequences and 3D models along with image sequences of multiple views of the subjects. The database is comprised of digital video of 295 subjects recorded over a period of five months with one month intervals. It was captured by a Sony VX1000E digital cam-corder and DHR1000UX digital VCR under controlled conditions with uniform illumination and blue background in order to facilitate face segmentation. Some sample images of this corpus are given in Fig. 2.1.

2.3.2 CMU-PIE

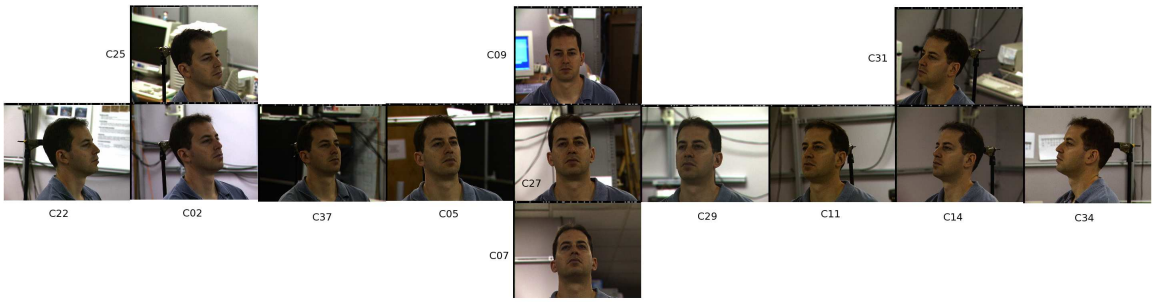


Figure 2.2: Illustration of pose variation in the CMU-PIE database.

The Carnegie Mellon University's Pose, Illumination and Expression (CMU-PIE) database [115] is one of the most challenging databases for 2D face recognition. It was collected between October and December 2000 and contains more than 40,000 images of 68 people. Each subject is imaged from 13 different viewpoints with 4 different expressions and under 43 different illumination conditions. The pose changes vary from full left profile to frontal and on to

full right profile. This database is potentially useful for the assessment of head pose estimation algorithms and for evaluating the robustness of face recognition algorithms against pose, illumination, and expression variations. Some example images of this database are shown in Fig. 2.2.

2.3.3 FERET

Figure 2.3: Sample images from the FERET database.

The FERET database [103] was collected at George Mason university and the US Army Research Laboratory facilities. It was collected as part of the FERET program which ran from 1993 through 1997, sponsored by the Department of Defense's Counterdrug Technology Development Program through the Defense Advanced Research Products Agency (DARPA). The primary goal of the program was to build face recognition algorithms which could be used in security, intelligence and law enforcement of personnel in the execution of their duties. This image corpus consists of 14051 eight-bit gray scale images of faces. 3,816 images of the database are frontal and the rest have poses ranging from frontal to left and right profiles. Sample images of this database are presented in Fig 2.3.

2.3.4 SOIL47

Figure 2.4: Illustration of pose variation in the SOIL database.

Although we are mainly concerned with the face recognition problem, we provide some experimental results on a general object recognition database (SOIL) [1]. The Surrey Object Image

Library (SOIL) is a database of household objects. The objects are captured over a considerable portion of the viewing sphere. The database collected at the University of Surrey shows mainly multi-colored objects, many of which have planar surfaces. The database includes images of 24 objects with approximately planar surfaces and 22 complex scenes. 21 images for each object were acquired by a robot arm moving around the object at intervals of approximately 9 degrees spanning a range of -90 to $+90$ degrees. Appearance variations are caused by 3D viewpoint changes and self-occlusion. Some example images of this database are shown in Fig. 2.4.

2.4 Summary

In this chapter we reviewed different approaches for pose-invariant recognition of faces. Different methods were categorized into four different classes and most prominent methods in each category were briefly reviewed. Our method, which will be introduced in the next chapters, falls into the graph-based class. These approaches have the capability to handle various imperfections of image capture while requiring minimum training data for recognition.

Next, we introduced databases which will be used in our experiments in the following chapters. Some aspects of the specifications of the databases were also provided.

Chapter 3

Background

3.1 Introduction

In this chapter we provide some background on graphical models and their application in image analysis. After reviewing some of the basic definitions of graph theory in section 3.2, the relation between graphical models and Gibbs distributions and their interpretation as Markov models are discussed in section 3.3. The discussion is then followed in section 3.4 by the the motivation behind employing these models for image analysis where we also describe how image analysis problems are posed in this framework. Section 3.5 discusses inference in graphical models and provides some insights into a family of inference algorithms in graphical models, known as *graphical decomposition* methods. Section 7.6 brings the chapter to a conclusion.

3.2 Graphs and Hypergraphs

Although in order to understand the thesis a thorough description of graph theory is not required, a review of the basics and the terminology of graphs would help to understand some aspects of the work. We provide a summary of the relevant definitions and set up some conventions used in different chapters. Additional material on graphical models can be found in references [22, 29].

A graph is comprised of a set of vertices (also referred to as nodes/sites) and its set of edges. Edges are defined as subsets of nodes. Most of the time, edges are subsets of pairs of vertices.

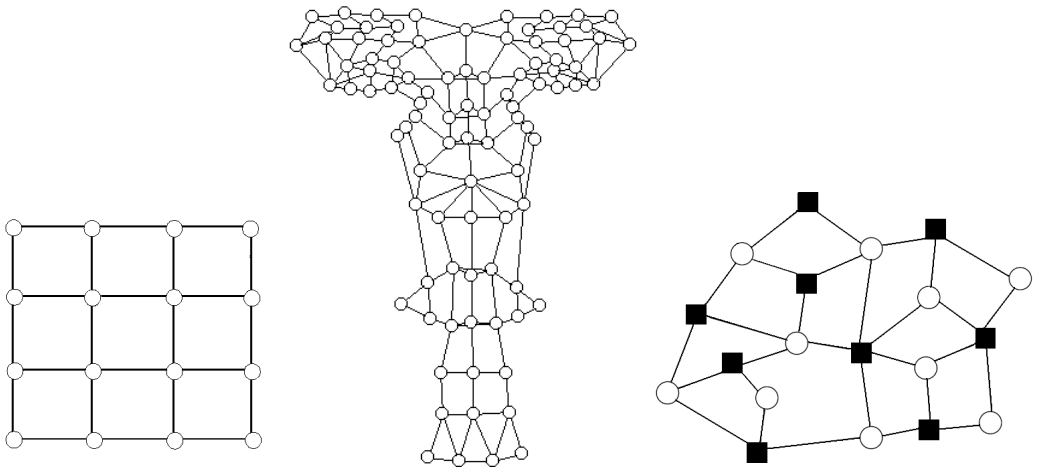


Figure 3.1: Different graphical models: from left to right: grid pairwise graph (2D lattice), irregular pairwise graph, hypergraph represented as factor graph.

A graph whose edges are subsets of *pairs* of nodes is called a *pairwise* graph. A more general definition of graph is the *hypergraph*, for which edges (also called hyperedges) are considered as subsets of one, two or more nodes. A hyperedge is also called a *clique* of the graph. Figure 3.1 illustrates different graphs and a hypergraph where each circle stands for a vertex. Edges in pairwise graphs are illustrated as lines connecting a pair of nodes whereas black square markers are used as symbols for hyperedges. The symbolization used here for the hypergraph is called *factor graph* in the graphical modeling literature [49, 92]. Through the thesis we will be concerned with graphs having at most pairwise relations. An exception to this is Chapter 7 where we use the hypergraph and hyperedge to underscore the existence of hyperedges.

3.3 Factorization and Gibbs Distribution

Assume $x = (x_1, \ldots, x_n) \in \mathbb{X}^n$ to be a group of variables each of which having the set $\mathbb{X}$ as its domain. Here we have made the assumption that all variables $x_i, i = 1 \ldots n$ have the same domain $\mathbb{X}$ and used the notation $\mathbb{X}^n = \mathsf{X}_{i=1}^n \mathbb{X}$ (that is the Cartesian product of the variables' domains) to denote their joint domain. In general each variable may have a different domain.

The domain of a binary variable may be represented by $\mathbb{X} = \{0, 1\}$ and a continuous variable's domain may for example be the set of real numbers, *i.e.* $\mathbb{X} = \mathbb{R}$. Throughout the thesis we will

be only concerned with the discrete case.

A probabilistic graphical model is defined as a probabilistic model symbolized by a graph $\mathscr{G}$ with the vertex set $\mathscr{V} = \{1,\ldots,n\}$ and the associated variables $x_1,\ldots,x_n$, of its vertex set, its edge set $\mathscr{E}$ and a probability distribution in the following form:

$$P(x) = \frac{1}{Z(\psi)} \prod_{E\in\mathscr{E}} \psi_E(x_E) \tag{3.1}$$

where E stands for an edge/hyperedge and each function $\psi_E : \mathbb{X}^{|E|} \to \mathbb{R}$, called a factor of the graphical model, is a non-negative function of variables $x_E = \{x_v | v \in E\}$. $Z(\psi)$ is a constant, normalizing the probability distribution. In the factor graph representation of Fig. 3.1, each circle node stands for a variable x_v and each square black node $E \in \mathscr{E}$ stands for one of the factors ψ_E. The probability distribution can also be represented as a *Gibbs distribution* if it is strictly positive, *i.e.* $P(x) > 0, \forall x$:

$$P(x) = \frac{1}{Z(f,\beta)} \exp\left\{-\beta \sum_{E\in\mathscr{E}} f_E(x_E)\right\} \tag{3.2}$$

The quantity $En(x) = \sum_{E\in\mathscr{E}} f_E(x_E)$ is called the *energy* and the terms $f_E(x_E)$ are called *potentials* of the model. The parameter β is the inverse of the so-called *temperature* of the Gibbs distribution and $Z(f,\beta)$ defined as

$$Z(f,\beta) \triangleq \sum_{x\in\mathbb{X}^n} \exp\left\{-\beta \sum_{E\in\mathscr{E}} f_E(x_E)\right\} \tag{3.3}$$

is called the *partition function* serving to normalize the probability distribution. The probability distributions defined by 3.1 and 3.2 become equivalent if we take $\psi_E(x_E) = \exp\{f_E(x_E)\}$ (and $\beta = 1$).

It is worth noting that the choice of potential functions ($f_E(x_E)$) that form a specific distribution $P(x)$ is not unique and different choices of potential functions can lead to the same probability distribution. Two reasons for this degeneracy are as follows. First, due to the normalization of $P(x)$, one can add a constant to the energy $En(x)$ and it does not change $P(x)$. Moreover, for a fixed energy $En(x)$ there are various ways one can split it into a set of potentials $f_E(x_E)$ so that the sum would not change $En(x) = \sum_{E\in\mathscr{E}} f_E(x_E)$. For instance, if two edges $A,B \in \mathscr{E}$ share nodes $S = A \cap B \neq \emptyset$, then one can add an arbitrary function $\lambda(x_S)$ to one potential $f'_A(x_A) = f_A(x_A) + \lambda(x_S)$ and subtract the same function from the other $f'_B(x_B) = f_B(x_B) - \lambda(x_S)$

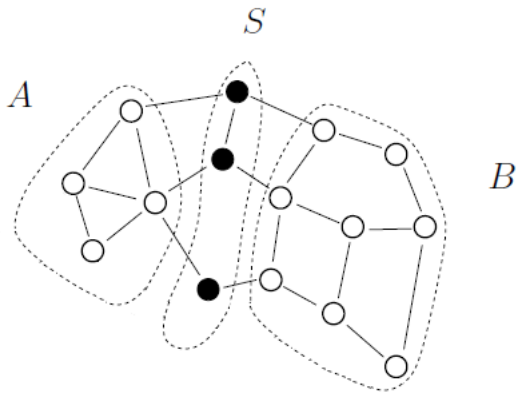


Figure 3.2: Example of a separator (filled circles), the variables A and B are conditionally independent given S [72].

which leaves the overall potential unchanged as $f'_A(x_A) + f'_B(x_B) = (f_A + \lambda) + (f_B - \lambda)$. Such form of changes in representation do not influence the distribution and are sometimes called *reparameterizations* or *equivalent transformations* of the model [79, 134]. This degeneracy is essentially very useful and serves as a basis in designing inference algorithms for graphical models, to be discussed more in the forthcoming sections and chapters.

3.3.1 Conditional Independence and Markovianity

The structure of a graphical model entails a set of conditional dependence/independence properties between its vertices. A *separator* of the graph $\mathscr{G}$ is defined as a subset of vertices $S \subset \mathscr{V}$ if by removing its nodes and the edges that contain any of its nodes, some part of the graph cuts off and the total number of connected components of the graph is increased. It is said that S separates two vertex sets $A, B \subset \mathscr{V}$ if there is no path connecting A and B that does not pass through S. The definition is illustrated in Fig. 3.2.

A probability distribution $P(x)$ is said to be *Markov* with respect to the graph $\mathscr{G}$ if for all (S, A, B), where S separates A from B, we have $P(x_A, x_B|x_S) = P(x_A|x_S)P(x_B|x_S)$. This property for a graphical model defined on $\mathscr{G}$ can be simply verified . Such graphical models are called *Markov models* or *Markov random fields* (MRFs). For a set of vertices $A \subset \mathscr{V}$, its *Markov blanket* denoted by ∂A is the set of nodes that are not included in A but are linked to A through some edges. The Markov property then implies $P(x_A|x_{\mathscr{V}\setminus A}) = P(x_A|x_{\partial A})$. The *Hammersley-Clifford*

theorem [32, 58, 62] states that all probability distributions that are *Markov* with respect to a graph can be represented as a Gibbs distribution on the same graph. This means that if $P(x) > 0$ for all $x \in \mathbb{X}^n$ and $P(x)$ is Markov on $\mathscr{G}$, then there exists a set of potential functions such that $P(x)$ can be represented as a Gibbs distribution in the form of equation 3.2 (above the critical temperature). If $P(x)$ is Markov on $\mathscr{G}$ and E is not a clique of $\mathscr{G}$ then the conditional independence property implies that $f_E(x_E) = 0\ , \forall\ x_E \in \mathbb{X}^{|E|}$.

3.3.2 Exponential Families

Graphical distributions can be also considered as parameterized families of graphical models in the form of an *exponential family* [16, 45]:

$$P(x;\theta) = \exp\left\{\theta^T \phi(x) - \Omega(\theta)\right\} \tag{3.4}$$

Where, $\phi : \mathbb{X}^n \to \mathbb{R}^d$ are called *sufficient statistics*, used to define the family. Given the parameter θ, the set of sufficient statistics ϕ fully characterize the probability distribution $P(x;\theta)$. $\Omega(\theta)$ is the *cumulant function* of the family, which serves to normalize the distribution (analogous to the partition function in the previous section). In discrete models, we have

$$\Omega(\theta) = \log \sum_{x \in \mathbb{X}^n} \exp\left\{\theta^T \phi(x)\right\} \tag{3.5}$$

In practice, only members of set $\Theta = \left\{\theta \in \mathbb{R}^d | \Omega(\theta) < \infty\right\}$ are considered. That is, only those θs for which $\Omega(\theta)$ is well-defined so that we may define a normalized probabilistic model. If we use α to index sufficient statistics then the complete vector is $\phi = \{\phi_\alpha, \alpha \in \mathscr{I}\}$. Here each α corresponds to the family of sufficient statistics corresponding to a clique of the graphical model and $\mathscr{I}$ represents the union of all such families of sufficient statistics over all cliques of the graph. With this notation, the probability distribution $P(x;\theta)$ is:

$$P(x;\theta) = \exp\left\{\sum_{\alpha \in \mathscr{I}} \theta_\alpha \phi_\alpha(x) - \Omega(\theta)\right\} \tag{3.6}$$

3.3.3 Sufficient Statistics in Discrete Graphical Models

We shall now discuss how the sufficient statistics and the index set $\mathscr{I}$ are chosen in a graphical model. For the sake of convenience we consider a pairwise graph here but generalization to a

hypergraph is straightforward. The main building blocks of sufficient statistics in this case are the indicator functions, showing whether a particular node has been assigned a special label and also indicating whether an edge of the graph has a specific state. For a pairwise graph the set $\mathscr{I}$ is:

$$\mathscr{I} = \{(u;x_u)|u \in \mathscr{V}, x_u \in \mathbb{X}_u\} \cup \{(uv;x_u x_v)|(u,v) \in \mathscr{E}, (x_u,x_v) \in \mathbb{X}_u \times \mathbb{X}_v\} \tag{3.7}$$

where we have used the notation $\mathbb{X}_u$ to denote the domain of the variable x_v, assuming that in general each variable may have a separate domain. The set $\mathscr{I}$ here (in a pairwise graph) has a total number of $d = \sum_{u\in\mathscr{V}} |\mathbb{X}_u| + \sum_{(u,v)\in\mathscr{E}} |\mathbb{X}_u| \times |\mathbb{X}_v|$ members. The functions $\phi_\alpha(x)$ are indicator functions defined as:

$$\begin{aligned} \phi_{u;y}(x) &= [\![x_u = y]\!] \\ \phi_{uv;yy'}(x) &= \phi_{u;y}(x)\phi_{v;y'}(x) \end{aligned} \tag{3.8}$$

where symbol $[\![\zeta]\!]$ equals 1 if expression ζ is true and 0 otherwise. The energy of the distribution is: $En(x;\theta) = \sum_{E\in\mathscr{E}} f_E(x_E)$. Usually one uses the notation $\theta_E(x_E)$ to denote the enumeration of all values of f_E in terms of sufficient statistics, *i.e.* $\theta_E(x_E) = \sum_{\tilde{x_E}} \theta_{E;\tilde{x_E}} \phi_{E;\tilde{x_E}}$. Here $\tilde{x_E}$ denotes possible configurations of a hyperedge. With this change of notation, the energy in a pairwise graphical model may be represented as:

$$En(x;\theta) = \sum_{u\in\mathscr{V}} \theta_u(x_u) + \sum_{(u,v)\in\mathscr{E}} \theta_{uv}(x_u,x_v) \tag{3.9}$$

For reasons to be clarified later where we discuss inference for graphical models, it is desirable to compute marginal probabilities of sufficient statistics for a graphical distribution $P(x;\theta)$. Roughly speaking, the marginal probabilities of sufficient statistics say how much a particular label is likely to occur in a graphical model representing a probability distribution $P(x)$. The set of marginals of each node $u \in \mathscr{V}$ in an MRF are calculated by taking expectations of nodewise sufficient statistics:

$$\mu_{u;y} = \mathbb{E}_P[\phi_{u,y}] = \sum_{x\in\mathbb{X}^n} P(x;\theta)\phi_{u;y}(x) \tag{3.10}$$

and similarly for an edge $(u,v) \in \mathscr{E}$

$$\mu_{uv;yy'} = \mathbb{E}_P\left[\phi_{u;y}\phi_{v;y'}\right] = \sum_{x\in\mathbb{X}^n} P(x;\theta)\phi_{u;y}(x)\phi_{v;y'}(x) \tag{3.11}$$

The set of all $\mu_{u;y}$ and $\mu_{uv;yy'}$ define a d-dimensional vector $\mu = \{\mu_\alpha | \alpha \in \mathscr{I}\}$ of marginal probabilities which are indexed by the elements of the set $\mathscr{I}$. The set of all such marginal probabilities are referred to as the *marginal polytope*, denoted by $\mathbb{M}(\mathscr{G})$:

$$\mathbb{M}(\mathscr{G}) = \left\{ \mu \in \mathbb{R}^d | \mu_{u;y} = \mathbb{E}_P [\phi_{u;y}] \text{ and } \mu_{uv;yy'} = \mathbb{E}_P \left[\phi_{u;y}\phi_{v;y'}\right] \right\} \tag{3.12}$$

The set $\mathbb{M}(\mathscr{G})$ is characterized by a number of linear constraints in the d-dimensional space [131]. Although the number of constraints which characterize the marginal polytope is always finite, it grows rapidly in number of nodes in a general graph with cycles [131]. As a result, it is common practice to characterize a *subset* of constraints that any $\mu \in \mathbb{M}(\mathscr{G})$ has to satisfy.

First, since the elements of $\mathbb{M}(\mathscr{G})$ are marginal probabilities, we must have $\mu \geq 0$. Secondly, as local marginals, the elements of $\mathbb{M}(\mathscr{G})$ must satisfy the normalization constraints:

$$\sum_{x_u \in \mathbb{X}_u} \mu_{u;x_u} = 1; \forall u \in \mathscr{V} \tag{3.13}$$

$$\sum_{(x_u,x_v) \in \mathbb{X}_u \times \mathbb{X}_v} \mu_{uv;x_u x_v} = 1; \forall (u,v) \in \mathscr{E} \tag{3.14}$$

Third, as the single-node marginal over x_u must be consistent with the joint marginal on (x_u, x_v), the marginalization constraint must also hold:

$$\sum_{x_v \in \mathbb{X}_v} \mu_{uv;x_u x_v} = \mu_{u;x_u}; \forall (u,v) \in \mathscr{E} \tag{3.15}$$

Based on these constraints, the set of all $\mu \in \mathbb{R}^d$ that satisfy constraints 3.13, 3.14 and 3.15 is denoted as $\mathbb{L}(\mathscr{G})$. The set $\mathbb{L}(\mathscr{G})$, defined by only a subset of the constraints characterizing the set $\mathbb{M}(\mathscr{G})$, specifies an outer bound on it, *i.e.* we have $\mathbb{M}(\mathscr{G}) \subseteq \mathbb{L}(\mathscr{G})$. In contrast to $\mathbb{M}(\mathscr{G})$, it is specified by only a number of inequalities that is polynomial in number of variables. More specifically, the set $\mathbb{L}(\mathscr{G})$ is characterized by $\mathscr{O}(m|\mathscr{V}| + m^2|\mathscr{E}|)$ constraints, m being defined as $m = \max_u |\mathbb{X}_u|$.The set $\mathbb{L}(\mathscr{G})$ is sometimes called the *pseudo-marginal polytope*.

3.4 Graphical Models in Image Analysis

Visual information in images is usually exposed to noise which results in various uncertainties and ambiguities. The noise and uncertainty may be introduced by different sources and have

different natures such as sensing device noise, illumination or pose changes, occlusion, inherent changes of a deformable object *etc*. It is commonly believed that holistic approaches are more susceptible to such changes in the appearance of objects. A common approach to deal the problem is to model an object as a collection of smaller parts and consider image features at a lower level of representation. This allows partly to recover from the omnipresent effects of unwanted variations in images but does not solve the problem completely. In fact, the problem of visual recognition in the presence of noise is ill-posed. There are too many degrees of freedom in the problem solving because the observation (image data) only partially constrains the solution.

One way to overcome the problem is to impose some sort of prior knowledge on the solution so that the hypothesized model would favor some part of the solution space over another. This kind of prior knowledge can be forced on the solution by incorporating conditional dependencies between different entities of an image. The injection of this kind of conditional dependencies is not far from reality and can be observed in the staistics of natural images. As a result, contextual relationships and constraints become not only helpful but ultimately indispensable in any visual recognition algorithm. In this context, objects are explained in terms of features at a lower level of representation and the features themselves are defined in terms of lower level primitives. Typically, such primitives are extracted at the lowest scale of representation, *i.e.* image pixel level.

Context-dependent and correlated entities can be conveniently modeled and handled by describing their interactions using graphical distributions and Markov random fields (MRFs). As discussed earlier, in this framework the state of each variable is only dependent on its Markov blanket and independent of the state of the rest of variables. From a computational viewpoint, the local properties of Markov random fields and their sparsity in conditional dependencies are appealing since they support algorithms that can be implemented in efficient ways. In addition, such representations provide a suitable basis for multi-resolution analysis of images to achieve further speed-up in inference. The above characteristics have motivated the application of MRFs to vision problems at all levels. While most of the applications were initially limited to low-level vision problems, such as image restoration, edge detection, segmentation *etc.*, their use in high-level vision *e.g.* object matching and recognition has also been considered. In [87] a unified framework for solving image analysis problems from low to high level is discussed.

The general interest in this field has been increasing during past couple of years, as manifested in the growing number of the publications in this area.

In computational vision problems, this philosophical approach has been advocated in [51]. It facilitates the development of algorithms for a variety of problems in a systematic way rather than relying on ad hoc heuristic methods. The framework is rather general. Each separate problem involves two important tasks. First, one needs to decide on the form of the posterior distribution. Second, the parameters of the distribution have to be determined. Another major ingredient is the optimization method for inferring the mode of the posterior distribution (MAP inference).

3.4.1 Image Labeling

An image analysis problem may be posed as one of finding the most probable configuration of a set of variables. This problem is often referred to as image labeling. The solution to the problem is a set of states/labels assigned to the variables/sites, *i.e.* image primitives. A site represents a point or a region in the image frame such as an image pixel or a corner point, a segment of line *etc.* In edge detection, for example, the sites are image pixels and the label set is

$$\mathbb{X} = \{edge, nonedge\} \tag{3.16}$$

In a graphical representation, each site is modeled as a node of a graphical model. Then the labeling problem is to assign each site in an image (nodes of the model graph) a label from its admissible label set. In image restoration, for example, $\mathbb{X}$ contains admissible pixel values that are common to all pixel sites, and $\mathbb{X}^n$ defines all feasible images. In certain conditions, admissible labels (domain of each variable) may not be common to all the sites, such as in feature-based object matching.

3.4.2 Conditional Dependencies in Image Analysis

Assume we have a set of measurements over the members of set $\mathscr{V}$ in an image, *i.e.* image sites. These measurements may for example correspond to a filtering operation on pre-specified locations of an image. Considering conditional independence between observations at each site,

finding the most likely configuration of these sites then amounts to the maximum likelihood estimation. In fact, when the knowledge about the data distribution is available but the prior information is not, the maximum likelihood (ML) estimate can be used to find the most likely configuration:

$$x^* = \arg\max_x \prod_{u \in \mathscr{V}} \Psi_u(x_u) \tag{3.17}$$

where $\prod_{u \in \mathscr{V}} \Psi_u(x_u)$ represents the joint likelihood of unary/independent measurements obtained directly from the image under consideration. On the other hand, when both the likelihood and prior distributions of observations are available, these two sources of information can be combined according to Bayes law and maximized to find the configuration with maximum a posteriori probability (MAP estimate). Very often in image analysis, the factors defining a graphical distribution are grouped into two categories: factors corresponding to the cliques of cardinality one and those of higher cardinalities:

$$P(x) = \frac{1}{Z(\Psi)} \prod_{u \in \mathscr{V}} \Psi_u(x_u) \prod_{E \in \mathscr{E}} \Psi_E(x_E) \tag{3.18}$$

where we have used the notation E to denote cliques constituting at least two nodes. The term $\prod_{u \in \mathscr{V}} \Psi_u(x_u)$ in the equation above is called the *likelihood* of the configuration x. Likelihood of MRF models in image analysis corresponds to unary factors, derived from image measurements. The likelihood is often referred to as *data term*. The term $\prod_{E \in \mathscr{E}} \Psi_E(x_E)$ is called the *prior* of the configuration. For image analysis using graphical models, by *prior* one refers to the factors of the graphical distribution including at least two nodes. The theorem of equivalence between Markov random fields and the Gibbs distribution discussed in previous sections provides a convenient platform to combine these two sources of information in terms of a joint probability distribution.

The idea of contextual information and conditional dependence in pattern recognition and image analysis dates back to the works in [36] and [3]. In [36], the problem of character recognition is formulated as a statistical decision making problem. By going beyond the hypothesis of statistical independence, a nearest neighborhood dependence of pixels in a lattice structure is achieved. Then, the information in the neighborhood is exploited to estimate conditional probabilities. The work in [3] is one of the earliest works in pattern recognition taking advantage of the Markovian assumption. Using contextual constraints in a Markov model, the

number of parameters needed for the analysis is reduced. A further milestone in context-based models is relaxation labeling (RL) proposed in [109]. Relaxation labeling is a class of iterative procedures which takes into account contextual constraints in order to reduce ambiguities in data analysis. In terms of probabilities, contextual relationships are interpreted locally using conditional probabilities.

In reality, the Markov blanket of a node can include many other nodes. In other words, measurements obtained from an object in very large neighborhoods may be conditionally dependent on each other. However, applying models based on such assumptions leads to a very densely connected graphical model for which inference becomes nearly intractable. As a result, in practice one makes simplifying assumptions on the inter-relationships between sites. Very often in practice, these dependencies are limited to pairwise interactions. However, one of the major recent research lines in MRF analysis is the development of efficient algorithms for inference in models which include groups of more than two variables as factors of the graph. We will employ one such representation in Chapter 7 where conditional dependencies among different components of an object are jointly modeled as high order factors of a graphical model.

3.5 Inference in Graphical Models

The problem of finding the maximum a posteriori probability of a distribution $P(x;\theta)$ (MAP inference) using exponential family representation is written as:

$$x^* = \arg\max_{x \in \mathbb{X}^n} \theta^T \phi(x) \tag{3.19}$$

Since all $\phi(x)$ are integers, that is, either 0 or 1, the problem is called an *integer program*. The problem in equation 3.19 has an alternative representation as a linear program over the marginal polytope (see [130] for a proof) that is:

$$\mu^* = \arg\max_{\mu \in \mathbb{M}(\mathscr{G})} \theta^T \mu \tag{3.20}$$

As noted earlier, the set $\mathbb{M}(\mathscr{G})$ is hard to be characterized thus in practice instead of this set, the set $\mathbb{L}(\mathscr{G})$ is considered as the constraint set:

$$\tau^* = \arg\max_{\tau \in \mathbb{L}(\mathscr{G})} \theta^T \tau \tag{3.21}$$

The objective function in equation 3.21 is linear and so is the constraint set ($\mathrm{L}(\mathscr{G})$, characterized by linear equations). As a result, this formulation is called the *linear programming relaxation* as opposed to the problem in equation 3.19 which is called integer program. In fact the problem in equation 3.21 is what most of inference algorithms for MRFs try to solve. It should be noted however that the optimal values of the two problems in equations 3.20 and 3.21 are not always equal. As discussed in the following sections, their optimal values only coincide for certain types of graphs.

3.5.1 Trees

In order to define a *tree* we need to define *path* and *cycle* of a graph. A *path* of a graph is a sequence of its nodes so that from every one of its nodes there exists an edge to the next node in the sequence. A *cycle* is a path in which the start node and the end node of the sequence are the same. A graph without any cycles is called a *tree* (also called acyclic graph). In contrast, a graph containing at least one cycle is referred to as a *loopy graph* (also called cyclic graph). In MAP inference for MRFs, trees are treated separately due to their appealing properties. In tree-structured distributions, the set $\mathbb{L}(\mathscr{G})$ is exactly equal to $\mathbb{M}(\mathscr{G})$ [131]. This means that local consistency also guarantees global optimality and leads to realizable and globally consistent solutions. There are exact algorithms for inference in the acyclic graphs, such as max-product belief propagation [100]. The algorithm is essentially a message passing algorithm which works by propagating information between adjacent nodes. This method is guaranteed to converge after a finite number of iterations and exact solution can be extracted. However, the situation in loopy graphs is not as simple as in trees. For example, if max-product algorithm is applied to loopy graphs, messages keep circulating between neighboring nodes without resulting in the optimal solution. As a result, the problem of mode finding in general graphs needs a particular treatment and specialized algorithms. The problem has motivated lots of research over the past decades.

3.5.1.1 Min-marginals

In this section, we first consider the definition of *min-marginals* and next discuss their role in finding the MAP solution on tree structured distributions. We note that in the rest of thesis we

will mainly consider the energy rather the probability. In order to simplify the definition, we consider a graphical model with at most pairwise potentials, then the *min-marginal* associated with each node is defined as:

$$\Phi_{u;y}(\theta) = \min_{\{x \in \mathbb{X}^n | x_u = y\}} En(x;\theta) + \text{const}_u \tag{3.22}$$

where const_u is a constant independent of y. $\Phi_{u;y}(\theta)$ is proportional to the minimum energy of the configuration with x_u set to the value y; or alternatively it is proportional to the maximum probability of the configuration with x_u fixed to the value y.

In a similar way, the edge min-marginal is defined as

$$\Phi_{uv;yy'}(\theta) = \min_{\{x \in \mathbb{X}^n | x_u = y, x_v = y'\}} En(x;\theta) + \text{const}_{uv} \tag{3.23}$$

where const_{uv} is a constant independent of y and y'. Similarly, the edge min-marginal is proportional to the minimum energy of the configuration under the constraint $x_u = y$ and $x_v = y'$. If one can compute such min-marginals, then for a tree structured distribution, they characterize an exact solution to the global MAP problem. Hence, in order to find the MAP solution on trees, one needs to compute the min-marginals. Dynamic programming algorithms are widely applied for this task. The general idea is based on the generalized distributive law [7] and the method is called the *max-product* algorithm. We show the procedure by a simple example. Suppose we want to the compute the minimum of a function F:

$$F^* = \min_x f(x_1, x_2, \ldots, x_6), \qquad x = \{x_1, x_2, \ldots, x_6\} \tag{3.24}$$

where the function F itself factorizes into smaller functions over subsets of variables:

$$F(x_1, x_2, \ldots, x_5) = f_3(x_4, x_5, x_6) f_2(x_2, x_3) f_1(x_1) \tag{3.25}$$

The minimization in equation 3.24 can alternatively be written as:

$$F^* = \min_{x_4, x_5, x_6} \left\{ f_3(x_4, x_5, x_6) \min_{x_2, x_3} \left\{ f_2(x_2, x_3) \left\{ \min_{x_1} f_1(x_1) \right\} \right\} \right\} \tag{3.26}$$

where we have split the minimization operation into minimization over subsets of variables and pushed the min operation as far as we could into the equation. This way, min-marginalization is performed locally and more efficiently. Each of the marginalizations passed to another variable is called a message. The max-product algorithm works in a similar way and passes messages

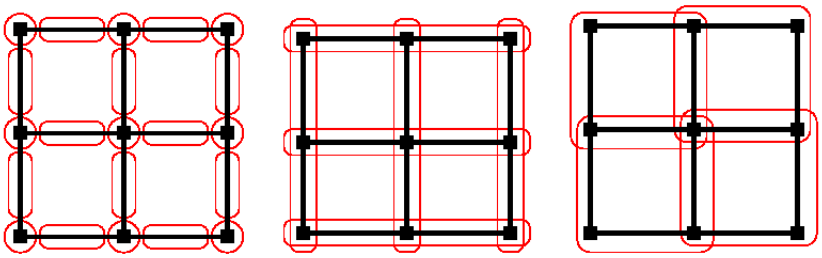


Figure 3.3: Different decompositions of a grid graph [136]. From left to right: the decomposition used in maxsum diffusion[134], the decomposition used in [79], decomposition to short cycles.

between neighboring nodes. The procedure is as follows. First one chooses a node as the *root* node. The choice is arbitrary but otherwise fixed. Based on this choice and starting with the leaf nodes, the method passes messages from leaves down to the root. This is called forward pass. After the forward pass, the min-marginal of the root node can be computed. Another pass of messages in the reverse direction is required to compute min-marginals on all other nodes. This pass is called backward pass.

This means that instead of globally optimizing the distribution, one can perform local operations on nodes and edges to find a globally optimal solution. In fact, if local minimizers of each node's min-marginal are unique, a globally optimal solution can be found by choosing the state at each node which minimizes the min-marginal [131]. In the case that the local minimizers are not unique, one needs to take a different approach. The solution deriving process starts with the root node, in which a local minimizer is chosen as the solution. Traversing the tree from the root up to the leaves, the cost-to-go functions are used for choosing the solution at a child node. Further details can be found in [129].

3.5.2 Loopy Graphs

The max-product algorithm may be viewed as an algorithm to solve the LP relaxation for a loopy graph. However, it is not guaranteed to converge in this case. In addition, in the case of convergence it may not produce correct min-marginals of a loopy graph. In spite of these facts, max-product algorithm has been used on loopy graphs and produced solutions useful in practice. Such methods are generally known as *belief propagation*. There has been some effort

to provide convergent forms of these algorithm such as the work in [79]. The approach taken in such methods is to define a lower bound on the original objective function. Then the goal becomes maximizing this lower bound. The problem of maximizing the lower bound turns out to be the lagrangian dual of solving the original relaxed problem. In order to define a lower bound, first the original problem is *decomposed* into a smaller set of subproblems on each of which exact inference is tractable. A criterion for the decomposition is that every node and edge should be at least covered by one subproblem. A requirement for the applicability of the approach is that exact inference on each of subproblems should be possible efficiently. Then by combining the solutions obtained from each subproblem, a solution to the primal problem is obtained. If the values of the dual and the primal programs (*i.e. the lower bound and the energy functional*) coincide then the optimal solution is found.

As an example of the methods in this category is the max-sum diffusion approach [134]. In this method the subproblems are assumed to be the smallest possible subproblems, that is individual nodes and edges. This iterative method seeks to find a consistent labeling by maximizing a lower bound on the objective function. It works by performing equivalent transformations on the nodes and edges connected to them so that the min-marginals of nodes coincide with the edge min-marginals. An equivalent transformation is basically a transformation which leaves the objective function unchanged but forces the local marginals to be consistent with each other. The method terminates at a point characterized by arc-consistency which simply means that in the final solution, single node and edge min-marginals agree with each other. As pointed out in [134], this condition although necessary, is not sufficient for optimality. The exception applies to the convex problems having binary labels ($\{0, 1\}$) in which arc-consistency is also sufficient for optimality.

Inference in the decomposition framework can be speeded-up by considering larger sub-problems as shown for example in [79]. There, the subproblems are assumed to be monotonic chains with respect to a pre-specified ordering on nodes. We will discuss and explain this approach and also another approach based on sub-gradients [81] in more detail during our discussion in the following chapters where they are used for inference.

The differences between the various algorithms based on graphical decomposition stems from two factors: how the subproblems are chosen and how the dual objective function is solved.

The first problem is not fully addressed in the literature. The common idea is that the larger the subproblems, the faster the convergence will be. Regarding the second issue, there are different methods for bound maximization. We will employ two of them for solving the dual objective function in the next chapters. One uses a fixed point update whereas the other uses sub-gradients of the dual function to solve it.

As a final comment on inference, we note that exact inference on loopy graphs with low tree-width is feasible by converting the loopy graph into an equivalent tree and using junction tree algorithms [130]. The tree-width of a loopy graph is defined as the highest cardinality of the cliques of its equivalent junction tree.

3.6 Summary

In this chapter we provided some background material on graphical modeling. The relation of graphical distributions to Gibbs distributions and their interpretation as Markov random fields were discussed. The driving forces behind the application of such tools in image analysis problems were explained. The discussion was then followed by inference methods on acyclic and cyclic graphs, providing some intuition on how a group of the methods for MAP inference in graphical models attempts to solve a labeling problem.

Chapter 4

Image Matching

4.1 Introduction

Estimation of a deformable mapping between a pair of images is useful in many image analysis problems. Such mappings find applications for example in joint segmentation-classification [139] or joint segmentation-deformation [75] approaches. In the context of face recognition, estimating a deformable mapping can be beneficial from different viewpoints. First of all, finding an alignment between images is an integral part of an object matching and recognition process. For face recognition, the procedure requires the images to be registered using at least two points corresponding to the eye coordinates. Any errors in this registration can seriously degrade the performance. In addition, because faces are non-rigid objects, inclusion of a dense correspondence information in recognition can enhance the performance by providing better means for comparison of different parts of faces subject to intra-class deformations and may be also used to decouple shape and texture information. On the other hand, in the case of in-depth rotation of head, in presence of such information the similarity criterion can be based on the similarities of visible regions of both faces which facilitates a direct comparison which focuses on observed evidence. This is in contrast to generative models in which the methodology is based on inferring new information using the training data. In the latter scenario, the method generates a new face (or equivalently infers the parameters or features of a new face) in a desired condition. The accuracy of this reconstruction is compromised by the information available for learning which in turn imposes a limitation on what these approaches can infer. A

further benefit of having dense correspondence information is that the recognition method can work with 2D images, obviating the need for multiple gallery images for training or the need for 3D data to handle out-of-plane rotation of the head.

In this chapter, we will review an image matching method [114] formulated on MRFs. We will use this method as part of our approach for face recognition to be discussed in the following chapters. After reviewing related approaches in section 4.2, the deformable image matching method is discussed in section 4.3. Section 4.4 explains the optimization algorithm adopted for inference and provides some image matching examples. We bring the chapter to a conclusion in section 4.5.

4.2 Related Approaches

The use of MRF models for image matching and optical flow estimation dates back to decades ago. In [83, 63] the authors considered coupled vector-binary MRF models for optical flow estimation in the presence of discontinuities and occlusion. There, the vector field represented displacement vector and the binary field modeled discontinuities in the displacement field, often occurring at boundaries of objects. The prior employed for discontinuities imposed linearity assumptions on object boundaries and prevented them from intersecting. The prior model employed on the motion vector imposed smoothness in regions where discontinuities were not present. For the optimization of the cost function, in [83] simulated annealing and in [63] iterated conditional modes are employed. In [145] the authors in addition to the smoothness and line constraint for discontinuities, introduced an additional binary field component into the model representing a segmentation, identifying areas of uncovered background due to object displacement for which the correct motion field cannot be computed. In order to find the MAP solution mean field approximation was used.

In [31] magnitude and orientation of the optical flow are modeled using separate Markov models. First, the problem of estimating the orientation of flow is solved and then the magnitude of flow is estimated. The authors formulate the estimation of orientation and magnitude of optical flow as max-flow problems. In [84], the authors introduced swap and expansion algorithms considering motion estimation as an example for their approach. Labels were allowed to take

their values from a product set of admissible displacements in horizontal and vertical directions. As a result, in this approach the optimization became intractable as soon as disparities in each direction grew. In another work [71] the authors presented a method for computing piecewise rigid deformation in a video sequence. Local rigid transformations were modeled by similarity transformations. For the optimization of the proposed MRF model first the sum-product Belief Propagation algorithm (BP) was applied. The initial estimate was then refined using graph cuts based methods. In the works in [54, 82] towards estimating a dense registration between image pairs, the authors tackle the problem by using a set of control points and an interpolation strategy to establish correspondences between control points. In order to handle large deformations the approach takes advantage of a multi-scale approach. The optimization of the cost function is performed via the primal-dual schema. In [48] max-product Belief Propagation is used to estimate optical flow between image pairs. The authors proposed efficient algorithms for message computation along with a multi-grid approach to enhance efficiency of their algorithm.

Apart from the MRF-based approaches, other methods for deformable image matching exist in the literature. The method in [50] proposes a direct method for non-rigid image registration with occlusion reasoning. The method uses the color discrepancy between images and a regularization term for obtaining smooth deformations. For the minimization of the cost function the Gauss-Newton algorithm is used. Other work in [70] proposes a framework for non-rigid registration based on free-form deformations and a multiresolution approach to speed up registration. The works in [121] and [17] are some other examples of non-MRF based methods.

4.3 Deformable Image Matching

The problem of deformable matching involves estimating a deformation map between a pair of images subject to maximizing/minimizing a criterion function value. The problem is illustrated in Figure 4.1. In an MRF formulation, the criterion function is defined as the probability of a configuration on a considered graphical model or inversely as the energy of the match. In the method to be discussed, the probability function involves unary and pairwise factors. The unary factors measure the fidelity of the estimated deformation to the data (color map).

In general, it is assumed that the deformation does not exhibit sharp discontinuities. In order to

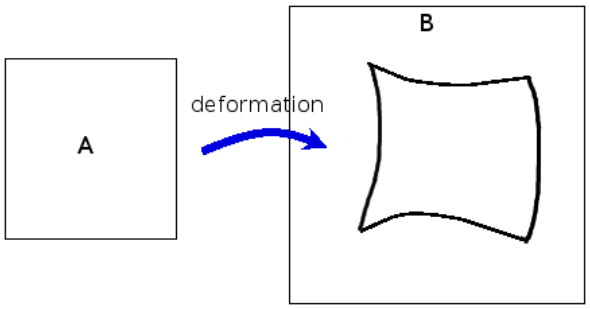


Figure 4.1: Estimating a deformation which maps image A onto (a sub-image of) B.

impose such a constraint, pairwise smoothness terms are included in the model. The advantages of the approach over some other methods are two fold. First, it allows for a considerable amount of image variations while giving a compact representation of the optimization task in an efficient fashion. The efficiency draws on the idea of decomposing 2D displacement vectors into two 1D disparities. Secondly, the method uses a successful method for optimization, *i.e.* TRW-S [79], which compared to some other approaches for optimization in Markov random fields achieves lower energy values with certain optimality guarantees[119].

4.3.1 Deformation Model

The matching problem can be formulated in a maximum a posteriori inference framework on a probabilistic graphical model. In such a formulation, the likelihood encodes the fidelity of the mapping in terms of pixel values and the prior imposes a smoothness on the deformation. Let A and B be two images for which we want to estimate a relative deformation. Also let x be the configuration of the graphical model corresponding to an injective mapping, meaning that not all pixels from B will have a corresponding region in image A. The situation is illustrated in Fig. 4.1 where the image A is mapped into a subimage of B. Using equation 3.18 the probability of a labeling x , representing a deformation between images A, B, can be written as

$$P(x) \propto \prod_{u \in \mathscr{V}} \Psi_u(x_u) \prod_{(u,v) \in \mathscr{E}} \Psi_{uv}(x_u, x_v) \tag{4.1}$$

where $\mathscr{V}$ denotes the node set of the graph and $\mathscr{E}$ denotes the edge set. The edge set $\mathscr{E}$ in the considered model consists of pair of nodes in an immediate four-connected neighborhood system in a lattice structure. In the above formula we have omitted the normalization constant

since it does not have any role in the maximization process. The likelihood term, $\prod_{u\in\mathscr{V}}\Psi_u(x_u)$, is based on the assumption that pixel values are conditionally independent. This means that when a site u in image A is mapped to the site $d(u)$ in image B, its signal, A_u, is only dependant on the corresponding signal $B_{d(u)}$ and independent of the rest of signals in B. The conditional distribution $p(A_u|B_{d(u)})$ is assumed to be the same and fixed over the whole configuration and is given by a Gaussian noise model. That is, the subimage of B to which A is mapped is assumed to be an observation of image A under the Gaussian noise model.

The prior of the distribution represented by $\prod_{(u,v)\in\mathscr{E}}\Psi_{uv}(x_u,x_v)$ makes the model favor smooth mappings. Here, the prior is modeled in terms of pairwise potentials and consequently the probabilistic graphical model is also a pairwise MRF.

Let $\mathbb{X}=\{1,2,\ldots,L\}$ denote a discrete set of admissible states/labels. Let $\mathscr{G}$ be a graph with the node set $\mathscr{V}$ and the edge set $\mathscr{E}$. The goal is to assign each node of the graph $\mathscr{G}$ a label from the set $\mathbb{X}$. The configuration of the MRF (*i.e.* the labeling) is denoted by $x=(x_1,x_2,\ldots,x_n)\in\mathbb{X}^n$, where n is the number of nodes in the graph. For deformable image matching, the label of a node denotes a 2D displacement vector such that when added to the coordinates of the site under consideration in image A, it yields the coordinates of its corresponding site in the target image, *i.e.* image B. If we denote the unary and pairwise potentials of the model by θ_v and θ_{uv} respectively, then the energy of a configuration is:

$$En(x;\theta)=\sum_{u\in\mathscr{V}}\theta_u(x_u)+\sum_{(u,v)\in\mathscr{E}}\theta_{uv}(x_u,x_v) \tag{4.2}$$

As noted earlier, the minimum of the energy above corresponds to a maximum a posteriori probability of the a Gibbs distribution for the configuration. Unary θ_u and pairwise potentials θ_{uv} of the model will be discussed more explicitly in the following sections.

4.3.1.1 Product Model

Assume that the configuration x with components x_u is a 2D displacement vector denoted by $x_u=(x_{u1},x_{u2})$. With a little abuse of notation let u correspond to the coordinates of a site in image A. Then deformation d implied by the configuration x, maps u into the position $u+x_u$ in image B.

In such formulation of the energy functional, considering for example L pixels of disparity in each direction, each node would have L^2 admissible states. The model constructed this way is called the *product model* since the admissible state set for each node is the Cartesian product of the two sets representing allowed displacements in each direction. Because of the high computational complexity, inference in this model becomes intractable as soon as L becomes large. The limitation is posed by the complexity of the optimization process (in particular the message passing operation in [79]) which is proportional in complexity to $\mathcal{O}(L^4)$. The message computation involves a min-marginalization over the state space of two neighboring nodes. As a result, because the cardinality of the state space is quadratically proportional to the dimension of disparities in each direction, as soon as the disparity in each direction grows, the min-marginalization becomes highly inefficient.

In order to make the method more efficient, the authors in [114] proposed the so-called *decomposed model*. The idea is to consider the disparity in each direction separately, in other words they *decompose* the label set into two sets, each representing displacement in one direction. This then requires to consider two MRFs (one for each direction) which should be optimized at the same time. In this case the computational complexity of message computation in the model reduces from $\mathcal{O}(L^4)$ to $\mathcal{O}(L^2)$. The complexity of message passing operation can be further reduced to $\mathcal{O}(L)$ for special types of potentials. Thus, a larger range of displacements can be handled more efficiently.

The other important factor in the complexity of optimization is the number of variables. The complexity of inference in an MRF grows with the increasing number of discrete variables. Thus, reducing the number of nodes in a model can reduce the computational cost leading to further efficiency. Hence, instead of assuming each pixel as a node, the authors in [114] group pixels together in the form of blocks in their decomposed model.

4.3.1.2 Decomposed Model

In the *decomposed* model the idea is to model the displacements in each direction (*i.e.* horizontal and vertical) separately. Clearly, the estimation of displacement vectors cannot be decoupled completely. In other words, the two models representing displacements, should interact with each other.

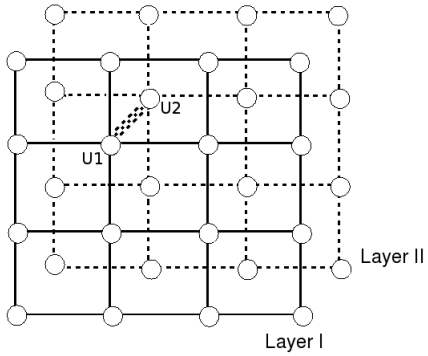


Figure 4.2: Two MRFs used in the decomposed model along with a sample inter-layer edge.

Each of these MRFs are called a *layer*. In each layer, the continuity term between every two neighboring nodes in the lattice is imposed using intra-layer edges. Another set of potential functions are responsible for the interaction between the two layers. Figure 4.2 shows the two layers interacting through one sample inter-layer edge. The inter-layer edges in this representation encode the data term, that is the unary cost of matching a site into its corresponding region in the other image. The graph $\mathscr{G}$ in this representation is constructed as follows. The node set is comprised of two sets of nodes as

$$\mathscr{V} = \mathscr{V}_1 \cup \mathscr{V}_2 \tag{4.3}$$

The set $\mathscr{V}_i$ represents the node set in layer $i \in \{1,2\}$. The edge set of the graph $\mathscr{G}$ is defined as

$$\mathscr{E} = \mathscr{E}_1 \cup \mathscr{E}_2 \cup \mathscr{E}_{12} \tag{4.4}$$

the sets $\mathscr{E}_1$, $\mathscr{E}_2$, $\mathscr{E}_{12}$ correspond to edges in layer 1, edges in layer 2, and edges between the two layers, that is, inter-layer edges.

4.3.2 Unary Potentials

The unary potentials penalize assigning a site in the model image into its corresponding site in the test image based on color deviation. In [114] a block model is employed to construct the unary potentials. In this model, pixels are grouped into non-overlapping blocks of pixels. The blocks are of size 4×4. Then the unary costs are defined as a sum of the color deviations between every pixel in the model block and the test block. Considering a block model is advantageous from two points of view. First, the computational cost of inference can be reduced as a

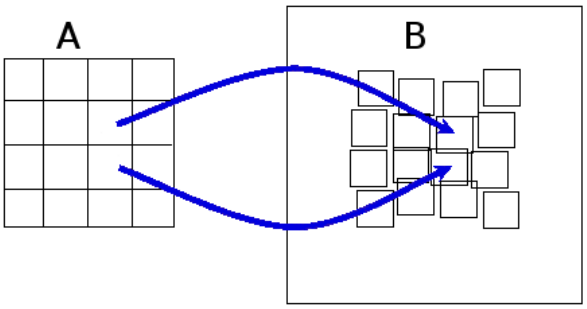


Figure 4.3: Mapping image A to image B on a block-by-block basis [114].

result of decreasing the number of sites involved in the MRF. Grouping every 16 pixels into a block and considering it as a single site, reduces the computational complexity of inference in each layer by a factor of 16.

The grouping also offers another advantage in practice. The motivation behind trying to design an image matching method is geometric distortion and texture distortion. That is, the target image, B, is assumed to be a noisy observation of the template image, A, both in terms of its texture and shape. The noise present in the texture can be partly corrected by using a mean filter. Grouping pixels together as blocks essentially acts as a mean filter on the likelihood of the model. Considering the block model, the data term is formed by the contribution of single pixels inside a block:

$$\theta_{u_1u_2} = \sum_{s\in u} -\log p(A_s, B_{s+(x_{u_1}, x_{u_2})}) = \frac{1}{2\sigma}\sum_{s\in u}(A_s - B_{s+(x_{u_1}, x_{u_2})})^2, \qquad (u_1,u_2)\in \mathscr{E}_{12} \tag{4.5}$$

where u_1 and u_2 are two isomorphic nodes in layers 1 and 2 corresponding to a single block u in image A. s denotes the coordinates of a pixel inside block u and (x_1, x_2) denotes the labels of nodes u_1 and u_2, that is a 2D displacement vector.

4.3.3 Pairwise Potentials

The pairwise terms impose continuity over the deformation field. All deformations in this approach are supposed to be representable by local displacements. Relative displacements by one pixel between blocks of pixels are penalized less whereas larger than one pixel relative displacements are penalized more. This way, neighboring pixels are encouraged to be assigned to nearby positions so that the deformation becomes smooth. The pairwise potentials are iden-

tically defined in each layer:

$$\theta_{uv}(x_u, x_v) = \begin{cases} 0, & x_u = x_v, \\ c_s & |x_u - x_v| \leq 1, \\ c_g & |x_u - x_v| > 1, \end{cases} \tag{4.6}$$

where $(u,v) \in \mathscr{E}_1 \vee (u,v) \in \mathscr{E}_2$ and $c_g \gg c_s$.

The subclass of the employed pairwise terms with low penalty (c_s) leaves one pixel overlaps/separations of neighboring 4×4 blocks less penalized. When two neighboring blocks overlap by one pixel the scale change would be $1 - \frac{1}{4}$. In contrast when they separate by one pixel from each other the change in scale is $1 + \frac{1}{4}$. As a result, the model accommodates scale changes between $[1 - \frac{1}{4}, 1 + \frac{1}{4}]$ and also a certain degree of flexibility.

4.4 Optimization

In order to infer the most likely configuration of the model, the sequential tree re-weighted message passing algorithm (TRW-S for short) [79] is used. The method is a graph decomposition method which essentially tries to solve the lagrangian dual of the original relaxed problem. In this method, the overall distribution is decomposed via a *convex combination* into a series of smaller distributions, *i.e.* trees. Considering the exponential family representation for the graphical model and a probability for the parameters of each its component probability distributions (trees), ρ_T, the original probability distribution with potential parameter θ is expressed as a convex combination of the potentials of the subproblems (θ_T):

$$\theta^T \phi(x) = \sum_T \rho_T {\theta_T}^T \phi(x) \tag{4.7}$$

The above combination of tree-structured distributions is called a convex combination if all $\rho_T \geq 0$ and $\sum_T \rho_T = 1$. Although the decomposition is not unique and the final solution does not depend on the choice of subproblems (trees), the convergence rate of the algorithm may be affected by different choices. One requirement for the decomposition is that every node and every edge should be covered by at least one tree. A good choice as suggested in [79] is to use the rows and columns of a 2D lattice as subproblems. The MAP problem involves finding the minimum of the LHS of the Eq. 4.7, or equivalently solving for the minimum of the RHS:

$$\min_x \theta^T \phi(x) = \min_x \sum_T \rho_T {\theta_T}^T \phi(x) \geq \sum_T \rho_T \min_x {\theta_T}^T \phi(x) = LB \tag{4.8}$$

where the last inequality follows from Jensen's inequality [136]. This essentially means that the original parameter θ is decomposed among a set of trees according to the probability ρ_T. From the above formula it can be seen that the last term provides a lower bound for the minimum energy. The tree reweighted message passing methods try to solve the problem by maximizing this lower bound. The problem to be maximized is defined as

$$\max_{\sum \rho_T \theta_T \equiv \theta} LB = \max_{\sum \rho_T \theta_T \equiv \theta} \sum_T \rho_T \min_x {\theta_T}^T \phi(x) \tag{4.9}$$

In the above formula, the sign "$\equiv$ "in the constraint set corresponds to reparameterizations such that the energy remains unchanged. This does *not* necessarily mean $\theta = \sum_T \rho_T \theta_T$ as illustrated in [79]. Maximization of the above function turns out to be the lagrangian dual of minimization of the original relaxed problem [131]. The maximization problem above is performed via reparameterization of the original energy functional using max-product message passing [79].

4.4.1 Min-marginals of the Loopy Graph

Let us have a high level look at the solution finding process. We have an original problem which is difficult to solve. This problem is then represented as the convex combination of a series of easier problems, that is tree distributions. As discussed earlier in Chapter 3, the solution on tree distributions can be found exactly via max-product algorithm requiring only two passes of messages, forward and backward. Hence, we solve each subproblem resulting from a convex combination using the max-product message passing algorithm. If the solutions on each tree found in this way do agree over all trees for any common node between them, then the final MAP solution is easily extracted and is the global minimum of the energy functional. However, in order for this to happen the min-marginals on all trees should agree with each other. In order to make these min-marginal equal for the subproblem one averages min-marginals in each iteration over the trees sharing them. When the solutions obtained on different trees agree on common nodes, the *strong tree agreement* condition has been achieved and the solution is the global minimum [131]. However, as shown in [79] this condition is not always achieved. The stopping criterion of the algorithm in general is characterized by the *weak tree agreement* condition which implies local consistency of solutions which is only necessary and not sufficient

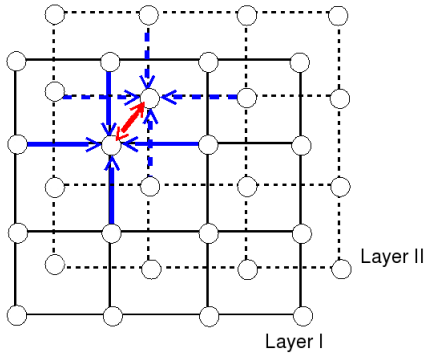


Figure 4.4: Message passing on the two interconnected MRFs.

for global optimality. A special case when the algorithm is guaranteed to find the exact solution is the family of binary problems, that is with label set $\{0,1\}$ and the energy being convex. The method is essentially an iterative one where in each step the univariate potentials associated with a vertex are distributed among subproblems in order to make the min-marginals equal for all trees. In [79] the tree subproblems are selected as chains and the computation of min-marginals is combined with updating the univariate potentials. The chains considered have to be monotonic with respect to an ordering on the vertex set $\mathscr{V}$. Let the orientation of the edges in our graph represented as $(u,v) \in \mathscr{E}$, imply $u < v$. Let each edge of $\mathscr{G}$ be covered by exactly by one chain. Also let the auxiliary variables $M_{uv}^{fw}(x_v)$ denote the forward messages and $M_{uv}^{bw}(x_u)$ stand for backward messages. In this case the univariate update of potential at each node of a subproblem T, $\theta_{T;u}$, is defined as:

$$\theta_{T;u}(x_u) = \frac{1}{n_u}\left(\theta_u(x_u) + \sum_{(v,u)\in\mathscr{E}^T} M_{vu}^{fw}(x_u) + \sum_{(u,v)\in\mathscr{E}^T} M_{uv}^{bw}(x_u)\right) \tag{4.10}$$

where n_u is number of trees containing the node u. A complete description of the algorithm is given in Table 4.1.

In Table 4.1 n_{term} is the number of chains in which node u serves as the last node, that is $\nexists(u,v) \in \mathscr{E}$. The messages M^{fw} are updated during step 3. After performing step 4, the next sweep will update values M^{bw}.

Table 4.1: TRWS on Monotonic Chains [79]

1. $M_{uv}^{fw}(x_v) = 0$ and $M_{uv}^{bw}(x_u) = 0, \forall (u,v) \in \mathscr{E}, x_u, x_v \in \mathbb{X}$
2. $LB = 0$
3. Select u in the increasing order and perform
   (a) average min-marginals $\Phi_u(x_u) = \frac{1}{n_u}\left(\theta_u(x_u) + \sum_{(v,u)\in\mathscr{E}} M_{vu}^{fw}(x_u) + \sum_{(u,v)\in\mathscr{E}} M_{uv}^{bw}(x_u)\right)$
   (b) Compute $M_{uv}^{fw}(x_v) = \min_{x_u}\left\{\Phi_u(x_u) - M_{uv}^{bw}(x_u) + \theta_{uv}(x_u, x_v)\right\}\ \forall v, (u,v) \in \mathscr{E}$
   (c) $LB = LB + n_{term}\min_{x_u}\Phi_u(x_u)$
4. Reverse the ordering on $\mathscr{V}$ and swap M^{fw} and M^{bw} and go to step 2.

Table 4.2: Comparison of the complexity of product and decomposed models.

	Product Model	Decomposed Model
Complexity of passing one message	$\|L\|^4$	$\|L\|^2$
Number of relaxed variables	$\|\mathscr{V}\|\|L\|^2$	$2\|\mathscr{V}\|\|L\|$
Number of discrete variables	$\|\mathscr{V}\|$	$2\|\mathscr{V}\|$

4.4.2 Comparison of the Decomposed vs. Product Model

A comparison of the decomposed model to the product model is provided in Table 4.2. In both models, certain types of messages can be computed faster using distance transform [47]. In the case that pairwise potentials are separable in the product model, the complexity of one message computed using distance transform reduces from $\mathscr{O}(L^4)$ to $\mathscr{O}(L^2)$. In the decomposed model, two types of pairwise potentials are employed: inter-layer and intra-layer. The computational complexity of intra-layer messages can be reduced from $\mathscr{O}(L^2)$ to $\mathscr{O}(L)$. In addition computing the relaxation of the product model has been found to be more memory-demanding and therefore the decomposed model is preferred.

4.4.3 Extracting the MAP Solution

If for each node, the associated min-marginal attains its minimum at a unique value, and all such min-marginals are equal for all trees containing that node, then a MAP configuration can be extracted by choosing local minimizers. It is shown that upon reaching this condition, the solution is the true MAP estimate of the problem [131]. However as shown in [79] this condition cannot always be achieved. The fixed point of the algorithm is characterized by the weak tree agreement condition. This condition states that the fixed point of the algorithm reaches a point which is locally consistent. On the other hand local consistency is only necessary and not sufficient for global optimality. This means that upon convergence there might be multiple local minimizers at each node. This can be partly compensated for by the following heuristic [79]: using the same ordering as in the message passing operation for each node, choose the label which minimizes the following quantity:

$$\theta_u(x_u) + \sum_{(v,u)\in\mathscr{E}} \theta_{uv}(x_u, x_v) + \sum_{(u,v)\in\mathscr{E}} M_{uv}^{bw}(x_u) \tag{4.11}$$

This procedure reduces the occurrence of multiple minima but does not solve it completely.

Figure 4.5 illustrates some results of employing the image matching technique we reviewed. We provide examples of using the approach on both synthetically and naturally deformed images. Also, we visualize the results of applying the method on noisy images. The reason for this is that in the experiments to be presented in the next chapters we will employ the method to match facial images of the same or different subjects to each other. The process can be alternatively considered as matching noisy images of the same subject to each other. Apparently, the degree of assumed noise would often be much greater when matching different subjects compared to the case of matching images of the same subject. In the figure in each row we match the template image to the target image. The results are illustrated by warping the template images according to the found displacement vectors. Visually, the more similar the warped template to the target image, the better the match is. In cases where the template image is smaller than the target image, one can visualize the dark lines between neighboring blocks in the warped template images which shows separation of blocks to accommodate scale changes. For a more detailed experimental evaluation of the algorithm the reader is referred to [114]. Quantitative results of using the mentioned method is reported in Table 5.4 in an identification scenario on the CMU-PIE database.

Upon establishing the correspondences one may employ different measures such as sum of squared differences, correlation or mutual information in order to quantitatively measure the match quality.

4.5 Summary

In this chapter we overviewed the deformable image matching method in [114]. The method, can handle a reasonable degree of deformation between a pair of images. The technique has two outstanding characteristics. First, it employs a successful optimization technique which outperforms others in MAP inference in MRFs. Secondly, by applying the idea of label decomposition the method is made more efficient, enabling it to handle a larger range of deformations and displacements. The optimization method employed was briefly overviewed. Finally, we provided some examples of matching using the matching technique. As illustrated by the examples, the method provides very good results for this application.

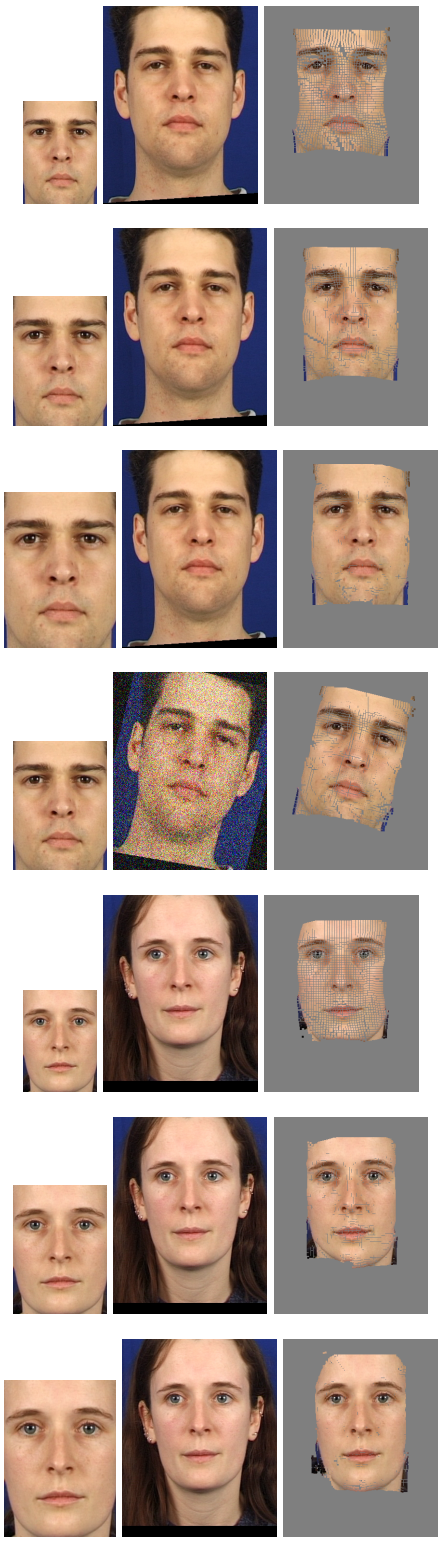


Figure 4.5: In each row from left to right: template, target and deformed template.

Chapter 5

Face Recognition Based on Image Matching

5.1 Introduction

In this chapter a face recognition method using the image matching algorithm in Chapter 4 is studied. Roughly speaking, the method uses the energy required to transform one image into another (both in terms of its texture and its shape) for classification. In general, in a recognition scenario using graphical models, the goodness of a match is often gauged in terms of the *maximum a posteriori probability* of the corresponding configuration of the Markov random field (MAP-MRF). Or inversely, the posterior energy considered is taken as the cost of matching. The method proposed in this chapter follows this general framework for recognition/verification of faces under arbitrary pose with the restriction that only frontal images are available as class exemplars.

A general frontal face recognition algorithm consists of different stages. After enrollment, the image is geometrically normalized using a number of landmark points. Usually eye coordinates (detected automatically or manually) are used for this purpose and the face image is geometrically normalized so that the landmark points are transformed to pre-specified locations. Next, photometric normalization is applied to reduce the effects of uneven illumination conditions. The next steps are feature extraction and classification where in feature extrac-

tion one uses characteristics of facial images which can provide maximum between-class to within-class scatter and for classification a suitable decision rule is employed to discriminate between classes. In a more general case where change of pose is present additional steps are required. As noted earlier in the literature review section these steps differ from one method to another. Generally in all but graph-based methods, before classification, one may employ pose detection, pose correction, feature transformation based on non-frontal training data or other additional stages. As a result non-frontal training data is required for classification. In addition, if pose is required to be known beforehand then the performance of such methods would be very dependent on the accuracy of the pose estimation module. On the other hand, graph-based approaches use the maximum a posteriori probability/minimum energy of a match for decision making.

In comparison to the existing approaches some of the distinguishing characteristics of the proposed method can be outlined as below:

- The proposed method circumvents the need for geometric pre-processing of face images (often done manually) by encapsulating an image matching technique as part of face recognition. As a result it can cope with moderate translation, in and out of plane rotation, scaling and perspective effects. This is very important as residual misalignments, remaining after geometric normalization of face images based on automatic face detection and localization, can seriously degrade the performance of face recognition systems. The misalignment problem is particularly pertinent as the automatic detection of facial landmarks used for geometric normalization is specially challenged by pose, lighting or expression changes.

- Non-frontal images are not needed for training. This is particularly advantageous for a number of reasons. First, it sidesteps the time consuming learning processes using images under different poses. Second, in applications where there is not enough data for training, the performance of the systems using large training sets degrade. Last but not least, with a limited number of training data, modeling the nonlinear structure of face images under pose variations is difficult. As a result, the learned feature transformations are not always applicable to new images or new databases. This makes the performances of such systems very dependent on the particular database used for training. Although

the proposed approach can achieve comparable performance on extreme poses without using non-frontal images for training, the application of such models can enhance the performance.

- In the proposed method, no strict assumption is made about the pose of the subject prior to matching and hence the system is better suited to more realistic scenarios. In contrast to other solutions, where separate models are used for each pose and the test image needs to be first aligned with the appropriate model, this eliminates the dependency of face recognition system on the accuracy of the pose estimation module.

- In order to reduce the problems introduced by self occlusion in the case of a pan movement only half of the face is used for matching and recognition. The decision whether a pan component is present or not is made by comparing the normalized energies of the full-face vs. half-face matches.

- In comparison to the state-of-the-art approaches based on 3D models the proposed approach operates on 2D images which bypasses the need for 3D face training data and the vagaries of 3D face model to 2D face image fitting.

- Last but not least, from the point of view of object recognition, the matching energy in MRF-based approaches (using at most pairwise potentials) has certain drawbacks and should not be used as a similarity criterion for hypothesis selection directly. The main shortcomings of the energy are identified and a plausible energy normalization scheme is proposed and discussed. In fact, one may directly incorporate a global interaction potential into the underlying MRFs *e.g.* as in [135, 80, 110] and optimize the energy including the higher order potential. However, in our case, because of the huge configurational space, which results in inefficient marginalization over the higher order cliques, we propose to match the images using at most pairwise potentials and then normalize the underlying energy for recognition which is more efficient. Clearly, the gained efficiency may come at the risk of the quality of match being partially compromised. However, the choice between viability and perfection seems to be rather stark and we have opted for the former.

The chapter is organized as follows: In section 5.2, we first consider important modifications

to the matching method introduced in Chapter 4. The modifications include using edge maps instead of color/greyscale in order to reduce the effects of uneven illumination and replacing the crisp penalty functions with soft quadratic potentials in order to achieve more flexibility in the deformation. Next, a deformable block matching method is proposed in which instead of using similar-shaped blocks, a global transformation is used to modify block shapes to cope better with spatial transformations. Finally, a heuristic label pruning is introduced in order to speed-up inference on the graphical model.

Having established a suitable method for face matching, a classification method is proposed for decision making in section 5.3. The method is essentially an *energy normalization* method [8, 9] which brings the energy of an established match to a state where it can serve better as a criterion for decision making. Evaluations of the proposed method on different databases are presented and discussed in section 5.4. The chapter is brought to a conclusion in section 5.5.

5.2 Modifications to the Matching

5.2.1 Unary Potentials

Unary potentials measure the degree of similarity/dissimilarity of the template and the target images in terms of their textural content. Any unwanted changes and noise introduced into the texture can adversely affect the accuracy of a match. Edge maps have been extensively used in image analysis for feature extraction and matching [90, 21]. This is due to the fact that most of the time the discriminative features of an object lie near the edges. In addition, edge maps provide invariance properties against illumination conditions to some extent. Motivated by these observations, for the computation of the data term we use horizontal and vertical edge maps obtained by Sobel edge detector instead of color or grey scale images. Horizontal and vertical edges are scaled to the range [-1,1] and combined to form the data term:

$$\theta_{u1u2}(x_{u1},x_{u2}) = \frac{1}{2\sigma^2}[\mathrm{Dis}(I_u^{1h}, I^{2h}_{u+(x_{u1},x_{u2})}) + \mathrm{Dis}(I_u^{1v}, I^{2v}_{u+(x_{u1},x_{u2})})], u1 \in \mathscr{V}^1, u2 \in \mathscr{V}^2$$

where I_u^{1h} and $I^{2h}_{u+(x_{u1},x_{u2})}$ denote a block in the horizontal edge map of the first image and its corresponding block in the horizontal edge map of the target image, respectively. I_u^{1v} and $I^{2v}_{u+(x_{u1},x_{u2})}$ are defined in a similar way. Dis(.,.) stands for the sum of squared differences of pixels inside a block.

Since we are interested in comparing configurational arrangements of the entities of the model and scene images, it is desirable to rely more on the common features of the two images and bypass the atypical features which appear only in one image. This can be achieved by ignoring weak edges and setting them to zero and at the same time by truncating the data term. The truncation makes the matching more robust to outliers, occlusions or spurious features and structures.

5.2.2 Pairwise Potentials

Remember the method introduced in Chapter 4 uses high penalties for relative displacements of more than one pixel between neighboring blocks. These penalties limit the range of deformations which can be handled with the method. In order to achieve more flexibility in deformation, we replace the hard continuity term by a quadratic penalty function:

$$\theta_{uv}(x_u, x_v) = q(x_u - x_v)^2 \tag{5.1}$$

where $(u,v) \in \mathscr{E}_1 \vee (u,v) \in \mathscr{E}_2$ and q is a normalizing constant. In [114], by restricting the neighboring blocks (blocks are of size 4×4) to have relative displacements of no more than one pixel, the scale changes were limited to [.75,1.25] of the model image size, whereas by replacing the hard constraints by a quadratic term a much greater range of scales and deformations can be accommodated. The constant, q, is a parameter to control the trade-off between the data term and the smoothness of the deformation. The model is not very sensitive to the value of q if chosen in the range of $[10^{-2}, 10^{-3}]$. In our experiments, setting the value of the constant term, q, in the pairwise potentials to 5×10^{-3} was found to give reasonable results. This value depends on the range of input data (normalized to [-1,1] in our case) and controls the elasticity of the model.

5.2.3 Block Adaptation

Using the method described in Chapter 4, for each block of pixels in the template image, a corresponding block with a similar size and shape in the target image is found. This approach is not realistic when pose differences between the faces lead to different contractions, expansions or deformations of individual blocks in different parts of the face. In fact, such an approach

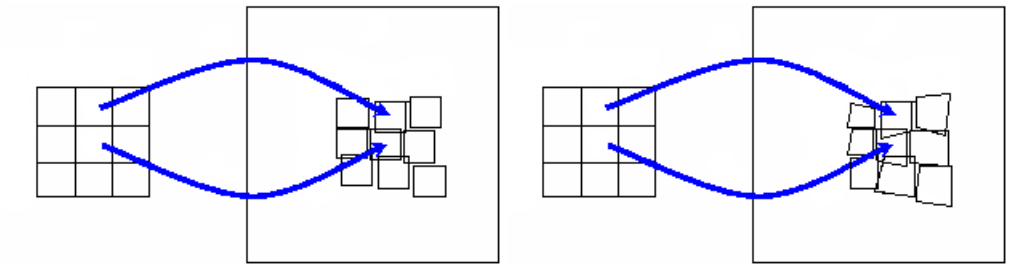


Figure 5.1: Left: blocks in [114], Right: blocks in the new deformable block scheme.

ignores any global geometric transformation between the template and the target images which is one of the omnipresent factors when matching objects viewed from different angles. Obviously, those parts of the object closer to the sensing device appear larger than the parts further away. In order to handle this effect, it seems appropriate to have denser sampling (smaller blocks) in the areas of contraction while coarser sampling (larger blocks) would be sufficient in areas of expansion. In the following, a method for handling this effect is proposed. Unlike some other approaches [13] which use training data, specific for each pose in order to estimate the deformations of local patches, the proposed method does not require training for estimating block deformations.

Essentially, we control the variation in block size and shapes by a global projective transformation. In order to estimate a global geometric transformation between two images we use the method described previously to find a set of corresponding points between the two images. Assuming that the underlying transformation between individual blocks is projective, a global spatial transformation is estimated using the Levenberg-Marquardt method [105]. Since there might be a few mismatches in some parts of the images, we use RANSAC to exclude outliers. In the second round of matching, each block on model image is warped according to the estimated transformation and the corresponding patch on the target image is sought. The proposed block adaptation method supports a more realistic sampling of signals subject to a global transformation while reducing the possibility of mismatches. Fig. 5.1 illustrates the block shape and size adaptation in comparison with the original method.

Considering Tr:

$$Tr = \begin{pmatrix} t_1 & t_2 & t_3 \\ t_4 & t_5 & t_6 \\ t_7 & t_8 & 1 \end{pmatrix} \tag{5.2}$$

as the estimated projective transformation between the images, the 2D spatial mapping of blocks can be interpreted as a combination of projective mapping and translational motion:

$$x_{u1} = (\frac{t_1x+t_2y+t_3}{t_7x+t_8y+1}) + \hat{x}_{u1}, x_{u2} = (\frac{t_4x+t_5y+t_6}{t_7x+t_7y+1}) + \hat{x}_{u2} \tag{5.3}$$

where x_{u1} and x_{u2} stand for horizontal and vertical displacements and x and y are coordinates of the block center. $\hat{x}_{u1}$ and $\hat{x}_{u2}$ are labels which are inferred in the second stage of matching. Since the projective transformation captures the dominant part of motion, the potential range of $\hat{x}_{u1}$ and $\hat{x}_{u2}$ can be reduced during the second round of matching, thus reducing the computational cost. Another advantage of the deformable-block matching method is its enhanced robustness against outliers in matching. In practice, the matching is not perfect and there might be parts of the model image which are not matched correctly to the unknown image. By reducing the search region in the second stage of matching and allowing the estimated global spatial transformation to carry the dominant part of the motion, this shortcoming is partly corrected.

5.2.4 Speeding up Inference by Label Pruning

As noted earlier for extracting the MAP solution the local minimizers are employed. Although the label with the minimum cost at each node might not correspond to the best solution when the number of iterations is limited (the inference is not exact and multiple minima exists), it is unlikely for a label with a high cost at a node in an intermediate iteration of the algorithm to correspond to the optimal solution at the end of optimization. Based on this observation, one can prune out labels which are unlikely to be optimal at each node (labels with larger costs) and meet only remaining admissible labels at each node during optimization. Pruning unlikely labels reduces the configurational search space, hence speeds up the method. In practice the following heuristic pruning is found to result in reasonable solutions:

After n_1 iterations, prune out up to n_2 least probable labels at each node based on their corresponding min-marginals ensuring that there are at least n_3 labels left at each node.

The choice of n_1, n_2 and n_3 depends on the difficulty of a specific task. The easier the problem, the smaller n_1 and n_3 and larger n_2. While the inference using tree reweighted message passing is based upon linear programming *relaxation*, the heuristic pruning method acts as a *hard propagator*. Although the pruning might sometimes lead to better results compared to the original method, in a limited number of iterations, it may sometimes introduce a trade-off between speed and accuracy. Care must be taken not to prune out the correct solution which may result in an erroneous final solution. Using the method described, each probe image is matched against all frontal gallery images. Some examples of matching using the deformable block matching method are illustrated in Fig. 5.2

5.3 Classification

Measuring the similarity between a pair of images contains two stages: matching the model image to the unknown probe image and then computing a similarity/cost function invariant to unwanted global spatial transformation and illumination variations. More explicitly, the problem can be described as follows: let A be the gallery image of a subject. Let B be the image of an unknown subject which depends on its geometrical parameters such as scale s, displacements d_x and d_y, rotation ϕ, perspective p and illumination conditions g, so that $B = B(s,d_x,d_y,\phi,p,g)$. Let $\mathrm{D}(A,B)$ be a dissimilarity function between the ideal image A and unknown image B. For a multi-class recognition task, the decision rule is:

$$assign \quad B \rightarrow \Upsilon_r \quad if$$

$$\Upsilon_r = \underset{all\ classes}{\operatorname{argmin}} \{ \min_{s,d_x,d_y,\phi,p,g} \mathrm{D}(A_r, B(s,d_x,d_y,\phi,p,g)) \} \tag{5.4}$$

and in a hypothesis verification (two class) problem the decision rule is:

$$assign \quad B \rightarrow \Upsilon_r \quad if$$

$$\min_{s,d_x,d_y,\phi,p,g} \mathrm{D}(A_r, B(s,d_x,d_y,\phi,p,g)) < thresh_r \tag{5.5}$$

where A_r is the template for the r^{th} class and $thresh_r$ is the dissimilarity threshold for the r^{th} class. In the context of recognition using MRFs, a cost function corresponding to the unary

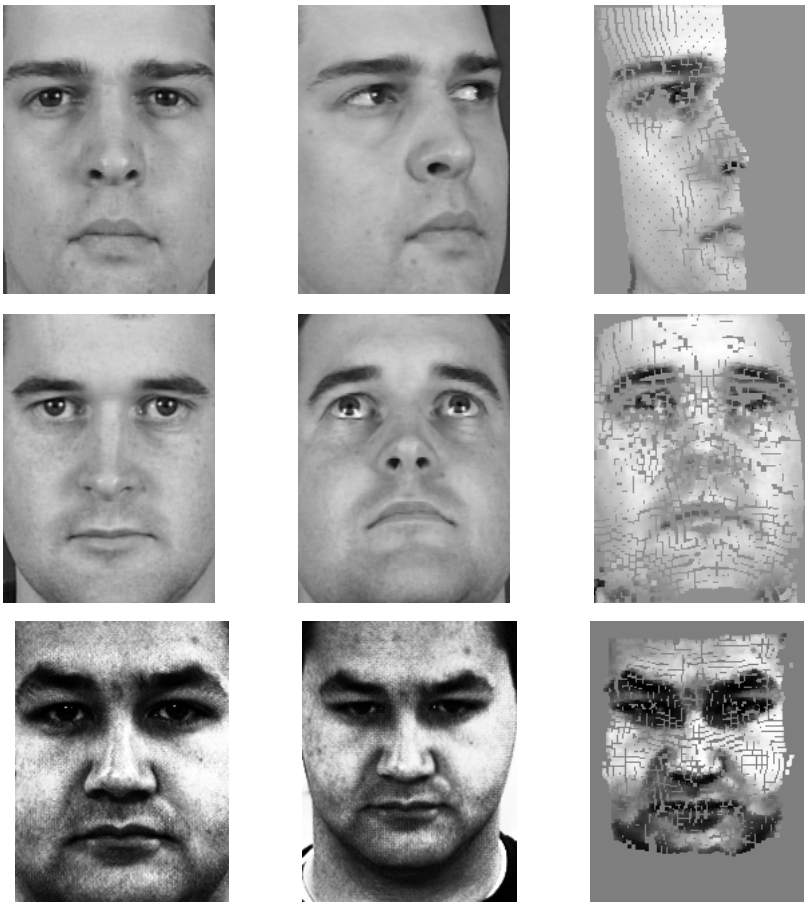


Figure 5.2: In each row from left to right: template image, target image and deformed template image. (In the first row, half of the template image is used for matching)

and pairwise terms is defined and optimized which is then used in the decision rule. However, the energy obtained in this way has been found not to have enough discriminatory capacity for classification. The factors which unfavorably affect the energy functional are identified as follows:

- The matching criterion, which partly gauges the geometric distortion, includes a global rigid transformation as well as local object shape deformations. For object recognition, only the latter is of importance.
- Restricting both, the neighborhood system of sites in an MRF to a limited spatial range, as well as clique cardinality, is an essential prerequisite of efficient optimization. However, this compromises the capacity to capture longer range interactions of object primitives.
- Measuring structural deformation as a function of the regularization term implicitly assumes a simple sum (Euclidean distance) as a measure of similarity. This assumption completely ignores any statistical dependencies between deformations of different sites.
- Last but not least, the goodness of match tracked down by the data term in the matching criterion can be dramatically influenced by environmental changes, such as changes in illumination.

We will not try to reformulate the energy itself. Instead, we provide possible ways of *normalizing* the cost of matching so that it can serve as a more suitable similarity measure. First, by removing the global geometric transformation between the unknown image and the target, we are left with the residual distortion map, which has a much better capacity to gauge the true structural differences between the matched images. Next, by inferring the statistical dependencies of deformations one can measure structural differences more accurately. The benefits of statistical modeling of deformations of all sites, which is performed by employing covariance matrices, is two fold. As previously noted, long-range interactions are not effectively modeled in an MRF with limited proximity of sites and clique cardinalities of up to two. In an ideal case, all edges (or hyperedges) of higher cardinalities should be incorporated into the energy term. One way to incorporate the effect of correlations of local deformation into the underlying MRFs is to incorporate a global interaction as in [137] with the cost given by the Mahalanobis

distance for local distortions. Unfortunately this makes the optimization very hard and inefficient. Instead, by modeling the statistical dependencies of deformation between all sites, one would not only take into account the inherent correlation of neighboring sites but can also partly make up for the weakly modeled interaction of sites which do not lie in a predefined limited neighborhood of one another and hence normalize the energy to offer a more meaningful comparison and ranking of competing candidates.

Finally, by using a photometrically invariant representation of the image content, rather than pixel intensities, one can suppress the corrupting influence of any photometric changes. Another observation in matching an unknown image to the class templates when the data has a deformable nature, is that they might differ slightly and hence an exact definition of class template is not available. We estimate the ideal distortion-free class exemplars which can then be used as prototypes for each class. The overall effect of these modifications to the matching process is to normalize the matching criterion values and render them comparable in absolute sense. In the following sections we study each of these modifications separately.

5.3.1 Structural Dissimilarity

After matching the two images, one expects small deformations for the objects of the same class whereas large deformations are expected when the gallery and the test images do not belong to the same category. Fig. 5.3 shows the distortion fields obtained for characters belonging to the same and different classes. In Fig. 5.3 the deformation of the grid is depicted on the unknown object. Clearly, objects of different classes tend to have a larger shape variation compared to the variation of object of the same class. Although face images compared to typewritten characters look more similar in terms of their shapes, as will be shown in the experiments, the shape differences between faces offer a useful discriminative feature.

Expanding the pairwise interaction term of the energy functional in the quadratic penalty function defined in one of the layers (*e.g.* layer one) of the four-connected neighborhood system considered using Eq. 5.1 yields:

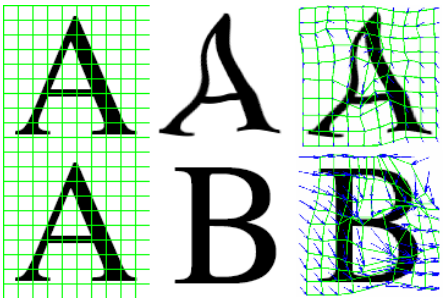


Figure 5.3: Distortion maps: upper row: distortion maps for objects of the same class, bottom row: distortion maps for objects of different classes.

$$\sum_{(u^1,v^1)\in\mathscr{V}^1} \theta_{u1v1}(x_{u1},x_{v1}) = 4q \sum_{u1\in In^1} {x_{u1}}^2 + 3q \sum_{u1\in B^1} {x_{u1}}^2 + 2q \sum_{u1\in C^1} {x_{u1}}^2 - 2q \sum_{(u1,v1)\in\mathscr{V}^1} x_{u1}x_{v1} \tag{5.6}$$

where $\mathscr{V}^1$ denotes the sites of layers one and In^1, B^1 and C^1 stand for internal nodes, nodes on the boundaries and nodes on the corners of layer one, respectively. In the above equation, the deformations of the nodes on the boundaries are weighted less as a result of having fewer neighboring nodes. Considering the deformation in each layer as an interpolation surface which maps each grid location into its corresponding location on the target image, the first three terms in the RHS of (5.6) can be considered as a weighted measure of deformation of the interpolation surface. A physical measure of similarity between the two objects can be defined as the deformation of the interpolation surface on each layer. This measure is represented by the first three terms in RHS of (5.6). In order to measure the dissimilarity, we omit the last term in (5.6) and weight the disparities of all the nodes in the MRF equally. We define the overall distortion energy of the interpolation surfaces on two layers as:

$$E_{distortion} = \sum_{u1\in\mathscr{V}^1,u2\in\mathscr{V}^2} ({x_{u1}}^2 + {x_{u2}}^2) \tag{5.7}$$

where x_{u1} and x_{u2} are labels of two isomorphic nodes corresponding to a single block.

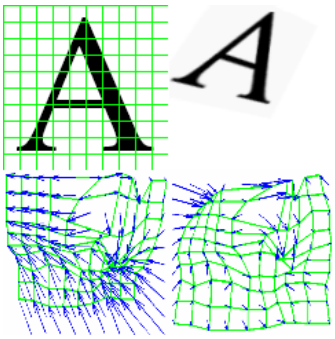


Figure 5.4: Distortion maps for objects of the same class when unknown object has undergone geometrical transformation, bottom left: distortion map before eliminating the effect of global geometric transformation, bottom right: distortion map after subtracting the effect of global geometric transformation.

5.3.1.1 Pose Estimation

Defining the distortion energy as in Eq. 5.7 implicitly assumes that objects are geometrically normalized prior to matching. In other words it has been assumed that the unknown object is not perturbed by a global geometric transformation. In order to remove the effect of rigid motion, we fit a global spatial projective transformation to the set of 2D corresponding points using the Levenberg-Marquardt method [105] and RANSAC to exclude mismatched parts, if any. The displacements resulting from the rigid motion are subtracted from the displacement vectors. The resulting distortion map leads to the computation of the distortion energy which is invariant under the assumed global spatial transformation. Fig. 5.4 shows an example where the template and an unknown object belong to the same class but a global geometrical transformation applied to the unknown object has resulted in distortion maps with large deformation magnitudes. The distortion map after subtracting the global transformation is also shown in the bottom right corner of the figure. Once local geometrical distortions have been estimated, the local distortion energy invariant to the rigid motion can be expressed as:

$$E_{distortion}^{local} = \sum_{u1 \in \mathcal{V}^1, u2 \in \mathcal{V}^2} (x_{u1} - x_{u1}^g)^2 + (x_{u2} - x_{u2}^g)^2 \tag{5.8}$$

where x_{u1}^g and x_{u2}^g represent vertical and horizontal displacements associated with the estimated global spatial transformation in site u. x_{u1} and x_{u2} correspond to vertical and horizontal dispar-

ities inferred at nodes $u1$ and $u2$.

5.3.1.2 Estimating Ideal Prototypes

In applications where the model objects have a non-deformable nature and an ideal template of each class is available (*e.g.* typewritten character recognition) one can match unknown objects to ideal distortion-free instances of different classes and classify the unknown pattern based on the similarity criterion adopted. However, in an application where the underlying object is deformable (*e.g.* face image data), because of the non-rigid nature of the object, a distortion-free class prototype is not available. One way around this problem is to match a number of different exemplars of each class one to another and compute the average distortion map (Fig. 5.5). The average distortion map obtained in this way gives an estimate of the deformation needed to warp an instance of each class to the ideal distortion-free prototype of the same class. The effective distortion energy for non-rigid objects can be computed in the following way: an unknown object is matched to one instance of the target class and the average distortion of the target class is subtracted from the estimated local distortion map. The new distortion energy can be expressed as the squared Euclidean distance between the structure of the unknown object and the average structure of the target class as:

$$E_{distortion}^{Euc} = (X_1 - \bar{X}_1)^T (X_1 - \bar{X}_1) + (X_2 - \bar{X}_2)^T (X_2 - \bar{X}_2) \tag{5.9}$$

X_1 and X_2 are the residual raster scanned disparity vectors on the two layers after matching the model to the unknown object, respectively. $\bar{X}_1$ and $\bar{X}_2$ denote the average displacement vectors on the two layers. T denotes matrix transpose.

5.3.1.3 Statistical Dependencies in Local Deformations

Statistical dependencies between different parts of a signal have been well studied before in speech recognition [78]. The problem under consideration was that of evaluating the similarity of an unknown pattern, representing a segment of a speech signal, to a set of reference prototypes producing a criterion function value which was then used to quantify goodness of

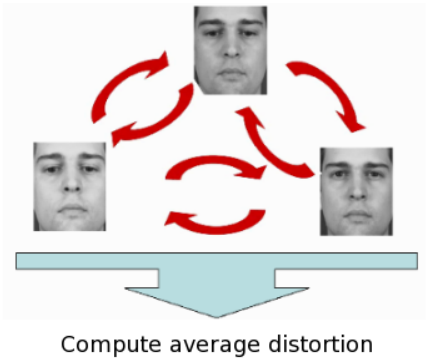


Figure 5.5: Estimating the average distortion for a non-rigid object.

match. Speech signals are processed on a frame-by-frame basis. The conventional methods assume that after establishing correspondence, the residual errors associated with the respective frames of speech signal are statistically independent. However, it was observed that taking into account statistical dependencies between different frames can improve the performance.

The problem under investigation in this work is the 2D counterpart of the classical matching problem in speech. In the previous section, a function of the binary term of the energy was normalized to resemble the Euclidean distance between the structure of the model and the unknown object. Because any changes in an object shape are usually smooth, deformation is prohibited from having sharp variations. In other words, the deformation of a part of an object causes neighboring regions to be deformed in a similar way. The larger the spatial separation between two regions, the lower the correlation between the deformations will be. Therefore, by modeling statistical dependencies in deformations, better estimates of similarity are expected. In order to take this effect into account we make use of covariance matrices. As stated previously, by considering correlation properties of local distortions between all sites instead of only four-connected neighbors, one can also partly compensate the weakly modeled long-range interaction of sites in computing the cost function. The estimated covariance matrices for the vertical and horizontal directions estimated on the XM2VTS [95] database are visualized as images in Fig. 5.6. In the figure the brighter areas correspond to higher correlation while darker areas represent low correlation. If there were no correlations between deformations of different sites, the correlation coefficients would be non-zero only on the main diagonal. Clearly, this is not the case. Remember that, in the cases where the head motion contains a pan component, only half of the face is used for matching and recognition, hence, in the figure, four covariance

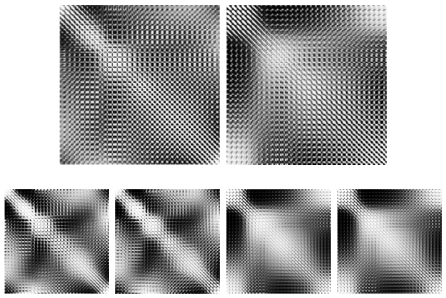


Figure 5.6: Covariance matrices of distortions: up row left: full face covariance matrix for vertical direction, up row right: full face covariance matrix for horizontal direction, bottom row from left to right: half face covariance matrices for left half of face in vertical direction, right half of face in vertical direction, left half of face in horizontal direction, right half of face in horizontal direction.

matrices are visualized, two for the full face and four for the half face matching. Having computed the covariance matrices for horizontal and vertical directions, the structural differences between a pair of images takes the following form:

$$E^{Mah}_{distortion} = (X_1 - \bar{X}_1)^T \sum_1^{-1} (X_1 - \bar{X}_1) + (X_2 - \bar{X}_2)^T \sum_2^{-1} (X_2 - \bar{X}_2) \tag{5.10}$$

where $\sum_1^{-1}$ and $\sum_2^{-1}$ represent inverse covariance matrices of the target class for residual distortions in layers 1 and 2, respectively. The statistical modeling of the correlations between local deformations proved to be useful and led to more than 8% improvement in performance in a verification test on frontal images of the XM2VTS [95] database. Although the common practice is to consider the shape deviations in two directions jointly, it is useful to consider them separately in the presence of severe pose changes of an object which may make one component of shape deviation (horizontal or vertical) less useful. In these cases, one may ignore the one in the direction parallel to the angle of pose deviation since it does not offer very useful information.

5.3.2 Textural Content

So far we have only considered the shape dissimilarities between faces. The spatial distortion measure can be complemented by a measure of quality of the match conveyed by the

data in order to refine the cost of match. However, the data term should not be sensitive to unwanted changes in lighting conditions during image capture and should be able to capture rich texture statistics. Although we use facial images captured almost under constant illumination conditions, in order to remove any residual effects of uneven illumination conditions we apply photometric normalization before measuring the texture similarities. In [120] a photometric preprocessing method based on a series of steps is introduced. The method is designed to decrease the effects of changes in illumination conditions, highlights and local shadowing, while keeping the fundamental visual information. The strategy of the approach is based on a gamma correction and then applying selective filtering. The image is first gamma corrected by a nonlinear gray level transformation replacing the pixel value I with I^{γ} where $\gamma > 0$. The purpose of this process is to improve the local dynamic range of the image in shadow and dark regions, while suppressing the bright region. The image is then processed using a band-pass filter defined as a difference of Gaussian filters, given by Eq. (5.11) to eliminate the impact of intensity gradients. The reason of choosing the band-pass filter is that it not only attenuates low frequency content caused by illumination gradient, but also reduces the high frequency noise caused by the aliasing artifacts.

$$DoG = (2\pi)^{-\frac{1}{2}}[\sigma_1^{-1}e^{-\frac{i^2+j^2}{(2\sigma_1)^2}} - \sigma_2^{-1}e^{-\frac{i^2+j^2}{(2\sigma_2)^2}}] \tag{5.11}$$

Then, the two stage contrast equalization given in Eq. (5.12) and Eq. (5.13) is employed to further re-normalize the image intensities and standardize the overall contrast.

$$J(i,j) = \frac{I(i,j)}{(\text{mean}(|I(i,j)|^a))^{\frac{1}{a}}} \tag{5.12}$$

$$\hat{J}(i,j) = \frac{J(i,j)}{(\text{mean}(\min(|I(i,j)|, \kappa)^a))^{\frac{1}{a}}} \tag{5.13}$$

a is used to reduce the influence of large values and κ, is a threshold used to truncate large values after the first stage of normalization. Finally, a hyperbolic tangent function in Eq. (5.14) is applied to suppress the extreme values and limit the pixel values in normalized image, $\hat{I}$, to a range between $-\kappa$ and κ.

$$\hat{I}(x,y) = \kappa \tanh(\frac{\hat{J}(x,y)}{\kappa}) \tag{5.14}$$

Figure 5.7: left: initial image, right: image after photometric normalization.

The parameters of the process are set according to [35] and are as follows: γ is set to 0.2, σ_1 is set to 1 and σ_2 is set 2. a, is set to 0.1 and κ, set to 10. Fig. 5.7 shows a result of applying the method on a sample face image. In the next step in order to extract features we use a local binary pattern operator [98]. The LBP operator is known to be one of the best performing texture descriptors which has been widely put into use in several applications. It is proved to be highly discriminative while being invariant to monotonic gray-level changes as well as computationally efficient. The original LBP operator assigns a label to every pixel of an image by comparing its 3×3-neighborhood with the value of the pixel under consideration and treating the results as a binary number. It is defined at a given pixel location (x_o, y_o) as:

$$LBP(x_o, y_o) = \sum_{n=0}^{7} S(i_n - i_o)2^n \tag{5.15}$$

where i_o is the gray value of the pixel under consideration and i_n is the gray value of its 8 neighboring pixels. Function $S(x)$ is defined as:

$$S(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases} \tag{5.16}$$

As long as the intensity order of the pixels in a neighborhood is preserved, LBPs are known to be unaffected by monotonic gray scale changes.

One extension to this operator proposes to use neighborhoods of different sizes in order to deal with textures at different resolutions [98]. Defining the local neighborhood as a set of sampling points which are evenly spaced on a circle centered at the pixel to be labeled, enables the radius and the number of sampling points to vary. The second extension to the original operator is the notion of *uniform patterns*. By considering the bit pattern circular, a local binary pattern is called uniform if the pattern has at most two bitwise transitions from 1 to 0 or vice versa. In the

experiments in [98], it has been noticed that for facial images uniform patterns explain almost 90 percent of all patterns when using (8,1) neighborhood and nearly 85 percent of the patterns in the (8,2) neighborhood. The advantage of using uniform patterns lies in their compactness in representation and reduced dimensionality which makes them less sensitive to noise and redundant information.

For face description in [98] the face image is divided into different subregions in order to extract local histograms. Once local histograms are extracted from each window, a spatially enhanced histogram is constructed by concatenating local histograms to form a global face descriptor. This histogram has three scales of description: in the lowest level are the LBP labels which include information about the patterns on the pixel-level, local histograms describe the image content in a regional level and finally the local histograms are concatenated to form a global description of the face image. While for the extraction of LBP histograms in [5] a regular division is used for both the gallery and test images, in the case of pose variation, this approach is not very effective since different regions of the test image may correspond to different facial features, compared to the frontal gallery image. But this problem has been circumvented by matching with deformable blocks. Using the information available from image matching, for each window in the gallery image the corresponding region in the test image can be identified (on a block by block basis) and the matched region in the test image can be used to extract LBP histogram. Extracted histograms from each region are normalized and concatenated into a single vector and compared using the χ^2 distance:

$$\chi^2(\eta,\xi) = \sum_{b,w} \frac{(\eta_{b,w} - \xi_{b,w})^2}{\eta_{b,w} + \xi_{b,w}} \tag{5.17}$$

where η and ξ are the normalized histograms of gallery and test images and b and w stand for the b^{th} bin of the histogram of the w^{th} window in the images. Recalling the definition of the energy of a match:

$$En(x;\theta) = \sum_{u\in\mathscr{V}} \theta_u(x_u) + \sum_{(u,v)\in\mathscr{E}} \theta_{uv}(x_u,x_v) \tag{5.18}$$

The overall energy is defined as the weighted sum of the data term (represented by ULBP histograms here) and the binary term (formulated as the Mahalanobis distance). So it makes

sense to combine these two terms after normalization to obtain the final distance measure as a weighted measure of shape and texture distances:

$$D(A_r, B) = \Delta\chi^2 + (1-\Delta)E^{Mah}_{distortion} \tag{5.19}$$

for $\Delta \in [0,1]$. $E^{Mah}_{distortion}$ corresponds to the structural distortion given in Eq. 5.10 and χ^2 represents the textural difference, given in Eq. 5.17.

5.4 EXPERIMENTAL EVALUATION

Upon the arrival of an unknown probe image, the method matches the probe image to the frontal gallery images of all classes and the similarity criterion in Eq. 5.19 is used in a nearest neighbor classifier for classification. The performance of the proposed methodology for pose-invariant face recognition is evaluated on two publicly available databases in two different scenarios, described next.

5.4.1 Verification Test on the XM2VTS Database

In the XM2VTS data set the evaluation protocol is based on 295 subjects consisting of 200 clients, 25 evaluation imposters and 70 test imposters. Two error measures defined for a verification system are false acceptance and false rejection given below:

$$FA = EI/I * 100\%, \qquad FR = EC/C * 100\% \tag{5.20}$$

where I is the number of imposter claims, EI the number of imposter acceptances, C the number of client claims and EC the number of client rejections. The performance of a verification system is often stated in *Equal Error Rate* (EER) in which the FA and FR are equal and the threshold for acceptance or rejection of a claimant is set using the true identities of test subjects. Consistent with the definition of EER, the parameter Δ in 5.19 is set using the true identities of test subjects. In the experiments to follow, we use the rotation shots of the database and do not provide results on the frontal images of this corpus.

Table 5.1: The effect of block adaptation and covariance estimation on equal error rates obtained on the XM2VTS corpus using shape information. Euc.: Euclidean distance, Mah.: Mahalanobis distance

Pose	Euc.	Euc. with block adap.	Mah. with block adap.
Pan	9.1%	7.5%	5.24%
Tilt	18.5%	16.5%	13.8%

5.4.1.1 Effects of the Proposed Modifications

Analyzing the effects of error correlation modeling using the covariance matrices we find an 8% improvement in error rate on the frontal images of the XM2VTS database. On the rotation shots of the same database the effect of block adaptation and correlation modeling on the error rates using different components of shape distance are reported in Table 5.1.

From the results it can be observed that the block adaptation decreases the overall error obtained using Euclidean shape distance by 3.6%. By exploiting the covariance information, a further 4.96% improvement in error rate is achieved. In total, block adaptation and covariance modeling reduce the *EER* by 4.28% using only shape information.

For texture modeling, we use the Uniform LBP operator with radius 2 and use the smallest resolution available for constructing local histograms (4×4 blocks and their corresponding patches in the test images). The overall average performance of the system improves with decreasing window size. This is understandable as severe pose changes in the image make different parts of the face undergo different appearance variations and hence more localized features can provide more discriminatory information. In the case of texture, block adaptation improves error rates by 1.28%.

5.4.1.2 Comparison of Shape and Texture

Table 5.2 provides a comparison between the discriminatory capability of the different components based on Mahalanobis distance and LBP histograms. Compared to the error rates obtained using shape, one observes that texture seems to be more discriminative. This can be explained from two points of views. First, shapes of the faces in the database are more similar. Secondly, a partial contradictory factor is the inadequacy of a planar transformation (*i.e.* pro-

Table 5.2: Comparison of shape and texture information on the XM2VTS corpus.

	Ver.	Hor.	Ver.& Hor.	Texture
Pan	5.74%	10.29%	5.24%	1.0%
Tilt	15.75%	15.99%	13.8%	9.0%

Table 5.3: Comparison of performance of the proposed method to the method in [122] on the XM2VTS database.

Method	FAR	FRR	HTER	EER
3D pose correction [122]	0.59	23.25	11.92	7.12
The proposed approach	4.99	11.62	8.30	4.85

jective) in modeling rigid motion of the head. Because the face is not planar, one can expect some errors as a result of the planarity assumption being made. Also from Table 5.2 it can be concluded that, the verification of faces subject to pan movement is more accurate than that of tilt, because in the case of tilt motion, the self occlusion problem can not be compensated for by exploiting symmetry. Inevitably this decreases the quality of the match and hence the performance.

5.4.1.3 Comparison to a 3D Geometric Normalization-based Method

In practical applications the thresholds for acceptance or rejection of a claimant are set on the evaluation set. The performance measure in this case is the *Total Error Rate* (TER) as below.

$$TER_{FAE=FRE} = FA_{FAE=FRE} + FR_{FAE=FRE} \tag{5.21}$$

where $FAE = FRE$ corresponds to the case when false acceptance and false rejection errors on the evaluation set are equal. In this case, the parameter Δ in Eq. 5.19 is set on the evaluation set and used on the test set. In [122], the authors use a 3D morphable model for geometrically normalizing the rotated images and then use LBP histograms in the 2D geometrically normalized images. The results obtained in [122] and the proposed approach are compared in Table 5.3. For comparison, the EERs are also included in the same table. In the table we report half TER (HTER) instead of TER for it to be comparable with the EER. From the table, it is observed that the proposed method outperforms the geometric normalization approach using 3D morphable model in [122], both in terms of *EER* and *TER*.

5.4.2 Identification Test on the CMU PIE Database

5.4.2.1 Test on Images with Neutral Illumination

In this test we use images captured under almost the same illumination conditions with neutral expression consisting of 884 images of 68 subjects viewed from 13 different angles. Frontal views of subjects (pose 27) are considered as gallery images while all the rest (12 different poses) are used as test images. We consider recognition results using shape and texture separately. The weighting of shape and texture scores in Eq. 5.19 is the same over all poses and is done in such a way that the average overall performance of the system is maximized. The results are reported in Table 5.4. From Table 5.4 the following conclusions can be drawn. The horizontal distance measure can be beneficial in poses where a large pan component is not present (poses C05, C07, C09 and C29). In contrast, the vertical distortion measure is more useful when the head movement contains a pan motion. It can be concluded that the two components complement each other and result in an average identification rate of nearly 70% for all poses in the database using only shape information. In Table 5.4 we also present the results of fusing texture and shape scores and compare our results to some other approaches using the original results reported in the literature. Some relevant details of the approaches are reported in Table 5.5. The identification rates reported correspond to using frontal images (pose C27) as gallery images. It can be observed that the proposed technique outperforms most other approaches, and is less restrictive in terms of assumptions. In order to show the merits of the modifications to the matching method we have included the results obtained using the matching method of Chapter 4 [114] for a number of poses. These results are obtained using the method in [114] for matching and keeping all other texture and shape representations similar to the current work. From the results it can observed that the modifications improved the performance significantly, specially in extreme poses. From the results it can also be observed that the performance of the proposed method is not completely symmetrical with respect to deviations from the frontal pose. This effect is partly due to inconsistencies in imaging and illumination conditions.

The best performing method in Table 5.4 among other approaches is the method in [146], with an average overall performance of 93.45%. Interestingly, the proposed method achieves the same average performance (excluding pose C22), but one needs to take into account the fol-

Table 5.4: Comparison of the performance of the proposed approach to the state-of-the-art methods on the CMU-PIE database.

Pose	**C02**	**C05**	**C07**	**C09**	**C11**	**C14**	**C22**	**C25**	**C29**	**C31**	**C34**	**C37**
Horizontal deviation angle	$-44°$	$-16°$	$0°$	$0°$	$32°$	$47°$	$-62°$	$-44°$	$17°$	$47°$	$66°$	$-31°$
Vertical deviation angle	$0°$	$0°$	$-13°$	$13°$	$0°$	$0°$	$1°$	$11°$	$0°$	$11°$	$1°$	$0°$
eigenlight-fields Complex [59]	58	94	89	94	88	70	38	56	57	56	47	89
PDM [56]	72	**100**	na	na	94	62	na	na	98	na	20	97
AA-LBP [146]	**95**	**100**	**100**	**100**	**100**	**91**	na	89	**100**	80	73	**100**
3D morphable model [108]	76	99	99	99	93	87	50	75	97	78	49	94
Matching [114]	na	85	91	na	na	na	22	na	79	na	25	na
Hor. Mah. distortion	35	73	85	100	58	26	13	35	65	44	10	58
Ver. Mah. distortion	54	72	44	57	66	60	44	60	67	61	47	70
Hor. & Ver. Mah. distortions	66	75	85	100	70	61	48	66	72	64	52	76
Texture	94	97	95	100	86	88	76	94	88	85	76	100
Shape & Texture	**95**	98	98	**100**	89	**91**	**79**	**95**	91	**88**	**83**	**100**

Table 5.5: Some specifications of the methods in Table 6.3 and test details.

Method	**Non-frontal training image**	**no. of landmark points used**	no. of subjects used for test
eigenlight-fields Complex [59]	**Y**	**39-54 depending on pose**	**34**
PDM [56]	**Y**	**62**	**68**
AA-LBP [146]	**Y**	**80**	**68**
3D morphable model [108]	**3D data**	> 6	**68**
The proposed approach	**N**	**None**	**68**

lowing considerations. The method in [146] uses non-frontal gallery images as well as frontal images for training, whereas we do not use any non-frontal training images. Also, the method in [146] uses 80 manually labeled landmark points, whereas the method proposed here does not need any manually annotated landmarks and only requires the face to be detected in a bounding box which is much easier than providing landmarks automatically. Other advantages of the method proposed here is that it can also cope with moderate global spatial transformation (*e.g.* projective) between the images. In conclusion, the method compares very favorably with most of the existing approaches in spite of its less restrictive assumptions and minimal injection of prior information. The main drawback of the algorithm proposed here is the computational complexity of the optimization stage which is a common characteristic of MRF-based approaches. However, this issue can be addressed by employing multi-resolution analysis or by using a sparse MRF model instead of dense image matching methods and also by taking advantage of parallel processing hardware such as GPUs.

Table 5.6: Comparison of performance of the proposed method under neutral lighting and variations in lighting on PIE database.

Pose	05	22	27
Neutral illum.	98	79	na
Varying illum.	71.5	40.2	95.6

5.4.2.2 Test on Images under Different Lighting Conditions

In order to determine the failing modes of the algorithm and to evaluate the degradation in system's performance under uneven illumination conditions, in this section we provide the results of a test on a subset of the PIE database consisting of images of 68 subjects captured in three different poses and three different lighting conditions. The images are captured under full profile (pose 22), 3/4 profile (pose 05) and full frontal (pose 27). In each pose, there are images captured with 21 different flashes for each subject of which we randomly select three different flash conditions for our test. The same set of gallery images is used as in the previous section. Table 5.6 reports the average recognition rates over different illumination conditions for each pose. The results obtained under neutral illumination conditions are also included for a comparison. In comparison with the recognition rates under neutral illumination conditions, a drop in system's performance is observed. This is caused by the matching being imperfect due to shadowing effects and also by the inadequacy of the photometric normalization method under severe illumination changes. One option is to use the method on near infra-red images which are known to be less affected by illumination changes.

5.4.3 Evaluation on the SOIL Database

In order to show that the proposed approach is equally applicable to other object matching, retrieval and recognition problems, where change of viewing angle and scale is even more severe than for the face data set, we test the same methodology on the SOIL database [1] for recognition. For each object we use the frontal image as the object model and 20 other views as test images. The objects images are scaled down to 50 for computational efficiency. The method outperforms the state of the art algorithm using a graph matching approach [4] on the same database. The results are illustrated in Fig 5.8. It is worth noting that in [4] the authors use color whereas we limit ourselves to using grey scale images. The reason the recognition rates

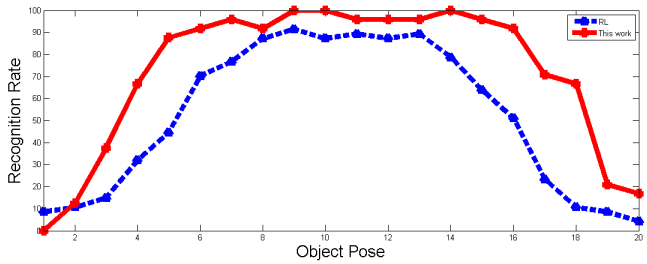


Figure 5.8: Comparison of the performance of the proposed approach to the one in [4] denoted by RL

drop for extreme poses is that as the pose changes, the objects' scales change dramatically and using 4×4 blocks in the model image is insufficient for capturing structural and also textural content of the images.

5.5 Summary

In this chapter, we proposed important modifications to the original matching method in Chapter 4 [114] in order to increase the accuracy of the match. The modifications included using edge maps for the computation of the data term, using soft penalty functions for pairwise potentials, deformable block matching and label pruning. Next, a method for normalizing the energy of an established match was proposed to make the energy a better measure of dissimilarity. The normalization procedure included normalizing both unary and pairwise terms of the energy functional. We next evaluated the method on two face databases. First, it was evaluated in a verification scenario on the rotation shots of the XM2VTS corpus. The results compared to another method, employing 3D model for pose normalization, showed the effectiveness of the methodology. Next, the method was tested on the CMU-PIE database in an identification scenario, achieving identification performance on par with the state-of-the-art methods and in some poses obtaining better recognition rates. In order to show the applicability of the method in a general object recognition scenario we provided the results of a recognition test on the SOIL database and compared the results to another method, with a favorable outcome.

Chapter 6

Multi-scale Image Matching

6.1 Introduction

The computational complexity of inference in graphical models can be considered as one of the bottlenecks of these approaches. The method we used in the previous chapter grouped pixels into non-overlapping blocks and estimated a single displacement vector for each block. The advantages of the using the block model as noted earlier was decreasing the computational complexity in addition to robustness against noise. One drawback of this approach is the error induced by assuming that all pixels inside a block have similar displacement vectors. A naive way to obtain displacements for each pixel is to assume each pixel as an individual node of the graph and construct the data term using single pixels instead of the block model. However, this approach prohibitively increases the computational burden as a result of increasing the number of variables and also makes the method more vulnerable to noise. The robustness against noise is highly desirable as, for recognition, one needs to match facial images which can potentially be taken with different devices and correspond to different subjects viewed from different angles. The other drawback is the increased probability of the optimization method to get stuck in a local minimum as a result of increased dimensionality of the configuration space. Multi-resolution analysis has successfully been employed for avoiding these problems.

In a multi-scale approach for motion estimation one uses relatively larger groups of nodes in the coarser scales of the hierarchy and groups with smaller number of nodes in the finer scales.

The idea is that the coarser levels provide a rough estimate of the displacements and finer levels serve to fine-tune the result of the previous coarser level.

Motivated by the success of multi-scale analysis of MRFs, in this chapter two methods are proposed for multi-level matching of images. The first one is based on a heuristic search in the original configuration space giving initially rough estimates by using larger blocks and then refining the estimated motion using smaller blocks with reduced search region in the configuration space. The second method starts with larger blocks but in a *coarser* configuration, compared to the original one. As the method proceeds, successively smaller blocks and a *finer* scale of the configuration is used for inference. The differences between the two methods are as follows. In the first approach, there is no explicit consistency between the energies being optimized in different levels whereas the consistency of the energy functional between different scales in the second method is maintained using the super-coupling transform. Also, in the second approach one achieves further speed-up compared to the first method by introducing a lumpiness into the configuration. This effectively means that the displacement range is subsampled in the coarser levels thus reducing the complexity of message passing. This is in contrast to the first approach in which we only use larger blocks without subsampling the displacements. Moving to the finer levels, in the first approach we reduce the search region whereas in the second method one only ends up with the labels consistent with the solution in the previous coarser resolution.

The chapter is organized as follows: In section 6.2 we describe our heuristic multi-level matching [12]. Section 6.3 introduces the multi-resolution method based on the super-coupling transform [11]. In section 6.4 we show how a statistical shape prior can be used to minimize the matching errors. Section 6.5 formulates a nearest neighbor classifier based on shape and texture similarities. We next provide some experimental evaluation of the two approaches in section 6.6. The chapter is brought to a conclusion in the section 6.7.

6.2 Heuristic Multi-level Matching

The general idea in this approach is illustrated in Fig. 6.1. Starting with larger blocks, one finds a coarse estimate of the motion. In a next finer level, each block is divided into four

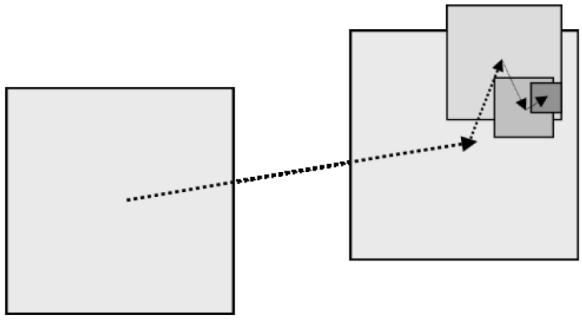


Figure 6.1: multi-level search for correspondences.

smaller blocks and the optimization is performed again. Th efficiency comes from having fewer discrete variables in coarser scales and reducing the search region in the finer levels. For example, if we are searching in a neighborhood of $[-40,40]$ in the coarse scale, we set the search region in the next fine level to a neighborhood of $[-20,20]$ of the obtained solution in the previous coarse level. It should be noted that what differs in different levels is the relative size of the block to the image and not the absolute sizes. Thus, in principle instead of low-pass filtering and sub-sampling the image, we only low-pass filter the image and retain the original size of the image but use different block sizes in different levels. In a method in which higher levels of pyramid are low-pass filtered and sub-sampled versions of the original image, fine scales of disparities cannot be detected as a result of sub-sampling. In contrast, by retaining the original image size, even fine disparity information is not lost in the higher levels of the hierarchy. In this approach all images corresponding to different scales are of the same size, but higher levels are more blurred versions of the original image. The method is relatively more robust to noise by virtue of matching more blurred versions of the images at higher levels. Apart from the speed gain in the hierarchical scheme, comparably better solutions can be obtained by successive matching and refinement of the match. This property comes partly from the fact that at higher levels, larger blocks capture relatively longer range interactions and the rough estimate of the disparities makes the optimization bypass local minima and converge closer to the true solution while finer scales fine-tune the obtained solution. Typical values for the Gaussian filter order (with binomial coefficients), block sizes and disparity search ranges are given in Table 6.1.

Figure 6.2 shows examples of matching images using the method described. We call the

Table 6.1: Typical values for block size, disparity search range and Gaussian filter order in hierarchical image matching.

Level	3	2	1	0
Block size	8×8	4×4	2×2	1×1
Maximum displacement	40	20	10	5
Filter order	20	15	5	-

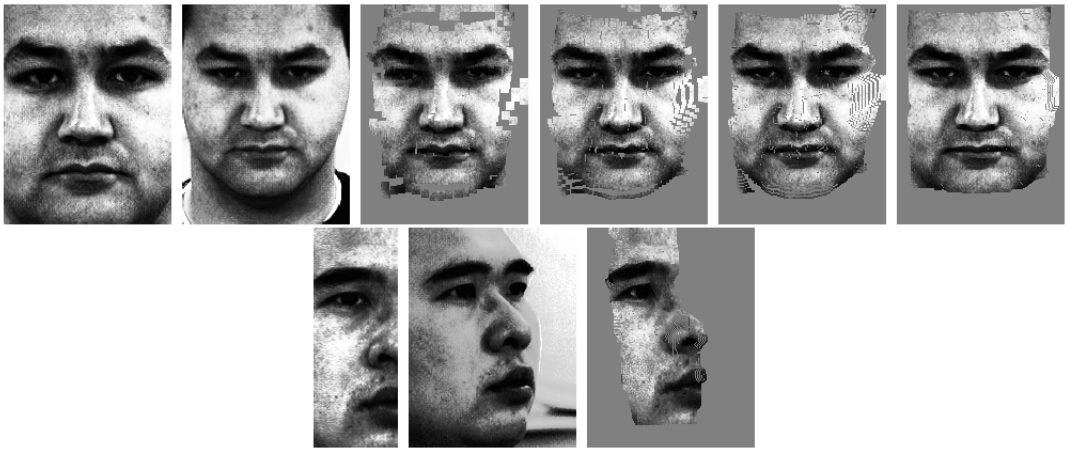


Figure 6.2: Top row from left to right: template image, target image and the results of warping the template in four consecutive scales; bottom row from left to right: template image, target image and the result of matching half of the template to the target image.

method heuristic in the sense there is no consistency between the energies being optimized in different scales. The method only facilitates and accelerates establishing pixel wise correspondences by initializing the search region and then refining it.

On the other hand, it is desirable to maintain consistency between the energies we minimize. This can be achieved via the Renormalization Group Transform (RGT) [53], discussed in the following.

6.3 RGT for Multi-resolution Analysis

One of the common approaches in multi-resolution optimization is based on the Renormalization Group Transform (RGT) [53, 102]. The algorithm consists of two main steps: *renormalization* and *processing*. In the renormalization step, one iteratively constructs finer and finer

grids of nodes and a corresponding sequence of energy functionals. Suppose we have an original grid of size $2^N \times 2^N$. Then in a finer level one obtains a coarser grid by grouping every 4 nodes together and identifying them as a single node. In order to define the energy functionals, one needs to choose a probability function measuring how likely is a coarse configuration (X') given a finer configuration $\bar{X}$:

$$\exp\left\{En(X')\right\} = \sum_{\bar{X}} \left\{P(X'|\bar{X}) \exp\left\{En(\bar{X})\right\}\right\} \tag{6.1}$$

In the *processing* step one performs a multi-scale coarse-to-fine optimization starting from the coarsest scale moving towards the finest one. More specifically, one performs optimization in a coarse scale and then moves into the next finer scale where only those configurations which are *constrained* by the previous coarse scale solution are considered. Searching in the subspace of the next finer configuration reduces the computational complexity of each level and as a result the complexity of the whole optimization. It is shown that if the conditional probabilities $P(X'|\bar{X})$ are chosen to be delta functions, then the procedure finds the global minimum of the energy [53].

The renormalization group transform is known to preserve the whole structure of the probability distribution. However, in most cases, one is not interested in preserving the full structure of the probability distribution, but only in preserving its maxima (just as in our task). In these cases, a potential-based coarsening technique (super-coupling transform [28]), which is known to be order preserving, in the sense that the inequalities obeyed by the original distribution remain true after coarsening, is employed. A natural consequence of order preserving property is that the mode of the original distribution maps exactly to the mode of the coarsened distribution. In fact, the multi-resolution approach we employ here is intended to accelerate the optimization process by coarsening the configuration space so that long range jumps that would lead faster to the global minimum are possible. Other potential benefit of the multi-resolution approach applied here can be considered as reducing the number of message passing operations and as a result achieving accelerated convergence.

One of the considerations in multi-resolution analysis is the transformation of the solution obtained in one level to a finer level. According to the theory of RGT, this solution should be used to restrict the solution in the finer level and only those configurations which are consistent with the coarser level should be considered. A common practice is to use a block-flat assump-

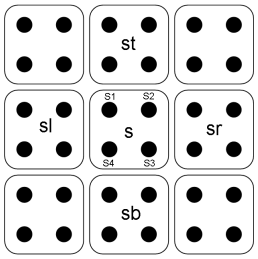


Figure 6.3: Geometry of sites in the coarse and fine lattice under consideration.

tion and assign all nodes inside a block the same label. In this way, the solution obtained in the coarser level serves as a starting point for the optimization in the finer level. Under the block-flat assumption, the super-coupling transform requires that the value of the cost function when going from one level of resolution to another should remain the same. It is shown that this transformation at the zero temperature limit is identically the same as RGT [28]. There are some other multi-scale approaches such as those in [48, 124] in the literature. The technique employed here applies multi-resolution ideas in a principled way, based on the super-coupling transform. Also, in the core of the method, TRW-S as it exhibits better convergence properties than belief propagation used in some other works is used for inference.

6.3.1 The Optimization Process

For brevity we will consider only two levels of resolution referring to them as the coarse and the fine level. We assume that images are of size $2^N \times 2^N$. We coarsen the image lattice by replacing every four nodes by one node in the coarse lattice. Thus, each node in the coarse lattice (denoted by s) corresponds to four nodes in the finer lattice (denoted by s_1, s_2, s_3 and s_4) as illustrated in Fig. 6.3. We use the symbol $\bar{X}$ to denote the fine configuration that can be produced under the block-flat assumption from the coarse configuration X'. The theory of super-coupling transform then says that the parameters of the energy should be chosen in such a way that

$$En(\bar{X}) = En(X'). \tag{6.2}$$

For each site in the coarse lattice and its four corresponding sites in the fine level the following

equation must hold:

$$\theta_{s_1}(\bar{x}_{s_1}) + \sum_{(s_1,u_1)\in\mathscr{E}_f} \theta_{s_1u_1}(\bar{x}_{s_1},\bar{x}_{u_1}) + \theta_{s_2}(\bar{x}_{s_2}) + \sum_{(s_2,u_2)\in\mathscr{E}_f} \theta_{s_2u_2}(\bar{x}_{s_2},\bar{x}_{u_2}) + \\ + \theta_{s_3}(\bar{x}_{s_3}) + \sum_{(s_3,u_3)\in\mathscr{E}_f} \theta_{s_3u_3}(\bar{x}_{s_3},\bar{x}_{u_3}) + \theta_{s_4}(\bar{x}_{s_4}) + \sum_{(s_4,u_4)\in\mathscr{E}_f} \theta_{s_4u_4}(\bar{x}_{s_4},\bar{x}_{u_4}) = \\ \theta'_s(x'_s) + \sum_{(s,u)\in\mathscr{E}_c} \theta'_{su}(x'_s,x'_u) \tag{6.3}$$

here $\mathscr{E}_f$ and $\mathscr{E}_c$ correspond to the edge set in the fine and coarse scales, respectively. Considering the relative positions of the sites illustrated in Fig. 6.3 we have

$$\begin{aligned} \bar{x}_{s_1} = \bar{x}_{s_2} = \bar{x}_{s_3} = \bar{x}_{s_4} = x'_s, \\ \bar{x}_{s_1t} = x'_{st}, \bar{x}_{s_1r} = x'_s, \bar{x}_{s_1b} = x'_s, \bar{x}_{s_1l} = x'_{sl}, \\ \bar{x}_{s_2t} = x'_{st}, \bar{x}_{s_2r} = x'_{sr}, \bar{x}_{s_2b} = x'_s, \bar{x}_{s_2l} = x'_s, \\ \bar{x}_{s_3t} = x'_s, \bar{x}_{s_3r} = x'_{sr}, \bar{x}_{s_3b} = x'_{sb}, \bar{x}_{s_3l} = x'_s, \\ \bar{x}_{s_4t} = x'_s, \bar{x}_{s_4r} = x'_s, \bar{x}_{s_4b} = x'_{sb}, \bar{x}_{s_4l} = x'_{sl}. \end{aligned} \tag{6.4}$$

where we have used the subscripts *t,b,l,r* to denote the top, bottom, left or the right neighbor of a site in an immediate four-connected neighborhood system. For the data term separately we can write

$$\theta'_s(x'_s) = \theta_{s_1}(x'_s) + \theta_{s_2}(x'_s) + \theta_{s_3}(x'_s) + \theta_{s_4}(x'_s) = \sum_{i=1}^{4} \theta_{s_i}(x'_s). \tag{6.5}$$

hence the data term associated with a block in the coarse level can be computed as the sum of its four corresponding nodes in the finer level which are themselves defined as the squared difference of magnitudes of the relevant elements on the normalized horizontal and vertical edge maps.

The pairwise potentials should satisfy:

$$\theta_{s_1,s_1t}(\bar{x}_{s_1},\bar{x}_{s_1t}) + \theta_{s_1,s_1r}(\bar{x}_{s_1},\bar{x}_{s_1r}) + \theta_{s_1,s_1b}(\bar{x}_{s_1},\bar{x}_{s_1b}) \\ + \theta_{s_1,s_1l}(\bar{x}_{s_1},\bar{x}_{s_1l}) + \theta_{s_2,s_2t}(\bar{x}_{s_2},\bar{x}_{s_2t}) + \theta_{s_2,s_2r}(\bar{x}_{s_2},\bar{x}_{s_2r}) \\ + \theta_{s_2,s_2b}(\bar{x}_{s_2},\bar{x}_{s_2b}) + \theta_{s_2,s_2l}(\bar{x}_{s_2},\bar{x}_{s_2l}) + \theta_{s_3,s_3t}(\bar{x}_{s_3},\bar{x}_{s_3t}) \\ + \theta_{s_3,s_3r}(\bar{x}_{s_3},\bar{x}_{s_3r}) + \theta_{s_3,s_3b}(\bar{x}_{s_3},\bar{x}_{s_3b}) + \theta_{s_3,s_3l}(\bar{x}_{s_3},\bar{x}_{s_3l}) \\ + \theta_{s_4,s_4t}(\bar{x}_{s_4},\bar{x}_{s_4t}) + \theta_{s_4,s_4r}(\bar{x}_{s_4},\bar{x}_{s_4r}) + \theta_{s_4,s_4b}(\bar{x}_{s_4},\bar{x}_{s_4b}) \\ + \theta_{s_4,s_4l}(\bar{x}_{s_4},\bar{x}_{s_4l}) = \\ \theta'_{st}(x'_s,x'_t) + \theta'_{sr}(x'_s,x'_r) + \theta'_{sb}(x'_s,x'_b) + \theta'_{sl}(x'_s,x'_l). \tag{6.6}$$

From (6.4) we have:

$$\begin{aligned}\theta_{s_1,s_1t}(x'_s,x'_t)+\theta_{s_1,s_1l}(x'_s,x'_l)+\theta_{s_2,s_2t}(x'_s,x'_t)+\\ \theta_{s_2,s_2r}(x'_s,x'_r)+\theta_{s_3,s_3r}(x'_s,x'_r)+\theta_{s_3,s_3b}(x'_s,x'_b)+\\ \theta_{s_4,s_4b}(x'_s,x'_b)+\theta_{s_4,s_4l}(x'_s,x'_l)=\\ \theta'_{st}(x'_s,x'_t)+\theta'_{sr}(x'_s,x'_r)+\theta'_{sb}(x'_s,x'_b)+\theta'_{sl}(x'_s,x'_l).\end{aligned} \tag{6.7}$$

hence

$$\begin{aligned}\theta'_{st}(x'_s,x'_t) &= \theta_{s_1,s_1t}(x'_s,x'_t)+\theta_{s_2,s_2t}(x'_s,x'_t)\\ \theta'_{sr}(x'_s,x'_r) &= \theta_{s_2,s_2r}(x'_s,x'_r)+\theta_{s_3,s_3r}(x'_s,x'_r)\\ \theta'_{sb}(x'_s,x'_b) &= \theta_{s_3,s_3b}(x'_s,x'_b)+\theta_{s_4,s_4b}(x'_s,x'_b)\\ \theta'_{sl}(x'_s,x'_l) &= \theta_{s_1,s_1l}(x'_s,x'_l)+\theta_{s_4,s_4l}(x'_s,x'_l)\end{aligned} \tag{6.8}$$

Adopting the quadratic pairwise potential

$$q'(x'_s-x'_t)^2=q(x'_s-x'_t)^2+q(x'_s-x'_t)^2. \tag{6.9}$$

from Eq. 6.9 we find $q'=2q$, which means that the model prescribes a stronger interaction between sites in the higher levels of resolution. This is intuitive because in coarser resolutions, the sites represent larger groups of pixels which require stronger interaction with each other. A comment regarding the multi-resolution technique employed here is that the coarser scales considered here do *not* correspond to any coarser grids of the original image in a physical sense. The method is essentially a mathematical trick in which the model is changed so that its optimal solution maps to the optimal solution of the original problem, as discussed in [102].

6.4 Statistical Shape Prior

Deformable models can broadly be classified into two categories: free-form and parametric. In the free-form models (*e.g.* snake) only general continuity and smoothness constraints are considered [116]. As a result, these models can be matched to an arbitrary shape. In contrast, parametric models incorporate a general shape of the object of interest. They encode special

attributes of an object and its variations and hence are more robust to occlusions and spurious structures (as G-Snake [85], Active Shape Model [38], [37] and [144]).

The quadratic pairwise potentials used so far only impose a smoothness prior into the deformation model. It is natural to expect that by injecting an object specific prior into the matching, better results can be obtained. This is even more important in the multi-resolution approaches because one of the drawbacks of multi-resolution approaches is that if an error occurs in a coarser scale of resolution, it will affect the solution in the finer scales. We constrain the solution in the coarsest scale using a shape prior so that the errors are minimal.

In order to construct a statistical shape model, we use frontal face images and match them one to another. Once a global spatial transformation (*e.g.* projective) is fitted to the set of corresponding points, the residual distortions are computed. The distribution of residual errors can then be approximated reasonably well by a Gaussian distribution [96]. In order to capture the main modes of variation in the underlying distribution, the first M eigenvectors of the covariance matrix are used as the basis vectors to construct a shape space. We consider the deformation in each direction separately in the the same way they are treated in the decomposed model. The distribution of the local deformations in each direction in the training set is expressed as

$$p(X') = \frac{\exp(-\frac{1}{2}\sum_{i=1}^{M}\frac{w_i^2}{e_i})}{(2\pi)^{M/2}\prod_{k=1}^{M}e_k^{1/2}} \tag{6.10}$$

where

$$\omega = \Gamma_M^T(X' - \bar{X}'_{mean}) \tag{6.11}$$

ω is the vector of shape parameters determining the coordinates of the projected point in the shape space with the elements, w_i. X' is the local deformation vector and $\bar{X}'_{mean}$ is the mean deformation for the target class. Γ_M is the matrix of the first M principal eigenvectors of the covariance matrix of local deformations and $e_k, k = 1 \ldots M$ denote the M largest eigenvalues of the covariance matrix. Any local deformation X' in each direction can then be approximated by a linear combination of the M eigenvectors corresponding to the largest eigenvalues as

$$X' \approx \bar{X}'_{mean} + \Gamma_M\omega \tag{6.12}$$

Projecting the local deformations into the shape space while retaining most of the variation compatible with the shape space, discards the variation of the deformation which is inconsistent with those reflected by the samples in the training set. The energy of the obtained local deformation in each direction projected into the shape space is denoted as

$$en(X') = \sum_{i=1}^{M} \frac{w_i^2}{e_i} \tag{6.13}$$

6.4.1 Regularizing and Constraining the Solution

Having defined a statistical energy term for the local deformations, the new energy takes the form:

$$En(X;\theta) = \sum_{s\in\mathscr{V}} \theta_s(x_s) + \sum_{(s,u)\in\mathscr{E}} \theta_{su}(x_s,x_u) + \theta_g(X) \tag{6.14}$$

where

$$\theta_g(X) = \sum_{i=1}^{M} (\frac{w_{ix}^2}{e_{ix}} + \frac{w_{iy}^2}{e_{iy}}) \tag{6.15}$$

where w_{ix}, w_{iy}, e_{ix}, e_{iy} denote respectively shape parameters in the horizontal direction, shape parameters in the vertical direction and the first M principal eigenvectors in the horizontal and vertical directions. We treat horizontal and vertical distortions separately (by having one eigenspace for each one) the same as in the decomposed model in which separate models were used for displacements in each direction. This is advantageous in terms of identification performance when severe head pose changes exist. It is apparent that the corresponding min-sum task incorporates cliques of size 1 (nodes), size 2 (ordinary edges) and a hyperedge containing all nodes (global interaction) represented by the data term, pairwise smoothness term and the local deformation energy term, respectively. The corresponding minimization problem needs two kinds of updates: between the unary and binary constraints and between the unary and the global constraint. This kind of minimization has been addressed for example in [135] for a special type of global constraints in the case that each site has two admissible states. In our task, minimizing the energy in Eq. 6.14 incorporates min-marginalization of a function (shape prior energy) over the whole configuration space with the complexity which is exponential with the size of the full MRF model. Since the method would eventually be used in a recognition scenario, it needs to be efficient. Instead of directly solving the problem in Eq. 6.14, we propose a recursive two-stage approach:

- Solve the task without considering the prior energy term.

- Estimate the local deformations and project them into the shape space using Eq. (6.12). Add the global rigid transformation back to the local deformations and reduce the local neighborhood of each site in which correspondences are sought for the next round.

We repeat step one and two until the global shape parameters and the prior shape energy term do not change beyond a specified precision. In practice we found that it is not needed to repeat the two steps more than two times if the TRW-S method is iterated sufficiently. For pose estimation we use Levenberg-Marquardt [105] method. While the first round of projection into the shape space makes the current solution compatible with those available in the training set, the next round refines the most up-to-date solution. In practice, as a result of reducing the searched neighborhood which effectively prunes the configurational space, the matching after the first iteration is more computationally efficient. The method is most similar to the eigen-snake model [143]. However, here we employ an efficient multi-scale optimization approach for the minimization of the energy functional using the tree reweighted message passing method. Also, in the eigen-snake model the prior information is used to deform the template whereas we use it to constrain the solution and prune the configuration space. Using the proposed method for deformable registration, we obtain pixelwise correspondences between a pair of images. Figure 6.4 shows some examples. In the figure, the template, the target image, the deformed template and also the deformed template superimposed on the target image are illustrated for two pairs of images.

6.5 Classification

Once the correspondences are established between a pair of images, in order to judge the goodness-of-match, we compute a similarity measure considering both textural and structural similarities between a pair of images. The structural similarity between the images, invariant to a global transformation, is best represented by the prior shape energy for the estimated match. In order to measure the texture similarities we employ a discriminative texture descriptor (LBP) similar to Chapter 5.

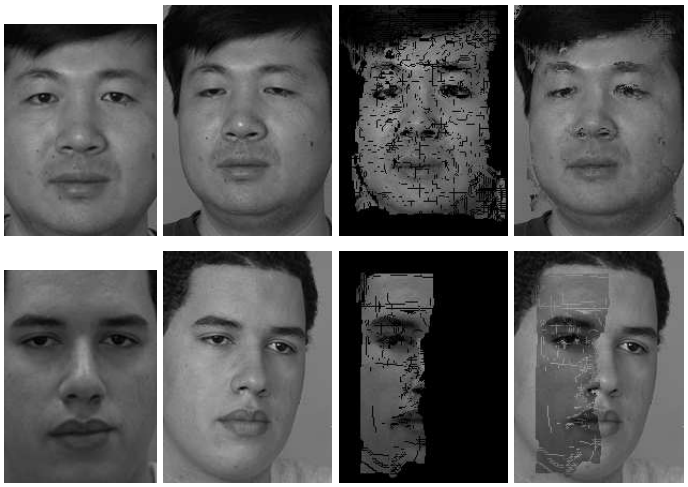


Figure 6.4: In each row from left to right: template, target, deformed template and deformed template superimposed on the target image.

6.5.1 Textural Similarity

We use the same texture representation we employed in the previous chapter. That is, we first apply a photometric normalization step [120] and then use uniform LBP features with radius 2 in circular neighborhoods as features. Using the information available from image matching, for a window in the gallery image the corresponding region in the test image can be identified and the matched region in the test image can be used to extract LBP histogram. Extracted histograms from each region are normalized and concatenated into a single vector and compared using the χ^2 distance:

$$\chi^2(\eta,\xi) = \sum_{b,w} \frac{(\eta_{b,w} - \xi_{b,w})^2}{\eta_{b,w} + \xi_{b,w}} \tag{6.16}$$

where η and ξ are the normalized histograms of gallery and test images and b and w stand for the b^{th} bin of the histogram of the w^{th} window in the images. For texture modeling, we use the Uniform LBP operator with radius 2 and use (16×16) windows in the gallery images and their corresponding patches in the probe images.

6.5.2 Structural Similarity

Different structures of faces can offer a discriminative measure for classification. Apparently this measure should be independent of the spatial transformation parameters of the probe image while at the same time should take into account the correlations of local deformations. These properties are met by the shape parameters discussed in Section 6.4. We use the prior energy term as a measure of structural similarity between a pair of facial images. Two observations about the imposition of the shape prior energy are as follows: because $w = 0$ and consequently $\theta_g = 0$ when $X' = \bar{X}'_{mean}$, w represents the local deformation parameters and as a result the rigid transformation parameters (*e.g.* projective) and the local deformation parameters are separated. Second, the imposition of the shape space constraint based on the prior knowledge of local deformations makes the matching more robust against spurious structures and outliers. The shape distance between a pair of images is defined as:

$$D_{Strcutural}(I,J) = \sum_{i=1}^{M} \left(\frac{w_{ix}^2}{e_{ix}} + \frac{w_{iy}^2}{e_{iy}}\right) \tag{6.17}$$

The overall dissimilarity measure between a pair of images is defined as the weighted sum of the data term (represented by ULBP histograms here) and the binary term (formulated in the PCA space). We combine these two terms after normalization in order to obtain the final distance measure as a weighted measure of shape and texture distances between a test image (J) and i^{th} gallery image (I_i):

$$D(I_i,J) = \Delta\chi^2 + (1-\Delta)D_{Strcutural} \tag{6.18}$$

for $\Delta \in [0,1]$. $D_{Strcutural}$ corresponds to structural distance and χ^2 represents the textural differences of the images being compared.

6.6 Experimental Evaluation

In this section we provide the results of an experimental evaluation of the multi-resolution matching method in various scenarios. For the heuristic multi-stage search we only provide the results of an identification test on the CMU-PIE database. In this case, it is not feasible to compare the energies with the original method in [114] since there is no consistency between the energies in different levels.

In the case of the multi-resolution matching based on the super-coupling transforms we consider the following evaluations. First, we compare the multi-scale matching method with the method of Chapter 4 [114] in terms of their running time and quality of the established match. Next, the proposed methodology for image matching is compared with that of [114] in a face identification scenario.

Next, the performance of the proposed face recognition algorithm is compared with some state-of-the-art algorithms on the challenging CMU-PIE database including a wide range of pose deviations. Finally, we provide the results of an identification test on the rotation shots of the FERET database for frontal and near frontal poses and compare the proposed approach to state-of-the-art methods in pose-invariant and frontal face recognition.

Throughout our experiments, the coarsest scale we start the optimization in is constructed by nodes of size 4×4 pixels and the finest scale is 1×1, that is at the pixel level. Hence, the multi-resolution analysis is performed in three scales. The range of displacements we search for correspondences is $[-36, +36]$ pixels in both horizontal and vertical directions. Gallery images we use in the experiments are normalized to the size of 160×208 in the CMU-PIE database and to the size 140×196 in the FERET database. In the experiments on the CMU-PIE database where a severe head pose deviation exists in horizontal direction, only half of the face is used for matching and recognition. Deciding whether a severe pan movement exists in the image or not is made by comparing full vs. half face energies of the match.

6.6.1 Computational Efficiency

In order to compare the running times of the proposed multi-scale method based on super-coupling transform with that of Chapter 4 [114] fairly, we consider the method of Chapter 4 using pixelwise nodes instead of the block model. We let the two algorithms iterate till both reach an equal energy value. In Figure 6.5 the energy plots of the two algorithms are visualized per Log number of iterations. As it is evident from the plots, the multi-resolution analysis achieves lower energy in fewer number of iterations. Of primary importance is that the computational complexity of one iteration for the method in [114] is much higher than that of the proposed multi-scale method. The efficiency in one iteration of the multi-scale scheme comes from two sources. First, the complexity of one message passing operation in each scale

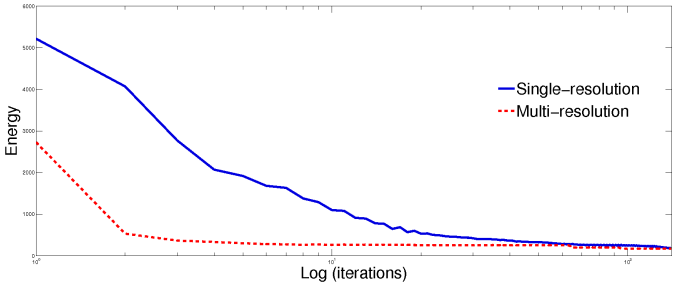


Figure 6.5: Comparison of multi-resolution vs. single-resolution matching in terms of energy of the match.

of the multi-resolution scheme is considerably less than that of the single-scale analysis due to the decimation of the disparity in coarser scales and the constraints imposed by a coarser scale onto the next finer scale which effectively prunes out the configurations that are not consistent with the previous coarse scale. Secondly, the number of nodes in all scales but the finest is less than that of [114]. These two factors lead to a considerable speed-up in matching by a factor of more than 10. But as discussed in Chapter 5 we do not use the matching method of [114] to establish pixelwise correspondences. Instead, the block model is employed. The proposed multi-resolution approach when compared to the method of Chapter 4 constructed using block model (instead of pixel level matching) is up to three times faster.

Next, we visually compare the quality of match obtained by the two methods by running the two algorithms for an equal number of iterations. As before, the method in [114] is constructed using pixelwise nodes rather than block model. Figure 6.6 illustrates a pair of images to be matched. In the same figure, the results of matching the template to the target image by warping the template image according to the found deformation are shown. As can been seen from the results, the multi-scale method leads to better results visually. This is due to the fact that in the multi-scale analysis, the method becomes less prone to be affected by noise in images as a result of using groups of pixels as nodes in the coarser scales and also coarsening the configuration space which enables the method to take big jumps towards the minimum.

6.6.2 Performance Gains in Face Identification

A question that might be asked is that whether the proposed multi-scale analysis and shape prior constraint lead to any improvement in terms of identification rate when compared to

Figure 6.6: Comparison of multi-resolution vs. single-resolution matching in terms of quality of the match. From left to right: template, target, multi-resolution result, single-resolution result

the matching method of Chapter 4 [114]. In this section we show the benefits of using the proposed method in this work over that of [114]. We run the method in [114] using the block model. In order to compare the improvements over the method in [114] we keep texture and shape modeling the same for the two methods and just compare the two different matching methods. In our experiments we use a subset of the CMU-PIE database [115]. The subset is used to evaluate the performance of the system subject to pose changes, under almost constant illumination conditions with neutral expression. This subset, consists of 884 images of 68 subjects viewed from 13 different angles. Nine images of each subject are captured roughly at the head height, spanning a range of angles from approximately full right profile to full left profile. Two images are captured from the corners of the room while the other remaining two are captured from above and below the central camera. For this comparison, we use the images corresponding to frontal pose (pose 27) as gallery images and three different poses corresponding to full profile (pose 22), images with slight pose deviation in horizontal direction (pose 05), and images with slight pose deviation in vertical direction (pose 07) as test images. The performance of the proposed method with that of Chapter 4 [114] is reported in Table 6.2. The method constantly outperforms the one in [114] in terms of identification performance. As evident from the results, the proposed approach can be particularly beneficial where there are severe head pose changes (*e.g.* pose 22).

Table 6.2: Comparison of the performance of the proposed matching method to the method in [114] in terms of Identification Rate.

Pose	c05	c07	c22
The method in [114]	85	91	22
The proposed multi-resolution matching	98	98	79

6.6.3 Comparison with Other Face Recognition Algorithms

In this section we compare the performance of the proposed heuristic approach (denoted by HM) and the one employing RGT (denoted by SM) with some other state-of-the-art methods for face recognition on the CMU-PIE database. In this experiment, as in the previous section we use frontal views of subjects (pose 27) as gallery images while the all the other poses are considered as probe images.

Table 6.3 presents the results and compares them to some other approaches. The results for the other approaches are taken from the literature. We also provide relevant details about the approaches in Table 6.4. In these methods we have provided the identification rates reported using frontal images (pose C27) as gallery images. It can be observed that the proposed techniques achieve comparable or sometimes better performance compared to most of the other approaches with less restrictive assumptions. In the comparison one needs to take into account the following considerations. The best performing method among other approaches in Table 6.3 is [146] which uses non-frontal gallery images as well as frontal images for training, whereas we do not use any non-frontal training images. Also, the method in [146] uses 80 manually labeled landmark points, whereas the methods proposed here does not need any landmark points and only require the face to be detected in a bounding box which is much easier than providing landmarks automatically. Other advantages of the methods proposed here is that they can also cope with moderate global spatial transformation (*e.g.* projective) between the images. The performance of the proposed approach for matching in terms of face identification on the CMU-PIE database is also compared with another registration method [46]. The method in [46] proposes a face recognition pipeline composed of face detection, landmark localization, feature extraction and identification. As we want to compare the performance of the registration method in [46] to our methods, we feed the systems with the same face images, bypassing the detection method in [46]. For this experiment we use the online code provided

Table 6.3: Comparison of the performance of the proposed approach to the state-of-the-art methods on the CMU-PIE database.

Pose	**C02**	**C05**	**C07**	**C09**	**C11**	**C14**	**C22**	**C25**	**C29**	**C31**	**C34**	**C37**
Pan deviation	$-44°$	$-16°$	$0°$	$0°$	$32°$	$47°$	$-62°$	$-44°$	$17°$	$47°$	$66°$	$-31°$
Tilt deviation	$0°$	$0°$	$-13°$	$13°$	$0°$	$0°$	$1°$	$11°$	$0°$	$11°$	$1°$	$0°$
L-fields [59]	58	94	89	94	88	70	38	56	57	56	47	89
PDM [56]	72	**100**	na	na	94	62	na	na	98	na	20	97
AA-LBP [146]	95	**100**	**100**	**100**	**100**	91	na	89	**100**	80	73	**100**
3D model [108]	76	99	99	99	93	87	50	75	97	78	49	94
Norm. [46]	11	51	48	54	26	8	3	8	51	13	4	22
Deformable block matching (Chapter 5)	95	98	98	**100**	89	91	**79**	**95**	91	88	**83**	**100**
SM	95	98	98	**100**	91	91	**79**	**95**	92	88	**83**	**100**
HM	**98**	98	**100**	**100**	97	**100**	77	**95**	**100**	**95**	72	98

Table 6.4: Some specifications of the methods in Table 6.3 and test details.

Method	**Non-frontal training image**	**no. of landmark points used**	no. of subjects used
L-fields [59]	**Y**	**39-54 depending on pose**	**34**
PDM [56]	**Y**	**62**	**68**
AA-LBP [146]	**Y**	**80**	**68**
3D model [108]	3D data	> 6	**68**
This work	**N**	**None**	**68**

by the authors. We use the landmarks obtained by running the method in [46] to normalize the image using an affine transformation. The rest of the procedure is kept the same as advocated in this work. The results obtained are reported in Table 6.3, denoted as *Norm*. One observes that both of the proposed approaches can outperform the method in [46] by a large margin. This can be attributed to a number of reasons. First, the method in [46] fails for large pose deviations from frontal. This leads to a large error in geometric normalization. Second, even in near frontal poses when the method in [46] gives reasonable results, it is outperformed by the method in this work , this is because our method establishes dense correspondence which leads to accurate comparison of the two faces in terms of their shape and texture. In conclusion, the proposed approaches are especially beneficial when only one gallery is available per subject and test data includes images with large deviations from frontal.

6.6.3.1 Discussion

From Table 6.3 one observes that the heuristic algorithm (HM) performs slightly better than the one using the super-coupling transform (SM). In order to make the comparison fair one needs to take into account the following issues. First, we experimentally observed that SM is faster than HM. This can be justified as in SM one deals with subsampled displacements in the coarsest scale and only consistent labels in the finer levels. As a result, the message passing is less expensive. On the other hand, subsampling displacements has a counter-effect by making fine disparities undetectable in the coarser levels and hence introducing some noise into the registration. Also, objects we are dealing with are non-planar and using block-flat assumption does not seem to be the best choice to transfer labels from one level to the next finer resolution. Further investigation is required on transforming labels from one level of resolution to another for face matching.

A further point is the performance of the multi-resolution methods in comparison with the matching method of Chapter 5. The SM method achieves the same identification performance as that of Chapter 5 but as noted earlier, it is up to three times faster. The HM achives better identification performance compared to the method of Chpater 5 but the computational effieciency is marginal.

6.6.4 Identification Test on the FERET Database

In this experiment we investigate the applicability of the SM approach to the problem of recognition of frontal and near frontal face images on a larger database and compare our results with the state-of-the-art methods. For this experiment we use images which have been captured at viewpoints *bf, ba* and *be* of FERET database which roughly correspond to $-15^\circ, 0^\circ, +15^\circ$. We use eye coordinates to crop frontal gallery images but for non-frontal images we use the face detection method in [73] to detect the face and enclose it in a bounding box. The results of the comparison with two other approaches on poses *be,bf* are provided in Table 6.5. It is worth noting that both methods, [13] and [74], use eye coordinates to crop the face out of the image whereas we use a face detector. Also, we used 200 subjects in the test and do not use non-frontal images in training whereas other methods use 100 subjects and use non-frontal images

Table 6.5: Comparison of the performance of the proposed SM approach to the state-of-the-art methods on the FERET database.

Pose	Kanade and Yamda [74]	Online Flow [13]	Stack Flow [13]	SM
bf	**85**	**80**	**90**	**88.5**
be	**80**	**70**	**82**	**89**

Table 6.6: Comparison of the performance of the proposed approach to the state-of-the-art methods on the FERET database.

Pose	bf	be
The method in [35] using manually labeled eye coordinates	93	92
The method in [35] using face detection data	54.5	56
SM using face detection data	88.5	89

for training. In spite of the aforementioned factors, it can be observed that the proposed SM method compares very favorably with other approaches.

In another experiment we compare the performance of the system with a state-of-the-art method developed for frontal images [35]. The method in [35] achieves 99.5% recognitions rate for the frontal pose *ba* and the performance of the proposed SM approach is 97% which is not far from the optimal performance. The results for the two near frontal poses are as in Table 6.6. For these poses, we report the results obtained using the method in [35] in two different situations. The first row in Table 6.6 reports the results using manually annotated eye coordinates. The second row of Table 6.6 reports the results obtained by the method in [35] without using the eye coordinates but using the face detection data as used in the SM method. From the table it can be observed that in a more realistic situation where the detection of eye coordinates becomes difficult (as a result of pose or expression change) while the performance of the state-of-the-art method in [35] drops significantly, the SM method performs reasonably well. The reason why the proposed approach does not outperform the method in [35], which uses the eye coordinates is that we use only one gallery image for texture comparison whereas in [35] the method uses large training data for multi-resolution LBP feature representation.

6.7 Summary

In this chapter we considered performing the MRF optimization for image matching in a multi-resolution framework. Two methods were proposed for multi-stage matching of images, one based on a heuristic search in the original configurational space and the other based on the super-coupling transform. The heuristic method was found to be slower than the one based on supercoupling transform. This was resulted by subsampling disparities in coarser scales of the SM approach and also considering only consistent solutions with the previous coarser scale in each level.

In terms of running times of the algorithms, the heuristic approach took approximate six minutes on an Intel Core(TM)2 Duo 3.0 GHz CPU for a pair of images of size 160×210, the same as the method of Chapter 5 but could establish *pixelwise* correspondences compared to block-wise matching in Chapter 5, resulting in superior recognition performance. By employing the multiresolution method based on supercoupling transform, we obtained considerable speed-up for the matching process by an average factor of three.

We provided experimental evaluation of these approaches with other methods on the FERET and the CMU-PIE databases. On the CMU-PIE database, in the presence of a large pose variation, the proposed methods compared very favorably with other approaches. On the FERET database, we simulate a real-world recognition system consisting of face detection, matching and then recognition. The proposed approach was then compared to some other approaches designed for pose-invariant recognition and also frontal face recognition and obtained superior results.

Chapter 7

Face Representation using a Sparse MRF Model

As denoted earlier in the previous chapters, the idea of dividing an object into its constituting parts and modeling their spatial interaction is established as an effective approach for object modeling and is commonly preferred [126] over holistic approaches [127] or part-based approaches which ignore configurational arrangements of object primitives [43]. The foregoing idea has been reflected in various applications, including image alignment for object recognition which can be considered as an integral part of all object recognition methods. Graphical models are widely applied in this context. However, one of the drawbacks of these models is the computational complexity of the inference in these models. In the previous chapter, two approaches were proposed to reduce the complexity of inference using multi-resolution analysis. The complexity of the optimization can further be reduced using a sparser model with reduced number of nodes.

Very often, the dependencies between object primitives are limited to pairwise interactions to simplify the model. However, recent works have shown that priors of higher order can provide better means to capture statistics of objects variations and hence result in superior performance in practice. Recently, the application of higher order priors has been at the center of attention as recent advances in MRF optimization provided the necessary tools for incorporating such information into the hypothesized models. Although higher order priors have been known to be useful, their applicability in the context of MRFs has been limited due to the curse of

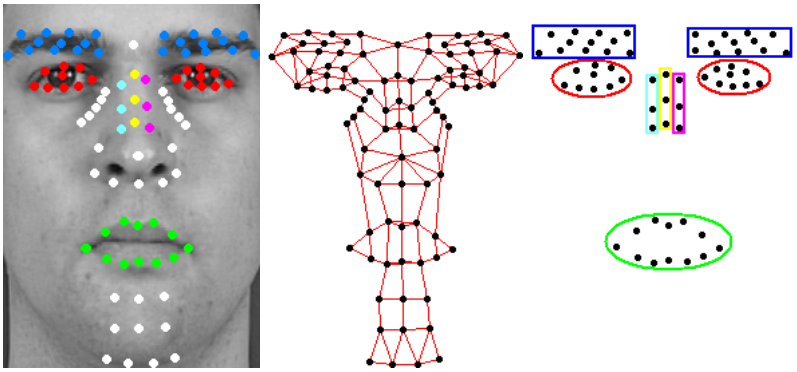


Figure 7.1: From left to right: landmark points used for constructing the face graph superimposed on a sample face image, graph illustrating binary connectivities, higher order cliques used for different face components.

dimensionality. Various approaches have recently addressed this issue and proposed different frameworks to incorporate a global prior into the model using various global optimization algorithms [135, 110, 80, 111]. In this chapter a graphical model using higher order shape priors for deformable face matching and feature localization is presented [10]. The proposed method learns the appearance and structure of faces in a probabilistic framework using a set of frontal training images in an unsupervised fashion and reduces the computational cost of MRF optimization by exploiting sparseness of facial features. Another novelty of the approach in the context of MRFs, is the incorporation of models of shape variation of the different components of face (*e.g.* eyes, mouth, *etc.*), based on point distribution models (PDM), as higher order clique potentials and formulate them as convex quadratic programming instances for which a variety of different approaches exist [18]. This is significantly important since it extends the application of higher order priors considered in specific forms [80, 110] to a more common and widely applied statistical model (PDM) for shape representation. While the application has been directed towards faces, the framework is more general and provides a principled way for incorporating such higher order statistical shape priors into graphical models. In the context of face recognition, the proposed methodology can be used in a variety of applications such as alignment of faces for geometric preprocessing or initialization of other approaches like 3D morphable model and dense image matching methods which are either computationally intractable or very expensive without proper initialization.

The chapter is organized as follows: In section 7.1 we briefly review the literature on methods

for face alignment. Section 7.2 explains the structure of the proposed model. In section 7.3 the energy functional combining texture and shape information is formulated. Section 7.4 discusses the approach we take for minimizing the energy. The experimental evaluation of the method on images collected from Google and on the rotation shots of XM2VTS database [95] is presented in section 7.5. Section 7.6 brings the chapter to a conclusion.

7.1 Related Work

Object matching and alignment has been studied extensively *e.g.* in [23, 113, 123, 144] and face matching and alignment is no exception. One popular category is based on the statistical models built from a set of representative samples [60, 93, 104]. Two primary examples of such methods are active appearance models and their extensions [37, 144, 15] and 3D morphable model [27]. The matching is usually expensive and often needs a good initialization. Another drawback of 2D statistical models based on AAMs is the lack of generalization capability in extreme poses, in spite of the view-based versions [40] and also the need for manual annotation of the training set. Other work in [39] employs a group-wise objective function to estimate non-rigid deformation. Other examples of the works based on shape constraints are [133, 112]. The work in [88] employs a component based discriminative approach using probabilistic shape constraints for face alignment. A similar approach to the current work is the elastic bunch graph matching (EBGM) approach [140]. Our approach differs from EBGM in various respects like node attributes, the geometric relations in the graphical structure and the optimization approach employed for inference. Also, the proposed method is completely unsupervised compared to EBGM for which manual intervention is needed for a number of images.

Another interesting group of methods [66, 86, 125] employ a cost function defined as sum of entropies and a sequential method to find the transformation parameters. The work in [41] is an extension of previous approaches which uses a sum of squared error functions for performing the alignment using the Lucas-Kanade's algorithm [91].

The work in [65] uses an elastic matching approach albeit discarding the relations between face image patches. Other works in [14] and [74] learn the changes in appearance of different patches of the faces using training data.

7.2 Graph Structure

Structural methods are based on the definition of a morphological model which is then combined with image measurements for object matching. In a sparse representation, the object is modeled using a number of landmark points. Then, given a statistical model, one learns from the training set independent and covariant probability distributions of the object appearance variations. These densities then enable one to describe the information contained in a new image based on the information observed in the training set. In practice, the training set is constructed by either manually labeling the landmarks for each instance of the object, or by inferring the landmark points by registering a labeled object to a set of un-labeled objects. In order to annotate our training set we used the method in [12] to register frontal images of 200 clients in the XM2VTS [95] dataset. In total we obtain a set of 1600 annotated images (8 images per each client) to train our model. It is worth reiterating that the proposed approach is completely *unsupervised* without any need for manual annotation of training images.

7.2.1 Selection of Landmark Points

Different approaches for representing and modeling faces use slightly different landmark points but the common characteristic is that the feature points are located around facial components *i.e.* eyes, eyebrows, nose, mouth *etc.* as in the active appearance models [37]. We discard the points on the contour of the face from our model since these regions lack distinctive features for matching. Assuming that important features of the face lie around the edges of facial components, after aligning a set of training images and averaging them, we chose a set of 92 landmark points based on the magnitudes of the edge map in the average face image manually as follows: 9 points for each eye, 12 points for each eyebrow, 12 points for the mouth, 10 points around the chin and finally 28 points for nose and surrounding regions. The set of adopted landmark points, superimposed on a sample face image, is illustrated in Fig.7.1.

7.2.2 Edges

Inclusion of an edge (connecting at most two nodes) between any two nodes of the graph is executed manually, based on the Euclidean distance. The aim is to ensure that a path exists

between every two nodes of each component of face without the need for traversing from the nodes which do not belong to the same component. A graph illustrating the binary relations is depicted in Fig. 7.1 (middle).

7.2.3 Hyperedges

As noted earlier, although limiting the cardinality of cliques produces computational efficiency, it may compromise the quality of the match. Point distribution models are useful for modeling shape variations. One can model the co-dependence of the positions of the nodes jointly as in non-MRF based approaches [38]. In the MRF framework, this can be achieved by incorporating a hyper-edge containing *all* nodes of the graph, the potential/cost of which is determined by the degree of deviation from the mean shape in the shape space constructed using PCA. However, we do not use such a clique and make use of *component-wise* hyper-edges in our model for the following reasons:

- Variations in the shape of different components of faces are almost independent of each other. In other words, representing deformations of all different parts of faces jointly using a unimodal distribution (*e.g.* Gaussian) is neither realistic nor sufficient.
- From an optimization point of view, marginalizing over a very large clique including all nodes is computationally inefficient.

We also include a set of higher order priors in our model to constrain certain nodes on the nose to lie on straight lines. The assumption is that if we select landmarks on the nose in a way that they lie on a straight line when viewed from frontal pose, then in presence of pose changes they will still almost remain on a straight line.

7.3 Energy Functional

Matching the model to an image involves maximizing the *a posteriori probability* of observing the model given the image. In the MRF context, the *a posterior probability* is defined in terms of a Gibbs distribution and one usually minimizes the $-log$ of a posteriori probability of the

model called the energy. Our energy functional comprises different terms, representing different aspects of shape and texture variations. The energy functional in the proposed graphical model has the following form:

$$En(x;\theta) = \sum_{v \in \mathscr{V}} \theta_v(x_v) + \sum_{E \in \mathscr{E}} \theta_E(\mathbf{x_E}) \tag{7.1}$$

$\mathscr{V}$ corresponds to the set of nodes of the graph and $\mathscr{E}$ to the set of cliques including higher order relations. θ_v and θ_E stand for the potentials of the nodes and edges/hyperedges. The cliques used in the current work have three different natures and cardinalities:

$$\mathscr{E} = \{BI, L, PDM\} \tag{7.2}$$

BI represents binary relations, *L* represents third-order relations and *PDM* stands for higher than third-order cliques, discussed in the following.

7.3.1 Unary Potentials

In the graphical model different features may be used as unary costs such as SIFT [90], shape contexts [19], Gabor features [69] and geometric blur [21]. Geometric blur features are known to be affine-invariant [21] and are used for shape matching and recognition [20].

The geometric blur descriptor is a blurred form of a signal around an interest point. The kernel used for blurring is partially varying, *i.e.* the amount of blur is assumed to be smaller in a close neighborhood of a reference point and larger in the regions further away. The idea behind employing such a varying kernel is that under an affine transform, the distance of a transformed signal sample point from a reference point is in linear proportion to the distance between these two points before applying the affine transformation. The feature is constructed around a neighborhood of an interest point. As a result, using the descriptor an affine-invariant comparison of the neighborhoods of feature points is possible. The geometric blur descriptor is extracted from sparse channels such as oriented edges. Because the signals this feature is computed from are sparse and the signals far away from a point under consideration is smooth, one can use a sub-sampled version of this feature in a neighborhood of a reference point. In our experiments we use positive and negative edge channels in horizontal and vertical directions to

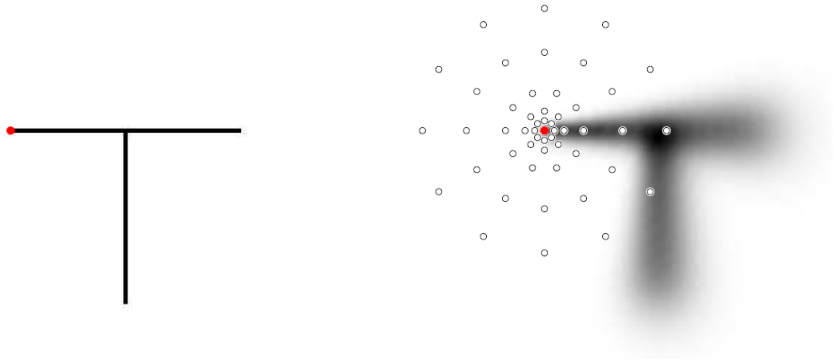


Figure 7.2: left: A sparse signal; right: the geometric blur around a feature point (red); image from [20]

extract geometric blur features. In order to obtain a blurred version of edge channels we use a spatially varying Gaussian kernel the standard deviation of which is linearly proportional to the distance from a feature point. We extract features in an area of 35 pixels radius around the feature points, sampled in 7 different radii and 18 different orientations.

7.3.1.1 Learning the Main Modes of Texture Variation

After extracting the geometric blur features for the 92 control points from our training set, we learn 92 probability density functions of texture variations. In the next step in order to remove redundancy and correlation effects and capturing the main modes of texture variation we apply PCA to the set of features at each node separately. Then the similarity between each candidate point in the test image and each control point in our model is measured in the PCA feature space.

7.3.2 Pairwise Potentials

The shape of an object can be modeled locally using binary relations. Using our training set, for each pair of nodes connected by an edge in the model graph, we estimate the mean and covariance of a 2D gaussian distribution for the relative positions of the nodes after aligning the training set using an affine transformation. We define the pairwise potentials of a pair of nodes in the graph in terms of the Mahalanobis distance from the mean configuration of the corresponding nodes:

$$\theta^{BI}_{(u,v)\in BI}(x_u,x_v) = [d_{u,v}(x_u,x_v) - m_{u,v}]^T C^{-1}_{u,v}[d_{u,v}(x_u,x_v) - m_{u,v}] \tag{7.3}$$

where $d_{u,v}(x_u,x_v)$ denotes the Euclidean distance between nodes u and v being assigned labels x_u and x_v. $m_{u,v}$ and $C^{-1}_{u,v}$ stand for the mean difference vector of coordinates and the inverse covariance matrix of their deviations obtained from the training set.

7.3.3 Higher-order Potentials

In order to model the variations of face shape more accurately, we make use of point distribution models and include them in the face model as higher order cliques. These cliques capture higher order statistical variations of shapes of different components of face. In our model, we consider one such clique for each eye, one per each eyebrow, three third-order cliques on the nose and one for the mouth, shown in Fig. 7.1. In addition we use another form of third order priors on the nose. In the following we first describe the clique potentials based on point distribution models and then the third-order potentials used for the nose.

7.3.3.1 Point Distribution Models

The shape of each facial component can be represented in a covariance space using a point distribution model. In order to construct a point distribution model for each component, we first align the training images using an affine transformation. The positions of all nodes contained in the clique under an affine transformation are used to estimate a mean shape for the corresponding component. Then in order to capture the main modes of shape variation, we apply PCA to the normalized positions of the nodes of the corresponding component in the training set. Then a configurational arrangement of a set of points in a facial component can be represented as:

$$Af(\mathbf{x}_{PDM}) = \Lambda + \Gamma \mathbf{w}(x_{PDM}) \tag{7.4}$$

where $\mathbf{x}_{PDM}$ and Λ correspond to the configuration of nodes of the clique in test image and their mean shape coordinates, respectively. Γ is the matrix of M principal eigenvectors ($M <$

2×cardinality of the clique) of the covariance matrix of the vectors of coordinates and $\mathbf{w}(x_{PDM})$ is the vector of weights. Af is a transformation, mapping the configuration (labels of nodes in the MRF) of nodes included in the clique into the corresponding spatial coordinates in the image frame under an affine transformation. The clique potential is then defined as:

$$\theta^{PDM}(\mathbf{x}_{PDM}) = \sum_{i=1}^{M} w_i^2(x_{PDM})/e_i \tag{7.5}$$

where e_i is the i^{th} eigenvalue. In practice one needs to minimize this potential in order to make the shape of the component under consideration as close as possible to those in the training set [143].

7.3.3.2 Linearity-based Priors

For the nose, we selected the landmarks in a way that every triplet of them lies on a line. The potentials of these cliques are defined differently from other higher order potentials. For these cliques, since we have selected them in such a way that all the nodes in a clique lie on a straight line, we impose a prior representing an error function measuring the offsets of the nodes from a line obtained by least square fitting. The above linearity assumption remains almost true even under pose and expression changes, since, compared to other facial components the nose is less deformable. In order to impose such priors we use regression to minimize the vertical offsets of points from the best fitted line. In this case the quality of the fit is given by a quantity known as *correlation coefficient* [55]:

$$r^2 = \frac{SS_{ii}SS_{jj}}{SS_{ij}^2} \tag{7.6}$$

where the quantities SS_{ij}, SS_{ii} and SS_{jj} are given by:

$$SS_{ii} = \sum_{n=1}^{N} (i_n - \bar{i})^2 \tag{7.7}$$

$$SS_{jj} = \sum_{n=1}^{N} (j_n - \bar{j})^2 \tag{7.8}$$

$$SS_{ij} = \sum_{n=1}^{N} (i_n - \bar{i})(j_n - \bar{j}) \tag{7.9}$$

$N = 3$ and i and j denote the horizontal and vertical axes and i_n and j_n stand for the horizontal and vertical coordinates of the n^{th} point. $\bar{i}$ and $\bar{j}$ stand for the average of the coordinates of

points in two directions. We take the linearity-based potentials to represent the overall quality of the fit:

$$\theta^L(\mathbf{x}_L) = \frac{SS_{ij}^2}{SS_{ii}SS_{jj}} \tag{7.10}$$

7.4 Minimizing the Energy Using Dual Decomposition

In order to minimize the energy we use the decomposition approach. We decompose the original problem (the so-called master problem) into several easier MRF subproblems (the so-called slave MRFs) on each of which exact inference is tractable and then extract a solution for the master by combining the solutions on the slaves which can be done based on an iterative projected subgradient approach. In other words, master acts as a coordinator which iteratively updates the costs of different configurations of each node in each subproblem separately so that the slaves agree on a common configuration at the end of the process. The method is essentially very similar to the TRW-S [79] of Chapter 4 but differs from that in certain ways. The similarities are that both methods use graph decomposition and the dual objective to find the MAP estimate. The difference is in the update of parameters in every iteration. While the TRW-S used a fixed point update (averaging min-marginals) to make the solutions consistent, the method we use here [81] uses sub-gradients of dual function to maximize it. The approach we take in this chapter enjoys better theoretical properties than the TRW family.

The work in [80] extended the above framework from pairwise case to minimize functions of arbitrary arities. The generic optimizer for the higher-order MRFs proposed in [80] decomposes the master problem into several slaves in such a way that a separate sub-problem exists for each higher order clique.

Similar to the TRW-S discussed in section 4.4, one requires $\hat{\theta} = \{\theta^\omega | \omega \in I\}$ to be a ρ-*reparameterization* of the original parameter vector θ [131] *i.e.*:

$$\sum_{\omega \in I} \rho_\omega \theta^\omega = \theta \tag{7.11}$$

Then for each subproblem a lower bound $LB_\omega(\theta^\omega)$ is defined which satisfies:

$$LB_\omega(\theta^\omega) \leq \min_x En(x, \theta^\omega) \tag{7.12}$$

It can be readily observed that the function

$$LB(\theta) = \sum_{\omega \in I} \rho_\omega LB_\omega(\theta^\omega) \tag{7.13}$$

is a lower bound of the original function in Eq. 7.1, *i.e.*

$$LB(\theta) \leq En(x^*; \theta) \tag{7.14}$$

where x^* is the optimal solution of Eq. 7.1.

Following the same framework, we decompose the original energy in such a way that a separate subproblem exists for each higher order clique of facial components including only the nodes of the corresponding component. Binary relations are decomposed into two edge-disjoint spanning trees. In the following we describe how to solve each of these subproblems.

7.4.1 Higher-order Subproblems

As noted earlier we associate one subproblem to each of our higher-order cliques. Two different kinds of higher order subproblems involved are solved as follows.

7.4.1.1 Higher-order Subproblems Based on PDM

In order to solve this kind of subproblem, one needs to optimize a slave problem having as prior the distance defined in Eq. 7.5. Hence the energy to minimize for each such clique is of the form:

$$\begin{aligned} En^{PDM}(\mathbf{x}_{PDM}) &= \theta^{PDM}(\mathbf{x}_{PDM}) + \sum_{u \in \mathscr{V}^{PDM}} \theta_u(x_u) \\ &= \sum_{u \in \mathscr{V}^{PDM}} \theta_u(x_u) + \sum_{i=1}^{M} w_i^2(x_{PDM})/e_i \end{aligned} \tag{7.15}$$

Considering a clique of n nodes, $\mathscr{V}^{PDM} = \{u_1, ..., u_n\}$, the problem of finding the optimum of the function in 7.15 is defined as

$$\min_{\mu} \{ \sum_{\mu_{u_1;j_1} ... \mu_{u_n;j_n}} \theta^{PDM} \mu_{u_1;j_1}(u_1; j_1) ... \mu_{u_n;j_n}(u_n; j_n) + \sum_{u;j} \theta_{u;j} \mu_{u_j;j}(u; j) \} \tag{7.16}$$

subject to

$$\sum_j \mu_{u_j;j}(u;j) = 1$$
$$\mu_{u_j;j}(u;j) \in \{0,1\}\,.$$

where μ represents all $\mu_{u_j;j}$.

Relaxing the integrality constraint we have $\mu_{u;j}(u;j) \in [0,1]$. In the following we show that the problem is in fact a *convex* quadratic programming problem. This then allows us to use convex optimization methods.

Let Y denote an assignment, that is a vector of binary values $\{0,1\}$ of dimension nL (n being the cardinality of the clique and L the number of admissible states for each node) which is obtained by concatenating n discrete sufficient statistics (μ) of dimensionality L, each of which has all its components zero except one component of value 1 indicating the state a node in the clique has been assigned. The problem of inferring a set of variables with a Gaussian prior can be written in a matrix form as [118]:

$$f(Y) = \frac{1}{2}Y^T HY + BY \tag{7.17}$$

s.t.

$$AY = I_n$$

A is a matrix of dimensionality $n \times nL$ which has all its elements zero except the elements in the range $[(row-1)L+1, row \times L]$ in each row. I_n is a vector of ones of dimension n. Under an affine transformation, matrix H and vector B are defined as:

$$\begin{aligned} H &= 2S^T M^T \Phi\Phi^T MS \\ B &= 2S^T M^T \Phi\Phi^T (t-\Lambda) + \theta^{PDM} \end{aligned} \tag{7.18}$$

where Φ and Λ are defined in Eq.7.4. S is a matrix of dimensionality $2n \times nL$ mapping an

assignment (Y) to its corresponding coordinates in the 2D image plane:

$$S = \begin{bmatrix} x_{1,1} & \dots & x_{1,L} & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ y_{1,1} & \dots & y_{1,L} & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & x_{2,1} & \dots & x_{2,L} & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & y_{2,1} & \dots & y_{2,L} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & x_{n,1} & \dots & x_{n,L} \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & y_{n,1} & \dots & y_{n,L} \end{bmatrix} \tag{7.19}$$

$x_{n,l}$ and $y_{n,l}$ denote horizontal and vertical coordinates of l^{th} candidate match for n^{th} node in the clique.

$M^{2n\times 2n}$ and $t^{2n\times 1}$ are the matrix (rotation and scale) and vector (translation) defining an affine transformation.

$$M = \begin{bmatrix} a & b & 0 & 0 & 0 & \dots & 0 & 0 \\ -b & a & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & a & b & 0 & \dots & 0 & 0 \\ 0 & 0 & -b & a & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & a & b \\ 0 & 0 & 0 & 0 & 0 & \dots & -b & a \end{bmatrix} \tag{7.20}$$

Assuming sf and r to be the scale factor and rotation angle respectively, $a = sf \times \cos(r)$ and $b = sf \times \sin(r)$

$$t = [t_x \;\; t_y \;\; t_x \;\; t_y \;\; \dots \;\; t_x \;\; t_y]^T \tag{7.21}$$

where t_x and t_y denote translations in two directions. It can be readily observed that matrix H in our problem is of the form DD^T hence positive semi-definite. Also, the points which satisfy the constraint form a convex set (any linear constraint defines a convex set). As a result, the quadratic program in Eq. 7.17 is a convex program. Quadratic programming is a well studied problem in nonlinear optimization field and many algorithms exist for optimization of such problems. In order to minimize the above function we use a method inspired by the work in [107] and use the following iterative algorithm:

Consider node $u1$ and suppose that values $\mu_{u_i;i}(u;i)$ are fixed for all other nodes $i \neq 1$, the optimal parameter $\mu_{u1;.}(u1;.)$ for node $u1$ is then given by:

$$\mu_{u1;.}(u1;.) = \underset{\mu_{u1;.}(u1;.)}{\arg\min}\{\sum_{\mu_{u_1;j_1}\ldots\mu_{u_n;j_n}} \theta^{PDM}\mu_{u_1;j_1}(u_1;j_1)\ldots\mu_{u_n;j_n}(u_n;j_n) + \sum_j \theta_{u_1;j}\mu_{u_1;j}(u_1;j)\}$$

subject to $\sum_j \mu_{u1;j}(u1;j) = 1$.

If for example we are looking for optimal $\mu_{u_1;j_1}(u_1;j_1)$, it can be obtained by taking:

$$j^*(u1) = \underset{j}{\arg\min}\{\theta_{u1;j} + \sum_{\mu_{u_1;j_1}\ldots\mu_{u_n;j_n}} \theta^{PDM}\mu_{u_1;j_1}(u_2;j_2)\ldots\mu_{u_n;j_n}(u_n;j_n)\} \tag{7.22}$$

and then setting $\mu_{u_1;j}(u_1;j) = [\![j^* = j]\!]$ where $[\![.]\!]$ is one if its argument is true and zero otherwise. The method above is iterated until none of the node changes its label.

The method is essentially a higher-order variant of the ICM (iterated conditional modes) approach. Theoretically the ICM algorithm terminates in a local minimum of the function. However because the objective function here is convex, a local minimum is a global one. Similar observations have been made in other works [107, 64]. The drawback of such method is that the speed of convergence is dependent on the initialization conditions. In this work, we initialize the above method using the configuration having minimum Euclidean distance to the mean shape which results in faster convergence.

7.4.1.2 Higher-order Subproblems Imposing Linearity Constraint

Solving these subproblems involves optimizing functions of the form:

$$En^L(\mathbf{x}_L) = \theta^L(\mathbf{x}_L) + \sum_{u\in\mathscr{V}^L} \theta_u(x_u) \tag{7.23}$$

Since the cardinalities of these cliques are not very large, we do an exhaustive search to find the optimum of these functions.

7.4.2 Binary Subproblems

In order to solve these subproblems we decompose the graph containing at most pairwise potentials into two spanning trees (Fig. 7.3). Exact inference on these trees is performed using

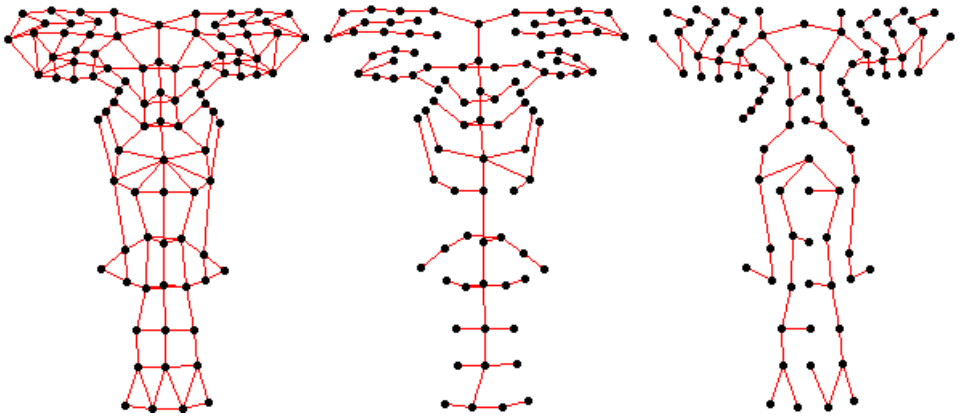


Figure 7.3: Decomposition of the pairwise loopy graph into two edge-disjoint spanning trees.

Table 7.1: Dual decomposition with sub-gradient updates [81]

1. Solve slave MRFs as described in the previous sections, *i.e.*

   compute $x^{slave} = \underset{x}{\text{argmin}} \; En(x, \theta^{Slave})$

2. Update parameters of each node in slave MRFs using sub-gradient updates [81]:

   $\theta_u^{slave} = \theta_u^{slave} + \delta_{iter}(x_u^{slave} - \frac{1}{N_u^{slaves}} \sum_{slaves} x_u^{slave})$

3. Repeat steps 1 and 2 till convergence.

max-product algorithm. The goal is to compute single-node and joint pairwise min-marginals of the energy ($En_T(X')$) associated with the tree distribution:

$$\begin{aligned} \Phi_{u;y}(\theta_T) &= \min_{\{X' | x'_u = y\}} En_T(X'; \theta_T) \\ \Phi_{uv;yy'}(\theta_T) &= \min_{\{X' | (x'_u, x'_v) = (y, y')\}} En_T(X'; \theta_T) \end{aligned} \tag{7.24}$$

the computation of which is done via max-product algorithm which facilitates computing MAP estimate by performing only local computations in tree structured distributions and message passing between adjacent nodes [131]. Having solved all the subproblems, the dual decomposition with subgradient updates works as in Table 7.1. In Table 7.1 N_u^{slaves} is the number of slaves containing node u and δ_{iter} is a positive diminishing step size.

7.4.3 Remarks

7.4.3.1 Interest Points

In order to match the model to an image we first sample the image coarsely. We first select 1000 interest points based on edge magnitudes. This is then followed by a regular sampling of the whole image in a coarser scale (*e.g.* 1 sample in every block of size 10×10) if no sample already exist in the block under consideration. The texture similarity of each node in our model is then compared to the samples and the most similar 50 samples are considered as admissible states for that node.

7.4.3.2 Visibility Assumption

In our experiments we have assumed that the feature points we are looking for are visible in the image and the model is forced to find a corresponding point for each node. Nevertheless, it is possible to include an occlusion label into the model along with a homogeneity constraint on the assignment of such label (as in [126]), if desirable.

7.4.3.3 Uniqueness Constraint

Another point to address is the uniqueness constraint which basically means that two nodes of the model cannot be matched to the same position. In order to impose such a constraint into the model, one option is to use a *linear assignment* subproblem. By solving this problem (*e.g.* by using the Hungarian algorithm [6]) uniqueness constraint is imposed on the linear subproblem and as a result on the optimal MAP solution.

7.4.3.4 Modeling Rigid Motion

Planar transformations are not the best choice for modeling the rigid motion of faces but they are simple and computationally efficient. An alternative can be to estimate local transformations for different parts of the face. However, we have not pursued this because in that case the estimated transformations would be more prone to errors before convergence and might cause the model to deviate from the true solution. In practice, we estimate an affine transformation

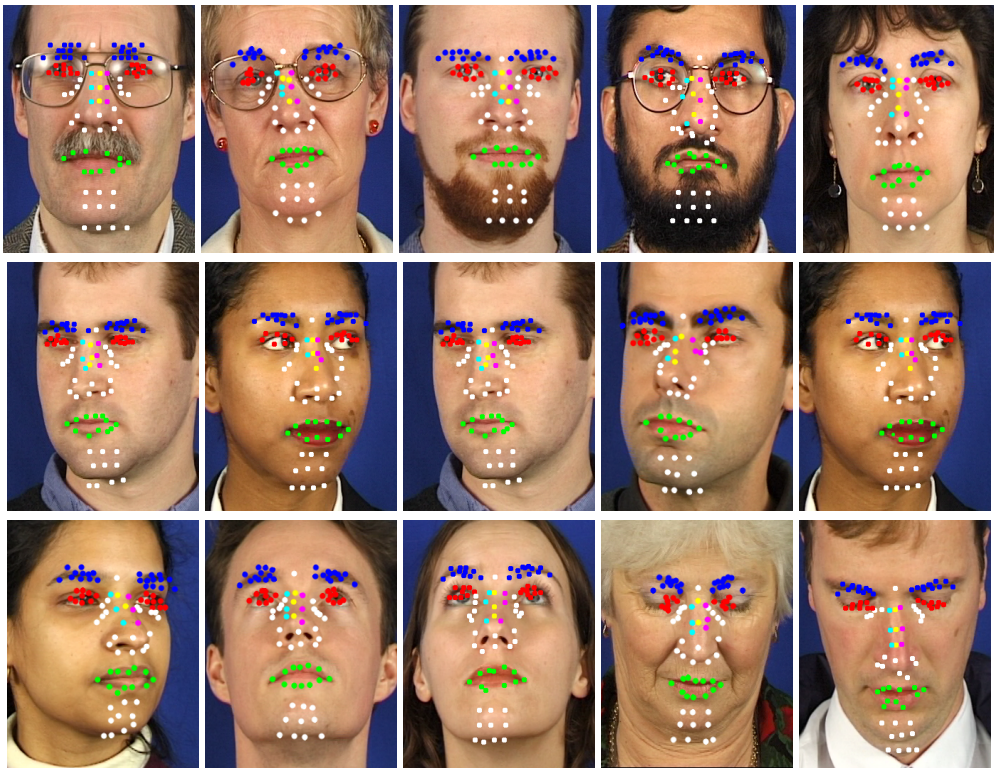


Figure 7.4: Facial feature localization on the XM2VTS images in presence of facial hair and glasses and in-depth rotation.

for the whole model using Levenberg-Marquardt method along with RANSAC and refine the transformation as the optimization proceeds. This transformation can then be used to update the binary relations according to the estimated transformation and make the model even more robust to global geometric transformations.

7.5 Experimental Evaluation

In this section we provide some experimental results first for matching the model to the different face images and illustrate the results. Next, we use the method in a verification scenario on the challenging rotation shots of the XM2VTS dataset [95]. In the experiments for landmark localization the same color coding of Fig. 7.1 is used.

7.5.1 Images Taken From XM2VTS Dataset

7.5.1.1 Frontal Images

We first evaluate the performance of the method on frontal images of the XM2VTS dataset. Some examples are shown in Fig. 7.4. As it is evident from the results, the method can detect landmarks very accurately.

7.5.1.2 Partial Occlusion due to Beard and Glasses

Next, we test the method against partial occlusion due to glasses and beard. As can be seen in Fig. 7.4, the method is quite robust to such occlusions and handles them very well even though some features are completely occluded due to beard.

7.5.1.3 Pose Variation

In order to assess the ability of the model (constructed purely from frontal images) to cope with non-frontal poses, we match the model to the rotation shots of the XM2VTS corpus. Some results are illustrated in Fig. 7.4 second and third rows. As can be seen from the examples, the model generalizes very well to non-frontal poses and performs reasonably well in the presence of severe pose changes.

7.5.2 Google Image Dataset

Next, in order validate the performance of the method across different databases, we use the images collected in [65]. These are images collected from Google taken in real world conditions. Some examples are illustrated in Fig. 7.5. In the figure we have also included the results obtained by the CMU's face alignment system [60] for comparison. We have taken the results of this approach provided in [2] and test our model on the same set of images.

The results of landmark localization are color coded the same as Fig. 7.1. The evaluation confirms that the proposed method can detect landmarks on these images accurately in spite of the facts that it has been trained on a seperate set of images which only contained frontal

pose. The variations observed in these images include scale changes, self-occlusion, variation in illumination, various deformations and severe pose changes. It can be observed that the method generalizes very well across different databases and can handle various appearance changes, such as pose differences, outperforming the CMU's approach which fails in cases where in-depth rotation is present in the image. In these cases when most of times the CMU's technique fails to operate properly and assigns some parts of the background to the face region, the proposed technique performs reasonably well. This is achieved in spite of the fact that we have trained our model using only frontal images.

7.5.3 Face Verification on the Rotation Shots of XM2VTS Dataset

In the XM2VTS data set, the evaluation protocol is based on 295 subjects consisting of 200 clients, 25 evaluation imposters and 70 test imposters. Two error measures defined for a verification system are false acceptance and false rejection given below:

$$FA = EI/I * 100\%, \qquad FR = EC/C * 100\% \tag{7.25}$$

where I is the number of imposter claims, EI the number of imposter acceptances, C the number of client claims and EC the number of client rejections. The performance of a verification system is often stated in *Equal Error Rate* (EER) in which the FA and FR are equal on the test set. After detecting the feature points using the proposed model, in order to find dense correspondences between gallery and test images one may take different approaches such as using a regularized thin plate spline or a dense matching algorithm initialized by the landmarks located by the proposed model. We take the second approach and employ the method in [11] initialized by the landmark points detected to establish pixel-wise correspondences between gallery and test images. Using the sparse matching as an initialization step, the runtime of the method in [11] reduces by a factor of 3. In order to extract features we use histograms of uniform LBP patterns in a circular neighborhood [98] and compare the histograms using the χ^2 distance. For shape comparison, the deformation of a pair of faces is measured as proposed in Chapter 5. The ROC curves using shape and texture information on the pan and tilt poses are illustrated in Fig. 7.6. The performance of the method is also compared to another approach in Table 7.2. The one in [122] is based on a 3D pose normalization with the recognition performed

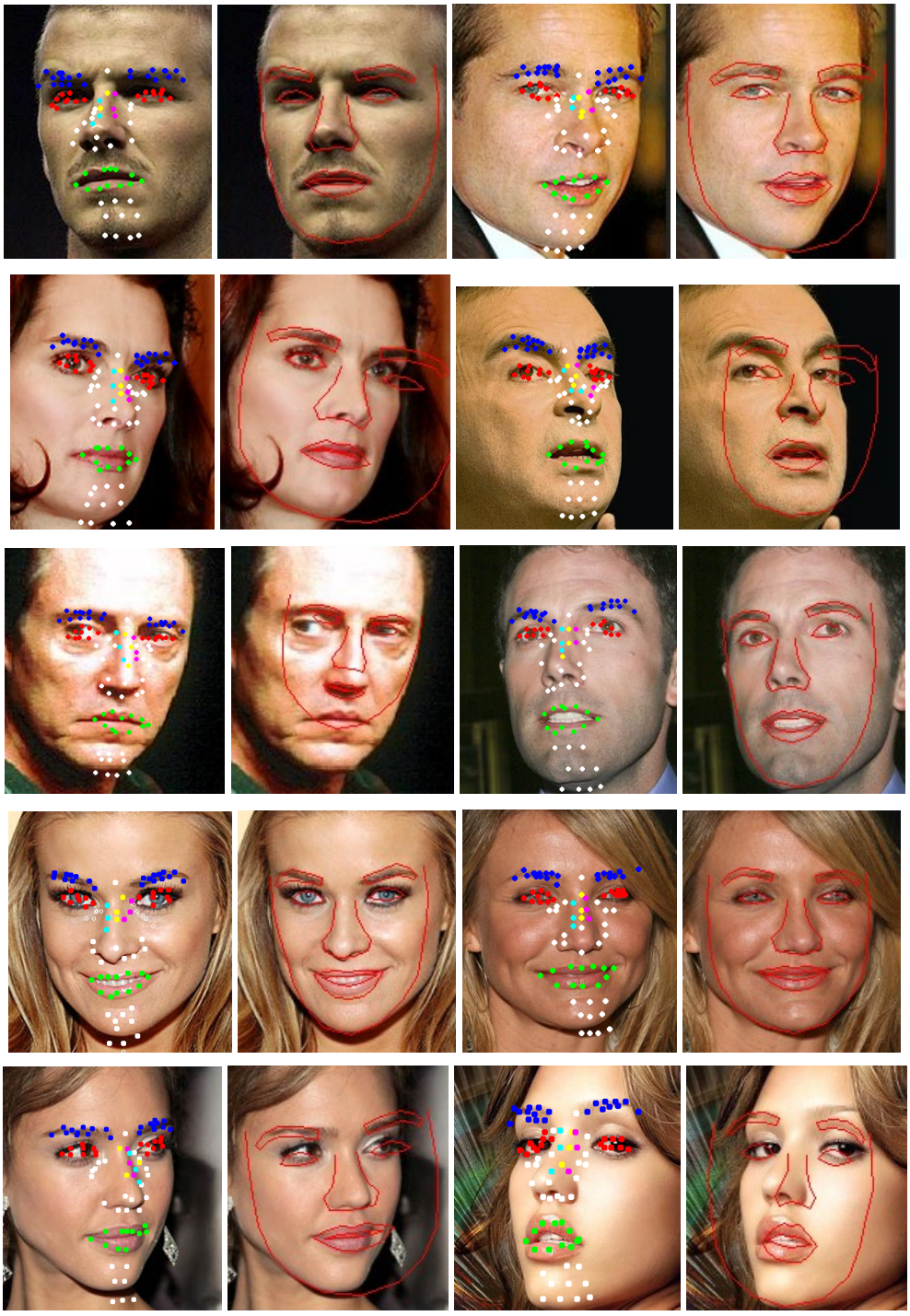
Table 7.2: Comparison of the current work to another method

Method	This work	3D pose correction [122]
EER	6.45	7.12

on the resulting 2D image. Note that the EER achieved in this work is lower than that of 3D pose normalization in [122].

7.6 Summary

The chapter presented an MRF model for deformable face matching. The approach used statistical models of texture and shape variations in building the model. Inference over the higher order statistical shape priors based on point distribution models were shown to be instances of convex quadratic programs and solved by an iterative primal algorithm. The evaluation of the approach for feature localization was performed both on the XM2VTS dataset images and images collected from Google. The method showed high performance in localization of facial features when applied to frontal face images and a very good generalization capability when applied to facial images with severe pose changes and partial occlusion. The method, was then used as an initialization step for a more costly dense matching approach, and was found to be instrumental in reducing runtime.



(a) Graph matching (b) CMU's method [60] (c) Graph matching (d) CMU's method [60]

Figure 7.5: Results on the images collected from Google compared to CMU's method. The results of using the proposed method are illustrated in columns (a) and (c) and compared to the CMU's approach on same set of images in columns (b) and (d).

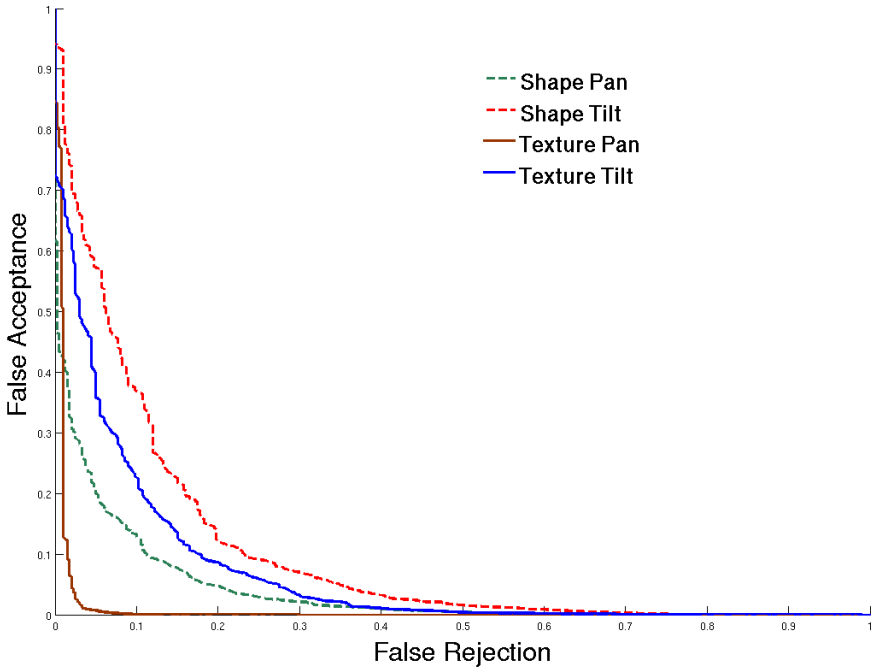


Figure 7.6: ROC curves on the rotation shots of the XM2VTS corpus.

Chapter 8

Conclusions and Future Work

In this thesis we considered the application of graphical models to face recognition. The motivation behind taking such an approach was two fold. First, these models are particularly useful in image modeling and analysis due to their inherent capabilities in handling various degradations and imperfections in imaging conditions. Graphical models also offer a suitable platform for designing recognition algorithms by fusing different sources of information in a probabilistic framework. In the thesis we used two different graphical models for face representation and recognition. The first one proposed in [114] was adapted to our problem of object instance recognition by a number of innovative modifications. We also proposed a sparse MRF model for deformable face matching and landmark localization. In the proposed model different aspects of shape and texture variations were considered and incorporated into the model in a probabilistic framework. Secondly, finding the maximum a posteriori probability configuration of these models is facilitated by the recent advances in the field of optimization and inference in MRFs. The techniques proposed in the past few years enjoy very good convergence and optimality properties compared to the earlier versions. We used two of them in the thesis. The methods we used belonged to the graphical decomposition family.

In the rest of this chapter and in section 8.1 we provide a summary of the conclusions drawn from the current work followed. Section 8.2 discusses possible future paths of research enabled by this study.

8.1 Conclusions

One of the paths of the study conducted in the thesis was the problem of measuring goodness-of-match in terms of the maximum a posteriori probability of a graphical model. Naturally the maximum probability of a match or equivalently the minimum energy which is optimized is expected to provide a good measure of similarity in a graphical modeling framework. However, as discussed in Chapter 5, various factors affect the energy terms, not all of which directly lead to a meaningful similarity for recognition.

The first variation which one requires the recognition system to be invariant to is the global geometric transformation, such as scale or rotation. In order to remove the effects of such perturbations two approaches were taken. First, in terms of shape distances we removed the effects of spatial transformations but fitting a projective transformation and subtracting the resulting displacements from the displacement vectors. The residual local distortion measure is then a better estimate of shape similarity between a pair of images. Second, as the two images to be matched are not geometrically aligned before the matching, we performed the matching two times. In the first round we estimated a global spatial transformation aligning the two images. We then used the estimated aligning transformation to deform the rectangular blocks on the frontal gallery images according to the estimated transformation to find their corresponding patches in a non-frontal image.

The other factor which adversely affects the enemy term is the clique cardinality. The limitation of the clique cardinality is often imposed by the computational cost of inference in graphical models. Such a limitation is useful in terms of efficiency but not advantageous in terms of match quality and recognition accuracy. A closely related factor to the cardinalities of the cliques in an MRF is the error correlation effect. This essentially translates to the dependence of local deformations of an object when converting it to another shape. The 1D case of the problem was studied before in speech recognition. There, it was shown that such local errors are very much correlated to each other, and considering such covariance effects could improve the recognition error. We treated both of the above problems (limited clique cardinality and correlation effects), up to certain extent, using covariance modeling of residual local distortions. The covariance matrices we employed effectively captured long and short range correlations between distortions of different parts of an image and enhanced the recognition

performance.

The other factor which can unfavorably influence the energy is the match similarity conveyed by the data term. This captures the unwanted environmental effects posed for example by illumination variation, sensing device noise *etc.* and translates into inconsistencies between the color/greyscale of similar objects. We removed the effects of such variations using a photometric normalization method followed by a texture descriptor which is invariant to monotonic changes in lighting conditions.

The proposed modifications to the energy term enhanced the method on par with the best performing approaches for pose-invariant recognition of faces using a simple nearest neighbor classifier.

We next considered the possibility of a multi-resolution MRF optimization. Inference in MRFs is computationally expensive and multi-resolution analysis can speed things up. We proposed two multi-stage methods for image matching. One heuristic and the other based on the super-coupling transform. The comparison with other face recognition approaches also verified the effectiveness of the proposed methodology.

Last but not least, we proposed a sparse graphical model for face graph matching and landmark localization. The model exploited the sparseness of facial features to reduce the complexity by decreasing the number of discrete variables in the model. The approach modeled shape variations of different facial components as higher order statistical cliques. It was shown that the inference over such cliques using a point distribution model under an affine transformation was essentially a convex quadratic program which could be easily solved with simple algorithms such as ICM (iterated conditional modes). We next used the sparse model to initialize the method used in previous chapters for speeding up matching. The extensive evaluation of the proposed approach showed that it had considerable merit.

8.2 Future Work

The research conducted during the course of this thesis suggests various avenues for future investigation. As a possible way of extension one may consider employing various classifiers in the space of the normalized energy of a match. Throughout the thesis we used a nearest

neighbor classifier designed in the space of shape and texture information. A further avenue of investigation is to consider other classifiers such as support vector machines which my provide better separation between the classes. In this respect one may also consider a multiple classifier fusion framework. Such an approach would be useful from two points of view. First, by using a faster classifier, one can reduce the number of competing hypothesises before extensive matching using an MRF methodology which is known to be more computationally expensive. Another advantage is to reduce error rates by combining independent information sources of information provided by different classifiers.

A further path of future research would be the implementation of the algorithms on parallel processing units such as GPU. One drawback of MRF-based approaches is their computational complexity which keeps these methods from being widely used in real-world commercial products. Without any doubt if one could perform classification in frame rate, such methods would be put into more practical use.

We considered matching full face or half face in the case of face partial self-occlusion. One possible direction of future investigation is to detect automatically occluded regions of the faces and only consider visible parts for recognition. Such an approach is more plausible using the sparser model of Chapter 7.

The current work could also be extended by considering different features as node potentials. For these potentials we used normalized edges and geometric blur descriptors. Future investigation is needed to find out the most suitable features for face representation. In addition, the priors one uses in an MRF are application specific. We used different forms of such priors as smoothness, linearity-based and point distribution models in our works. Further research is required to find the best suitable shape priors for pose-invariant matching and recognition of faces.

Also, one can enhance matching by combining top-down and bottom-up information concurrently. For example the detection and segmentation performed in the top layers can guide the bottom layer matching process and the bottom layer can enhance the segmentation in the upper layer. We believe that such an approach can make the algorithm to work better in realistic situations where detection, segmentation and matching are performed concurrently and improve each other.

Last but not least, in order to make the method less susceptible to unwanted lighting changes, it seems plausible to use the same methodology for matching faces captured via near infra-red imaging, since such images are believed to be less affected by illumination variations.

Bibliography

[1] http://www.ee.surrey.ac.uk/cvssp/demos/colour/soil47.

[2] http://www.vision.ee.ethz.ch/ zhuji/facealign/.

[3] K. Abend, T. Harley, and L. Kanal. Classification of binary random patterns. *Information Theory, IEEE Transactions on*, 11(4):538 – 544, Oct. 1965.

[4] A.R. Ahmadyfard and J. Kittler. Using relaxation technique for region-based object recognition. *Image and Vision Computing*, 20(11):769 – 781, 2002.

[5] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, Dec. 2006.

[6] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, February 1993.

[7] S.M. Aji and R.J. McEliece. The generalized distributive law. *Information Theory, IEEE Transactions on*, 46(2):325 –343, Mar. 2000.

[8] S.R. Arashloo and J. Kittler. Pose-invariant face matching using mrf energy minimization framework. In *EMMCVPR*, pages 56–69, 2009.

[9] S.R. Arashloo and J. Kittler. Energy normalization for pose-invariant face recognition based on mrf model image matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010,accepted.

[10] S.R. Arashloo, J. Kittler, and W. Christmas. Facial feature localization using graph matching with higher order statistical shape priors and global optimization. In *BTAS'10:*

Proceedings of the 4rd IEEE international conference on Biometrics: Theory, applications and systems. IEEE Press, 2010.

[11] S.R. Arashloo, J. Kittler, and W. Christmas. Pose invariant face recognition by matching on multi-resolution mrfs linked by supercoupling transform. *Computer Vision and Image Understanding*, ., 2010, under second review.

[12] S.R. Arashloo and J.V. Kittler. Hierarchical image matching for pose-invariant face recognition. In *BMVC*, 2009.

[13] A.B. Ashraf, S. Lucey, and T. Chen. Learning patch correspondences for improved viewpoint invariant face recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[14] A.B. Ashraf, S. Lucey, and C. Tsuhan. Learning patch correspondences for improved viewpoint invariant face recognition. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 1–8, 23-28 2008.

[15] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1380 –1384, Oct. 2004.

[16] O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. New York: John Wiley and Sons, 1978.

[17] A. Bartoli. Groupwise geometric and photometric direct image registration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2098 –2108, dec. 2008.

[18] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty. *Nonlinear Programming: Theory And Algorithms*. Wiley-Interscience, May 2006.

[19] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, April 2002.

[20] A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, pages 26–33, 2005.

[21] A.C. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, pages 607–614, 2001.

[22] C. Berge. *Graphs and Hypergraphs*. North Holland, 1973.

[23] A. Besbes, N. Komodakis, G. Langs, and N. Paragios. Shape priors and discrete mrfs for knowledge-based segmentation. pages 1295 –1302, Jun. 2009.

[24] D. Beymer and T. Poggio. Face recognition from one example view. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 500–507, Jun 1995.

[25] D.J. Beymer. Face recognition under varying pose. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 756–761, Jun 1994.

[26] V. Blanz, S. Romdhani, and T. Vetter. Face identification across different poses and illuminations with a 3d morphable model. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 192–197, May 2002.

[27] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003.

[28] M. Bober, M. Petrou, and J. Kittler. Nonlinear motion estimation using the supercoupling approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(5):550–555, May 1998.

[29] B. Bollobas. *Modern Graph Theory*. Springer, July 1998.

[30] K.W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+2d face recognition. *Computer Vision and Image Understanding*, 101(1):1 – 15, 2006.

[31] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 377 –384 vol.1, 1999.

[32] P. Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag New York Inc., corrected edition, February 2001.

[33] C.D. Castillo and D.W. Jacobs. Using stereo matching for 2-d face recognition across pose. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[34] X. Chai, S. Shan, X. Chen, and W. Gao. Locally linear regression for pose-invariant face recognition. *Image Processing, IEEE Transactions on*, 16(7):1716–1725, July 2007.

[35] C.H. Chan. *Multi-scale Local Binary Pattern Histogram for Face Recognition*. PhD thesis, 2008,University of Surrey.

[36] C.K. Chow. A recognition method using neighbor dependence. *IRE Trans. Electr. Computers*, EC-11(9):683– 690, Oct. 1962.

[37] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, Jun 2001.

[38] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[39] T.F. Cootes, C.J. Twining, V. Petrovic', R. Schestowitz, and C.J. Taylor. Groupwise construction of appearance models using piece-wise affine deformations. In *in Proceedings of 16 th British Machine Vision Conference*, pages 879–888, 2005.

[40] T.F. Cootes, K. Walker, and C.J. Taylor. View-based active appearance models. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 227–232, 2000.

[41] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least squares congealing for unsupervised alignment of images. pages 1 –8, jun. 2008.

[42] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[43] G. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, pages 634–640, 2003.

[44] G.J. Edwards, A. Lanitis, C.J. Taylor, and T.F. Cootes. Statistical models of face images – improving specificity. *Image and Vision Computing*, 16(3):203 – 211, 1998.

[45] B. Efron. The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376, 1978.

[46] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy automatic naming of characters in tv video. In *In BMVC*, 2006.

[47] P.F. Felzenszwalb, D.P. Huttenlocher, and J.M. Kleinberg. Fast algorithms for large-state-space hmms with applications to web usage analysis. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[48] P.F. Felzenszwalb and D.R. Huttenlocher. Efficient belief propagation for early vision. In *CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–261 – I–268 Vol.1, 27 2004.

[49] B.J. Frey. *Graphical models for machine learning and digital communication*. MIT Press, Cambridge, MA, USA, 1998.

[50] V. Gay-Bellile, A. Bartoli, and P. Sayd. Direct estimation of nonrigid registrations with image-based self-occlusion reasoning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):87 –104, jan. 2010.

[51] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721 –741, nov. 1984.

[52] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):643–660, Jun 2001.

[53] B. Gidas. A renormalization group approach to image processing problems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(2):164–180, Feb 1989.

[54] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through mrfs and efficient linear programming. *Medical Image Analysis*, 12(6):731 – 741, 2008. Special issue on information processing in medical imaging 2007.

[55] L. Gonick and W. Smith. *The Cartoon Guide to Statistics*. HarperResource, February 1994.

[56] D. Gonzalez-Jimenez and J.L. Alba-Castro. Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry. *Information Forensics and Security, IEEE Transactions on*, 2(3):413–429, Sept. 2007.

[57] D.B. Graham and N.M. Allinson. Face recognition from unfamiliar views: subspace methods and pose dependency. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 348–353, Apr 1998.

[58] G. Grimmett. A theorem about random fields. *Bulletin of London Mathematical Society*, 5(1):81–84, 1973.

[59] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(4):449–465, April 2004.

[60] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 413–426, Berlin, Heidelberg, 2008. Springer-Verlag.

[61] J-Y. Guillemaut, J. Kittler, M.T. Sadeghi, and W.J. Christmas. General pose face recognition using frontal face model. In *Progress in Pattern Recognition, Image Analysis and Applications: 11th Iberoamerican Congress on Pattern Recognition (CIARP 2006)*, LNCS 4225, pages 79–88, Cancún, Mexico, November 2006. Springer.

[62] J.M. Hammersley and P.E. Clliford. Markov models on finite graphs and lattices. *Unpublished manuscript*, 1971.

[63] F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using

markov random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(12):1217 –1232, Dec. 1993.

[64] K.C. Ho. Iterated conditional modes for inverse dithering. *Signal Process.*, 90(2):611–625, 2010.

[65] Gang Hua and A. Akbarzadeh. A robust elastic and partial matching metric for face recognition. pages 2082 –2089, sep. 2009.

[66] G.B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. pages 1 –8, oct. 2007.

[67] J. Huang, P.C. Yuen, W-S. Chen, and J.H. Lai. Choosing parameters of kernel subspace lda for recognition of face images under pose and illumination variations. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(4):847–862, Aug. 2007.

[68] R. Huang, V. Pavlovic, and D.N. Metaxas. A hybrid face recognition method using markov random fields. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 157–160 Vol.3, Aug. 2004.

[69] J. Ilonen, J.K. Kamarainen, P. Paalanen, M. Hamouz, J. Kittler, and H. Kälviäinen. Image feature localization by multiple hypothesis testing of gabor features. *IEEE Transactions on Image Processing*, 17(3):311–325, 2008.

[70] M. Quist J.M. Blackall1 A.D. Castellano-Smith T. Hartkens1 G.P. Penney W.A. Hall H. Liu C.L. Truwit F.A. Gerritsen D.L.G. Hill J.A. Schnabel1, D. Rueckert and D.J. Hawkes. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In Wiro Niessen and Max Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2001*, volume 2208 of *Lecture Notes in Computer Science*, pages 573–581. Springer Berlin / Heidelberg, 2010.

[71] H. Jiang, M.S. Drew, and Z-N. Li. Matching by linear programming and successive convexification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):959 –975, june 2007.

[72] J.K. Johnson. *Convex Relaxation Methods for Graphical Models: Lagrangian and Maximum Entropy Approaches*. PhD thesis, Sep. 2008.

[73] Z. Kalal, J. Matas, and K. Mikolajczyk. Weighted sampling for large-scale boosting. In *British Machine Vision Conference*, 2008.

[74] T. Kanade and A. Yamada. Multi-subregion-based probabilistic approach toward pose-invariant face recognition. In *Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on*, volume 2, pages 954–959 vol.2, July 2003.

[75] A. Kannan, N. Jojic, and B.J. Frey. Generative model for layers of appearance and deformation. In *In AIStats*, 2005.

[76] T.K. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):318–327, March 2005.

[77] D.R. Kisku, A. Rattani, M. Tistarelli, and P. Gupta. Graph application on face for personal authentication and recognition. In *Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on*, pages 1150 –1155, dec. 2008.

[78] J. Kittler and A. Lucas. A new method for dynamic time alignment of speech waveforms in pattern recognition and understanding. In P Laface and R De Mori, editors, *Speech Recognition and Understanding*, volume NATO ASI Series F - 75, pages 537–542. Springer-Verlag, 1991.

[79] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1568–1583, Oct. 2006.

[80] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. pages 2985–2992, 2009.

[81] N. Komodakis, N. Paragios, and G. Tziritas. Mrf optimization via dual decomposition: Message-passing revisited. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, 14-21 2007.

[82] N. Komodakis, G. Tziritas, and N. Paragios. Fast, approximately optimal solutions for single and dynamic mrfs. In *CVPR*, 2007.

[83] J. Konrad and E. Dubois. Bayesian estimation of motion vector fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(9):910 –927, sep 1992.

[84] M.P. Kumar, P.H.S. Torr, and A. Zisserman. Learning layered motion segmentations of video. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 33 – 40 Vol. 1, 17-21 2005.

[85] K.F. Lai and R.T. Chin. Deformable contours: modeling and extraction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(11):1084–1090, Nov 1995.

[86] E.G. Learned-Miller. Data driven image models through continuous joint alignment. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(2):236 –250, feb. 2006.

[87] S.Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[88] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 72–85, Berlin, Heidelberg, 2008. Springer-Verlag.

[89] X. Liu and T. Chen. Pose-robust face recognition using geometry assisted probabilistic modeling. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 502–509 vol. 1, June 2005.

[90] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[91] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI81*, pages 674–679, 1981.

[92] D.J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge Univ Press, 2003.

[93] I. Matthews and S. Baker. Active appearance models revisited. *Int. J. Comput. Vision*, 60(2):135–164, 2004.

[94] T. Maurer and C.V.D. Malsburg. Single-view based recognition of faces rotated in depth. In *Proceedings, International Workshop on Automatic Face and Gesture Recognition*, pages 176–181, 1995.

[95] K. Messer, J. Matas, J. Kittler, and K. Jonsson. Xm2vtsdb: The extended m2vts database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[96] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):696–710, Jul. 1997.

[97] H. Murase and S.K. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int. J. Comput. Vision*, 14(1):5–24, 1995.

[98] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, Jul 2002.

[99] B.G. Park, K.M. Lee, and S.U. Lee. Face recognition using face-arg matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12):1982–1988, Dec. 2005.

[100] J. Pearl. *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[101] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 84–91, Jun 1994.

[102] M. Petrou. Accelerated optimization via the renormalization group transform. *Complex Stochastic Systems and Engineering.London: Clarendon Press*, pages 105–120, 1993.

[103] P.J. Phillips, Hyeonjoon Moon, P. Rauss, and S.A. Rizvi. The feret evaluation methodology for face-recognition algorithms. pages 137 –143, jun. 1997.

[104] J. Pilet, V. Lepetit, and P. Fua. Fast non-rigid surface detection, registration and realistic augmentation. *Int. J. Comput. Vision*, 76(2):109–122, 2008.

[105] W. H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2 edition, October 1992.

[106] S.J.D. Prince, J. Warrell, J.H. Elder, and F.M. Felisberti. Tied factor analysis for face recognition across large pose differences. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):970–984, June 2008.

[107] P. Ravikumar and J. Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. In *ICML*, pages 737–744. ACM Press, 2006.

[108] S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3d morphable model using linear shape and texture error functions. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 3–19, London, UK, 2002. Springer-Verlag.

[109] A. Rosenfeld, R.A. Hummel, and S.W. Zucker. Scene labeling by relaxation operations. *Systems, Man and Cybernetics, IEEE Transactions on*, 6(6):420 –433, june 1976.

[110] C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1382–1389, 2009.

[111] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *CVPR 2006*, volume 1, pages 993 – 1000, 17-22 2006.

[112] J.M. Saragih, S. Lucey, and J.F. Cohn. Face alignment through subspace constrained mean-shifts. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1034 –1041, Sept. 2009.

[113] T. Schoenemann and D. Cremers. Globally optimal image segmentation with an elastic shape prior. In *ICCV*, pages 1–6, 2007.

[114] A. Shekhovtsov, I. Kovtun, and V. Hlavac. Efficient mrf deformation model for non-rigid image matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–6, June 2007.

[115] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 46–51, May 2002.

[116] Ajit Singh, Demetri Terzopoulos, and Dmitry B. Goldgof. *Deformable Models in Medical Image Analysis*. IEEE Computer Society Press, Los Alamitos, CA, USA, 1998.

[117] R. Singh, M. Vatsa, A. Ross, and A. Noore. A mosaicing scheme for pose-invariant face recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(5):1212–1225, Oct. 2007.

[118] T.P. Speed and H.T. Kiiveri. Gaussian markov distributions over finite graphs. *The Annals of Statistics*, 14(138-150), 1986.

[119] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):1068–1080, June 2008.

[120] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *AMFG*, pages 168–182, 2007.

[121] M. Taron, N. Paragios, and M.-P. Jolly. Registration with uncertainties and statistical modeling of shapes with variable metric kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):99 –113, jan. 2009.

[122] J.R. Tena, R.S. Smith, M. Hamouz, J. Kittler, A. Hilton, and J. Illingworth. 2d face pose normalisation using a 3d morphable model. In *Proceedings of the International Conference on Video and Signal Based Surveillance*, pages 1–6, September 2007.

[123] D. Terzopoulos, A.P. Witkin, and M. Kass. Constraints on deformable models: Recovering 3d shape and nonrigid motion. 36(1):91–123, August 1988.

[124] P. Tipwai and S. Madarasmi. A coarse-and-fine bayesian belief propagation for correspondence problems in computer vision. In *MICAI*, pages 683–693, 2007.

[125] Y. Tong, X. Liu, F.W. Wheeler, and P. Tu. Automatic facial landmark labeling with minimal supervision. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2097 –2104, 20-25 2009.

[126] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. In *ECCV '08*, pages 596–609. Springer-Verlag, 2008.

[127] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[128] M.A.O. Vasilescu and D. Terzopoulos. Multilinear image analysis for facial recognition. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 511–514 vol.2, 2002.

[129] M.J. Wainwright, T. Jaakkola, and A.S. Willsky. Tree consistency and bounds on the performance of the max-product algorithm and its generalizations. *Statistics and Computing*, 14(2):143–166, 2004.

[130] M.J. Wainwright and M.J. Jordan. *Graphical Models, Exponential Families, and Variational Inference*, volume 1. Now Publishers Inc., Hanover, MA, USA, 2008.

[131] M.J. Wainwright and T.S. Jaakkola A.S. Willsky. Map estimation via agreement on trees: message-passing and linear programming. *Information Theory, IEEE Transactions on*, 51(11):3697–3717, Nov. 2005.

[132] R. Wang, Z. Lei, M. Ao, and S.Z.Li. Bayesian face recognition based on markov random field modeling. In *ICB '09: Proceedings of the Third International Conference on Advances in Biometrics*, pages 42–51, Berlin, Heidelberg, 2009. Springer-Verlag.

[133] Y. Wang, S. Lucey, and J.F. Cohn. Enforcing convexity for improved alignment with constrained local models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, 23-28 2008.

[134] T. Werner. A linear programming approach to max-sum problem: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(7):1165–1179, July 2007.

[135] T. Werner. High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (map-mrf). In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[136] T. Werner. Revisiting the decomposition approach to inference in exponential families and graphical models. *Research report*, May 2009.

[137] T. Werner. Revisiting the linear programming relaxation approach to gibbs energy minimization and weighted constraint satisfaction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8):1474 –1488, aug. 2010.

[138] S.A.J. Winder and M. Brown. Learning local image descriptors. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, 17-22 2007.

[139] J.M. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *ICCV*, pages 756–763, 2005.

[140] L. Wiskott, J.M. Fellous, N. Kuiger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.

[141] J. Wright and H. Gang. Implicit elastic matching with random projections for pose-variant face recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1502 –1509, 20-25 2009.

[142] Zhong Wu, Qifa Ke, Jian Sun, and Heung-Yeung Shum. Scalable face image retrieval with identity-based quantization and multi-reference re-ranking. pages 3469 –3476, jun. 2010.

[143] Z. Xue, S.Z. Li, and E.K. Teoh. AI-EigenSnake: an affine-invariant deformable contour model for object matching. *Image and Vision Computing*, 20(2):77–84, February 2002.

[144] A.L. Yuille, D.S. Cohen, and P.W. Hallinan. Feature extraction from faces using deformable templates. In *Computer Vision and Pattern Recognition, 1989. Proceedings CVPR '89., IEEE Computer Society Conference on*, pages 104–109, 1989.

[145] J. Zhang and J. Hanauer. The mean field theory for image motion estimation. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 5, pages 197 –200 vol.5, 27-30 1993.

[146] X. Zhang, Y. Gao, and M. Leung. Recognizing rotated faces from frontal and side views: An approach toward effective use of mugshot databases. *Information Forensics and Security, IEEE Transactions on*, 3(4):684–697, Dec. 2008.

[147] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, 2003.