Learning Data Driven Representations from Large Collections of Multidimensional Patterns with Minimal Supervision

Parvez Ahammad

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-90
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-90.html

August 4, 2008

Learning Data Driven Representations from Large Collections of Multidimensional Patterns with Minimal Supervision

by

Parvez Ahammad

B.E. (Osmania University, India) 1998
M.S. (University of Central Florida, FL) 2002
M.S. (University of California at Berkeley, CA) 2005

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Engineering—Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor S. Shankar Sastry, Chair
Professor Jitendra Malik
Professor Sandrine Dudoit

Spring 2008

The dissertation of Parvez Ahammad is approved:

Professor S. Shankar Sastry, Chair	Date
Professor Jitendra Malik	Date
Professor Sandrine Dudoit	Date

University of California, Berkeley

Spring 2008

Learning Data Driven Representations from Large Collections of Multidimensional Patterns with Minimal Supervision

by

Parvez Ahammad

Abstract

Learning Data Driven Representations from Large Collections of Multidimensional Patterns with Minimal Supervision

by

Parvez Ahammad

Doctor of Philosophy in Engineering—Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor S. Shankar Sastry, Chair

Traditionally, taking experimental measurements of a physical or biological phenomenon was an expensive, laborious and very slow process. However, significant advances in device technologies and computational techniques have sharply reduced the costs of data collection. Capturing thousands of images of developing biological organisms, or recording enormous amounts of video footage from a network of cameras monitoring an observation space, or obtaining a large number of neural measurements of brain signal patterns via non-invasive devices are some of the examples of such data proliferation. Analyzing such large volumes of multi-dimensional data through expert supervision is neither scalable nor cost-effective. In this context, there is a need for systems that complement the expert user by learning meaningful and compact representations from large collections of multidimensional data (images, videos etc.) with minimal supervision. In this dissertation, we present minimally supervised solutions to two such scenarios generally encountered.

The first scenario arises when a large set of labeled noisy observations are available from a given class (or phenotype) with an unknown generative model. An interesting challenge here is to estimate the underlying generative model and the distribution

over the distortion parameters that map the observed examples to the generative model. For example, this is the scenario encountered while attempting to construct high-throughput data-driven spatial gene expression atlases from many thousands of noisy images of Drosophila melanogaster imaginal discs. We discuss improvements to an existing information theoretic approach for joint pattern alignment (JPA) in order to address such high-throughput scenarios. Along with the discussion of the assumptions, advantages and limitations of our approach (Chapter 2), we show how this framework can be applied to a variety of applications (Chapters 3, 4, 5).

The second scenario arises when there are observations available from multiple classes (phenotypes) without any labels. An interesting challenge here is to estimate a data driven organizational hierarchy that facilitates efficient retrieval and easy browsing of the observations. For example, this is the scenario encountered while organizing large collections of unlabeled activity videos based on the spatio-temporal patterns, such as actions of human beings, embedded in the videos. We show how some insights from computer vision and data-compression can be efficiently leveraged to provide a high-speed and robust solution to the problem of content-based hierarchy estimation (based on action similarity) for large video collections with minimal user supervision (Chapter 6). We demonstrate the usefulness of our approach on a benchmark dataset of human action videos.

Professor S. Shankar Sastry, Chair Date

Acknowledgements

Many people have helped me along the way to get to this point in graduate studies (and in life), and I owe my deep gratitude to them.

First, I would like to thank my academic advisor, Shankar Sastry, for taking me under his wing, for having faith in my ability, for his generous support, and for providing unmatched intellectual freedom to try new things out. His erudition, visionary insight into new directions, and dedication to standards of academic excellence are remarkable.

I have immensely enjoyed learning from Jitendra Malik, both through the classes he taught, and the insightful comments he made during vision group meetings. I thank him for being on my dissertation committee and providing useful feedback. I thank Sandrine Dudoit for reading my dissertation, and providing feedback. I thank Gerry Rubin for allowing me to work on the imaginal disc project with Cyrus. I thank Ruzena Bajcsy for being a wonderful human being, and for patiently listening to my ideas. I thank Kannan Ramchandran for feedback related to action recognition work, and Avideh Zakhor for supporting me during my first year at Berkeley. I also thank Michael Gastpar for being on my qualifying exam committee, and for a fun-filled signal processing class he taught.

Cyrus Harmon has been instrumental in introducing me to the psyche of a biologist, and helping me understand the biological aspects of gene expression atlas problem. Chuohao Yeo and I spent many hours brain-storming, playing pool, and discussing myriad topics. His criticism and feedback have been constructive and useful. Ram Vasudevan's enthusiasm kept my interest alive in the neuroscience related applications I discussed in this dissertation. While Erik Learned-Miller was a postdoc at Berkeley, he introduced me to the information theoretic ideas related to bias

removal in MRI images, and patiently guided me to the stage where I could extend the ideas and apply them into new domains on my own. I would like to thank Sue Celniker, Ann Hammonds, Richard Weiszmann and the Berkeley Drosophila Genome Project for help with data generation for the Drosophila imaginal discs, and Dr. Terrie Inder and Dr. Simon Warfield for providing the infant brain images (obtained under NIH grant P41 RR13218) used in this work. This research was partially supported by Army Research Office (ARO) grants DAAD 19-02-1-0383 and W911NF-06-1-0076.

Stéphane Coulombe, Sharon Core, Amar Mukherjee and David Chester played a pivotal role in my decision to apply for PhD; without their encouragement and support, I would not have made it to Berkeley. Some people have been instrumental in improving my understanding of various concepts, and in keeping my life sane when things weren't exactly going right. On this note, I want to thank (names listed in alphabetical order): Saurabh Amin, Alexander Berg, Mary Byrnes, Phoebus Chen, Alyosha Efros, Mikael Eklund, Padmapani Ganti, Christopher Geyer, Ruth Gjerde, Humberto Gonzalez, Maria Jauregui, Rajasekhar and Padma Kamada, Edgar Lobaton, Yi Ma, Carmel Majidi, Marci Meingast, Songhwai Oh, Bhanu Pappu, Pradeep Ragothaman, Jonathan and Mary Margaret Sprinkle, Todd Templeton, Claire Tomlin, Rene Vidal and Allen Yang.

Vivian Leung came into my life when I was badly stuck in a rut, and things were going astray; she managed to help me turn things around. I am incredibly thankful for her companionship, and for her presence in my life. I express my sincere gratitude to my parents (Abdul and Nasheela Rahiman) and my siblings (Shama and Firoz) for their constant love, support and faith. My journey toward PhD is built upon many a broken promise of returning home often, and I cannot thank them enough for their understanding and love throughout these years.

Dedicated to:

God

and

My beloved parents

Contents

Chapter 1

Overview

1.1 Motivation

The search for patterns in nature and the man-made world is an immensely interesting pursuit. If *a priori* knowledge is provided about the pattern that one is searching for, the task becomes that of aligning the prior with the observed patterns, and choosing one that satisfies the expectations. Approaches that follow this philosophy are denoted as the model-based approaches. In day-to-day life, there are many instances when we learn about the world around us even without having an *a priori* model. Given the proliferation of data in the modern world, there has been a renewed scientific interest in finding ways to automatically learn the underlying patterns simply by observing many examples.

One such scenario is illustrated by thinking about the task of teaching a child about some exotic fruit (say, a mango). This can be accomplished by showing a few examples of how a mango looks like: either by taking the child to a mango expo or by doing an image search on a reliable search engine and walking the child through the search results (see Figure 1.1). During the process of observing these examples,

somehow the child learns the implicit notion of what an innate representation of mango is, while also learning the range of variations that a mango can have. With enough examples, the child may also learn how to recognize a new instance of a mango, and reject an instance of a fruit that is not a mango. It is worthwhile to consider this scenario, and ask ourselves if a machine can be taught to do the same task as the child learning the compact representation of the space of mangoes; and if so, under what conditions. This scenario shows up in many different domains, such as high-throughput biology, medical imaging, surveillance, and many others. Our emphasis in this dissertation is on the notion of learning compact representations from examples, and not on using the learned representations for future decision-making. This difference leads to some key differences in the assumptions made in the mathematical formulation of the problem, as we will discuss later.

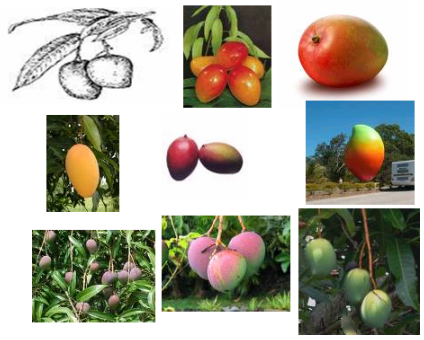

Figure 1.1: Sample image search results for mangoes

Another scenario is illustrated by thinking about walking into a newly constructed library, only to encounter a big heap of books and some empty shelves. When assigned the task of organizing these books into a meaningful order to facilitate easy access for the library users, we would naturally use our implicit understanding of how similar (or dissimilar) any two books are - and organize the books in such a way that similar books end up together on the shelves and dissimilar books end up far apart from each other. In other words, given a collection of books and the task of organizing them

(along with an implicit understanding of the similarity or dissimilarity of these books), the library user discovers a data-driven organizational representation that facilitates efficient use of the book collection. There are many scenarios in everyday world, where we encounter similar situations - for example, large collections (or databases) of images or video clips without a meaningful organizational structure. Being able to find a meaningful data-driven organizational representation simply based on content based similarity (or dissimilarity), but without any labels, would greatly facilitate the usability of such databases.

The two scenarios above (a child learning the compact representation of the space of mangoes, and a library user learning the data-driven organizational structure of a collection of books based on the implicit notion of similarity or dissimilarity) illustrate the general life examples that motivate the set of problems that we attempt to address in this dissertation.

1.1.1 Background

Mathematical characterization of the similarity of forms across different organisms was one of the subjects of exploration in the classic work of D'Arcy Thompson [Thompson, 1942] (first published in 1917). In the final chapter of his influential book, Thompson proposed the *theory of transformations* to show how the differences between the forms of related species can be represented geometrically. Figure 1.2 shows the transformations that are used to relate different species of fish to each other [Arthur, 2006]. While Thompson's work primarily dealt with geometric transformations in biological organisms, Grenander proposed significant generalizations in his *pattern theory* [Grenander, 1993]. Similar ideas have been extensively used in computer vision and medical imaging domains to solve alignment and recognition problems - especially from a geometric perspective [Thompson, 1942; Fischler and

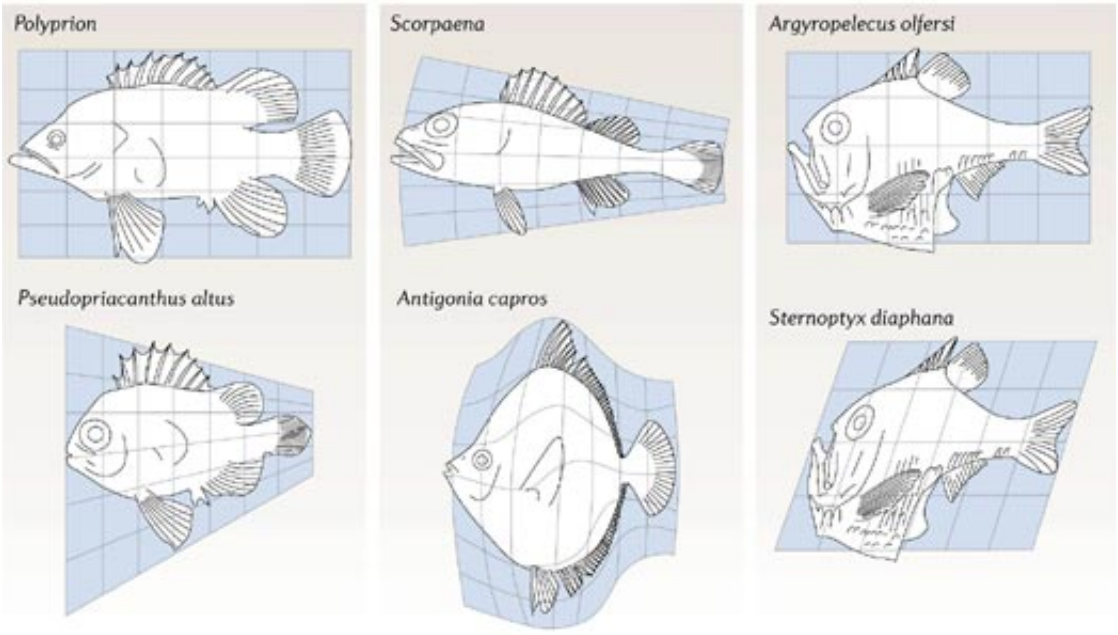

Figure 1.2: D'Arcy Thompson's Theory of Transformations
The deformed geometric coordinate frame around each fish shows the transformations that are used to relate different fish to each other.The simplest transformation is the 'shear' required to produce the form of *Sternoptyx diaphana* (bottom right) from that of *Argyropelecus olfersi*(top right)[Arthur, 2006].

Elschlager, 1973; Bajcsy and Kovacic, 1989; Yuille *et al.*, 1992; Grenander and Miller, 1994; Christensen *et al.*, 1996; Chui and Rangarajan, 2000; Belongie *et al.*, 2002; Berg, 2005].

1.2 Outline

1.2.1 Case A: Many Examples from One Class - Unknown Model

Let us assume that the given examples (multi-dimensional signals) are all various transformed instances of some unknown underlying model and assume that the structure of transformations is known. All examples are drawn from one class. Let us

denote the set of examples of a given class as $\Phi \doteq \{I^i\}_{i=1}^N$ where $N \in \mathbb{Z}_+$ is the cardinality of the set. Let I_l be the latent underlying model. $I_l(\cdot)$ and $I^i(\cdot)$ can be represented as: $I_l, I^i : \Omega \subset \mathbb{R}^P \mapsto \mathbb{R}$, where P indicates the dimension of the domain of I_l. Let g^i induce the mapping from $I^i(\cdot)$ to $I_l(\cdot)$ such that $g^i : \Omega \mapsto \Omega$. Thus, $I_l = I^i \circ g^i$. Now the challenge here is to estimate this unknown generative model I_l jointly from the set of examples Φ without making any prior assumptions on the model, while simultaneously estimating the various transformations $\{g^i\}_{i=1}^N$ affecting I_l. Given this formulation, we ask ourselves:

(1) *How can we infer the generative model*?

(2) *What real-world situations can we model using this formulation*?

(3) *Once these underlying model representations are learned, how can we adapt the learned representations to account for novel observed data*?

In Chapter 2, we choose one example scenario of aligning binary shape images (Section 2.1) and describe the problem formulation and algorithmic details for Joint Pattern Alignment (JPA) framework, that allows us to answer some of the questions above. In Section 2.5, we will describe some relevant work from the literature in the context of JPA. We discuss how useful representations (generative models) can be learned for multi-dimensional signals such as: (a) *Noisy shape templates undergoing geometric transformations (by combining segmentation and alignment)* (Chapter 3), (b) *Gray-scale anatomical images undergoing intensity variations* (Chapter 4), (c) *Spatio-temporal neural signals undergoing amplitude and phase variations* (Chapter 5), all using the JPA framework. We show how such systems can be used in practical applications such as: constructing high-throughput data-driven spatial gene expression atlases for Drosophila Melanogaster imaginal discs (Chapter 3), performing comparative gene expression analysis (Chapter 3), denoising the magnetic resonance (MR) images from random field (RF) bias (Chapter 4) and estimating the multicomponent event related potentials from single trial of neural signal recordings

(Chapter 5). In each chapter, we discuss the state of the art in the domain, as well as the future directions for the approach we have taken.

1.2.2 Case B: Many Classes, No Labels, Known Metric

In Chapter 6, given a large collection of videos containing activities, we investigate the problem of organizing it in an unsupervised fashion into a hierarchy based on the *similarity of actions* embedded in the videos. We use spatio-temporal volumes of filtered motion vectors to compute appearance-invariant action similarity measures efficiently (and robustly) - and use these similarity measures in hierarchical agglomerative clustering to organize videos into a hierarchy such that neighboring nodes contain similar actions. This naturally leads to a simple automatic scheme for selecting videos of representative actions (exemplars) from the database, for easily browsing the entire database and for efficiently indexing the whole database. We compute a performance metric on the hierarchical structure to evaluate goodness of the estimated hierarchy, and show that this metric has potential for predicting the clustering performance of various joining criteria used in building hierarchies. Our results show that perceptually meaningful hierarchies can be constructed based on action similarities with minimal user supervision, while providing favorable clustering performance and retrieval performance.

Finally, in Chapter 7, we discuss some future directions and potentially interesting extensions to the work we have presented in this dissertation.

1.3 Contributions

1) We have proposed a joint pattern alignment (JPA) algorithm that improves upon state of the art via principled choice of regularization in the optimization scheme. This allows users to encode domain knowledge systematically. This also makes JPA amenable to learning parameters of probability distributions in the situations where training data and user supervision are available (Chapter 2).

2) We have demonstrated a principled way of combining unsupervised alignment and segmentation into a semi-supervised algorithm for high-throughput joint pattern alignment that is tolerant to clutter or background noise. We demonstrated the applicability of this semi-supervised algorithm in geometric alignment of noisy shape templates used in constructing high-throughput spatial gene expression atlases (Chapter 3).

3) We have demonstrated the applicability of joint alignment approach for removing random field bias from magnetic resonance images (Chapter 4) and to denoise time-series signals such as event related potentials recorded from neuroscience experiments (Chapter 5).

4) We have demonstrated the applicability of unified framework of JPA for handling both spatial transformations (e.g., geometric variations) as well as non-spatial transformations (e.g., variable spatial bias, variations in latency or amplitude of signals), thereby addressing a variety of useful practical applications (Chapters 3, 4, 5).

5) We have proposed the idea of directly using compressed domain features (specifically, filtered motion vectors) as features for computing action similarity. A similarity measure built upon this idea has resulted in robust and efficient solution to compute the action similarity between any two given video clips. Due to the use of block-based sampling of motion field (inherent in compressed domain motion features) and a direct re-use of precomputed features from the compressed representation of the

videos, our solution to computing compressed domain action similarity is orders of magnitude faster than the state of art and robust to appearance variations, while providing high quality action recognition performance (Chapter 6).

6) We have proposed an unsupervised algorithm to efficiently organize large collections of high-dimensional data such as video clips based on content-based similarity (specifically, similarity of actions embedded in the videos). Our results show that perceptually meaningful hierarchies can be constructed based on action similarities with minimal user supervision, while providing favorable clustering performance and retrieval performance (Chapter 6).

Chapter 2

Joint Pattern Alignment via Entropy Minimization

The classical set-up of pattern matching is a pair-wise pattern alignment (PPA) formulation [Thompson, 1942; Fischler and Elschlager, 1973; Bajcsy and Kovacic, 1989; Yuille *et al.*, 1992; Grenander and Miller, 1994; Christensen *et al.*, 1996; Chui and Rangarajan, 2000; Belongie *et al.*, 2002; Berg, 2005]. In PPA, a model is either given or assumed, and a novel observation is warped through a pre-defined transformation space such that the warped observation matches the given model (Figure 2.1). One of the standard problems with pairwise pattern matching formulation is that the alignment result can suffer from local minima. In the joint pattern alignment formulation we only have access to a given set of examples - but no model is given. While the space of transformations is specified, the specific transformation parameters that map each example to the underlying model are also unknown. So joint pattern alignment procedure has to solve for both the model and the set of of warping transformations simultaneously (Figure 2.2).

In this chapter, we discuss an information theoretic formulation to solve the joint

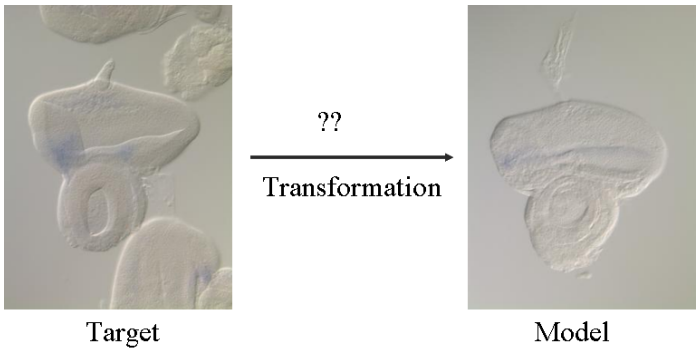

Figure 2.1: Pairwise Pattern Alignment Set-up
A model is given, and a novel observation (target) is warped through a pre-defined transformation space such that the warped observation matches the given model. The example images here are the eye/antennal imaginal discs of Drosophila melanogaster.

pattern alignment problem. Note that the idea of an information theoretic formulation for alignment problem is not new. Viola et al. and Collignon et al. introduced a novel information theoretic approach for solving pair-wise medical image registration problem [Viola and Wells, 1997; Collignon *et al.*, 1995]. Viola's method of alignment via maximizing mutual information has become the dominant method in the medical imaging community and has been widely adopted in many clinical applications. Learned-Miller et al. proposed a generalization of Viola's approach to joint group-wise (ensemble) alignment scenario (in the context of aligning and recognizing hand-written characters) [Miller *et al.*, 2000]. This joint ensemble alignment procedure called "congealing" [Miller *et al.*, 2000] utilizes the information across the whole ensemble, and is empirically shown to be robust against local minima. The purpose of this alignment procedure was to learn densities on transforms, so that these learned densities could be used for classification purposes. Their alignment method inspired the unsupervised version of the joint alignment algorithm that we present in Section 2.3. However, some differences exist between our approach (JPA) and that of Learned-Miller et al. (congealing). We will highlight and discuss the differences

between our approach and congealing in Section 2.5, after the problem formulation and technical details are introduced in Section 2.1 using a concrete example. We also proposed a semi-supervised version of JPA that is robust to noise, such as background clutter in biological images(see Figure 2.3), which we will present in Chapter 3.

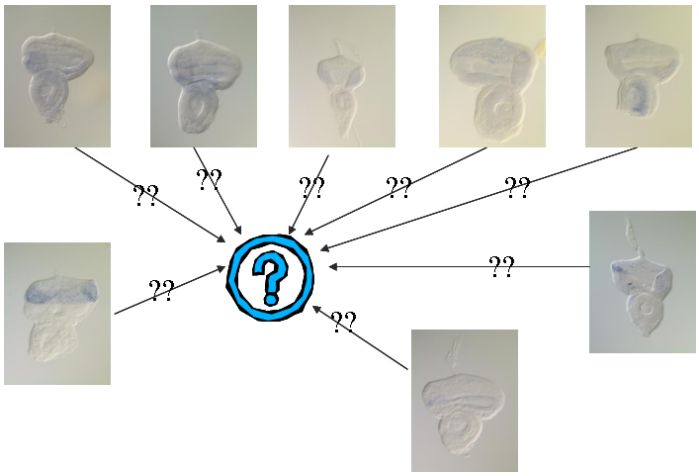

Figure 2.2: Joint Pattern Alignment Set-up. A set of examples (all from same class) are given but the model is unknown. The joint pattern alignment procedure has to solve for both the model and the set of of warping transformations simultaneously. The example images here are the eye/antennal imaginal discs of Drosophila melanogaster.

Let us assume that the given examples (multi-dimensional signals) are all various transformed instances of some unknown underlying model and assume that the structure of transformations is known. All examples are drawn from one class. Let us denote the set of examples of a given class as $\Phi \doteq \{I^i\}_{i=1}^N$ where $N \in \mathbb{Z}_+$ is the cardinality of the set. Let I_l be the latent underlying model. $I_l(\cdot)$ and $I^i(\cdot)$ can be represented as: $I_l, I^i : \Omega \subset \mathbb{R}^M \mapsto \mathbb{R}$, where M indicates the dimension of the domain of I_l. Let g^i induce the mapping from $I^i(\cdot)$ to $I_l(\cdot)$ such that $g^i : \Omega \mapsto \Omega$. Thus, $I_l = I^i \circ g^i$. Now the challenge here is to estimate this unknown generative model I_l jointly from the set of examples Φ without making any prior assumptions on the model, while simultaneously estimating the various transformations $\{g^i\}_{i=1}^N$ affecting I_l. In the following discussion (Section 2.1), we address this challenge using a joint

Figure 2.3: Background clutter in biological images. These typical images of Drosophila melanogaster imaginal discs show the biological clutter (such as parts of other disc tissues, trachea etc.) that is practically hard to remove while taking the measurements of gene expression patterns. Considering the significant costs of extracting these biological tissues, and of imaging them, it makes sense to use all the data instead of throwing away data that has such clutter.

pattern alignment algorithm, and explain the mathematical formulation using the example of joint alignment of binary shape templates.

2.1 An Example: Aligning Binary Shape Images

Let us consider the problem of aligning binary shape images of a given class. We want to derive the objective function directly from the relevant assumptions (on the probability density functions of the transformation vectors) instead of regularizing in an ad-hoc manner. The following derivation will make a specific (Gaussian) assumption on the probability density functions (p.d.f.'s) of transformation components, but this can be changed appropriately depending on domain knowledge or user-preference.

2.1.1 Notation

Let us denote the set of input binary shape images of a given class as $\Phi_I \doteq \{I^i\}_{i=1}^N$ where N is the cardinality of the set. Let I_l be the latent underlying binary shape image that generates the observed images. $I_l(\cdot)$ and $I^i(\cdot)$ can be represented as maps

from $\mathbb{Z}^2$ to the set $B = \{0, 1\}$ with a domain $\Omega \subset \mathbb{Z}^2$:

$$I_l, I^i : \ \Omega \ \mapsto B. \tag{2.1}$$

Typically, the domain Ω is a square or a rectangular window. Let $\boldsymbol{x} \in \Omega$ denote the pixel location in the image (in homogeneous coordinates) such that $\boldsymbol{x} = [x, y, 1]^T$. Let us assume that there exists g^i that is the one-to-one and invertible map from $I^i(\boldsymbol{x})$ to $I_l(\boldsymbol{x})$ such that $g^i : \Omega \ \mapsto \Omega$. Thus, for any given pixel location $\boldsymbol{x}$,

$$I_l(\boldsymbol{x}) \ = \ I^i(g^i(\boldsymbol{x})). \tag{2.2}$$

Let us parameterize each g^i using affine component transformations: x-translation (t_x), y-translation (t_y), rotation (θ), x-log-scale (s_x), y-log-scale (s_y), x-shear (h_x), and y-shear (h_y). Let us denote the set of given binary shape images as $\Phi_I \doteq \{I^i\}_{i=1}^N$, the set of transformations associated with the set of input images as $\Phi_g \doteq \{g^i\}_{i=1}^N$ and the set of transformed binary shape images as $\Phi_{I_g} \doteq \{I_g^i\}_{i=1}^N$ where N is the cardinality of the set. Let us denote $I_g^i(\boldsymbol{x}) = I^i(g^i(\boldsymbol{x})))$. Let us assume that the transform parameters are *i.i.d.* random variables. Fixing the order of composition to ensure unique mapping (since the matrix multiplication is not commutative), $\forall \boldsymbol{x} \in \Omega$, this can be written as:

$$g^i(\boldsymbol{x}) \ = \ F(\boldsymbol{x}; t_x^i, t_y^i, \theta^i, s_x^i, s_y^i, h_x^i, h_y^i) \tag{2.3}$$

$$g^i(\boldsymbol{x}) \ = \ F(\boldsymbol{x}; \{v_j^i\}_{j=1}^K) \tag{2.4}$$

$$\{v_j^i\}_{j=1}^K \ = \ [t_x^i, t_y^i, \theta^i, s_x^i, s_y^i, h_x^i, h_y^i] \tag{2.5}$$

where $1 \leq i \leq N$, $1 \leq j \leq K$(K is the number of parameters chosen), and $v \in \mathbb{Z}_+^N \times \mathbb{R}^K$. In this example, $K = 6$. Writing g^i out explicitly (fixing the order of

composition), we get:

$$
\begin{aligned}
g^i(\boldsymbol{x}) \quad = \quad & \begin{bmatrix} 1 & 0 & t_x^i \\ 0 & 1 & t_y^i \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} cos(\theta^i) & -sin(\theta^i) & 0 \\ sin(\theta^i) & cos(\theta^i) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \\
& \begin{bmatrix} e^{s_x^i} & 0 & 0 \\ 0 & e^{s_y^i} & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & h_x^i & 0 \\ h_y^i & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \boldsymbol{x}
\end{aligned} \tag{2.6}
$$

2.1.2 Problem Formulation

Let $v^i = \{v_j^i\}_{j=1}^K$ and $\Phi_v = \{v^i\}_{i=1}^N$. Let $P(I^i|v_1^i, ..., v_K^i; I_l)$ be some likelihood function such that,

$$P(\Phi_I|\Phi_v; I_l) = \prod_{i=1}^{N} P(I^i|v^i; I_l). \tag{2.7}$$

Let $\Theta \doteq \{I_l, \Phi_v\}$. We would like to infer Θ given the set of binary shape templates Φ_I. The graphical model shown in Figure 2.4 illustrates the generative model associated with JPA. Formulating our goal as a *Maximum a posteriori (MAP) estimation* problem, we want to estimate Θ by $\hat{\Theta}$ such that

$$\hat{\Theta} = \arg\max_{I_l, \Phi_v} P(\Phi_v|\Phi_I; I_l). \tag{2.8}$$

Using Bayes' rule and ignoring the constant denominator,

$$\hat{\Theta} = \arg\max_{I_l, \Phi_v} P(\Phi_I|\Phi_v; I_l) P(\Phi_v). \tag{2.9}$$

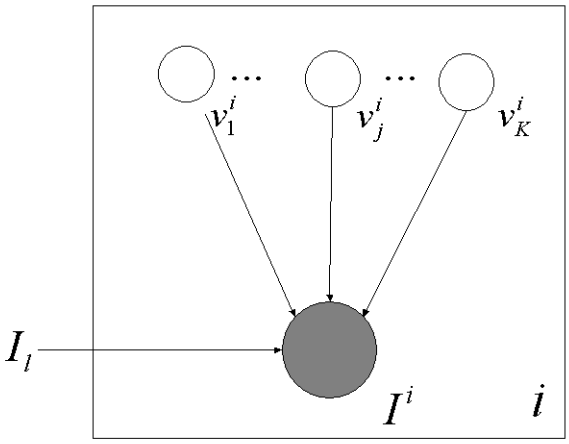

Figure 2.4: Graphical model illustrating the generative model associated with Joint Pattern Alignment (JPA).

Since we assume that the transformation parameters v^i and given binary shape images I^i are independent,

$$\hat{\Theta} = \arg\max_{I_l, \Phi_v} \prod_{i=1}^{N} P(I^i|v^i; I_l) P(v^i). \tag{2.10}$$

Using Equation (2.4), and noting that g^i is a bijective map such that $g^i : \Omega \mapsto \Omega$, we can write:

$$P(I^i|v^i; I_l) = P(I^i \circ g^i|v^i; I_l). \tag{2.11}$$

Let us make the assumption that the value of $I^i(g^i(\boldsymbol{x}))$ at pixel location $\boldsymbol{x}$ is independent of the other pixel locations. In other words, we assume that the probability distributions of values at each pixel location are *i.i.d.* Thus,

$$\begin{aligned} P(I^i \circ g^i|v^i; I_l) &= \prod_{\boldsymbol{x}\in\Omega} P(I^i(g^i(\boldsymbol{x}))|v^i; I_l) \\ &= \prod_{\boldsymbol{x}\in\Omega} P(I^i(g^i(\boldsymbol{x}))|v^i; I_l(\boldsymbol{x})). \end{aligned} \tag{2.12}$$

Thus,

$$\begin{aligned}\prod_{i=1}^{N} P(I^i|v^i;I_l) &= \prod_{i=1}^{N}\prod_{\boldsymbol{x}\in\Omega} P(I^i(g^i(\boldsymbol{x}))|v^i;I_l(\boldsymbol{x})) \\ &= \prod_{\boldsymbol{x}\in\Omega}\prod_{i=1}^{N} P(I^i(g^i(\boldsymbol{x}))|v^i;I_l(\boldsymbol{x})). \end{aligned} \tag{2.13}$$

Since we assumed that the transformation parameters v_j^i are independent,

$$P(v^i) = \prod_{j=1}^{K} P(v_j^i). \tag{2.14}$$

Hence,

$$\hat{\Theta} = \arg\max_{I_l,\Phi_v} \left\{ \left\{ \prod_{\boldsymbol{x}\in\Omega}\prod_{i=1}^{N} P(I^i(g^i(\boldsymbol{x}))|v^i;I_l(\boldsymbol{x})) \right\} \left\{ \prod_{i=1}^{N}\prod_{j=1}^{K} P(v_j^i) \right\} \right\}. \tag{2.15}$$

Now, let us assume (in this example) that $P(v_j^i)$ has a Gaussian distribution, such that

$$P(v_j^i) = N(v_j^i;\mu_j,\sigma_j^2). \tag{2.16}$$

Taking logarithm,

$$\begin{aligned}\hat{\Theta} &= \arg\max_{I_l,\Phi_v} \sum_{\boldsymbol{x}\in\Omega}\sum_{i=1}^{N} \log\{P(I^i(g^i(\boldsymbol{x}))|v^i;I_l(\boldsymbol{x}))\} \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{K} \log\{\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\{-\frac{(v_j^i-\mu_j)^2}{2\sigma_j^2}\}\}. \end{aligned} \tag{2.17}$$

Let us define $\alpha(\boldsymbol{x})$ to be the pixel stack in Φ_I at location $\boldsymbol{x}$ and $\alpha_g(\boldsymbol{x})$ as the pixel stack in Φ_{I_g} at location $\boldsymbol{x}$. Since Φ_I is a set of binary images, $\alpha(\boldsymbol{x}) \in B^N$ and $\alpha_g(\boldsymbol{x}) \in B^N$, where $B = \{0,1\}$. Writing this out explicitly:

$$\alpha_g(\boldsymbol{x}) = [I^1(g^1(\boldsymbol{x})), I^2(g^2(\boldsymbol{x})), ..., I^i(g^i(\boldsymbol{x})), ..., I^N(g^N(\boldsymbol{x}))]^T. \tag{2.18}$$

Also, define $H(\alpha_g(\boldsymbol{x}))$ as the empirical entropy of the pixel stack $\alpha_g(\boldsymbol{x})$. Noting that entropy is the expectation of negative log-likelihood, and expanding the logarithm in the second term (while ignoring the constant),

$$\hat{\Theta} = \arg\min_{I_l, \Phi_v} \left\{ \sum_{\boldsymbol{x} \in \Omega} H(\alpha_g(\boldsymbol{x})) - \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} \log\{\sigma_j\} + \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{(v_j^i - \mu_j)^2}{2\sigma_j^2} \right\}. \quad (2.19)$$

If we assume that $\sigma_j = \sqrt{K}, \forall j = 1, \ldots, K$ and ignore the constant term, then

$$\hat{\Theta} = \arg\min_{I_l, \Phi_v} \left\{ \sum_{\boldsymbol{x} \in \Omega} H(\alpha_g(\boldsymbol{x})) + \sum_{i=1}^{N} \frac{1}{2NK} ||v_j^i - \mu_j||_2^2 \right\} \quad (2.20)$$

where $||\cdot||_2^2$ represents L_2-norm. Since we model transformation parameters as the random variables causing $I^i(\boldsymbol{x})$ to vary from $I_l(\boldsymbol{x})$, we can see that these two will be the same when the randomness due to v^i is removed.

This Maximum a posteriori (MAP) estimation can be formulated as solving an optimization problem. The optimization objective function $\Psi \doteq \Psi(\Phi_v)$ is defined as

$$\Psi \doteq \{ \sum_{\boldsymbol{x} \in \Omega} H(\alpha_g(\boldsymbol{x})) \; + \; \sum_{i=1}^{N} \frac{1}{2K} ||v^i - \bar{v}^i||_2^2 \} \quad (2.21)$$

where $v \in \mathbb{Z}_+^N \times \mathbb{R}^K$ are the vectors of transformation parameters (Equation (2.5)).

The JPA algorithm for the alignment of binary shape templates proceeds as follows:

1. Maintain a transform parameter vector v^i (Equation (2.5)) for each shape image I^i.

2. Initialize all of the transformation matrices g^i to the identity matrix. (In the binary shape example discussed earlier, each parameter vector will specify a transformation matrix $g^i = F(v^i)$ according to Equation (2.6). Initialize all v^i to zero vectors. This means $\mu_j = 0$ for $1 \leq i \leq N$, $1 \leq j \leq K$.)

3. Choose an appropriate penalty term in Ψ (Equation (2.21)) based on the probability assumptions made on transformation parameters (Equation (2.16))

4. Compute Ψ for the current set of images from Equation (2.21).

5. Repeat until convergence:
   For $i = 1, ..., N$,

   (a) Calculate the numerical gradient $\bigtriangledown_{v^i}\Psi$ of Equation (2.21) with respect to the transformation parameters v^i_j's for the current image ($1 \leq j \leq K$).

   (b) Update v^i as: $v^i = v^i - \gamma \bigtriangledown_{v^i} \Psi$. (where the scaling factor $\gamma \in \mathbb{R}$).

   (c) Update γ (according to some reasonable update rule such as the Armijo rule [Boyd and Vandenberghe, 2004]).

Since $\Psi(\cdot)$ is a differentiable function and the level sets

$$\mathcal{A}(\{u^i\}_{i=1}^N) = \left\{\{v^i\}_{i=1}^N \in \mathbb{R}^{K\times N} \mid \Psi(\{v^i\}_{i=1}^N) \leq \Psi(\{u^i\}_{i=1}^N)\right\} \tag{2.22}$$

are bounded for all $\{u^i\}_{i=1}^N \in \mathbb{R}^{K\times N}$, then the JPA routine will at least reach an accumulation point such that $\nabla_{v^i}\Psi = 0$ for all $i = 1, \ldots, N$ [Polak, 1997], even though the optimization routine will generally converge to a local minimum. Note that at a local minimum the set of binary shape images $\Phi_I = \{I^i\}_{i=1}^N$ are reasonably aligned (but need not be perfectly aligned) and the set of transformations $\{g^i\}_{i=1}^N$ is properly described by the parameters $\{v^i\}_{i=1}^N$. I_l is estimated by choosing the medoid of the set of shapes, using an appropriate measure (such as the magnitude of transformation from one shape template to another based on the values of v^i. Note that the introduction of a penalty (regularization) function is critical in achieving the convergence of the optimization routine since this term diverges as the norm of v^i goes to infinity, thus making the level sets of $\Psi(\cdot)$ be bounded.

2.2 Appropriate Regularization

One of the key things to note from the discussion earlier, is the relationship between the priors imposed on the transformation parameters (Equation (2.16)) and the penalty term

in the final objective function that is minimized in the optimization part (Equation (2.21)). The previous section shows that when the prior has a Gaussian p.d.f. form (with a specific choice of value for the variance parameter), the penalty term will be of the form of an L_2-norm on $v^i \in \mathbb{R}^K$. Similarly, if the prior has a Laplacian p.d.f. form, the penalty term will be of an L_1-norm form on $v^i \in \mathbb{R}^K$. Various other penalties can be constructively computed, depending on the assumptions applied on the distributions of the transformation parameters. Since the goodness of the final outcome of the joint alignment is critically dependent on the appropriate choice of penalty in the objective function[1], this key observation provides us a structured way of choosing a penalty based on the probabilistic assumptions imposed on the transformation parameters.

Typically, the information in Equation (2.16) is defined by the user, based on empirical knowledge. Going back to the example of the child learning about the space of mangoes (as discussed in Section 1.1), this is similar to knowing that mangoes are not spherical and are restricted to a certain space of shapes. This information encodes knowledge about the range of transformations, which imposes an appropriate structure on the optimization path.

2.3 JPA Algorithm: General Setting

In a general setting, let us assume that the given examples (multi-dimensional signals) are various transformed instances of some unknown underlying model and assume that the structure of the transformations is known. All examples are drawn from one class. Let us denote the set of examples of a given class as $\Phi \doteq \{I^i\}_{i=1}^N$ where $N \in \mathbb{Z}_+$ is the cardinality of the set. Let I_l be the latent underlying model. $I_l(\cdot)$ and $I^i(\cdot)$ can be represented as: $I_l, I^i : \Omega \subset \mathbb{R}^M \mapsto \mathbb{R}$, where M indicates the dimension of the domain of I_l. Let g^i induce the mapping from $I^i(\cdot)$ to $I_l(\cdot)$ such that $g^i : \Omega \mapsto \Omega$. Thus, $I_l = I^i \circ g^i$. Let $g^i(\boldsymbol{x}) = F(\boldsymbol{x}; v^i)$ for $1 \leq i \leq N$, where the parametric structure of function $F(\cdot)$, and the

[1] See Section 5.3, where we show some example results that demonstrate that choosing appropriate regularization function based on the assumptions made on prior distributions of transformation parameters has a clear impact on the final result.

probability distribution function $P(v_j^i)$ for $1 \leq j \leq K$ are known. K denotes the number of components in the parameterization of transformation g^i.

The JPA algorithm proceeds as follows:

1. According to the problem at hand, set up the explicit parameterization for $g^i = F(v^i)$.
2. Maintain a transform parameter vector v^i for each example I^i.
3. Initialize all of the transformation matrices g^i, for $1 \leq i \leq N$, to the identity matrix.
4. Choose an appropriate penalty term in objective function Ψ. In the binary shape template alignment example, Ψ can be computed from Equation (2.21) based on the probability assumptions made on transformation parameters ($P(v_j^i)$) (See Section 2.2 for details).
5. Compute Ψ for the current set of examples ($\Phi \doteq \{I^i\}_{i=1}^N$).
6. Repeat until convergence:
   For $i = 1, ..., N$,
   (a) Calculate the numerical gradient $\bigtriangledown_{v^i}\Psi$ with respect to the transformation parameters v_j^i's for the current image.
   (b) Update v^i as: $v^i = v^i - \gamma \bigtriangledown_{v^i} \Psi$. (where the scaling factor $\gamma \in \mathbb{R}$).
   (c) Update γ (according to some reasonable update rule such as the Armijo rule [Boyd and Vandenberghe, 2004]).

2.4 JPA Algorithm: Convergence

Since $\Psi(\cdot)$ is a differentiable function and the level sets

$$\mathcal{A}(\{u^i\}_{i=1}^N) = \left\{\{v^i\}_{i=1}^N \in \mathbb{R}^{K\times N} \mid \Psi(\{v^i\}_{i=1}^N) \leq \Psi(\{u^i\}_{i=1}^N)\right\} \tag{2.23}$$

are bounded for all $\{u^i\}_{i=1}^N \in \mathbb{R}^{K\times N}$, then the JPA routine will at least reach an accumulation point such that $\nabla_{v^i}\Psi = 0$ for all $i = 1, \ldots, N$ [Polak, 1997], even though the optimization routine will generally converge to a local minimum. Note that at a local minimum the set of examples $\Phi_I = \{I^i\}_{i=1}^N$ are reasonably aligned (but need not be perfectly aligned) and the set of transformations $\{g^i\}_{i=1}^N$ is properly described by the parameters $\{v^i\}_{i=1}^N$. I_l is estimated by choosing the medoid of the set of examples, using an appropriate measure (such as the magnitude of transformation from one example to another based on the values of v^i. Note that the introduction of a penalty (regularization) function is critical in achieving the convergence of the optimization routine since this term diverges as the norm of v^i goes to infinity, thus making the level sets of $\Psi(\cdot)$ be bounded.

In a general setting, for non-differentiable (but continuous) Ψ, one can use coordinate descent (or any derivative free algorithm). If Ψ is not even continuous, then the optimization can be approached by using methods such as *simulated annealing* [Kirkpatrick *et al.*, 1983].

2.5 Related Work

As mentioned earlier, Viola et al. and Collignon et al. introduced a novel information theoretic approach for solving pair-wise medical image registration problem [Viola and Wells, 1997; Collignon *et al.*, 1995]. A generalization of Viola's pair-wise approach to joint group-wise (ensemble) alignment scenario (in the context of aligning and recognizing binary shape masks of hand-written characters) was proposed by Learned-Miller et al. [Miller *et al.*, 2000]. Their joint ensemble alignment procedure called "congealing" [Miller, 2002] utilizes the information across the whole ensemble, and is robust against local minima. The congealing method is the closest in spirit to the unsupervised version of the joint alignment algorithm that we discussed in Section 2.3. Geometric approaches to the joint pattern alignment problem also exist in literature, but since our approach is information-theoretic, the reader interested in geometric approaches is referred to [Pennec, 2006a; Pennec, 2006b] for an overview of these methods.

Vedaldi et al. pointed out that, from the point of view of learning data-driven representations, the regularization is ad-hoc in "congealing" algorithm [Vedaldi and Soatto, 2007]. Given a specific instance of joint pattern alignment problem, it is not clear how the penalty term needs to be chosen. In Section 5.3, we also show results that demonstrate that not choosing appropriate regularization can have adverse impact on the final result of joint alignment algorithms. This is an important practical issue for a system designer who wants to use an ensemble alignment algorithm. Since the emphasis of Learned-Miller et al. was to learn a density over the transformations for the purposes of classification, they do not impose any priors on the transformation parameters *a priori*. Our Joint Pattern Alignment approach (JPA) for learning the optimal representations from examples follows the information theoretic paradigm (similar to "congealing"), where we have shown a clear and principled connection between the priors on the transformation parameters and the regularization term used in the optimization stage. Since our focus is primarily on learning the data-driven representation of the ensemble, we assume that the user has some prior domain knowledge that is implicitly encoded in the densities imposed on the transforms. In the unsupervised version of the JPA, the densities on the transformation parameters are assumed to be known *a priori*. In practice, this approach allows the user to make use of the domain knowledge (in the form of probabilistic assumptions given on the transformation parameters) to derive the correct penalty term for the objective function Ψ. Another key problem with "congealing" is that it cannot handle noisy patterns (see Figure 2.3), because it is explicitly designed to only account for the registration errors. In our work, we have proposed a semi-supervised variation to JPA that addresses this issue[2] by combining segmentation and joint alignment for analyzing large datasets (high-throughput scenarios).

Frey and Jojic proposed the idea of transform invariant clustering by including clutter and transformation as unobserved, latent variables in a mixture model. Using this procedure, they obtained a new transformed mixture of Gaussians, which is invariant to a specified set of transformations. They also showed how a linear-time EM algorithm can be

[2]See Chapter 3 for details.

used to fit this model by jointly estimating a mixture model for the data and inferring the transformation for each image. In their work, a fixed set of transformations are used (as opposed to continuous set of transforms used in JPA or "Congealing"). The complexity of their algorithm is linear in the number of possible transformations (as opposed to the number of transformation parameters in JPA) per each iteration.

2.6 Discussion

It is useful to note that our approach does not make any parametric assumptions on the underlying model (in the current example, the model is I_l). The joint shape alignment approach requires no correspondence solving stage, since we attempt to align the entire template without choosing any landmark points (as done by approaches like shape context and geometric blur [Belongie *et al.*, 2002; Berg, 2005]). JPA is more robust to local minima compared to the standard pair-wise alignment formulations due to the data-driven smoothing provided by the ensemble of examples. Just like the other information theoretic approaches to alignment, our approach also gives a quantitative measure (Ψ) that can be used as a putative goodness measure to evaluate the quality of alignment.

On the other hand, convergence of the JPA algorithm can get very slow as the number of examples (N) increases. The convergence can potentially slow down significantly as the dimensionality of the data samples increases (due to the inherent increase in the number of associated transformation parameters). The unsupervised version of the joint alignment is not designed to handle errors other than the alignment and noise parameters that it explicitly models in the formulation. Landmark point based alignment algorithms may perform better in situations where occlusion is a serious practical issue, but they would require the model to be known *a priori*. An alternative for handling partial occlusions in the joint alignment framework is the semi-supervised variation of JPA that combines the JPA algorithm with segmentation (Section 3.3).

Chapter 3

Constructing Atlases of Gene Expression in Drosophila Imaginal Discs

3.1 Introduction

The central goal of this chapter is to discuss the semi-supervised framework introduced in [Ahammad *et al.*, 2005] for high-throughput joint pattern alignment problem and to show the application of the framework to construct data driven atlases of gene expression in Drosophila melanogaster imaginal discs [Harmon *et al.*, 2007]. It is worth noting that, while our early work [Ahammad *et al.*, 2005] used the coordinate descent based "congealing" approach [Miller, 2002] for the unsupervised learning part, our current pipeline uses the improved formulation of the joint alignment problem as discussed in Chapter 2. This allows us to make use of appropriate probabilistic assumptions given on the transformation parameters to derive the correct penalty term for the objective function Ψ, while also making it possible to learn the distribution over the alignment parameters using some manually segmented truth data. The goal of this chapter is to produce a map that provides

precise, detailed information about where particular genes are expressed, and to do so over a large number of genes. This atlas is represented in a format that facilitates computational analysis of these patterns, in addition to providing images of the maps of expression for human consumption.

The genome sequences of metazoan organisms contain information required to direct the development of the animal [Arnone and Davidson, 1997]. This information is encoded in the genome in the form of genes and regulatory sites which are bound by transcription factors that precisely control growth and development. In order to fully understand the developmental program, it is critical to know which genes are expressed during development, what sequences are responsible for activating or repressing these genes, what the functions of these genes are, and precise characterization of when and where these genes are expressed.

3.1.1 Characterizing Gene Expression Patterns

Advances in molecular biology over the past few decades have enabled researchers to determine the complete DNA sequence of the genomes of many organisms and to determine the sequence of the vast majority of genes expressed in these, and other, organisms [Adams *et al.*, 2000]. Microarray technologies have enabled researchers to measure the level of expression of large numbers of individual genes in a single experiment [Lipshutz *et al.*, 1999]. This has provided a greatly enhanced view of the genes that are active at specific times in specific tissues [Alizadeh *et al.*, 2000]. However, these techniques generally require a large tissue sample and, on their own, provide no information about the spatial patterns of expression of genes.

In situ hybridization of tissues with labeled probes for individual genes enables researchers to measure spatial patterns of expression at high resolution. However, in situ hybridization can only be used to measure a limited number of genes at a time, usually one. Ideally, one would be able to measure the spatial patterns of expression of large numbers of genes and be able to compare, cluster, and classify these patterns of expression. Current experimental strategy for such experiments entails setting up a high-throughput production

system for the generation of large numbers of images which can then be processed by either human or computer. In order to interpret and compare individual patterns, one must either identify individual morphological structures or bring the images of corresponding tissues into an alignment, or registration, such that they can be meaningfully compared.

3.1.2 Spatial Patterns of Gene Expression in Drosophila Imaginal discs

In holometabolous insects (insects that undergo a complete metamorphosis), also known as the members of the taxonomical subdivision endopterygota, or as endopterygotes, the wings and other appendages develop on the inside of the juvenile insect [Snodgrass, 1935; Beverley and Ponsonby, 2003]. In the hemimetabolous insects, on the other hand, the adult appendages grow out from the larval appendages directly. In Drosophila and other higher diptera, the primordial tissues of the adult appendages are formed during embryogenesis [Cohen, 1993]. For the fly, the benefit is that these tissues can be patterned at an early stage, allowing for complex simultaneous development both of the embryonic and larval insect, and of the eventual adult appendages. For the scientist, these primordial tissues afford a unique opportunity to study appendage development throughout the life-cycle of the fly. These primordial tissues are known as imaginal discs. For further details about imaginal discs, the reader is referred to [Held, 2002].

Drosophila have nine pairs of imaginal discs: the labial, clypeolabral, humeral, eye-antennal, first leg, second leg, third leg, wing, and haltere discs, as well as a single genital disc. In our work, we restrict out attention to the larger disc types: the wing, leg, haltere and eye/antenna imaginal discs. Imaginal discs have been well studied for decades and there is a large body of literature concerning the development and patterning of imaginal discs [Held, 2002]. Large-scale studies of patterns of gene expression in Drosophila have been performed using DNA microarrays both on whole organisms [Arbeitman *et al.*, 2002] and individual tissues such as imaginal discs [Klebes *et al.*, 2002; Butler *et al.*, 2003]. Klebes et al. compared differential gene expression in different imaginal discs and between imaginal

discs and non-disc tissue. Butler et al. manually dissected imaginal discs and were able to identify transcripts that were enriched in specific compartments of the wing discs [Klebes *et al.*, 2002]. However, these studies yield little information about the precise spatial patterns of gene expression.

In order to compute spatial patterns of gene expression, one must find the structure of interest in the image. For imaginal discs, this means recognizing and extracting the portion of the image that corresponds to the imaginal disc, separating this from the image background, and bringing this model into registration with a canonical model such that multiple patterns of expression can be spatially compared. Registration of the image to a canonical model requires that the shape of the disc be generated in a computable form. One can either manually specify the shape of the objects in question, or the shapes can be learned from the data. We present an information theoretic framework to joint pattern alignment (JPA) for aligning images of Drosophila imaginal discs that is robust, and requires minimal expert supervision.

3.1.3 Problem Definition

Given a large input ensemble of noisy Drosophila imaginal disc images of a given tissue class, our goal is to learn the underlying shape representation of the tissue nonparametrically while bringing the given images into alignment. This learned representation greatly facilitates quantitative stain scoring analysis on the imaginal disc images, as well as allowing comparative analysis of spatial gene expressions. We demonstrate the segmentation and alignment results for various tissue classes of Drosophila imaginal discs and discuss the results.

3.1.4 Challenges

The tissue structures in typical Drosophila imaginal discs have significant intra and inter-class variability in size, shape and stain patterns. Thus, shape models learned from one class cannot be used for another tissue class without making significant changes in the processing

pipeline if one were to use model based alignment algorithms. Furthermore, there are far fewer identified and named morphological parts or regions in Drosophila imaginal discs than in the developing Drosophila embryo, for example. It is difficult to make any parametric or model based assumptions for tissue shapes given these limitations. Manual annotation and curation are extremely time-consuming and costly in high-throughput spatial gene expression analysis experiments. It is highly desirable to have a processing pipeline that can operate for various imaginal disc tissues such as wings, halteres, legs, eyes, etc., (shown in Figure 3.1) without significant re-structuring. In our work, we propose a simple yet effective computational framework that addresses these demands of high-throughput systems for the analysis of spatial gene expression by combining segmentation and nonparametric alignment algorithms.

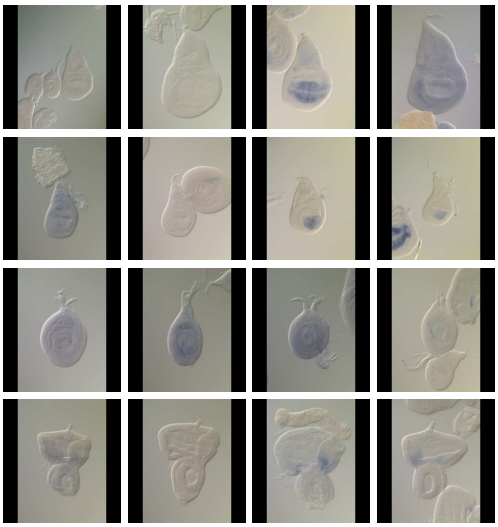

Figure 3.1: A typical set of in situ stained Drosophila imaginal disc images. Each row shows a different tissue class. First row: wing discs, second row: haltere discs, third row: leg discs, fourth row: eye discs. There is significant inter-class and intra-class variability both in the shapes, sizes and the stain patterns.

3.1.5 Related Work

Precise spatial patterns of expression of individual genes has been well studied for many years and recently the BDGP has studied the spatial patterns of gene expression of large numbers of genes in developing Drosophila embryos through in-situ hybridization to individual gene probes [Tomancak *et al.*, 2002]. The individual images are then manually curated with a list of tissues in which the gene of interest is expressed. This yields spatial information at the resolution of recognizable morphological structures in the embryo and this information is then clustered with DNA microarray data to yield clusters of genes with related spatial patterns of gene expression [Berman *et al.*, 2002]. The limiting factor for this approach is that the annotation of spatial expression pattern requires manual curation. Furthermore, this approach would likely be less successful in the imaginal disc where there are far fewer identified and named morphological parts or regions.

Large-scale studies of patterns of gene expression in Drosophila have been performed using DNA microarrays both on whole organisms [Arbeitman *et al.*, 2002] and individual tissues such as imaginal discs. Klebes et al. compared differential gene expression in different imaginal discs and between imaginal discs and non-disc tissue [Klebes *et al.*, 2002]. Butler et al. manually dissected imaginal discs and were able to identify transcripts that were enriched in specific compartments of the wing discs [Butler *et al.*, 2003]. However, these studies yield little information about the precise spatial patterns of gene expression.

Recent studies of precise spatial patterns of gene expression for large numbers of genes in developing Drosophila embryos through in situ hybridization [Tomancak *et al.*, 2002; Berman *et al.*, 2002] require annotation, and suffer from the fact that the annotation of spatial expression pattern requires manual curation. Kumar et al. [Kumar *et al.*, 2002] applied machine vision techniques to low-resolution images of in situ stained embryos and developed an algorithm for searching a database of patterns of gene expression in the embryos. Peng and Myers [Peng and Myers, 2004] have performed automated embryo registration and stain classification by using Gaussian mixture models. Yet, most of the previous work makes some parametric assumptions on the shape of the tissue, and the registration

techniques used are very simplistic (such as aligning the major and minor axes for embryo images [Peng and Myers, 2004]).

In the approach taken by Tomancak et al. [Tomancak *et al.*, 2002], genes are manually curated by an expert operator who labels images corresponding to individual genes as indicating in which specific tissues the gene of interest is expressed. This approach works well, but requires manual annotation and a precise (and consistent application of a) taxonomical classification of the tissues in the organism. Such a taxonomy does not exist for Drosophila imaginal discs, and we see this atlas of gene expression in imaginal discs as a tool that could be used to create such a taxonomy, rather than a process that requires that such a taxonomy exists. A similar approach has been taken by Baldock et al. for the analysis of the developing mouse embryo (as part of the Edinburgh Mouse Atlas Project -EMAP), although their approach yields a volumetric 3-dimensional representation and requires manual alignment and warping of the image to a reference model by the annotation of tie-points entered by an expert operator [Baldock *et al.*, 2003]. Our approach, while presently limited to 2-dimensional representations, automatically learns consensus imaginal disc shapes from a handful of manually segmented training examples and subsequently automatically aligns and extracts the imaginal disc shapes and produces a representation of the spatial extent of the expression of individual genes, in the context of a global reference map for each imaginal disc. There are also efforts to explore spatial patterns of gene expression in *Drosophila* embryos by automatically reconstructing volumetric models of transcript localization using confocal slices of embryos stained with fluorescent probes to a number of genes [Fowlkes *et al.*, 2005]. Fowlkes et al. use the method of "shape contexts" to perform pair-wise registration of landmark feature points across animals for a given reference gene pattern at cellular level, and allow the resulting geometric warp to align the rest of the embryo structure. While this approach has some advantages in handling occlusions robustly, choosing a reference gene's pattern as a basis for registering across various animals can be problematic in situations where either there is a variation in the number of cells due to further development or inherent variation of the chosen gene's expression across animals due to

natural evolutionary differences. Allen Institute for Brain Science has recently published a detailed 3-dimensional volumetric map of spatial patterns of gene expression of 20,000 genes in the Mouse brain [Lein *et al.*, 2007], thus demonstrating the feasibility of a systematic genome-wide approach to this problem. While this work is impressive for demonstrating a genome-wide systematic results for the organism, as opposed to a subset of approximately 130 genes analyzed in our work, their registration approach requires a manually constructed global reference atlas computed *a priori*. This can be a difficult requirement to satisfy in a lot of high-throughput biological studies. In contrast, our work discovers the atlas in a data-driven fashion.

3.2 Proposed Approach

Figure 3.2 illustrates the data flow in our approach. In this dissertation, we will mainly focus on the computational aspects related to the segmentation, alignment and shape learning.

Let us first provide a simple overview of the procedure. Our process for determining spatial patterns of gene expression begins with the mass-isolation of third instar larval imaginal discs which serve as the input both to a microarray-based gene selection procedure and to the subsequent labeling with probes that hybridize to individual genes. Probes are created and then hybridized to the mass-isolated discs and are then imaged with a light microscope. Images are captured into a digital format and an initial set of images is used for training purposes to learn the imaginal disc shapes. Starting with a set of imaginal disc images of a given tissue class, we find the structure of interest in the given image via segmentation. For imaginal discs, this means recognizing and extracting the portion of the image that corresponds to the imaginal disc and separating this from the cluttered image background. We use a combination of manual and automatic segmentation schemes to segment a set of these imaginal disc images and input these into the unsupervised joint pattern alignment (JPA) procedure to learn the shape prior. This data-driven procedure learns the canonical shape model of a given ensemble of shapes and the set of transformations associated with

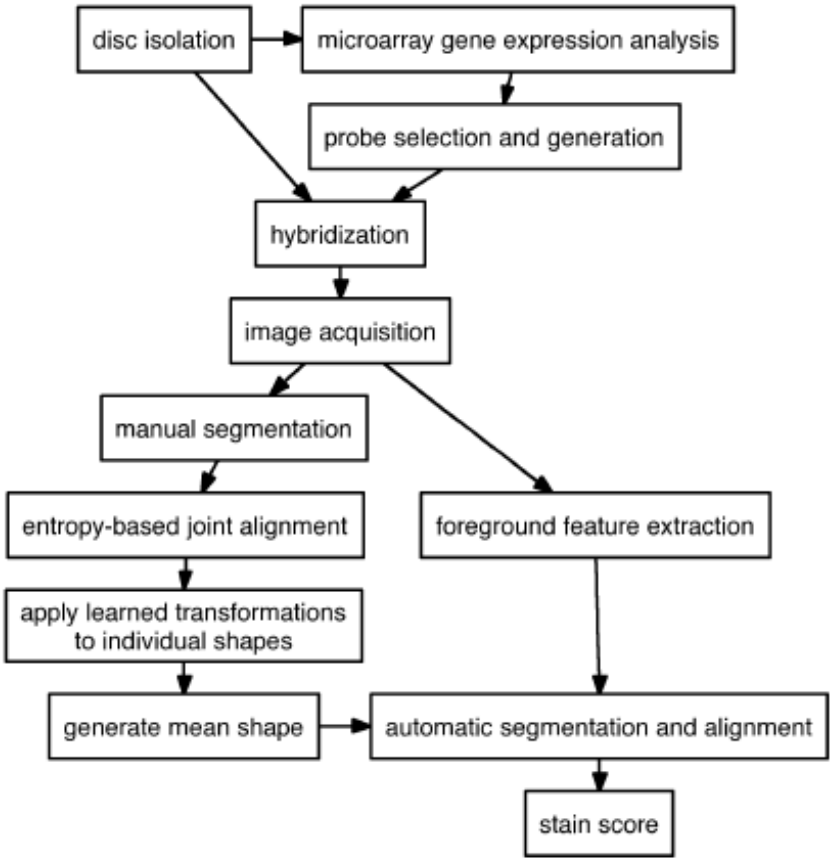

Figure 3.2: Overview of the high-throughput process for determining spatial patterns of expression of genes in Drosophila imaginal discs.

each shape in the ensemble simultaneously by solving a constrained optimization problem. JPA operates by minimizing the regularized sum of component-wise (pixel-wise) entropies over a continuous set of transformations on the image ensemble and is robust against local minima in the optimization. Once the underlying shape model is learned, it can be used subsequently as the canonical model. We also use this model as a shape template to improve our automatic segmentation algorithm. The learned transformations are applied back to the original imaginal disc images to bring them into alignment in one step. Once the images are aligned and the disc shapes are learned, the shapes are used to extract instances of these

shapes in the target images which are then stain scored and processed. These aligned stain scores are then used as descriptors to make comparative analysis of spatial gene expression across multiple genes.

3.2.1 Data Generation

While there are over 13,000 genes in the Drosophila genome, many of these genes are either not expressed at detectable levels in imaginal discs, or are expressed in all cells in the animal. We are interested in genes that have a characteristic pattern that is neither ubiquitous expression nor a lack of detectable expression. Our initial experiences with randomly chosen genes, suggested that less than five percent of the genes had non-trivial, detectable (by our methods) expression patterns in imaginal discs. Therefore we chose to enrich the set of genes we analyzed by using microarray-based gene expression data to identify candidate genes that are likely to have non-trivial spatial patterns. To identify genes that may be specifically up-regulated in individual imaginal discs, we compared the expression of genes in individual disc types to their expression in the other disc types and selected genes that were expressed at different levels in the five imaginal disc types we measured with expression microarrays. To perform our experiments in a high-throughput, parallel fashion, we used the mass-isolation procedure developed by Eugene et al. [Eugene *et al.*, 1979] to gather hundreds of thousands of discs. We used a protocol similar to that used by the Berkeley Drosophila Genome Project for the staining of large numbers of Drosophila embryos with individual probes in 96-well plates. Approximately 100,000 discs were used per 96-well plate, yielding on the order of 1000 discs per probe. Mass-isolated imaginal discs were placed in 96-well plates and stained with digoxigenin-labeled RNA or DNA complementary to genes of interest. Images were acquired using a light microscope equipped with Nomarski optics as described by Tomancak et al. [Tomancak *et al.*, 2002]. The local presence of stain results in the appearance of blue in the image; darker blue suggests a greater local concentration of the gene of interest. However, there is substantial probe-to-probe variability and these intensities should not be relied on as an

accurate quantitative measure of gene concentration. Nevertheless, the different intensity values can be used to suggest where local gene concentration is high. Images were acquired as 16-bit per channel RGB TIFF images using Nomarski optics. Images are stored in a database with metadata information about the preparation and image capture process. For a detailed discussion of the data acquisition procedure and the protocols used, please refer to [Harmon, 2007].

3.3 JPA for Learning Shapes of Drosophila Imaginal Discs

To determine the spatial pattern of expression for a given gene from an image containing a stained imaginal disc, we must identify the location of the disc within the image and must be able to segment out the portion of the image that corresponds to the disc from the background. Moreover, to perform meaningful comparative analysis across patterns, it is highly desirable to have a reference shape model (or shape prior) to which discs of a given type can be aligned. Given such a model (or shape prior), we can then perform quantitative analysis of the spatial patterns across multiple images and across multiple genes.

3.3.1 Image Model for Drosophila Imaginal Discs

Drosophila imaginal discs have a morphology similar to that of an uninflated balloon [Held, 2002]. Each imaginal disc type has a characteristic shape, although the T1, T2 and T3 (thoracic segments 1, 2 and 3) leg discs are rather similar in shape to one another. In addition, the haltere discs have a shape similar to that of the wing, but are much smaller than wing discs. One additional factor that induces variability in the sizes and shapes of the discs we measured, is the age of the larva from which the discs were recovered. A simple approach to representing a canonical shape would be to choose a single reference example for each disc and to use these as the canonicals shape to which other instances of

the corresponding disc types would then be aligned. This is the approach taken by [Lein *et al.*, 2007]. Albeit, it is not clear which example would serve as a suitable reference. In our work, we take the approach that the variation in the size and shape of the discs suggest that learning a consensus shape model from an ensemble of images will yield a single consensus image that is more appropriate both for representing the overall shape contained in the training images and for use in model-based identification of discs in new images. Our approach to bringing multiple discs into correspondence consists of manually segmenting a small number of training images, and simultaneously learning canonical imaginal disc shapes from these shapes while learning a set of affine transformations, with one transform for each image, that bring the images of the discs into alignment.

Imaginal discs do have substantial depth to them, but we image a single plane from the discs and consider an idealized 2-dimensional representation of a disc. A single focal plane is selected for each image to maximize visibility of any stain present in the disc and the choice of focal plane generally does not affect the perceived boundaries of the disc.

We denote the set of input imaginal disc images of a given class as $\Phi \doteq \{I^i\}_{i=1}^N$ where N is the cardinality of the set. Each image $I^i(\cdot)$ can be represented as a map from the image $\mathbb{R}^3$(in homogeneous coordinates) to the color space $C \subset \mathbb{R}^3$ with a small compact support $\Omega \subset \mathbb{R}^3$:

$$I^i(\boldsymbol{x}) \;:\; \boldsymbol{x} \in \Omega \;\mapsto\; \boldsymbol{c} = I^i(\boldsymbol{x}) \in C; \tag{3.1}$$

where $\boldsymbol{x} = [x, y, 1]^T \in \mathbb{R}^3$ (in homogeneous coordinates), $\boldsymbol{c} = [r, g, b]^T \in \mathbb{R}^3$ is a vector in the color space. In general, the domain Ω is a square or a rectangular window. This representation will be used throughout the rest of the discussion whenever the mathematical details of individual steps are explained.

All imaginal discs except the genital discs occur in pairs, one on each side of the body, yielding a left disc and a right disc of each type. When we image the discs they are either lying on the slide with the peripodial epithelium up or down. The combination of the handedness of the disc and the orientation of the peripodial epithelium gives us 4 possibilities

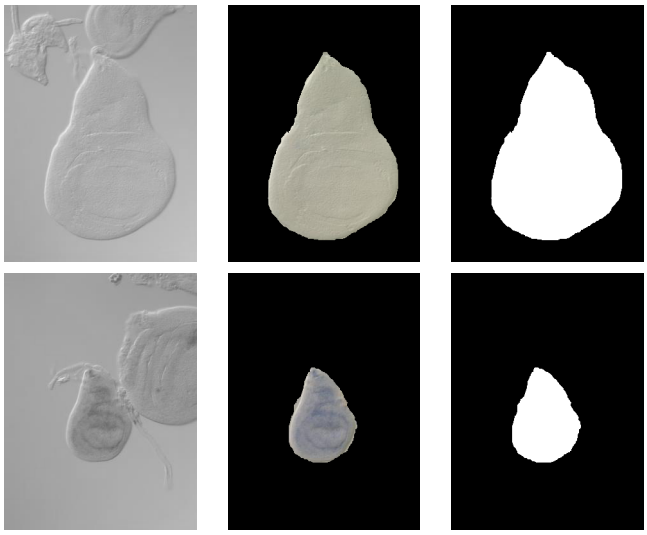

Figure 3.3: Example segmentation results for wing discs (first row) and haltere discs (second row) using the combined segmentation procedure. **Left Column**: Original image $I^i(\boldsymbol{x})$. **Middle Column**: Segmented tissue structure of interest $I^i_f(\boldsymbol{x})$. **Right Column**: Extracted binary shape $I^i_s(\boldsymbol{x})$.

for the combined state and orientation of the disc. We make a simplifying assumption and assume that the left and right discs are mirror images of each other and the stain pattern of a gene in a right disc will be equivalent to the mirror image of the stain pattern of that gene from the left disc of the same type. This assumption gives us two handedness/orientation combinations that are considered to be in the canonical orientation. The shapes corresponding to the other two handedness/orientation combinations are automatically flipped by the JPA pipeline after manual segmentation.

3.3.2 Extracting Tissue Shapes via Segmentation

Two salient features of our image dataset make the segmentation task relatively simple. First, Nomarski images of the discs [Tomancak *et al.*, 2002] yield substantial highlights and lowlights at the periphery of the discs. The background of the images is generally uniform and one can use the significant contrast generated at the edge of the discs to identify border regions. Second, compared to the background, the pixel intensities of the imaginal disc

tissues have much more variability than the background, even over a small window. While the mean pixel intensities of large regions of disc and background are very similar, the disc has a broader range of pixel intensities and, more importantly, the local derivative values are much higher for the disc than that of the background. This enables the use of either the magnitude of the gradient and/or the variance of a window around a pixel as a feature for distinguishing disc from non-disc tissue. This bimodal distribution lends itself to fitting a mixture of two Gaussian random variables followed by labeling of each pixel as disc or background. Furthermore, the second derivative of the image (the Laplacian) is also quite high at the edge of the disc and serves as a useful feature for identifying discs. Using these insights, we implemented a simple filter-and-threshold module for segmentation. It computes the *local variance* of the image in a small support region, estimates the bimodal distribution of variance in the filtered image and thresholds it appropriately to separate the disc region from the background. This process creates a binary shape mask that we use in following sections to learn the canonical shape model of the disc tissue.

This can be written as follows:

$$var(I^i(\boldsymbol{x})) = \int\int \mathbb{E}\{I^i(\boldsymbol{x})^2\} - (\mathbb{E}\{I^i(\boldsymbol{x})\})^2 dx\, dy \tag{3.2}$$

where $\boldsymbol{x} = [x, y, 1]^T$(in homogeneous coordinates), and $\mathbb{E}(.)$ is the expectation over the small square window centered at $\boldsymbol{x}$ such that $-\alpha \leq x \leq +\alpha$, $-\alpha \leq y \leq +\alpha$, $\alpha \in \mathbb{R}$. The extracted shape image I_s^i is then calculated as:

$$I_s^i(\boldsymbol{x}) = \begin{cases} 1, \textit{ if } \quad var(I^i(\boldsymbol{x})) \geq \delta \\ 0, \textit{ otherwise} \end{cases} \tag{3.3}$$

where $I_s^i(\boldsymbol{x})$ is a binary image of the extracted shape, $\boldsymbol{x} \in \Omega$ and δ is a threshold value where $\delta \in \mathbb{R}$. $I_s^i(\boldsymbol{x})$ is a map from the Ω to the set $B = \{0, 1\}$. The segmented structure of interest (or the foreground) $I_f^i(\boldsymbol{x})$ can be computed by point-wise multiplication of $I^i(\boldsymbol{x})$

and $I_s^i(\boldsymbol{x})$:

$$I_f^i(\boldsymbol{x}) = I^i(\boldsymbol{x}) \cdot I_s^i(\boldsymbol{x}). \tag{3.4}$$

where $I_f^i(\boldsymbol{x})$ is a map from Ω to C. In other words,

$$I_f^i(\boldsymbol{x}) = \begin{cases} I^i(\boldsymbol{x}),\ \textit{if} \quad var(I^i(\boldsymbol{x})) \geq \delta. \\ 0,\ \textit{otherwise.} \end{cases} \tag{3.5}$$

Sometimes this simple filter-and-threshold process results in unsatisfactory performance, for two reasons. First, the presence of non-disc tissue (such as the trachea which is attached to the wing disc), may interfere with the segmentation of the disc from the background, thereby corrupting the extracted shape with the presence of this additional biological material. Second, some regions inside the disc appear more homogeneous than others and sometimes get misclassified as the background. This is especially true when the imaginal discs are heavily stained. Both these problems can be addressed by performing a template matching operation to identify the disc. Given the rough shape template of the disc, the template matching operation is quite simple but the problem is that there is no such clean shape template to begin with. We address this issue as follows: Using manual segmentation on a set of disc images, we obtain relatively clean shapes of the tissue for each class. These relatively clean structures are then fed to the nonparametric shape learning algorithm to form a good canonical shape template. This learned shape template was used in conjunction with the simple filter-and-threshold algorithm to obtain better segmentation results automatically in cluttered images. The manually segmented shapes were also used as truth data for comparing the performance of our implementations of segmentation algorithms. Our current implementation gives satisfactory segmentation results. We show some sample segmentation results from our segmentation procedure in Figure 3.3 for wing and haltere discs.

3.3.3 Shape Learning

Once the shapes of the relevant disc tissues, $I_s^i(\boldsymbol{x})$ (Equation (3.3)), are extracted in binary image format, we use this set of binary shapes to learn the canonical underlying shape model of the given class of disc tissues using '*Joint Pattern Alignment*'(JPA). We denote the set of binary shape images as $\Phi_s \doteq \{I_s^i\}_{i=1}^N$ where N is the cardinality of the set. Let us denote the latent binary shape of the given class of disc tissues as I_l. We model each shape image in Φ_s as I_l transformed through a geometric transformation. Given a class, the latent shape and the transformation are conditionally independent [Miller, 2002]. We assume that the transformations are affine and model the affine parameters as i.i.d. random variables. We shall assume that the transformation is a one-to-one and invertible mapping between I_l and I_s^i. We make the further assumption that the probability distribution of pixel values at each pixel location are *i.i.d.* A thorough discussion of joint alignment procedure for the set of binary shape images as $\Phi_s \doteq \{I_s^i\}_{i=1}^N$ is provided in Section 2.1.

In our implementation, we parameterize the set of transformations g^i using the following component transformations: x-translation (t_x), y-translation (t_y), rotation (θ), x-log-scale (s_x), y-log-scale (s_y), x-shear (h_x), and y-shear (h_y). Clearly, this is an *over-complete* parameterization, but we made this specific choice following the efficiency arguments presented by Miller [Miller, 2002]. We experimented with different choices of parameterization, and we will show results based on the parameterization as shown in Equation (3.9).

Fixing the order of composition to ensure unique mapping (since the matrix multiplication is not commutative), this can be written as:

$$g = F(t_x, t_y, \theta, s_x, s_y, h_x, h_y) \tag{3.6}$$

$$g^i = F(\{v_j\}^i) \tag{3.7}$$

$$\{v_j\}^i = (t_x^i, t_y^i, \theta^i, s_x^i, s_y^i, h_x^i, h_y^i) \tag{3.8}$$

where $1 \leq i \leq N$, (N is the cardinality of the set Φ_s), $1 \leq j \leq K$, (K is the number of parameters chosen), and $v \in \mathbb{Z}^{N \times K}$.

Writing g out explicitly, we get:

$$g = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos(\theta) & -sin(\theta) & 0 \\ sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} e^{s_x} & 0 & 0 \\ 0 & e^{s_y} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & h_x & 0 \\ h_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.9}$$

The goal is to find the transformation g^i that brings the associated shape template I_s^i closest to I_l. We experimented with various choices for probability distribution on parameters $P(v_j^i)$, and found that the alignment error is minimum for the choice of Gaussian prior. However, the σ_j values for different transformation components are different - since some parameters are allowed to vary more than others (for example, we restrict the range of rotations that can happen). Hence, Gaussian prior (with appropriate σ_j value) is what we have used in our experiments for constructing the spatial gene expression atlases. As discussed in Section 2.2, the regularization term in the objective function Ψ[1] depends on the probabilistic assumptions made on the distribution of parameters.

The JPA algorithm for learning the shape template of imaginal discs (of a given class) proceeds as follows:

1. Maintain a transform parameter vector v^i (Equation (3.8)) for each shape image I_s^i. Each parameter vector will specify a transformation matrix $g^i = F(v^i)$ according to Equation (3.9). Initialize all v^i to zero vectors. This has the effect of initializing all of the transformation matrices g^i to the identity matrix.

2. Choose an appropriate penalty term in Ψ (Equation (2.21)) based on the probability assumptions made on transformation parameters.[2]

[1] For details, please refer to Section 2.1.
[2] See Section 2.2 for details.

3. Compute the regularized pixel-wise ensemble entropy Ψ for the current set of images from Equation (2.21).

4. Repeat until convergence:
   For $i = 1, ..., N$,

   (a) Calculate the numerical gradient $\bigtriangledown_{v^i}\Psi$ of Equation (2.21) with respect to the transformation parameters v^i_j's for the current image ($1 \leq j \leq K$).

   (b) Update v^i as: $v^i = v^i - \gamma \bigtriangledown_{v^i} \Psi$. (where the scaling factor $\gamma \in \mathbb{R}$).

   (c) Update γ (according to some reasonable update rule such as the Armijo rule [Boyd and Vandenberghe, 2004]).

Since $\Psi(\cdot)$ is a differentiable function and the level sets

$$\mathcal{A}(\{u^i\}_{i=1}^N) = \left\{\{v^i\}_{i=1}^N \in \mathbb{R}^{K\times N} \mid \Psi(\{v^i\}_{i=1}^N) \leq \Psi(\{u^i\}_{i=1}^N)\right\} \tag{3.10}$$

are bounded for all $\{u^i\}_{i=1}^N \in \mathbb{R}^{K\times N}$, then the JPA routine will at least reach an accumulation point such that $\nabla_{v^i}\Psi = 0$ for all $i = 1, \ldots, N$ [Polak, 1997], even though the optimization routine will generally converge to a local minimum. Note that at a local minimum the set of shape templates $\Phi_{I_s} = \{I^i_s\}_{i=1}^N$ are reasonably aligned (but need not be perfectly aligned) and the set of transformations $\{g^i\}_{i=1}^N$ is properly described by the parameters $\{v^i\}_{i=1}^N$. I_l is estimated by choosing the medoid of the set of shapes, using an appropriate measure (such as the magnitude of transformation from one shape template to another based on the values of v^i). Note that the introduction of a penalty (regularization) function is critical in achieving the convergence of the optimization routine since this term diverges as the norm of v^i goes to infinity, thus making the level sets of $\Psi(\cdot)$ be bounded.

To visualize the entropy of the transformed image set for a class at each step of the optimization, one can construct an image (Figure 3.4) in which each pixel is the mean of its corresponding pixel stack.

3.3.4 Joint Alignment

We apply the $\{v_j\}^i$ learned from the JPA process to the extracted structures I_f^i to bring all the images into alignment in one step.

$$I_a^i(\boldsymbol{x}) = I_f^i(g^i(\boldsymbol{x})). \tag{3.11}$$

where $1 \leq i \leq N$. We show our results in Figures 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, 3.14, 3.15, 3.16.

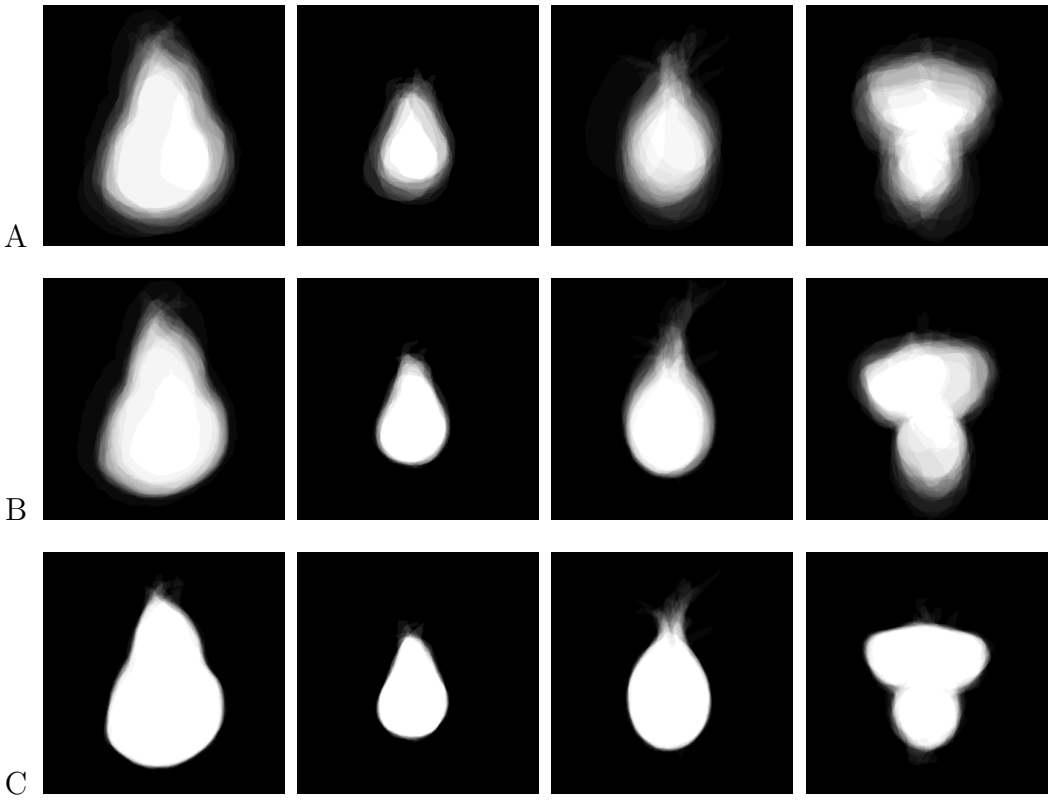

Figure 3.4: Mean shape images from the learning stage.
Mean images from optimization process during *JPA* for wing discs (*first column*), haltere discs (*second column*), leg discs (*third column*), eye discs (*fourth column*): **A**: Mean image of Φ_s *before* JPA. **B**: Mean image of Φ_s *after* JPA to convergence with only 3 parameters (t_x, t_y and θ). **C**: Mean image of Φ_s *after* JPA to convergence with 7 parameters (Equation (3.9)).

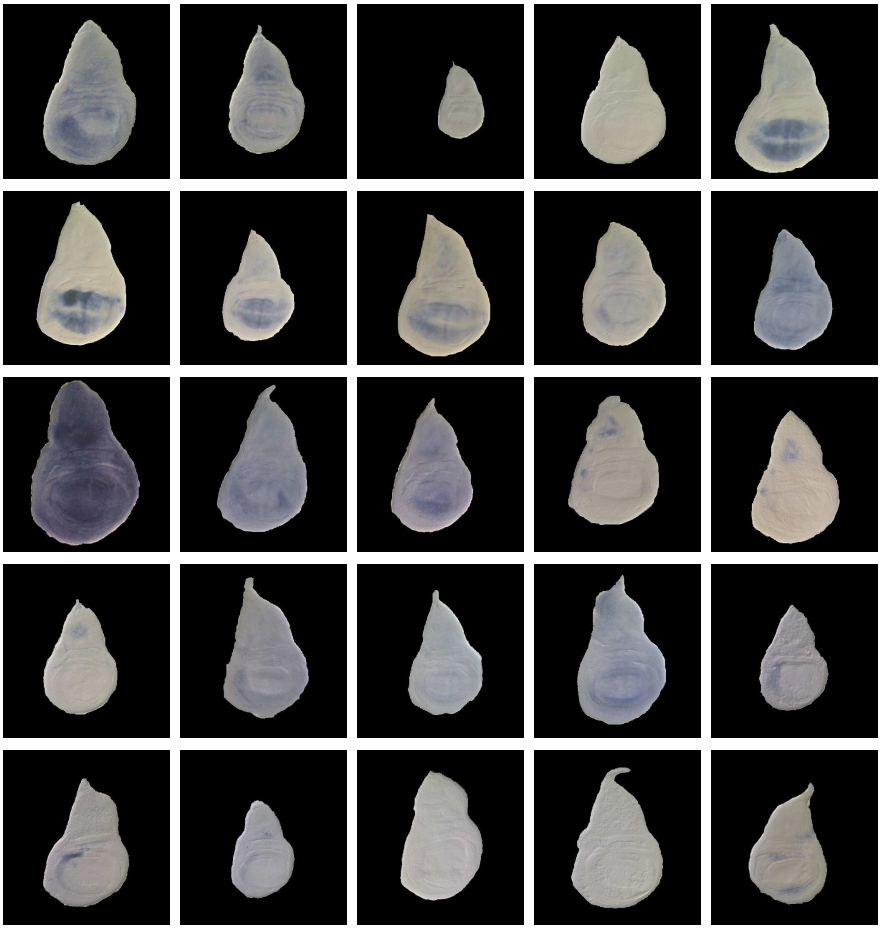

Figure 3.5: Segmented wing discs I_f^i before JPA.

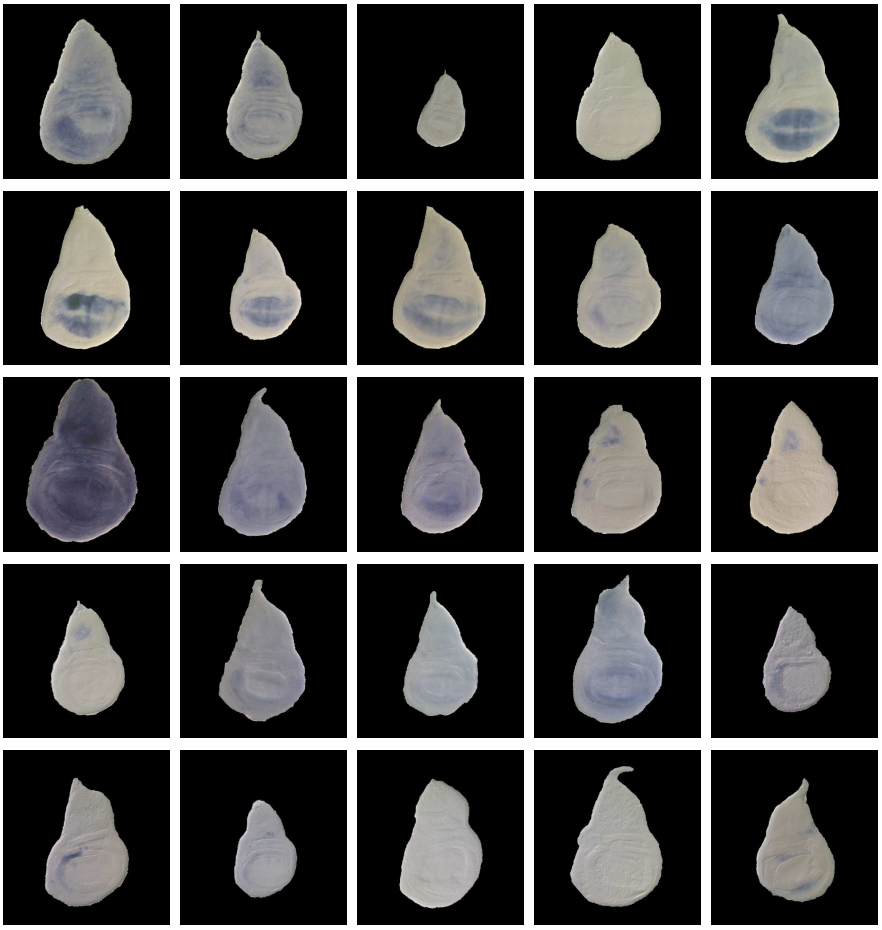

Figure 3.6: Alignment results for wing discs: Segmented wing discs I_a^i after applying the transformations learned by JPA with only 3 parameters (t_x, t_y and θ).

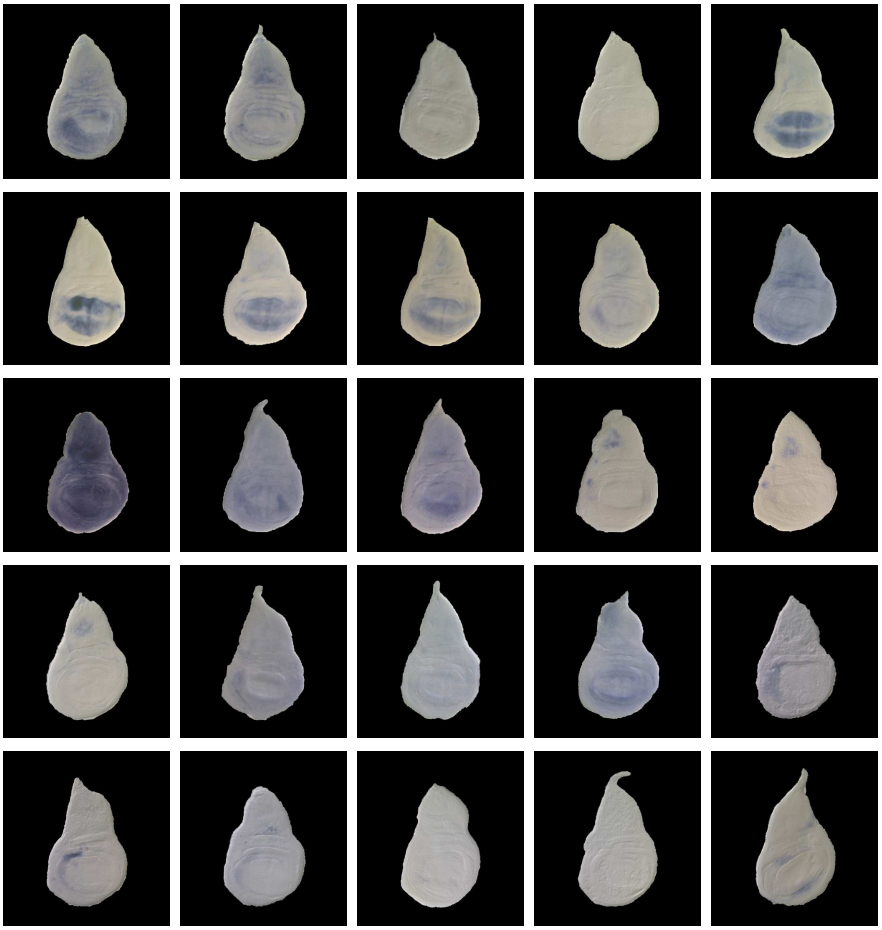

Figure 3.7: Alignment results for wing discs: Segmented wing discs I_a^i after applying the transformations learned by JPA with 7 parameters (Equation (3.9)). Note that the third image was taken with a 10x objective, while the rest were taken with a 20x objective. The reduced size of the disc is due to the way we captured the image, rather than due to biological variability. However, the algorithm was able to properly align the image.

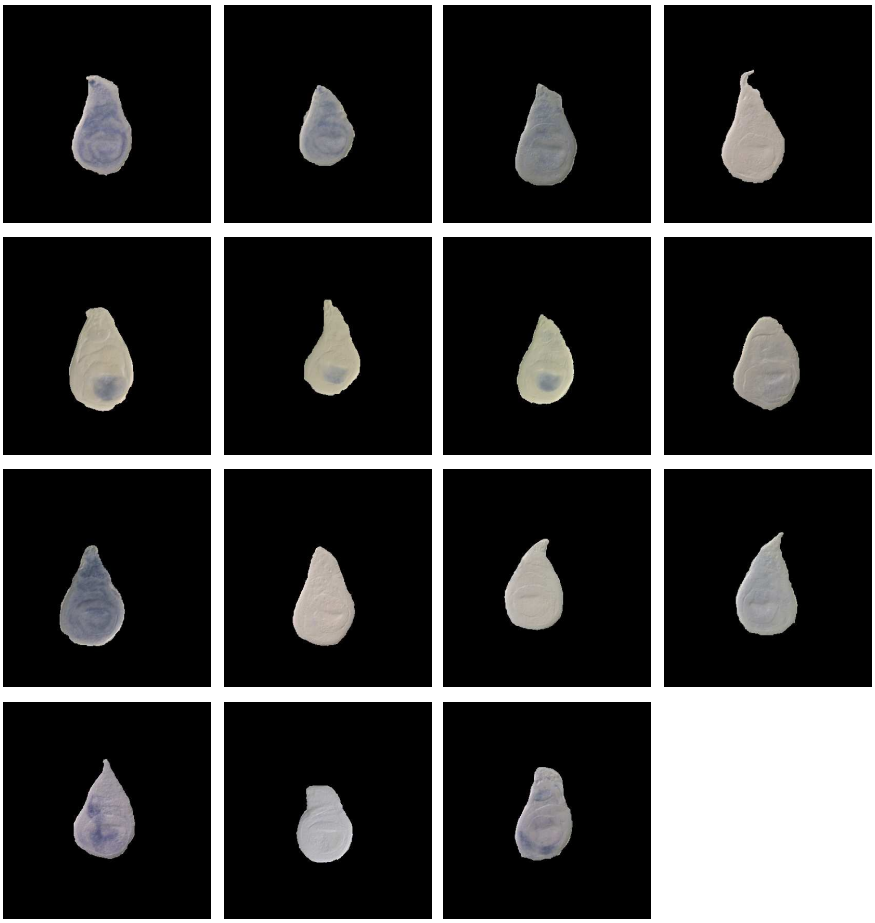

Figure 3.8: Segmented haltere discs I_f^i before JPA.

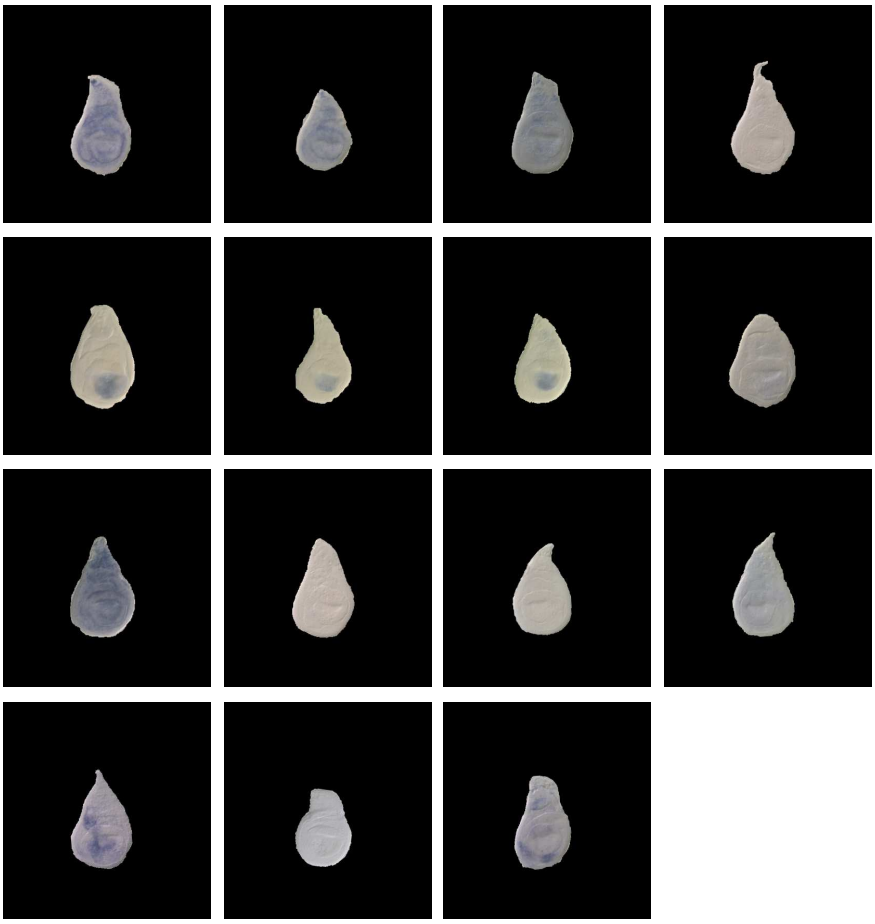

Figure 3.9: Alignment results for haltere discs: Segmented haltere discs I_a^i after applying the transformations learned by JPA with only 3 parameters (t_x, t_y and θ).

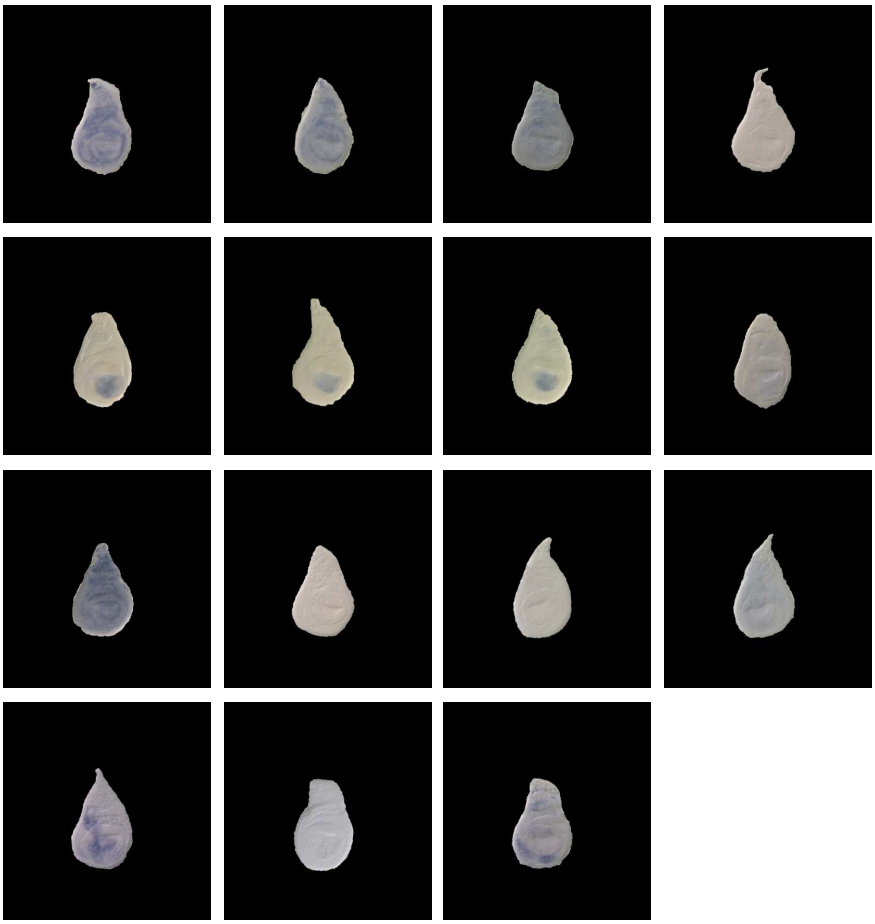

Figure 3.10: Alignment results for haltere discs: Segmented haltere discs I_a^i after applying the transformations learned by JPA with 7 parameters (Equation (3.9)).

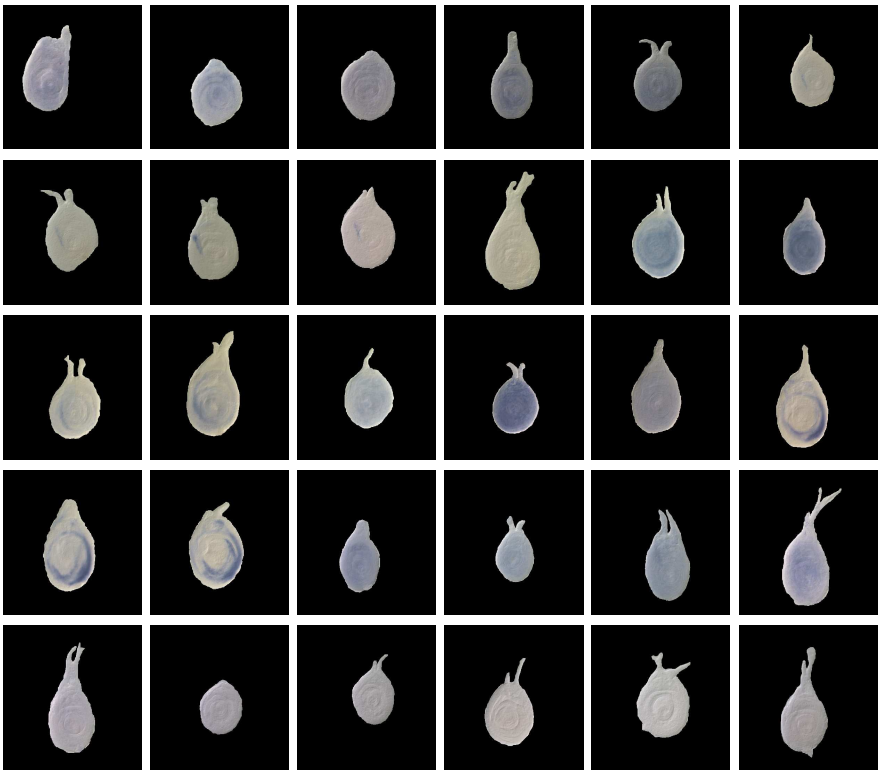

Figure 3.11: Segmented leg discs I_f^i before JPA.

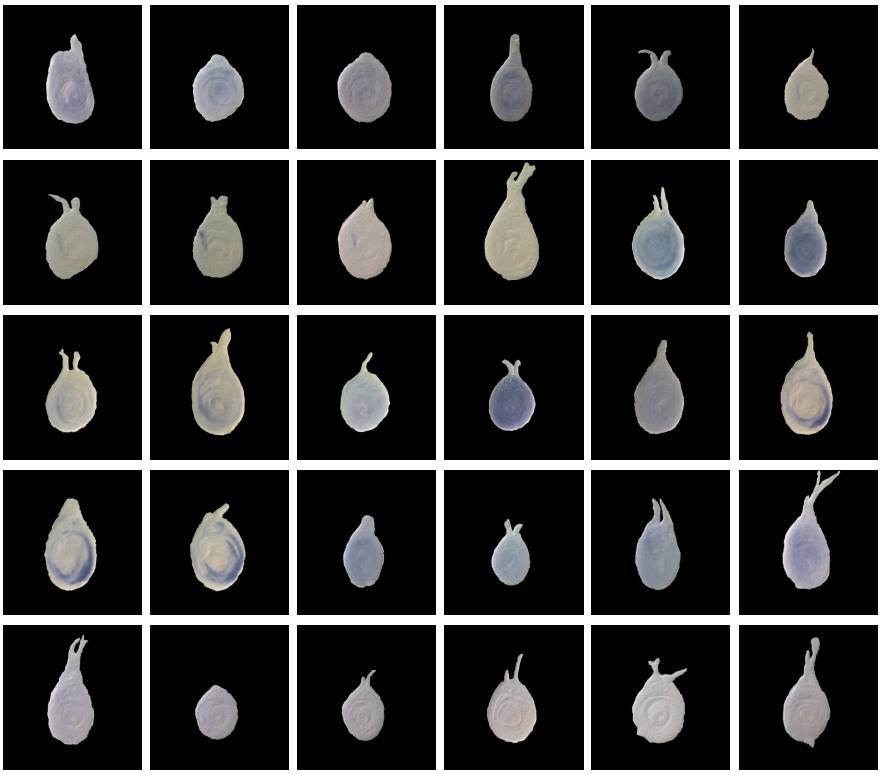

Figure 3.12: Alignment results for leg discs: Segmented leg discs I_a^i after applying the transformations learned by JPA with only 3 parameters (t_x, t_y and θ).

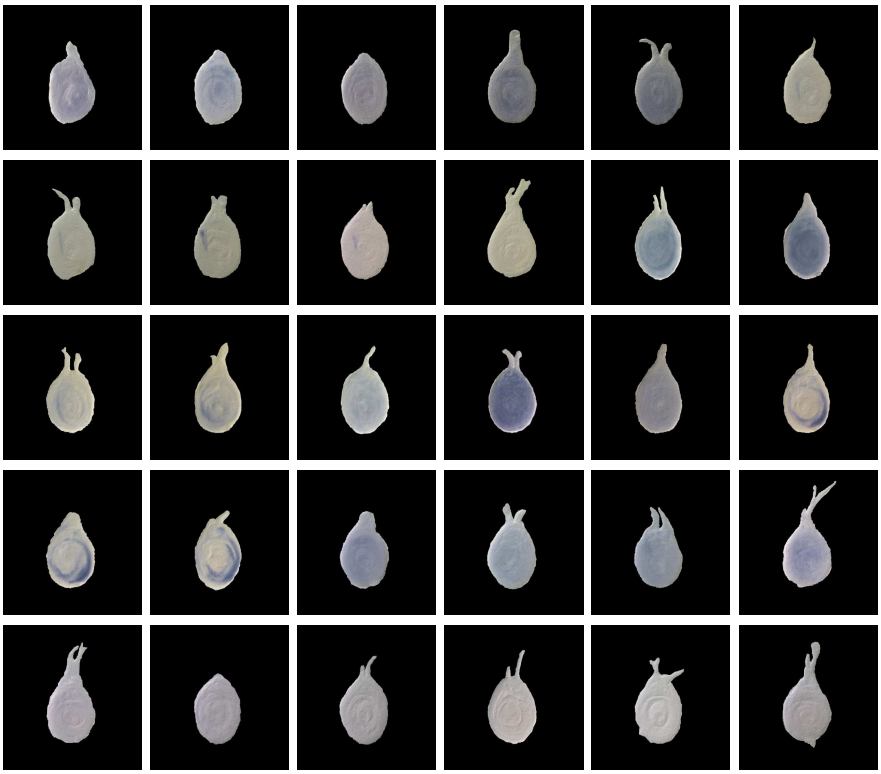

Figure 3.13: Alignment results for leg discs: Segmented leg discs I_a^i after applying the transformations learned by JPA with 7 parameters (Equation (3.9)).

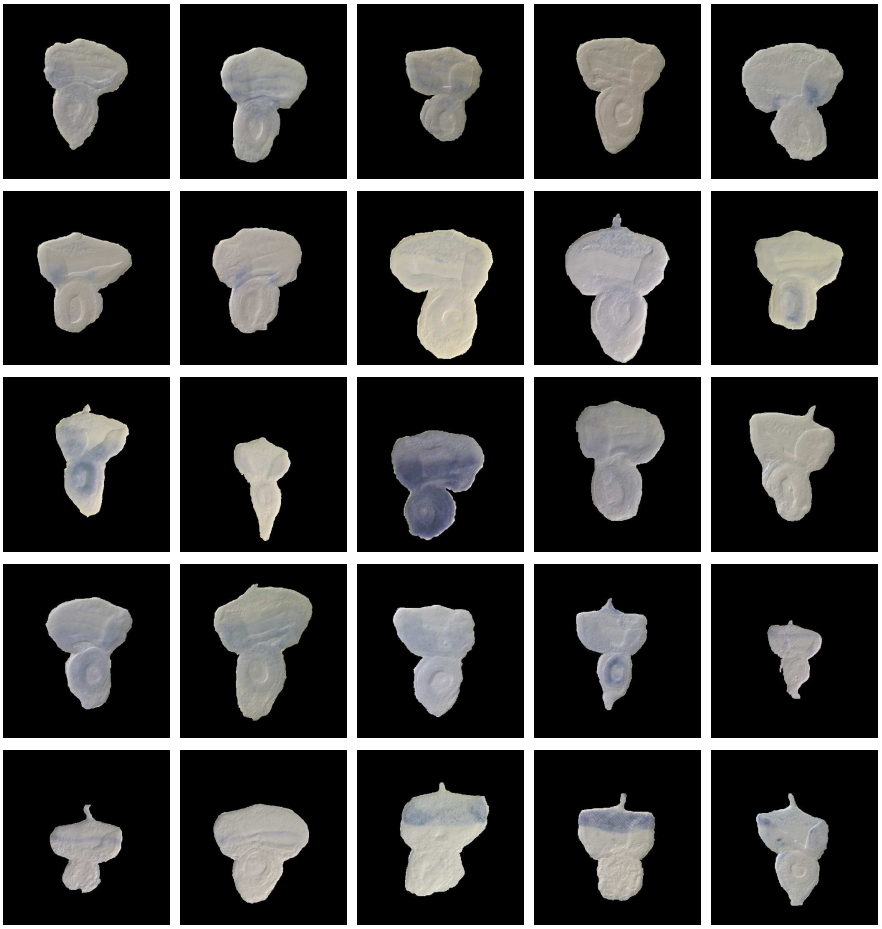

Figure 3.14: Segmented eye discs I_f^i before JPA.

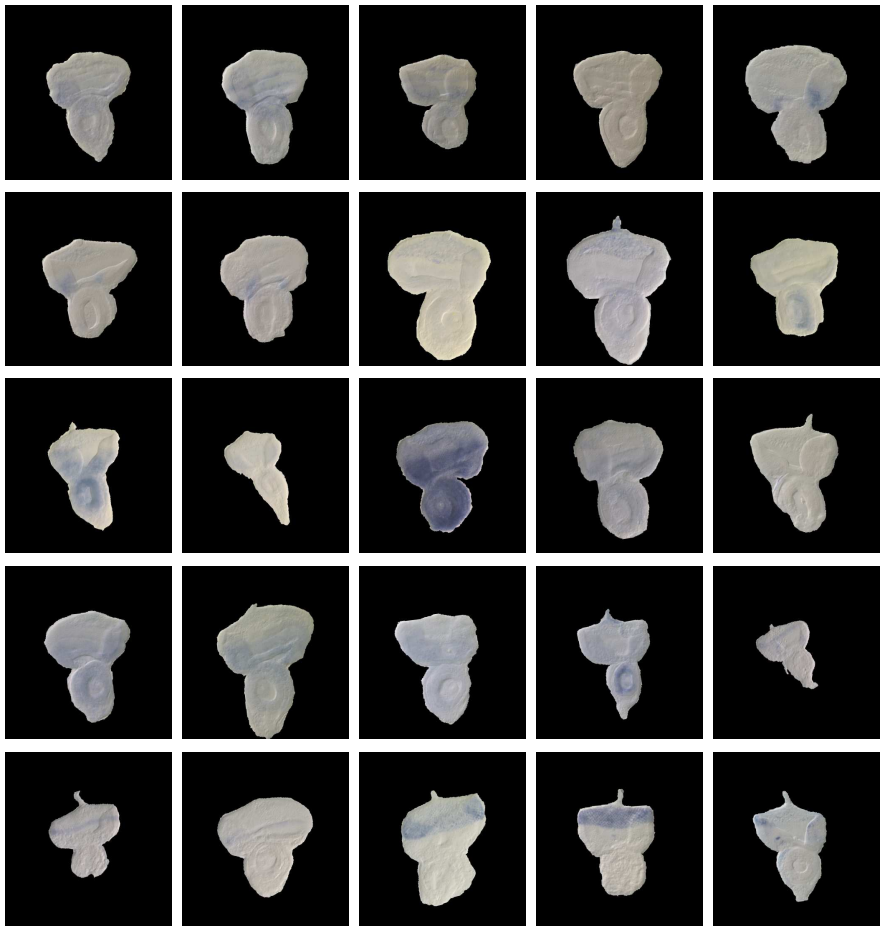

Figure 3.15: Alignment results for eye discs: Segmented eye discs I_a^i after applying the transformations learned by JPA with only 3 parameters.

Figure 3.16: Alignment results for eye discs: Segmented eye discs I_a^i after applying the transformations learned by JPA with 7 parameters (Equation (3.9)).

3.3.5 Semi-supervised JPA for Aligning Noisy Data

Using the learned shape prior from JPA procedure performed on the manually segmented disc images (approximately 15-20 examples), we use it as a template into the pipeline for searching through thousands of new input images to identify and extract instances of the given imaginal discs. In the *automated search and extract* stage, the shape model is geometrically transformed through affine space while searching through the new images, and recognition procedure is performed as a minimization of distance transform between the target and the model. Bootstrapping the learned shape prior from unsupervised JPA stage involving about 15-20 manually segmented disc examples, we can process thousands of images automatically using this semi-supervised approach. This variation allows a simple framework for handling partial occlusions and background clutter in the images. The data flow is illustrated in Figure 3.2. A detailed discussion of this segment of our pipeline can be found in [Harmon *et al.*, 2007]. Some example results of automatic segmentation and alignment procedure can be seen in Figures 3.17, 3.18, 3.19.

3.4 Stain Scoring

The local presence of stain results in the appearance of blue in the image; darker blue suggests a greater local concentration of the gene of interest. However, there is substantial probe-to-probe variability and these intensities should not be relied on as an accurate quantitative measure of gene concentration. Nevertheless, the different intensity values can be used to suggest where local gene concentration is high.

We developed a simple semi-quantitative metric that ranges from 0 to 5 where 0 indicates no expression of the gene of interest and 5 indicates high expression; the presence of blue stain causes a decrease in the intensities of the red and green channels in an RGB image. In unstained discs the Nomarski optics yield an imaginal disc image that is generally varying shades of gray, where the local features and folds of the tissue result in lighter and darker intensities. The intensities of the red, blue and green channels are generally in the same

range for a given pixel. We measured the intensity of the blue channel minus the average of the red and green channels to determine the level of staining. A small baseline value was subtracted from this number to reduce local noise due to the variability of the intensities of the channels of unstained images and the resulting value was then thresholded into six values with the lowest value suggesting no stain and, therefore, no or minimal gene of interest present, and the highest value suggesting strong expression of the gene of interest.

Examples of segmented, stain-scored wing discs, both unaligned and aligned, can be seen in Figure 3.20. Notice that the overall shape and size of the discs are more consistent in the aligned images and that a pixel-wise comparison of stain intensity of biologically similar patterns would appear more similar when comparing the aligned, stain-scored images than the unaligned stain-scored images.

3.5 Experimental Results

3.5.1 Consensus Maps of Gene Expression

Using the pipeline described in the previous sections, we have produced maps that represent the median and standard deviation maps of a given gene in a given imaginal disc tissue across all the analyzed images. Some examples are shown in Figures 3.21 and 3.22.

3.5.2 Reverse Look-up for Similar Gene Expression Patterns

Since all the images are now aligned to a consistent reference, it is straight forward to use simple metrics to come up with comparative analysis of various spatial gene expressions to identify genes that have similar spatial gene expression patterns. Some examples of such reverse look-up application using Normalized Cross Correlation (NCC) measure are shown in Figure 3.23.

3.6 Summary and Conclusions

The proposed overall methodology shown in Figure 3.2 operates without making any assumptions about the underlying structure of a given tissue class. It is semi-supervised (learning the shape prior is supervised via manually segmented example shapes, and the rest of the process is automated) and robust to biological clutter and noise in the data. Clearly, pattern alignment is a critical enabling step in building high-throughput spatial gene expression atlases. The proposed pipeline is highly amenable to large scale spatial gene expression analysis and needs no further tweaks from one class of tissue to another (under the assumption that all supplied images belong to one tissue class). The same framework could also be applicable to construct high-throughput data driven atlases of spatial expression for other biological or neurobiological substrates. It augments any model-based registration methods one may choose to apply by supplying the nonparametrically learned canonical structure model from the given ensemble of images for a given tissue class.

We have developed a method of generating a large number of spatial patterns of gene expression in Drosophila melanogaster imaginal discs, and for using shapes learned from the data, rather than using a single exemplar, as a global model of the shape of interest, to which a set of patterns can be aligned. We successfully implemented and demonstrated the applicability of this methodology using Drosophila imaginal discs. Using our methods, we have determined the patterns of over 130 genes in some or all of the four largest and most well-characterized imaginal disc types, the wing, leg, haltere and eye/antenna discs. Yet, our characterization of spatial gene expression in Drosophila imaginal discs is not exhaustive (mainly due to resource constraints related to data collection). Applying our pipeline to more genes could be potentially informative on a number of levels. Further analysis of genes known to play a role in the patterning and development of imaginal discs, and the quantification of the precise extent of spatial expression of these genes may provide a more detailed view of the roles of and interactions between these important genes.

Using the salient features of our images, we suggested a simple filter-and-threshold algorithm for segmentation which performs well once the learned shape template is supplied.

The discussion in this chapter is focused on the geometric transformations that can be approximated by an affine model in a two-dimensional plane since imaginal discs can be represented using a two-dimensional representation. One really interesting extension of our work would be in extending our system to work in the space of non-rigid transformations, since this could potentially provide a better alignment fit.

Our representation of stain patterns as a quantitative measure of gene expression, aligned to a global model, enables us to efficiently cluster both the patterns of the genes themselves, and the regions of the tissues, as represented by the pixels in the global model. Finally, we have developed a reverse-lookup procedure, that enables us to take a new image, stained for a gene of interest, and to search our database of patterns to find genes with similar spatial patterns of gene expression. We have also performed detailed comparative analysis of spatial patterns of gene expression in aligned imaginal discs using pixel-wise comparisons based on the approaches described in this work. In other datasets where shape prior is a strong cue, similar approach can be taken. Another natural direction to extend our approach is to apply this procedure to three-dimensional datasets such as image stacks from confocal microscopy studies of in situ stained tissues (such as Drosophila embryos) to construct data-driven gene expression atlases.

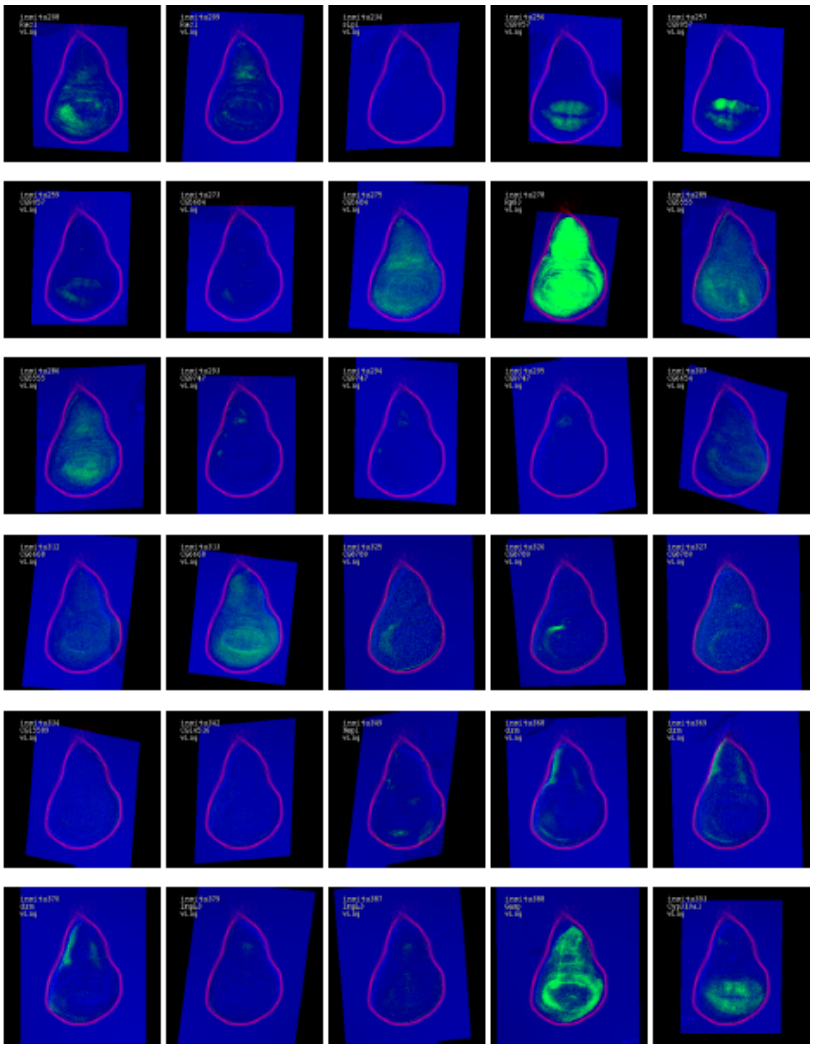

Figure 3.17: Example images of Drosophila melanogaster wing imaginal discs, automatically segmented and aligned to the model. The warping of the target frame is shown via the warped plane of each of the examples.

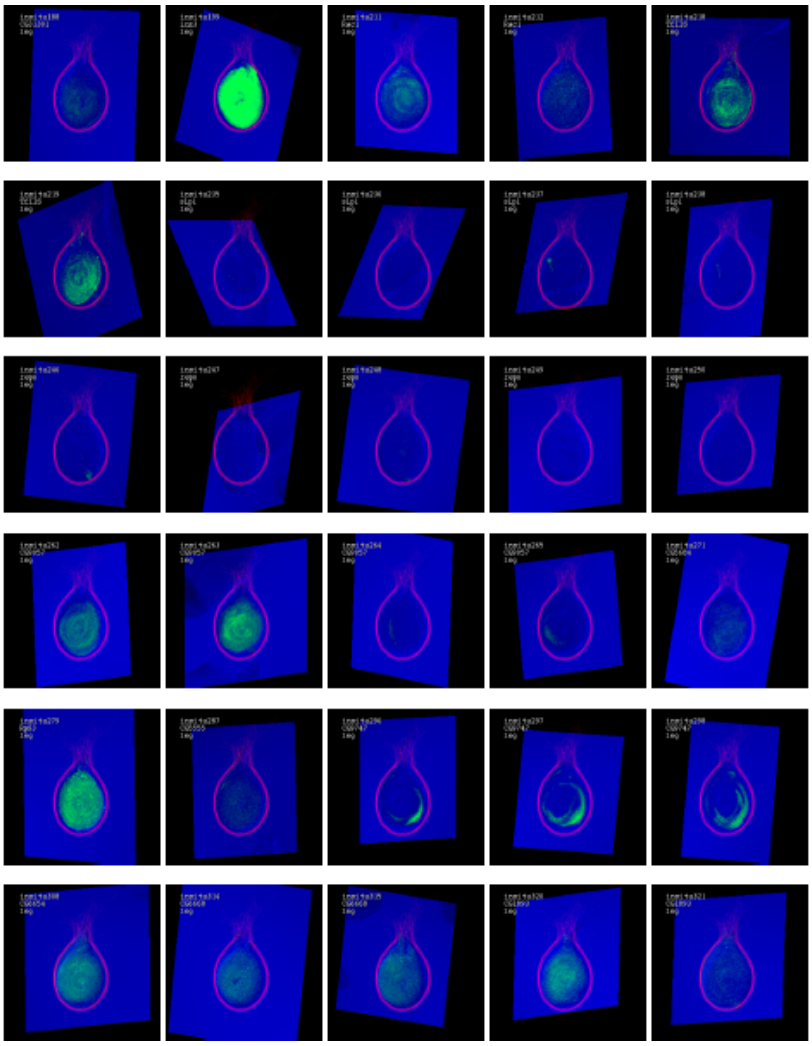

Figure 3.18: Example images of Drosophila melanogaster leg imaginal discs, automatically segmented and aligned to the model. The warping of the target frame is shown via the warped plane of each of the examples.

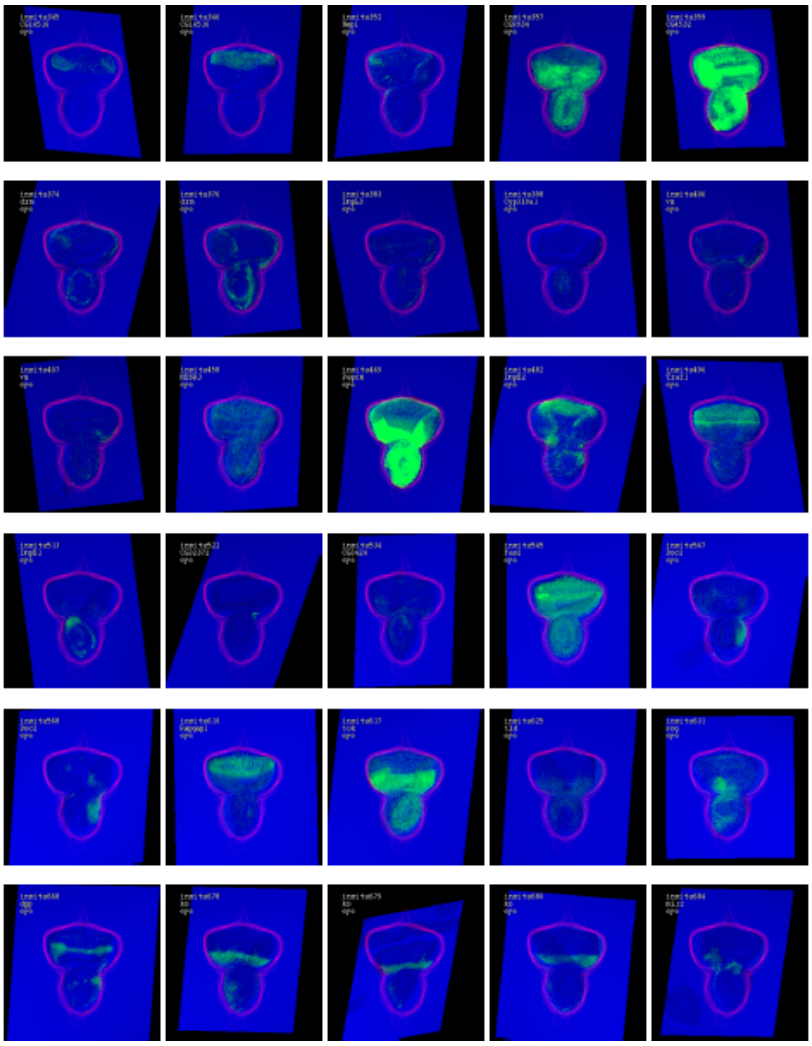

Figure 3.19: Example images of Drosophila melanogaster eye/antennal imaginal discs, automatically segmented and aligned to the model. The warping of the target frame is shown via the warped plane of each of the examples.

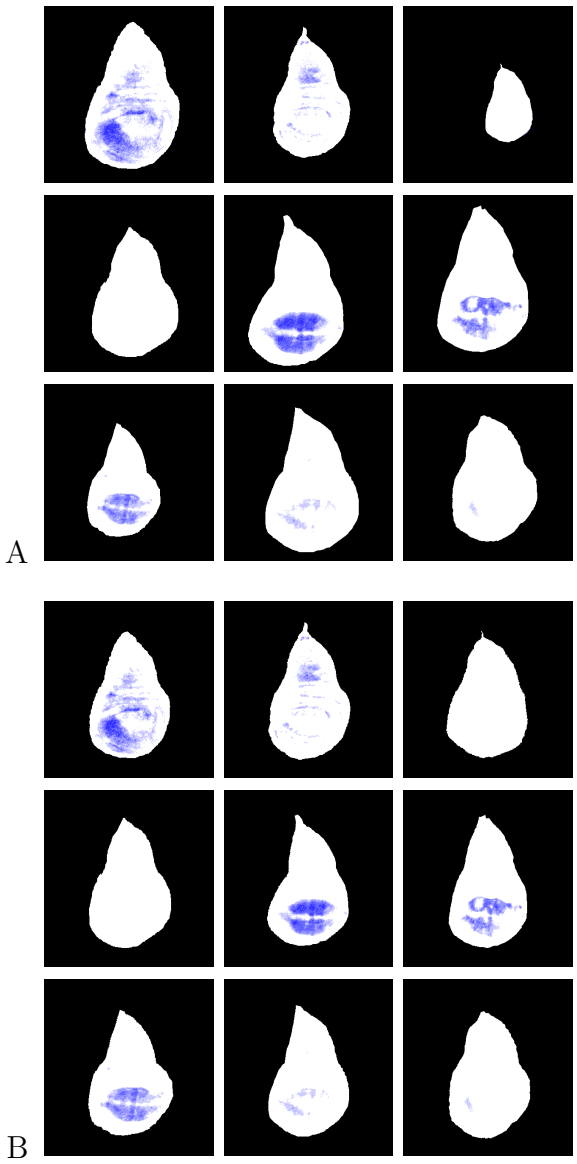

Figure 3.20: Stain patterns in Drosophila imaginal disc images: unaligned vs. aligned. **A**: Unaligned stain-scored wing disc images. **B**: Aligned stain-scored wing disc images after JPA. Black pixels have been segmented as background, white indicates no stain and shades of blue indicate stained pixels. Blue intensity values were calculated from the semi-quantitative stain scoring algorithm with the lightest blue representing value 1 and the darkest blue representing value 5. It can be noted that pixel-wise stain count is much more meaningful in these images. (This image is better viewed in color).

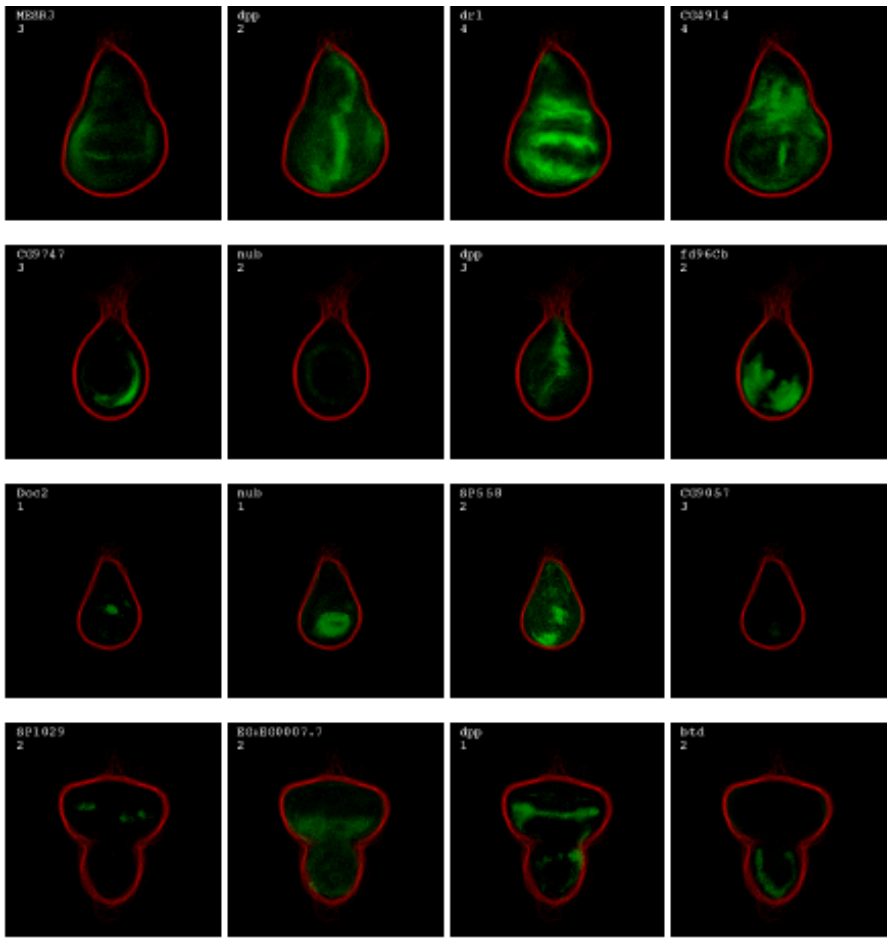

Figure 3.21: Median Gene expression maps of imaginal discs in Drosophila melanogaster. The maps are made by taking the median expression value at each pixel from the stack of images for each gene. (Top row: Maps of $MESR3$, dpp, drl and $CG4914$ in the wing. Second row: Maps of $CG9747$, nub, dpp and $fd96Cb$ in the leg. Third row: Maps of $Doc2$, nub, $SP558$, and $CG9057$ in the haltere. Fourth row: Maps of $SP1029$, $EG:EG0007.7$, dpp and btd in the eye/antenna disc).

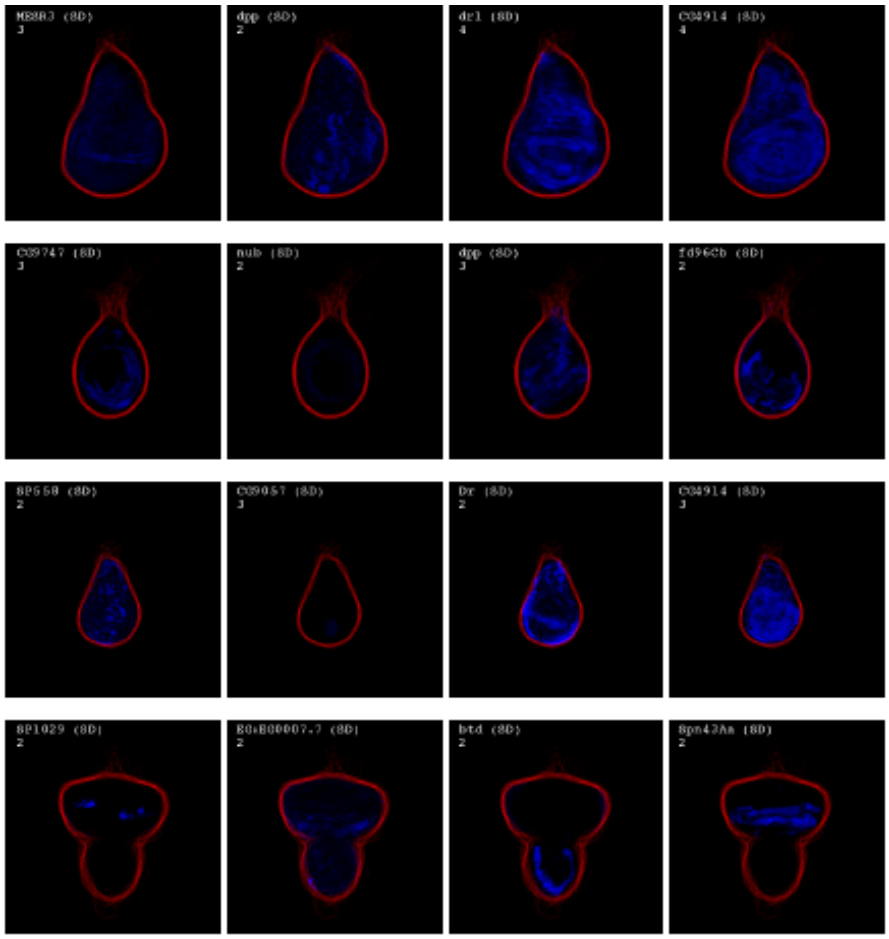

Figure 3.22: Gene expression standard deviation maps of imaginal discs in Drosophila melanogaster. The maps are made by taking the standard deviation at each pixel from the stack of images for each gene. (Top row: Maps of $MESR3$, dpp, drl and $CG4914$ in the wing. Second row: Maps of $CG9747$, nub, dpp and $fd96Cb$ in the leg. Third row: Maps of $SP558$, $CG9057$, Dr and $CG4914$ in the haltere. Fourth row: Maps of $SP1029$, $EG:EG0007.7$, btd and $Spn43Aa$ in the eye/antenna disc).

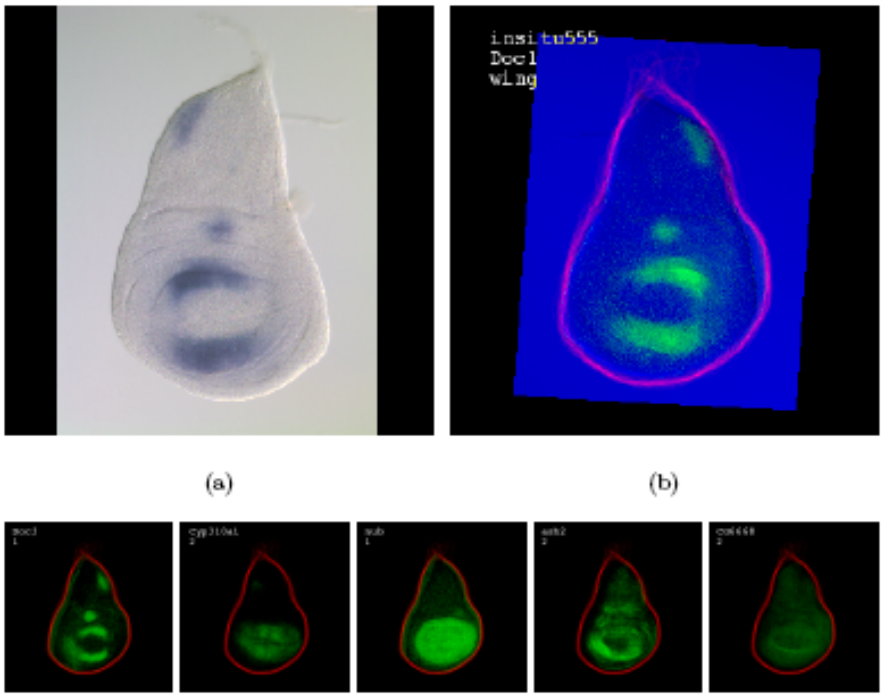

Figure 3.23: Reverse lookup for $insitu555$. (a) Original (resized and padded to 256×256 pixels) image $inistu555$. (b) Automatically aligned, extracted and stain scored $insitu555$. The extracted boundary is shown in red and the aligned and scored stain shown in green. (c) The 5 gene maps that most closely match the stain pattern derived from $insitu555$ using the NCC distance measure.

Chapter 4

Joint Random Field Bias Removal in MRI Images

4.1 Introduction

The central goal of this chapter [1] is to demonstrate the applicability of joint pattern alignment framework to the application of random field bias removal in Magnetic Resonance imaging[Learned-Miller and Ahammad, 2005]. This chapter also demonstrates that JPA can be extended to model non-geometric transformations such as intensity variations.

Magnetic Resonance (MR) imaging is a powerful noninvasive imaging modality that has experienced rapid growth over the past decade. Standard applications of MR include diagnostic imaging studies of the central nervous system and musculo-skeletal system [Nishimura, 1996]. There are a number of artifacts that can arise in the MR imaging process and make subsequent analysis very challenging. Possibly the most drastic visual effect is the intensity inhomogeneity caused by the spatially varying signal response of the electrical coil that receives the MR signal. This coil inhomogeneity results in a multiplicative gain field that biases the observed signal from the true underlying signal [Fan, 2003]. This problem is

[1] This chapter is a more detailed version of [Learned-Miller and Ahammad, 2005], with additional controlled experiments with BrainWeb data.

illustrated in Figure 4.1. When a patient is imaged in the MR scanner, the goal is to obtain an image which is a function solely of the underlying tissue (first image in Figure 4.1). However, typically the desired anatomical image is corrupted by a multiplicative bias field (first image in Figure 4.1) that is caused by engineering issues such as imperfections in the radio frequency coils used to record the MR signal. The result is a corrupted image (first image in Figure 4.1). Most of the diagnostic MR image processing procedures operate on

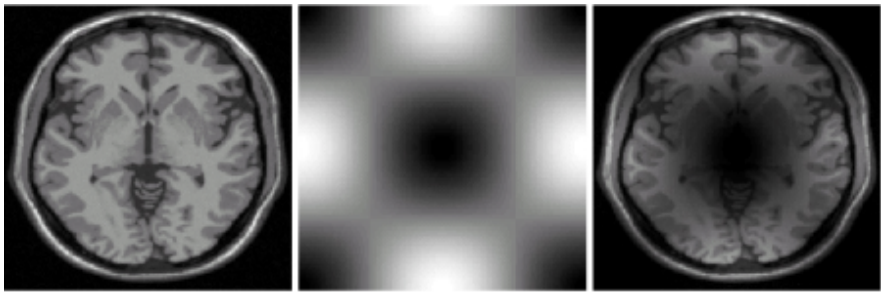

Figure 4.1: Illustration of the effect of bias in MR images. On the left is a mid-axial MRI scan of the human brain with little or no bias field. In the center is a simulated low-frequency bias field. It has been exaggerated for ease of viewing. On the right is the result of pixelwise multiplication of the image by the bias field. The goal of MR bias correction is to recover the low-bias image on the left from the biased image on the right.

the intensity values obtained in MR images. MR images are constructed from electromagnetic responses and are captured using coils of wire. These intensities are corrupted both by random noise as well as systematic electromagnetic effects. The latter are collectively known as bias fields or intensity inhomogeneities. The bias in this case is a multiplicative bias rather than an additive bias which is more common. The term bias is used because the intensity inhomogeneity is a systematic effect and not a random effect. Image processing in general relies on the intensity values and can be significantly impaired by imperfections in the image collection process. Both the noise and the bias can confuse automated image processing algorithms, and it is highly desirable to minimize both as much as possible. The goal of MR bias correction is to estimate the uncorrupted image from the corrupted image. The bias correction problem is currently a challenging one and is very widely studied. The importance of this problem will increase as MR magnets increase in strength and

electromagnetic effects become more and more pronounced.

4.1.1 Related Work

A variety of statistical methods have been proposed to address this problem. Wells et al. [Wells *et al.*, 1996] developed a statistical model using a discrete set of tissues, with the brightness distribution for each tissue type (in a bias-free image) represented by a one-dimensional Gaussian distribution. An expectation-maximization (EM) procedure was then used to simultaneously estimate the bias field, the tissue type, and the residual noise. While this method works well in many cases, it has several drawbacks:

1. Models must be developed *a priori* for each type of acquisition (for each different setting of the MR scanner), for each new area of the body, and for different patient populations (like infants and adults).

2. Models must be developed from *bias-free* images, which may be difficult or impossible to obtain in many cases.

3. The model assumes a fixed number of tissues, which may be inaccurate. For example, during development of the human brain, there is continuous variability between gray matter and white matter.

In addition, a discrete tissue model does not handle so-called partial volume effects in which a pixel represents a combination of several tissue types. This occurs frequently since many pixels occur at tissue boundaries. Non-parametric approaches have also been suggested, as for example by Viola [Viola, 1995]. In that work, a non-parametric model of the tissue was developed from a single image. Using the observation that the entropy of the pixel brightness distribution for a single image is likely to increase when a bias field is added, Viola's method postulates a bias-correction field by minimizing the entropy of the resulting pixel brightness distribution. This approach addresses several of the problems of fixed-tissue parametric models, but has its own drawbacks:

1. The statistical model may be weak, since it is based on data from only a single image.

2. There is no mechanism for distinguishing between certain low-frequency image components and a bias field. That is, the method may mistake signal for noise in certain cases when removal of the true signal reduces the entropy of the brightness distribution.

We shall show that this is a problem in real medical images. The method we present overcomes or improves upon problems associated with both of these methods and their many variations (see, e.g., [Fan, 2003] for recent techniques). It models tissue brightness non-parametrically, but uses data from multiple images to provide improved distribution estimates and alleviate the need for bias-free images for making a model. It is also conditional on spatial location, taking advantage of a rich information source ignored in other methods. Experimental results demonstrate the effectiveness of our method.

4.2 JPA for Bias Removal

4.2.1 The Image Model and Notation

We assume we are given a set I of observed images I_i with $1 \leq i \leq N$, as shown on the left side of Figure 4.2. Each of these images is assumed to be the product of some bias-free image L_i and a smooth bias field $B_i \in \Phi$. We shall refer to the bias-free images as latent images (also called *intrinsic images* by some authors). The set of all latent images shall be denoted L and the set of unknown bias fields B. $L_i(\cdot)$, $B_i(\cdot)$ and $I_i(\cdot)$ can be represented as maps from $\mathbb{R}^2$ to the set $\mathbb{R}$ with a domain $\Omega \subset \mathbb{R}^2$:

$$L_i, B_i, I_i : \quad \Omega \mapsto \mathbb{R}. \tag{4.1}$$

Typically, the domain Ω is a square or rectangular window. Let $\boldsymbol{x} \in \Omega$ denote the pixel location such that $\boldsymbol{x} = [x, y]^T$. Using a multiplicative bias model, each observed image can be written as the product $I_i(\boldsymbol{x}) = L_i(\boldsymbol{x}) * B_i(\boldsymbol{x})$, and $\boldsymbol{x} \in \Omega$.

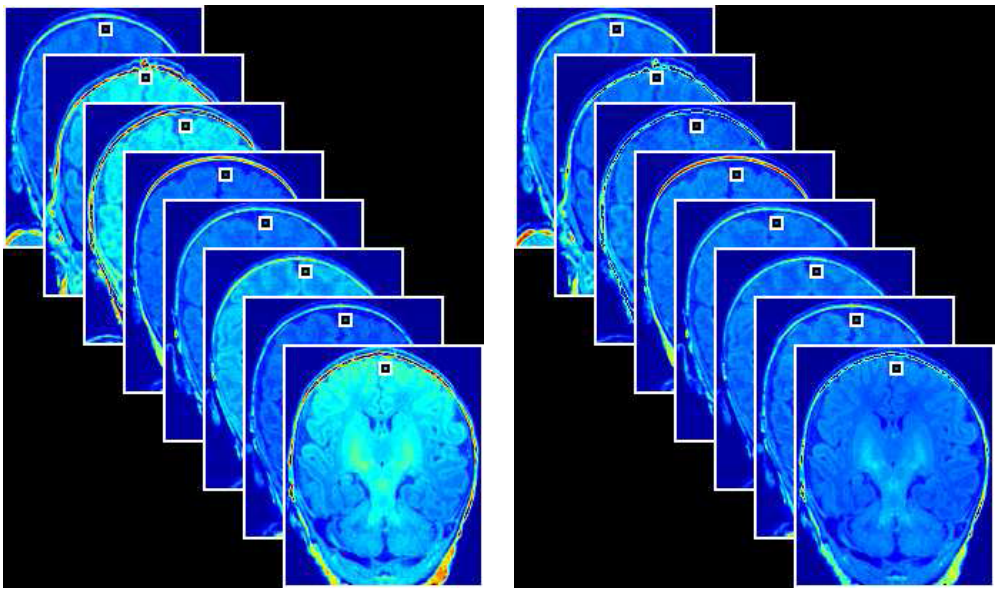

Figure 4.2: Pixel stacks in MR images. On the left are a set of mid-coronal brain images from eight different infants, showing clear signs of bias fields. A pixel-stack, a collection of pixels at the same point in each image, is represented by the small square near the top of each image. Although there are probably no more than two or three tissue types represented by the pixel-stack, the brightness distribution through the pixel-stack has high empirical entropy due to the presence of different bias fields in each image. On the right are a set of images that have been corrected using our bias field removal algorithm. While the images are still far from identical, the pixel-stack entropies have been reduced by mapping similar tissues to similar values in an unsupervised fashion, i.e. without knowing or estimating the tissue types.

Consider again Figure 4.2. A *pixel-stack* through each image set is shown as the set of pixels corresponding to a particular location in each image (not necessarily the same tissue type). Our method operates using the intuition that the pixel-stack values will have lower entropy when the bias fields have been removed.

The latent image generation model assumes that each pixel is drawn from a fixed distribution $p_{\boldsymbol{x}}(\cdot)$ which gives the probability of each gray value at the the location $\boldsymbol{x}$ in the image. Furthermore, we assume that all pixels in the latent image are independent, given the distributions from which they are drawn. It is also assumed that the bias fields for each image are chosen independently from some fixed distribution over bias fields. Unlike most

models for this problem which rely on statistical regularities within an image, we take a completely orthogonal approach by assuming that pixel values are independent given their image locations, and that pixel-stacks in general will have low entropy when bias fields are removed.

4.2.2 Problem Formulation

We assume Uniform prior over the basis fields in this derivation, but a different assumption can as well be used if the user has specific domain knowledge. We formulate the problem as a maximum a posteriori (MAP) problem, searching for the most probable bias fields given the set of observed images. Letting Φ represent the product space of smooth bias fields (corresponding to the $K = 25$ basis images of Figure 4.3), we wish to find $arg\ max_{B\in\Phi}P(B|I)$. We define $\hat{\Theta}$ as:

$$\hat{\Theta} = arg\ max_{B\in\Phi}P(B|I). \tag{4.2}$$

Using Bayes' rule and ignoring the constant denominator, we can write it as:

$$\hat{\Theta} \quad = \quad arg\ max_{B\in\Phi}P(I|B)P(B). \tag{4.3}$$

We assume Uniform prior over the basis fields in this derivation. Thus, we can write this as:

$$\hat{\Theta} \quad = \quad arg\ max_{B\in\Phi}P(I|B). \tag{4.4}$$

Our method can be easily altered to incorporate non-uniform prior as well[2]. The probability of an observed image given a particular bias field is the same as the probability of the latent

[2]Assuming a non-uniform prior will result in a penalized form of ensemble entropy in the optimization objective function. See section 2.2 for details.

image associated with that observed image and bias field. This can be expressed as

$$\hat{\Theta} = arg\ max_{B\in\Phi} P(L(I,B)) \tag{4.5}$$

$$= arg\ max_{B\in\Phi} \prod_{\boldsymbol{x}\in\Omega}\prod_{i=1}^{N} p_{\boldsymbol{x}}(L_i(\boldsymbol{x})) \tag{4.6}$$

Taking logarithm,

$$\hat{\Theta} = arg\ max_{B\in\Phi} \sum_{\boldsymbol{x}\in\Omega}\sum_{i=1}^{N} p_{\boldsymbol{x}}(L_i(\boldsymbol{x})) \tag{4.7}$$

At each pixel, the empirical mean of the log probability can be approximated with the negative entropy of the underlying distribution at that pixel. This can be written as:

$$\hat{\Theta} \approx arg\ min_{B\in\Phi} \sum_{\boldsymbol{x}\in\Omega} H(p_{\boldsymbol{x}}). \tag{4.8}$$

Here H is the empirical entropy of the pixel stack (in the case of Shannon entropy, it is defined as $H = \mathbb{E}_p(-\log\ p_{\boldsymbol{x}})$). We use the entropy estimator of Vasicek [Vasicek, 1976] to directly estimate this entropy from the samples in the pixel-stack, without ever estimating the distributions $p_{\boldsymbol{x}}$ explicitly. The approximation in Equation (4.8) becomes an equality as N grows large by the law of large numbers, while the consistency of Vasicek's entropy estimator [Beirlant *et al.*, 1997] implies that Equation (4.9) also goes to equality with large N [3].

$$\hat{\Theta} \approx arg\ min_{B\in\Phi} \sum_{\boldsymbol{x}\in\Omega} \widehat{H}_{Vasicek}(L_1(\boldsymbol{x}),\ldots,L_N(\boldsymbol{x})) \tag{4.9}$$

$$\hat{\Theta} = arg\ min_{B\in\Phi} \sum_{\boldsymbol{x}\in\Omega} \widehat{H}_{Vasicek}(\frac{I_1(\boldsymbol{x})}{B_1(\boldsymbol{x})},\ldots,\frac{I_N(\boldsymbol{x})}{B_N(\boldsymbol{x})}). \tag{4.10}$$

[3] Please refer to [Beirlant *et al.*, 1997] for a review of entropy estimators.

This estimation can be formulated as solving an optimization problem. We parameterize the set of bias fields using the sine/cosine basis images shown in Figure 4.3:

$$B_i(\boldsymbol{x}) \quad = \quad \sum_{1}^{K} v_j^i \phi_j(\boldsymbol{x}). \tag{4.11}$$

where v^i are the vectors of bias field parameters (Equation (4.11)). The objective function

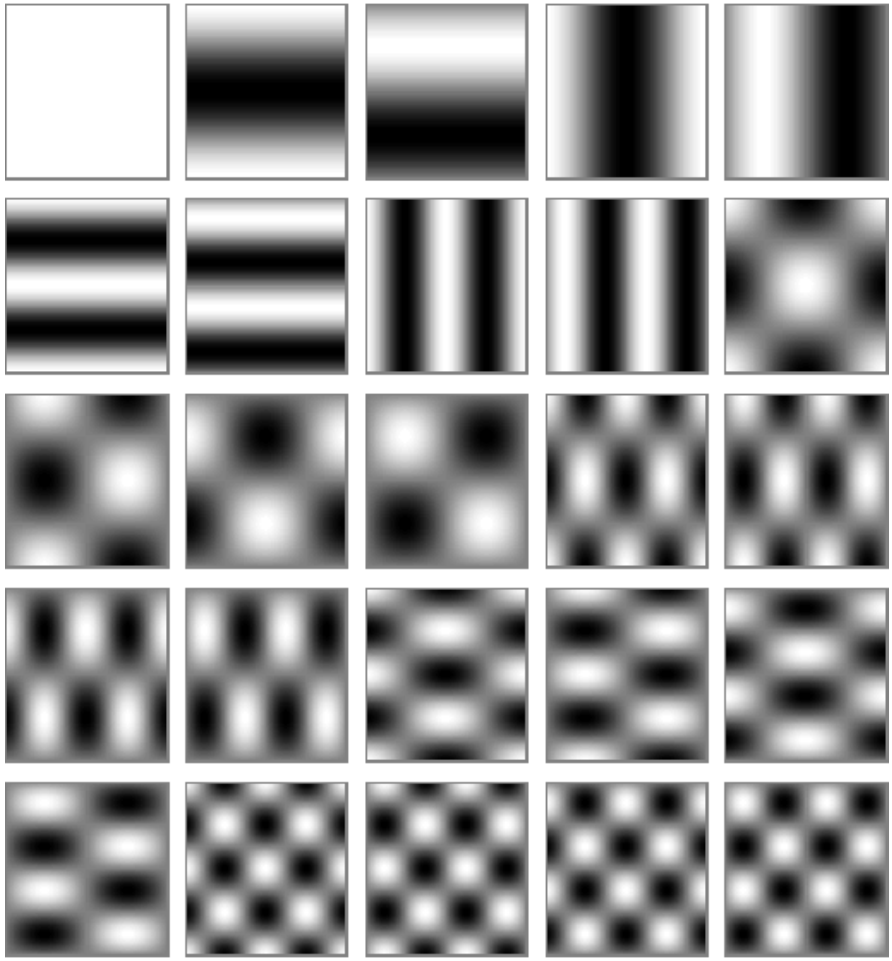

Figure 4.3: Sine/Cosine Basis Fields
$K = 25$ sine/cosine basis fields are shown here, that are combined to construct band-limited bias fields using Equation (4.11)

for our optimization Ψ defined as

$$\Psi \doteq \sum_{B \in \Phi} \sum_{\boldsymbol{x} \in \Omega} \widehat{H}_{Vasicek}(L_1(\boldsymbol{x}), \ldots, L_N(\boldsymbol{x})) \tag{4.12}$$

If the prior on the bias fields is non-uniform, as discussed in Section 2.2, the objective function will have penalty term that is a function of the logarithm of the prior on v_j^i. Ψ is called the regularized pixel-wise entropy [Learned-Miller and Ahammad, 2005], and since we assumed a Uniform prior in this case, there is no penalty term in Equation (4.12).

4.2.3 Joint RF Bias Removal Algorithm

Using the ideas discussed so far, it is straightforward to construct an algorithm for joint bias field removal. We chose to optimize Equation (4.10) over the set of band-limited bias fields. We optimize Equation (4.10) by simultaneously updating the bias field estimates (taking a step along the numerical gradient) for each image to reduce the overall entropy. That is, at time step t, the coefficients v_j for each bias field are updated using the latent image estimates and entropy estimates from time step $t-1$. After all the v have been updated, a new set of latent images and pixel-stack entropies are calculated, and another gradient step is taken. Though it is possible to do a full gradient descent to convergence by optimizing one image at a time, the optimization landscape tends to have more local minima for the last few images in the process. The appeal of our joint gradient descent method, on the other hand, is that the ensemble of images provides a natural smoothing of the optimization landscape in the joint process. It is in this sense that our method is *multi-resolution*, proceeding from a smooth optimization in the beginning to a sharper one near the end of the process.

We now summarize the JPA algorithm for estimating the RF bias fields from MR images:

1. Initialize the bias field coefficients for each image to 0, with the exception of the coefficient for the DC-offset (the constant bias field component), which is initialized to 1. Initialize the gradient descent step size δ to some value.

2. Choose an appropriate penalty term in objective function Ψ. Ψ can be computed based on the probability assumptions made on basis coefficients ($P(v_j^i)$) (See Section 2.2 for discussion). For Uniform prior over v_j^i, there is no penalty.

3. Compute Ψ for the set of images with initial *neutral* bias field corrections. (See below for method of computation.)

4. Iterate the following loop until no further changes occur in the images.

   (a) For each image:

      i. Calculate the numerical gradient $\bigtriangledown_{v^i}\Psi$ of equation (4.10) with respect to the bias field coefficients (v_j's) for the current image.

      ii. Set $\gamma = \gamma - \delta \bigtriangledown_{v^i} \Psi$.

   (b) Update δ (according to some reasonable update rule such as the Armijo rule [Boyd and Vandenberghe, 2004]).

Upon convergence, it is assumed that the entropy has been reduced as much as possible by changing the bias fields, unless one or more of the gradient descents is stuck in a local minimum. Empirically, the likelihood of sticking in local minima is dramatically reduced by increasing the number of images (N) in the optimization. In our experiments described below with only 21 real infant brains, the algorithm appears to have found a global minimum of all bias fields, at least to the extent that this can be discerned visually.

4.2.4 Discussion

Note that for a set of *identical* images, the pixel-stack entropies are not increased by multiplying each image by the same bias field (since all images will still be the same). More generally, when images are approximately equivalent, their pixel-stack entropies are not significantly affected by a *common* bias field, i.e. one that occurs in all of the images[4].

[4]Actually, multiplying each image by a bias field of small magnitude can artificially reduce the entropy of a pixel-stack, but this is only the result of the brightness values shrinking toward zero.

This means that the algorithm cannot, in general, eliminate all bias fields from a set of images, but can only set all of the bias fields to be equivalent. We refer to any constant bias field remaining in all of the images after convergence as the *residual bias field*. Fortunately, there is an effect that tends to minimize the impact of the residual bias field in many test cases. In particular, the residual bias field tends to consist of components for each v_j that approximate the mean of that component across images. For example, if half of the observed images have a positive value for a particular component's coefficient, and half have a negative coefficient for that component, the residual bias field will tend to have a coefficient near zero for that component. Hence, the algorithm naturally eliminates bias field effects that are non-systematic, i.e. that are not shared across images.

If the same type of bias field component occurs in a majority of the images, then the algorithm will not remove it, as the component is indistinguishable, under our model, from the underlying anatomy. In such a case, one could resort to within-image methods to further reduce the entropy. However, there is a risk that such methods will remove components that actually represent smooth gradations in the anatomy. This can be seen in the bottom third of Figure 4.6), and will be discussed in more detail below.

4.3 Experimental Results

To test our algorithm, we ran two sets of experiments, the first on synthetic images for validation (controlled set-up), and the second on real brain images.

4.3.1 Controlled Experiments with BrainWeb Data

We obtained synthetic brain images from the BrainWeb project [Collins *et al.*, 1998; Bra,] such as the ones shown in Figure 4.4. The top image is a clean brain phantom with no bias, and the bottom image is a brain phantom corrupted by some bias field. These images

Such artificial reductions in entropy can be avoided by normalizing each image distribution to unit variance between iterations of computing its entropy, as is done in this work.

can be considered *idealized* MR images in the sense that the brightness values for each tissue are constant (up to a small amount of manually added isotropic noise). That is, they contain no bias fields (the left image in Figure 4.4). The initial goal was to ensure that our algorithm could remove synthetically added bias fields, in which the bias field coefficients were known. Using N copies of a single *latent* image, we added known but different bias fields to each one.

In our experiments, for as few as $N = 5$ images, we could reliably recover the known bias field coefficients, up to a fixed offset for each image, to within 1% of the power of the original bias coefficients. We show the results on BrainWeb synthetic images in Figure 4.5. As expected, when the bias removal is done, the images look more like each other, since the latent image among all the images is the same in this experiment.

4.3.2 Experiments with Real Baby Brain Data

More interesting are the results on real images, in which the latent images come from different patients. We obtained 21 pre-registered [5] infant brain images (top of Figure 4.6) from Brigham and Women's Hospital in Boston, Massachusetts. Large bias fields can be seen in many of the images. Probably the most striking is a *ramp-like* bias field in the sixth image of the second row. (The top of the brain is too bright, while the bottom is too dark.) Because the brain's white matter is not fully developed in these infant scans, it is difficult to categorize tissues into a fixed number of classes as is typically done for adult brain images; hence, these images are not amenable to methods based on specific tissue models developed for adults (e.g. [Wells *et al.*, 1996]).

The middle third of Figure 4.6 shows the results of our algorithm on the infant brain images. (These results must be viewed in color on a good monitor to fully appreciate the results.) While a trained technician can see small imperfections in these images, the results

[5] It is interesting to note that registration is not strictly necessary for this algorithm to work. The proposed MAP method works under very broad conditions, the main condition being that the bias fields do not span the same space as parts of the actual medical images. It is true, however, that as the latent images become less registered or differ in other ways, that a much larger number of images is needed to get good estimates of the pixel-stack distributions.

are remarkably good. All major bias artifacts have been removed.

It is interesting to compare these results to a method that reduces the entropy of each image individually, without using constraints between images. Using the results of our algorithm as a starting point, we continued to reduce the entropy of the pixels *within* each image (using a method akin to Viola's method [Viola, 1995]), rather than across images. These results are shown in the bottom third of Figure 4.6. Carefully comparing the central brain regions in the middle section of the figure and the bottom section of the figure, one can see that the butterfly shaped region in the middle of the brain, which represents developing white matter, has been suppressed in the lower images. This is most likely because the entropy of the pixels *within a particular image* can be reduced by increasing the bias field *correction* in the central part of the image. In other words, the algorithm strives to make the image more uniform by removing the bright part in the middle of the image. However, our algorithm, which compares pixels across images, does not suppress these real structures, since they occur across images. Hence coupling across images can produce superior results.

4.4 Summary and Conclusions

The idea of minimizing pixelwise entropies to remove nuisance variables from a set of images is not new. In particular, Learned-Miller et al. [Miller *et al.*, 2000; Miller, 2002] presented an approach they call *congealing* in which the sum of pixelwise entropies is minimized by separate affine transforms applied to each image. Our method can thus be considered an extension of the congealing process to non-spatial transformations. Combining such approaches to do registration and bias removal simultaneously, or registration and lighting rectification of faces, for example, is an obvious direction for future work.

This work uses information unused in other methods, i.e. *information across images*. This suggests an iterative scheme in which both types of information, both within and across images, are used. Local models could be based on weighted neighborhoods of pixels, *pixel cylinders*, rather than single pixel-stacks, in sparse data scenarios. For *easy* bias correction

problems, such an approach may be overkill, but for difficult problems in bias correction, where the bias field is difficult to separate from the underlying tissue, as discussed in [Fan, 2003], such an approach could produce critical extra leverage.

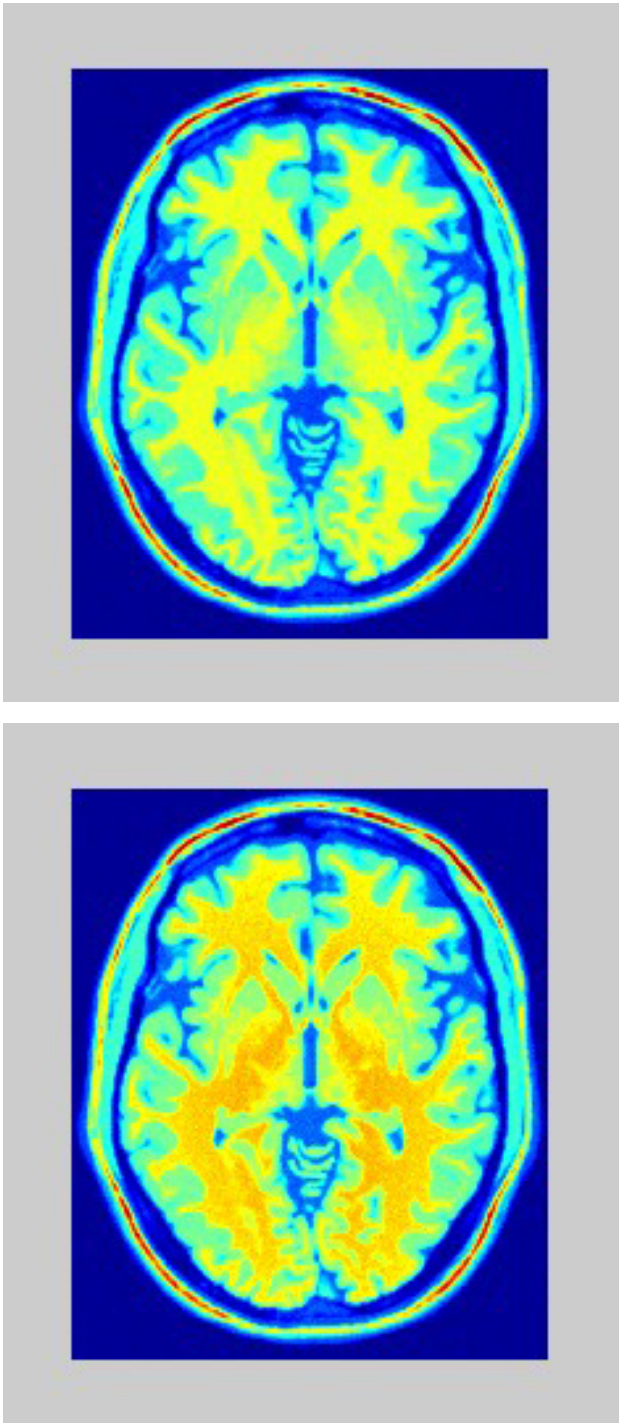

Figure 4.4: BrainWeb Sample Images: Each image shows a different RF bias field superimposed on the latent phantom.

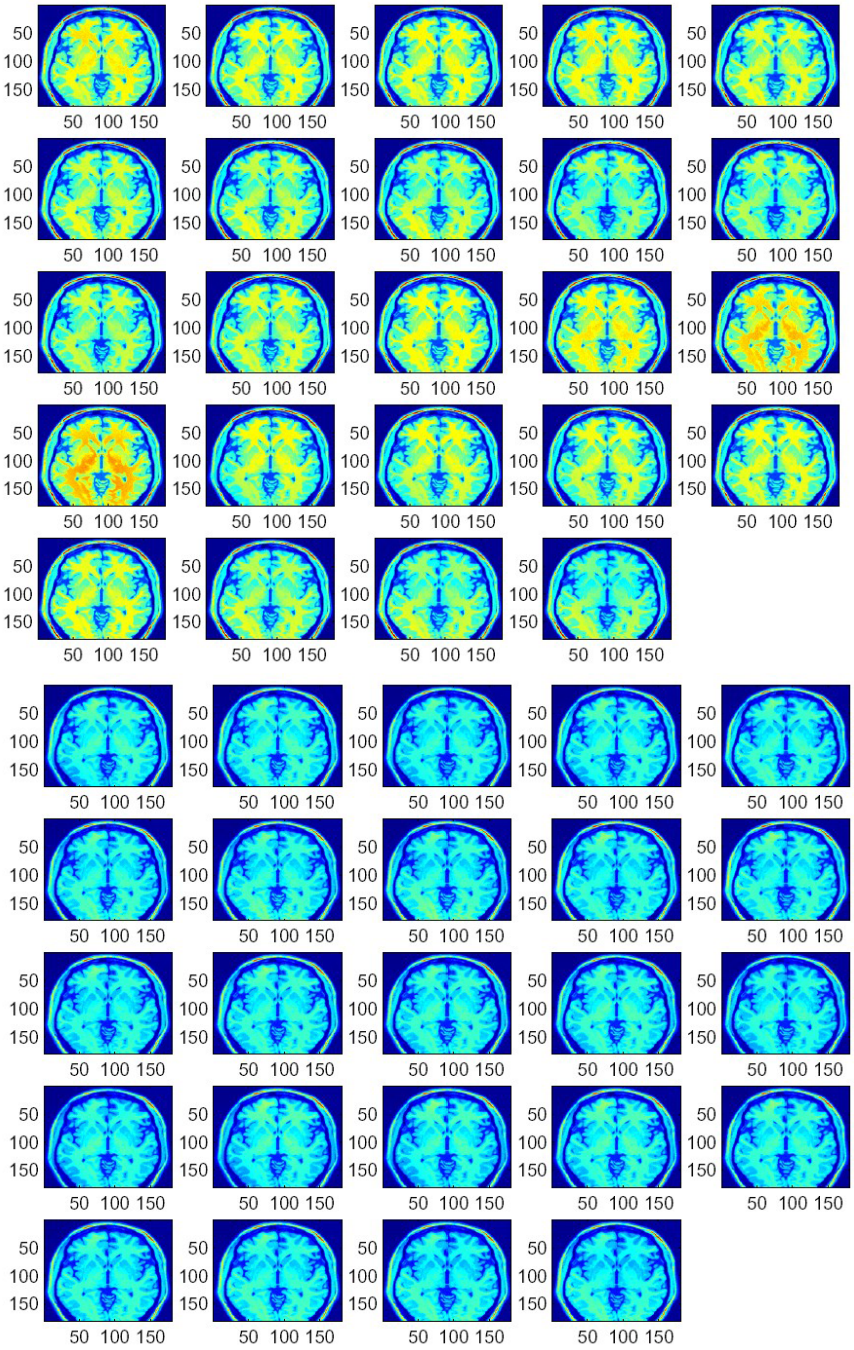

Figure 4.5: Experimental results for BrainWeb Images with different but known bias fields. **Top:** Brainweb images before bias removal. **Bottom:** Brainweb images after bias removal. **NOTE:** This image must be viewed in color (preferably on a bright display) for full effect.

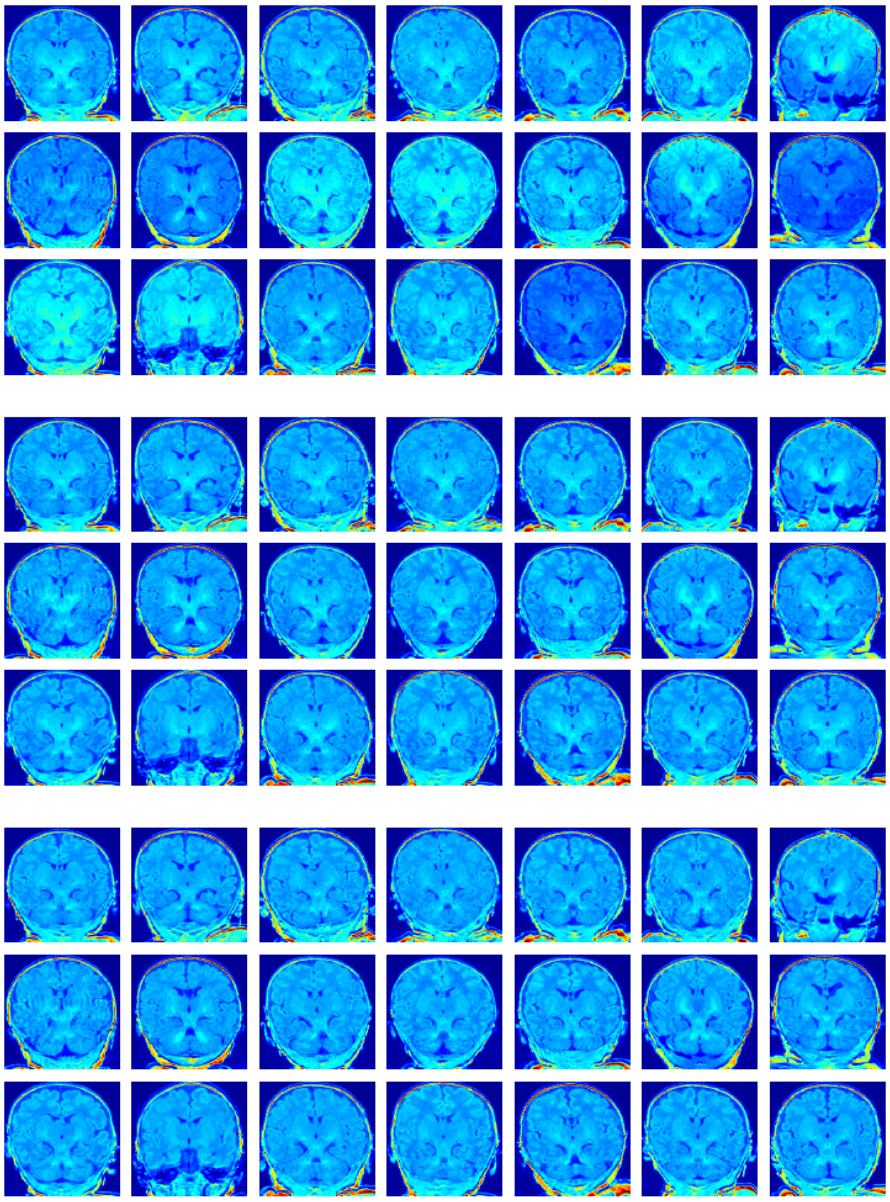

Figure 4.6: Results for Infant Brain MR image set. NOTE: This image must be viewed in color (preferably on a bright display) for full effect. **Top.** Original infant brain images. **Middle.** The same images after bias removal with our algorithm. Note that developing white matter (butterfly-like structures in middle brain) is well-preserved. **Bottom.** Bias removal using a single image based algorithm. Notice that white matter structures are repressed.

Chapter 5

Characterization of Event Related Neuronal Activity

5.1 Introduction

The human brain is an amazingly complex structure that plays a central role in our lives. The multitude of neurons in the outer layer of the human brain function as the active components in a vast signal processing network. Understanding the communication links in this network, and determining the functional mapping of brain have been the holy grails of neuroscience (Figure 5.1).

While evoked potentials reflect the processing of the physical stimulus, event-related

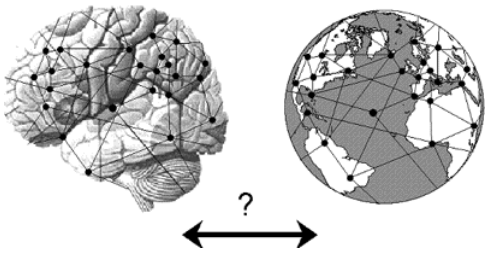

Figure 5.1: Holy grails in neuro science: functional mapping of the brain and understanding the communication links in brain's network.

potentials are caused by the "higher" processes, that might involve memory, expectation, attention, or changes in the mental state, among others.

Analysis of electric potentials or magnetic fields produced by the brain in response to sensory stimulation or in association with its cognitive and/or motor operations is critical in domains such as neurophysiology. These electric or magnetic fields are generated from transmembrane current flow produced by multiple ensembles of synchronously firing neurons. The underlying neural ensembles, also called generators or sources, are often dynamically coupled in unknown ways that are of interest to the experimentalists.

An event-related potential (ERP) is any stereotyped electrophysiological response to an internal or external stimulus. In other words, it is any measured brain response that is directly the result of a thought or perception [Handy, 2004]. In actual recording situations, it is difficult to quantify an ERP after the presentation of a single stimulus. Rather the ERPs are determined after many dozens or hundreds of individual presentations are averaged together. This averaging technique aims to filter out noise in the data, allowing only the voltage response to the stimulus to stand out clearly. Typically, such a simple averaging procedure does not perform very well. As a result of the property of linear superposition of electric currents and magnetic fields, both invasive and noninvasive electroencephalographic (EEG) recordings and magnetoencephallographic (MEG) recordings reflect linear mixtures of the activity from these sources in addition to ongoing background activity and sensor noise. Thus even in single-trial recordings, the individual responses of each of the sources are mixed within the recorded signal, making it difficult to identify them and to study their dynamical interactions. Furthermore, it is standard practice to enhance the signal-to-noise ratio by averaging event-related potentials (ERPs) or fields (ERFs) over experimental trials. However, implicit in this construction is the assumption that the evoked waveform is constant over trials and that any variability represents noise. The phase-locked signal may have trial-to-trial variability in amplitude and latency and may in fact be the superposition of multiple components with differential variability in their single trial amplitude scaling factors and latency shifts. Many ignore this fact and resort to extracting the event-related

signal as an average across the ensemble of trials, denoted by *Averaged Event Related Potential* (AERP). Using AERP makes it impossible to assess trial-dependent effects in the data. In this chapter, we describe an information theoretic formulation for obtaining a Bayesian estimate of the canonical response at a given vertex location in cerebral cortex based on the neuronal response signals recorded across a set of multiple observations (all from a single trial).

We propose an unsupervised nonparametric learning algorithm (based on joint pattern alignment framework) to extract the phase-locked component as a MAP estimate given the set of observed signals. The central goal of this chapter is to discuss how JPA can be applied in the context of single trial multicomponent estimation of event related potentials (ERP) in neuroscience domain [Ahammad *et al.*, 2006a; Ahammad *et al.*, 2006b; Vasudevan *et al.*, 2008].

5.1.1 Measuring Neural Signals

One of the ways of measuring an ERP is via electroencephalography (EEG), a procedure that measures electrical activity of the brain through the skull and scalp. One of the most robust features of the ERP response is a response to unpredictable stimuli. This response-known as the P300 (or simply "P3")-manifests as a positive deflection in voltage approximately 300 milliseconds after the stimulus is presented.

When decisions are made or information is processed in cortical network, small currents flow in the network and produce a weak magnetic field that can be non-invasively measured through external devices called SQUID (superconducting quantum interference device) magnetometers. These SQUIDs are placed outside the human skull and this form of recording neuronal activity signals is known as magnetoencephallography (MEG) [Hamalainen *et al.*, 1993]. The time resolution of MEG is better than 1 millisecond, and the spatial discrimination (under favorable circumstances) is around 2-3 millimeters for sources in cerebral cortex. State of the art MEG systems usually have around 300 SQUIDs situated around the skull for measuring the MEG signals. Since the magnetic signals induced by these currents in the

brain are very weak, shielding from external magnetic signals is necessary. The net currents can be thought of as current dipoles (known as *Equivalent Current Dipoles*: ECD) which are currents defined to have an associated position, orientation, and magnitude, but no spatial extent. According to the right-hand rule, a current dipole gives rise to a magnetic field that flows around the axis of its vector component. The magnetic field arising from the net current dipole of a single neuron is too weak to be directly detected. However the combined fields from a region of about 50,000 active neurons can give rise to a net magnetic field that is measurable. Since current dipoles must have similar orientations to generate magnetic fields that reinforce each other, it is often the layer of pyramidal cells in the cortex, which are generally perpendicular to its surface, that give rise to measurable magnetic fields.

5.1.2 Challenges in Functional Mapping of Human Brain

Given the measured ERPs, there are some interesting challenges in the way to obtaining the functional mapping of human brain. One such challenge is: given a certain stimulus, we would like to understand how the canonical response is at any given vertex in the cortical network. Canonical response is defined as the underlying response that is common to all signals in the set of observations at that given vertex location for a given stimulus (in other words, a template for a given stimulus). The measurements are generally modeled as a dynamical interaction between signals that are relatively phase-locked to a specific event onset and signals that are not phase-locked to the event, such as measurement noise or ongoing brain activity. Once inferred, this canonical phase-locked response signal can be used as a template to determine which vertices have similar kinds of neuronal responses, and potentially help establish the communication links between various vertices in the cortical network. Our proposed algorithm can be used to estimate this canonical response from multiple observations.

5.1.3 Related Work

In recent years, there have been great developments in blind source separation and Independent Component Analysis (ICA) techniques, such as Infomax ICA [Bell and Sejnowski, 1995], FastICA [Hyvrinen and Oja, 1997], and Second-Order Blind Identification (SOBI) [Belouchrani *et al.*, 1993]. These algorithms have been useful in identifying sources in EEG and MEG signals using both ensemble-averaged data [Makeig *et al.*, 1997; Vigario *et al.*, 2000] and single trials [Cao *et al.*, 2000; Jung *et al.*, 1999; Makeig *et al.*, 2002; Tang *et al.*, 2002]. Along with the respective strengths, each technique has its limitations, and often these limitations can be addressed. ICA, for example, allows reliable source (component) separation with minimal *a priori* assumptions and constraints. Its limitation is that although trial-to-trial variability can assist in separation, these effects are not explicitly considered and quantified, and these are substantial opportunities missed. Also, like many other techniques, ICA solves for maximal independence of components, despite the fact that components are often dynamically coupled. Thus although ICA may be reasonable for source separation per se, it is not explicitly designed to quantify the dynamical interactions between the neuronal ensembles that generate the components.

Jaskowski et al. present a solution in which the phase-locked signal has the same shape but may vary in its amplitude and latency from trial to trial [Jaskowski and Verleger, 1999]. The utility of the solution to estimate the P3 component in single trials was investigated both by extensive pseudo-real simulations and in an application to real data. Their simulations showed some advantage of the method over two other commonly used methods (Woody's method and peak-picking) in event-related potentials research. Quirogaa and Garcia present a denoising implementation based on the Wavelet Transform to obtain the ERPs at the single-trial level [Quiroga and Garcia, 2003]. Knuth et al. and Truccolo et al. describe a model of the sensory-evoked neural response that is more realistic than previous models in that it explicitly models trial-to-trial amplitude as well as latency variability in single-trial responses [Knuth *et al.*, 2006; Truccolo *et al.*, 2003]. Using this model, they derive, what they call, Differentially Variable Component Analysis (dVCA) algorithm, and

demonstrate how different variability patterns in neural ensemble activity can be used to separate and identify their component signals. Using simulations, they evaluated not only the ability of dVCA to characterize single-trial responses, but also its robustness to noise. While dVCA is more realistic than AERP, it suffers from two shortcomings. First, it requires an *a priori* knowledge of either the number of components within the phase-locked signal or of the ratio between the phase-locked and on-going process signals. Second, it assumes that the trial-to-trial variability in amplitude and latency is Gaussian. Our proposed approach based on JPA for estimating ERP improves upon the state of the art by making fewer assumptions on the composition of phase-locked ERP signal, and eliminating the need to have *a priori*knowledge of the number of components in phase-locked ERP signal.

5.1.4 Problem Statement

We define the problem of Event Related Potential estimation as: at a chosen vertex location in cortical network, for a given stimulus, extract the phase-locked ERP signal as a MAP estimate. No parametric assumptions are made about the ERP signal composition; but we do impose certain restrictions on the noise that could corrupt the canonical signal. Intuitively, this is similar to looking at a set of example pictures of apples or faces, and figuring out what a canonical apple or face looks like by jointly undoing the distortions.

5.2 JPA for Nonparametric Estimation of ERPs

5.2.1 Notation

Let us consider the problem of estimating the ERP signal given an ensemble of neural signal observations (or recordings) at a given location in brain. We want to derive the objective function directly from the relevant assumptions (on the p.d.f.'s of the transformation vectors) instead of regularizing in an ad-hoc manner. The following derivation will make a specific (Gaussian) assumption on the p.d.f.'s of transformation components, but this can

be changed appropriately depending on the user-preference or prior knowledge.

Let us denote the input set of neural signal observations (or recordings) as $\Phi_S \doteq \{S^i\}_{i=1}^N$ where N is the cardinality of the set. Let S_l be the latent underlying Event Related Potential (ERP) that is phase-locked to the stimulus which gets corrupted by variations in latency and scaling parameters. $S_l(\cdot)$ and $S^i(\cdot)$ can be represented as maps from the domain $\Omega \subset \mathbb{R}$ to the set $\mathbb{R}$:

$$S_l, S^i : \quad \Omega \mapsto \mathbb{R}. \tag{5.1}$$

Let $\boldsymbol{t} \in \Omega$ denote the time-point (in homogeneous coordinates) such that $\boldsymbol{t} = [t, 1]^T$. Let g^i induce the one-to-one and invertible map from $S^i(\boldsymbol{t})$ to $S_l(\boldsymbol{t})$ such that $g^i : \Omega \mapsto \Omega$. Thus, for any given time-point $\boldsymbol{t}$,

$$S_l(\boldsymbol{t}) \quad = \quad S^i(g^i(\boldsymbol{t})). \tag{5.2}$$

Let us denote the set of transformations associated with the set of observed signals as $\Phi_g \doteq \{g^i\}_{i=1}^N$ and the set of transformed neural signal observations as $\Phi_{S_g} \doteq \{S_g^i\}_{i=1}^N$ where N is the cardinality of the set and S_g^i are the transformed signals. Let us parameterize each transformation g^i using scaling (s) and latency (l) as the transform parameters. Let us assume that the transform parameters are *i.i.d.* random variables. This can be written as:

$$g^i(\boldsymbol{t}) \quad = \quad F(\boldsymbol{t}; s^i, l^i) \tag{5.3}$$

$$g^i(\boldsymbol{t}) \quad = \quad F(\boldsymbol{t}; \{v_j^i\}_{j=1}^K) \tag{5.4}$$

$$\{v_j^i\}_{j=1}^K \quad = \quad [s^i, l^i] \tag{5.5}$$

where $1 \leq i \leq N$, $1 \leq j \leq K$, (K is the number of parameters chosen - which is 2 in our

formulation), and $v \in \mathbb{Z}_+^N \times \mathbb{R}^K$. Writing g out explicitly, we get:

$$g = \begin{bmatrix} e^s & l \\ 0 & 1 \end{bmatrix} \tag{5.6}$$

5.2.2 Problem Formulation

Let us assume that the latent model of the class (S_l) and the set of transformations Φ_g are independent. Let $v^i = \{v_j^i\}_{j=1}^K$ and $\Phi_v = \{v^i\}_{i=1}^N$. Let $P(S^i|v_1^i, ..., v_K^i; S_l)$ be some likelihood function such that,

$$P(\Phi_S|\Phi_v; S_l) = \prod_{i=1}^{N} P(S^i|v^i; S_l). \tag{5.7}$$

Let $\Theta \doteq \{S_l, \Phi_v\}$. We would like to infer Θ given the set of neural signal observations Φ_S. Formulating our goal as a *Maximum a posteriori (MAP) estimation* problem, we want to estimate Θ by $\hat{\Theta}$ such that

$$\hat{\Theta} = \arg\max_{S_l, \Phi_v} P(\Phi_v|\Phi_S; S_l). \tag{5.8}$$

Using Bayes' rule and ignoring the constant denominator,

$$\hat{\Theta} = \arg\max_{S_l, \Phi_v} P(\Phi_S|\Phi_v; S_l) P(\Phi_v). \tag{5.9}$$

Since we assume that the transformation parameters v^i and given neural signal observations S^i are independent,

$$\hat{\Theta} = \arg\max_{S_l, \Phi_v} \prod_{i=1}^{N} P(S^i|v^i; S_l) P(v^i). \tag{5.10}$$

Using Equation (5.4), and noting that g^i is a bijective map such that $g^i : \Omega \mapsto \Omega$, we can write:

$$P(S^i|v^i; S_l) = P(S^i \circ g^i|v^i; S_l). \tag{5.11}$$

Let us make the assumption that the value of $S^i(g^i(\boldsymbol{t}))$ at time-point $\boldsymbol{t}$ is independent of the other time-point locations. In other words, we assume that the probability distributions of values at each time-point location are *i.i.d.* Thus,

$$\begin{aligned} P(S^i \circ g^i | v^i; S_l) &= \prod_{\boldsymbol{t}\in\Omega} P(S^i(g^i(\boldsymbol{t}))|v^i; S_l) \\ &= \prod_{\boldsymbol{t}\in\Omega} P(S^i(g^i(\boldsymbol{t}))|v^i; S_l(\boldsymbol{t})). \end{aligned} \tag{5.12}$$

Thus,

$$\begin{aligned} \prod_{i=1}^{N} P(S^i|v^i; S_l) &= \prod_{i=1}^{N}\prod_{\boldsymbol{t}\in\Omega} P(S^i(g^i(\boldsymbol{t}))|v^i; S_l(\boldsymbol{t})) \\ &= \prod_{\boldsymbol{t}\in\Omega}\prod_{i=1}^{N} P(S^i(g^i(\boldsymbol{t}))|v^i; S_l(\boldsymbol{t})). \end{aligned} \tag{5.13}$$

Since we assumed that the transformation parameters v^i_j are independent,

$$P(v^i) = \prod_{j=1}^{K} P(v^i_j). \tag{5.14}$$

Hence,

$$\hat{\Theta} = \arg\,\max_{S_l,\Phi_v} \left\{ \left\{ \prod_{\boldsymbol{t}\in\Omega}\prod_{i=1}^{N} P(S^i(g^i(\boldsymbol{t}))|v^i; S_l(\boldsymbol{t})) \right\} \left\{ \prod_{i=1}^{N}\prod_{j=1}^{K} P(v^i_j) \right\} \right\}. \tag{5.15}$$

Now, let us assume (in this example) that $P(v^i_j)$ has a Gaussian distribution, such that

$$P(v^i_j) = N(v^i_j; \mu_j, \sigma^2_j). \tag{5.16}$$

Taking logarithm,

$$\begin{aligned}\hat{\Theta} &= \arg\max_{S_l,\Phi_v} \sum_{\boldsymbol{t}\in\Omega}\sum_{i=1}^{N} \log\{P(S^i(g^i(\boldsymbol{t}))|v^i; S_l(\boldsymbol{t}))\} \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{K} \log\{\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\{-\frac{(v_j^i-\mu_j)^2}{2\sigma_j^2}\}\}. \end{aligned} \tag{5.17}$$

Let us define $\alpha(\boldsymbol{t})$ to be the time-point stack in Φ_S at location $\boldsymbol{t}$ and $\alpha_g(\boldsymbol{t})$ as the time-point stack in Φ_{S_g} at location $\boldsymbol{t}$. Since Φ_S is a set of neural signal observations, $\alpha(\boldsymbol{t}) \in \mathbb{R}^N$ and $\alpha_g(\boldsymbol{t}) \in \mathbb{R}^N$. Writing this out explicitly:

$$\alpha_g(\boldsymbol{t}) = [S^1(g^1(\boldsymbol{t})), S^2(g^2(\boldsymbol{t})), ..., S^i(g^i(\boldsymbol{t})), ..., S^N(g^N(\boldsymbol{t}))]^T. \tag{5.18}$$

Also, define $H(\alpha_g(\boldsymbol{t}))$ as the empirical entropy of the time-point stack $\alpha_g(\boldsymbol{t})$. Noting that entropy is the expectation of negative log-likelihood, and expanding the logarithm in the second term (while ignoring the constant),

$$\hat{\Theta} = \arg\min_{S_l,\Phi_v} \left\{ \sum_{\boldsymbol{t}\in\Omega} H(\alpha_g(\boldsymbol{t})) - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} \log\{\sigma_j\} + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} \frac{(v_j^i-\mu_j)^2}{2\sigma_j^2} \right\}. \tag{5.19}$$

If we assume that $\sigma_j = \sqrt{K}, \forall j = 1, \ldots, K$ and ignore the constant term, then

$$\hat{\Theta} = \arg\min_{S_l,\Phi_v} \left\{ \sum_{\boldsymbol{t}\in\Omega} H(\alpha_g(\boldsymbol{t})) + \sum_{i=1}^{N} \frac{1}{2NK} ||v_j^i - \mu_j||_2^2 \right\} \tag{5.20}$$

where $||\cdot||_2^2$ represents L_2-norm. Since we model transformation parameters as the random variables causing $S^i(\boldsymbol{t})$ to vary from $S_l(\boldsymbol{t})$, we can see that these two will be the same when the randomness due to v^i is removed.

This Maximum a posteriori (MAP) estimation can be formulated as solving an opti-

mization problem. The optimization objective function $\Psi \doteq \Psi(\Phi_v)$ is defined as

$$\Psi \doteq \{\sum_{\boldsymbol{t}\in\Omega} H(\alpha_g(\boldsymbol{t})) \; + \; \sum_{i=1}^{N} \frac{1}{2K} ||v^i - \bar{v}^i||_2^2\} \tag{5.21}$$

where $v \in \mathbb{Z}_+^N \times \mathbb{R}^K$ are the vectors of transformation parameters (Equation (5.5)).

The JPA algorithm for the estimation of ERP signal (that is phase-locked to the stimulus) proceeds as follows:

1. Maintain a transform parameter vector v^i (Equation (5.5)) for each shape image S^i.
2. Initialize all of the transformation matrices g^i to the identity matrix. Each parameter vector will specify a transformation matrix $g^i = F(v^i)$ according to Equation (5.6). Initialize all v^i to zero vectors. This means $\mu_j = 0$ for $1 \leq i \leq N$, $1 \leq j \leq K$.
3. Choose an appropriate penalty term in Ψ (Equation (5.21)) based on the probability assumptions made on transformation parameters (Equation (5.16))
4. Compute Ψ for the current set of images from Equation (5.21).
5. Repeat until convergence:

   For $i = 1, ..., N$,

   (a) Calculate the numerical gradient $\bigtriangledown_{v^i}\Psi$ of Equation (2.21) with respect to the transformation parameters v_j^i's for the current image ($1 \leq j \leq K$).

   (b) Update v^i as: $v^i = v^i - \gamma \bigtriangledown_{v^i} \Psi$(where the scaling factor $\gamma \in \mathbb{R}$).

   (c) Update γ (according to some reasonable update rule such as the Armijo rule [Boyd and Vandenberghe, 2004]).

Since $\Psi(\cdot)$ is a differentiable function and the level sets

$$\mathcal{A}(\{u^i\}_{i=1}^N) = \left\{\{v^i\}_{i=1}^N \in \mathbb{R}^{K\times N} \mid \Psi(\{v^i\}_{i=1}^N) \leq \Psi(\{u^i\}_{i=1}^N)\right\} \tag{5.22}$$

are bounded for all $\{u^i\}_{i=1}^N \in \mathbb{R}^{K\times N}$, then the JPA routine will at least reach an accumulation point such that $\nabla_{v^i}\Psi = 0$ for all $i = 1, \ldots, N$ [Polak, 1997], even though the optimization routine will generally converge to a local minimum. Note that at a local minimum the set of neural signal observations $\Phi_S = \{S^i\}_{i=1}^N$ are reasonably denoised (if not perfectly) and the set of transformations $\{g^i\}_{i=1}^N$ is properly described by the parameters $\{v^i\}_{i=1}^N$. S_l is estimated by choosing the medoid of the set of signals, using an appropriate measure (such as the magnitude of transformation from one signal to another based on the values of v^i. Note that the introduction of a penalty (regularization) function is critical in achieving the convergence of the optimization routine since this term diverges as the norm of v^i goes to infinity, thus making the level sets of $\Psi(\cdot)$ be bounded.

5.3 Experimental Results

We evaluated the JPA algorithm for estimating the ERP signal using synthetic data to demonstrate the applicability of our approach under various assumptions on the generator signal and the noises that corrupt the generator signal. Similar evaluation methodology was used, by Knuth et al., to evaluate the performance of dVCA algorithm [Knuth *et al.*, 2006]. In our experiments, we use dVCA as the baseline method for comparison. Our controlled experiments allow us to evaluate the performance of our algorithm in the presence of noise as well as variability in the assumptions made on generative model.

Several examples of sample observations are illustrated in Figure 5.2. The underlying signal generator is drawn in red on all plots. The generator signal for the three sub-plots on the left column contains only two components, while the generator signal for the three sub-plots on the right column contains five components. The top row models the variations in latencies and scaling using a Gaussian probability distribution. The middle row models the variations in latencies and scaling using a Laplacian probability distribution. The bottom row models the variations in latencies and scaling using a Uniform distribution. Note that even though the right column is generated by a generator signal with 5 components, it looks

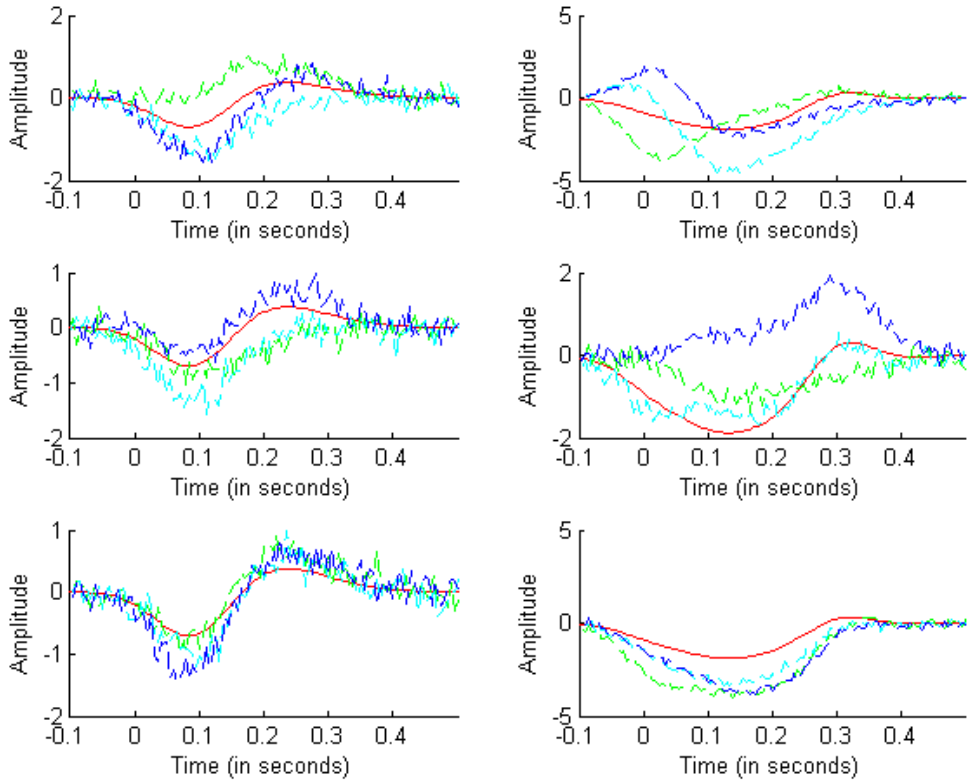

Figure 5.2: Several examples of sample observations. The underlying signal generator is drawn in red on all plots. The generator signal for the three sub-plots on the left column contains only two components, while the generator signal for the three sub-plots on the right column contains five components. The top row models the variations in latencies and scaling using a Gaussian probability distribution. The middle row models the variations in latencies and scaling using a Laplacian probability distribution. The bottom row models the variations in latencies and scaling using a Uniform distribution.

as if there are only two components. Since dVCA requires the user to choose the number of components, it is easy to choose the wrong number of components. JPA does not make such assumptions, so it is robust the number of components in the generator signal.

We show the results of the JPA algorithm when compared to dVCA in six instances in the accompanying Figures 5.3. The generator signal is created as a linear combination of a given number of Gaussian components. Note that this generator signal can be any signal for our purposes, since JPA does not make any assumptions on the generator signal. On the other hand, making the correct guess about the number of components in the generator signal is critical to the success of dVCA algorithm. The underlying signal generator is drawn in red on all plots. The generator signal for the three sub-plots on the left column contains only two components, while the generator signal for the 3 sub-plots on the right column contains 5 components. The top row models the variations in latencies and scaling using a Gaussian probability distribution. The middle row models the variations in latencies and scaling using a Laplacian probability distribution. The bottom row models the variations in latencies and scaling using a Uniform distribution.

In order to quantitatively evaluate the performance of our algorithm, we followed the paradigm used by [Knuth *et al.*, 2006], and used the root-mean-squared error (RMSE) measure[1]. First we wanted to test how JPA performed under several noise assumptions and varying regularization choices. Following the arguments we made about the importance of choosing appropriate regularization in JPA (Section 2.2), we expected to see that the best results for joint alignment occur when the regularization choice matches that of the assumptions made on the p.d.f. of the transformation parameters.

From Table 5.1 and Table 5.2, one can notice that JPA based multi-component ERP estimation performs best when L1-regularization is used in conjunction with Laplacian p.d.f. assumptions on the transformation parameters, while the L2-regularization works best in conjunction with Gaussian p.d.f. assumptions on the transformation parameters. We tested this regularizer assumptions under two different SNR regimes (SNR=0 dB, and

[1]Since we have access to the generator signal, RMSE is straight-forward to compute.

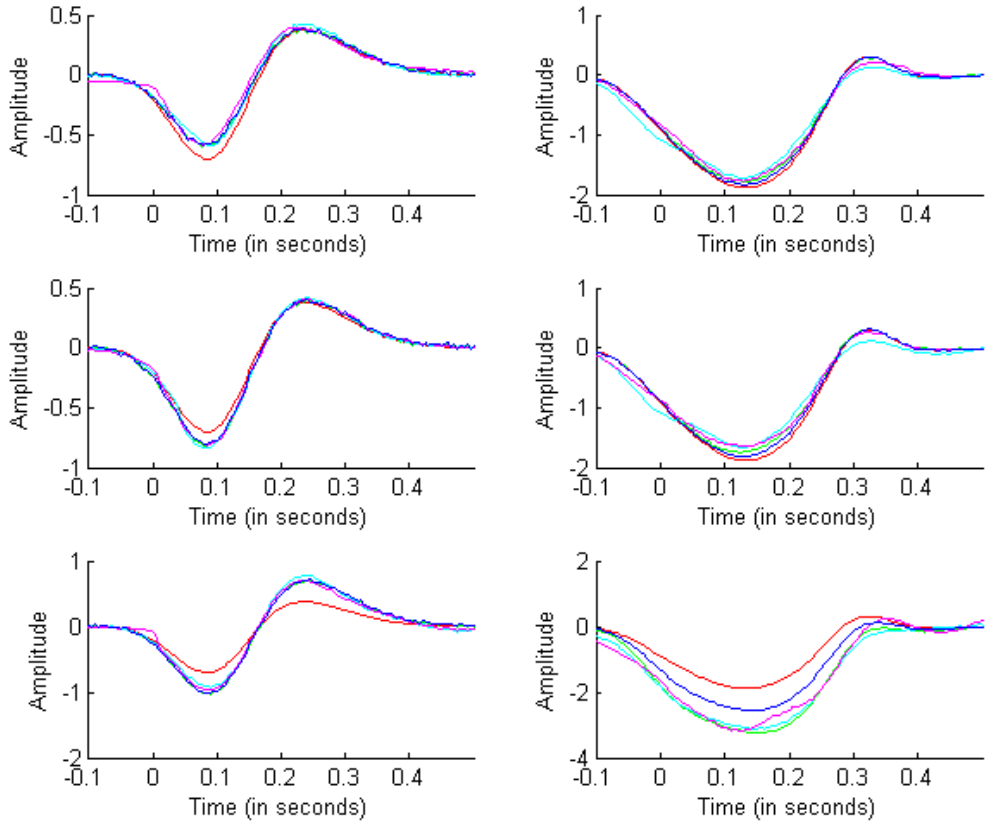

Figure 5.3: Results for estimating the phase-locked ERP signal. The generator ERP signal for the three sub-plots on the left column contains only two components, while the generator ERP signal for the three sub-plots on the right column contains five components. The top row models the variations in latencies and scaling using a Gaussian probability distribution. The middle row models the variations in latencies and scaling using a Laplacian probability distribution. The bottom row models the variations in latencies and scaling using a Uniform distribution. Note that even though the right column is generated by a generator signal with 5 components, it looks as if there are only two components. Colorcode is given as: Red for Generator ERP signal, Green for AERP, Light blue for dVCA=2, Pink for dVCA=5 and Blue for JPA.

Table 5.1: RMSE error rates at SNR= 20 dB: Varying reglurization assumptions

SNR=20dB	Gaussian p.d.f.	Laplacian p.d.f.
L2-regularizer	0.0277	0.0216
L1-regularizer	0.0416	0.0067

Table 5.2: RMSE error rates at SNR= 0 dB: Varying reglurization assumptions

SNR=0dB	Gaussian p.d.f.	Laplacian p.d.f.
L2-regularizer	0.1729	0.1719
L1-regularizer	0.1922	0.1564

SNR=20 dB), and we used the RMSE to measure the error of alignment (since we know the underlying generator signal in these experiments). Please note that SNR stands for signal-to-noise ratio.

From Table 5.3, it is clear that JPA based multi-component ERP estimation outperforms the other approaches across a variety of conditions, while making no assumptions on the structure of the underlying generating signal.

Table 5.3: RMSE error rates for ERP signal estimation

Generator Type	JPA	dVCA2	dVCA5	AERP
2comp-Gaussian	0.0277	0.0376	0.058	0.0268
2comp-Laplacian	0.0216	0.0269	0.0226	0.0243
2comp-Uniform	0.1843	0.3248	0.2754	0.2007
5comp-Gaussian	0.0009	0.0784	0.014	0.0039
5comp-Laplacian	0.002	0.0318	0.0161	0.0083
5comp-Uniform	0.1971	0.7634	0.6655	0.8004

5.4 Summary and Conclusions

We have proposed a nonparametric approach (based on JPA) for estimating an ERP signal given an ensemble of observations. We have provided a principled procedure for using the prior knowledge on the distribution of noise parameters (variations in latency and scaling) to arrive at the correct form of regularized objective function in the JPA's optimization routine. We have compared our algorithm, both qualitatively(Figure 5.3) and quantitatively(Table 5.3), with state of art approaches (dVCA and AERP) using various settings for the number of components in the generator signal, as well as the distributions of noise parameters.

Our results show that improved estimates of the phase-locked ERP component can be achieved in a completely unsupervised fashion, without making any parametric assumptions about the underlying signal. Since the structure of ERP signals recorded from brain are yet to be well understood, we believe that such a nonparametric approach to estimating ERP signals would be the least-biased. Note that the assumptions on the distributions of noise parameters do impose a certain structure on the regularization aspect of JPA, and these could be considered as a form of *a priori* knowledge available to the experimentalists, based on observing many experimental sessions.

Moving forward, we plan to apply our approach to characterize the responses of subjects in real neuroscience experiments, at various locations in cortical networks. Assuming that an appropriate metric is available to compare the ERP signals, this could potentially enable us to establish the communication links across the cortical networks. Ahammad et al. showed some preliminary demonstrations that a learned set of phase-locked ERP signals (using JPA) can used in a nearest-neighbor classification scheme to improve classification rates in a magnetoencephallography (MEG) experiment [Ahammad *et al.*, 2006b]. An interesting direction for future would be to investigate JPA based generative modeling based classification approach for signal-to-state conversions (for applications in BCI).

Chapter 6

Unsupervised Discovery of Action Hierarchies in Large Video Collections

6.1 Introduction

In this chapter, we focus on the task of learning a meaningful data-driven hierarchical organizational structure for large collections of videos containing activities. The videos are unlabeled, and the collections may contain examples from many different classes (categories). The representation we are attempting to learn (in an unsupervised manner) in this scenario is the data-driven organizational hierarchy that facilitates the most efficient retrieval.

Given the growing popularity of online video databases (such as YouTube^{TM}, Google Video^{TM}, Yahoo Video^{TM} and MSN Video^{TM}) and the ease of recording and storing videos, access to video data is bound to increase. Considering the low storage costs and ease of acquiring video data, the notion of *personalized* video databases is not that far from reality. Even in special scenarios such as surveillance or environmental habitat monitoring where networks of cameras are deployed, it is typical to record large amounts of video data. While it is becoming more common for cities to record video data in public places (for

example, in London, Washington, D.C., New York) , the recorded data typically resides in archives without much use (either due to some privacy concerns, or operator fatigue or the lack of technical tools to sift through such large amount of video data). A lot of interesting details in these video databases are related to dynamic patterns such as actions of human beings or objects, or to static patterns such as faces. The utility of these video collections can be immensely improved by the ability to efficiently organize videos using semantically meaningful cues. In our work, we are focused on video data that contain actions or movements of human beings or objects. While we are generally interested in all spatio-temporal patterns in video, human actions are specifically interesting since they are articulated motions and offer a different set of challenges than rigid body motions.

Our goal is to discover a meaningful data-driven hierarchical organizational structure for the given large collections of videos containing activities in an unsupervised manner. We aim to organize based on the similarity or dissimilarity of human actions embedded within the video clips. The notion of action similarity induces a perceptual hierarchy on the database of videos (see Figure 6.1 for example). Recalling the example of a library user attempting to organize an unlabeled collection of books using content based similarity (as discussed in Section 1.1), we would like to build a system that takes as input a collection of video clips of human actions, and outputs an organizational structure containing the videos that respects the content based notion of similarity or dissimilarity between video clips. Such a system would be very useful in facilitating efficient navigation of the database, thus improving its utility as well as providing the users with an efficient tool to index the large database.

6.1.1 Challenges

There are some significant challenges in building such an unsupervised organization system that we discussed. Firstly, measuring similarity or dissimilarity of the actions of mobile articulated structures like humans and animals is a very difficult task. In practice, the rate and the style of actions can be different from person to person. So, the measurement

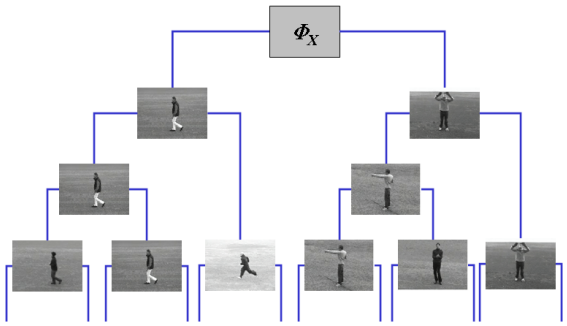

Figure 6.1: A qualitative example of an action hierarchy for the activity video collection Φ_X, with associated exemplars for the subtree under each node, shown up to 6 clusters. This was generated using our proposed approach with NCNC as the action similarity measure and Ward linkage as the neighbor-joining criterion. The 6 clusters from left to right: Jogging, Walking, Running, Boxing, Handclapping, Handwaving. See Section 6.3 for further discussion.

of action has to be robust to changes in the rate of actions as well as variations in the appearance of the individuals due to clothing or lighting variations. While there is significant amount of work that exists in literature on measuring action similarity, many proposed approaches make simplifying assumptions that make the approaches unusable in practice. It is preferable to assume *no metadata* (e.g. labels), *no segmentation* and *no prior alignment* for the video collections. In other words, we would like the system to simply take the videos as they are, and work with them without much preprocessing. Since we would like handle large collections of video clips, ability to organize in a short amount of time is also a very critical factor. A key point to note here is that in practice, large collections of videos are always stored in compressed format. So re-using these pre-computed features from compressed domain representation eliminates the step of feature computation and thus saves significant amount of time in estimating action similarity. For these reasons of efficiency, we proposed a similarity measure that uses compressed domain features for organizing the video collections in our work. Our proposed approach computes the action similarity directly based on compressed domain features (such as filtered motion vectors and DCT coefficients) that are pre-computed when the videos are compressed for storage purposes (Section 6.2.1). We also show that our proposed measure is quite robust with

respect to parameter settings in video coding as well as being robust to appearance and rate variations.

A second challenge is to estimate a figure-of-merit that tells us how good our data organization is, without having access to labels. Relating this situation back to the task of organizing books in a library without access to labels, we would like to know who organized the collection in the best way if there were a few different choices of organizational structure available, depending on the user who performed the task. In our approach, we use agglomerative hierarchical clustering for the purposes of building the hierarchy - and it is well known that different choices of neighbor-joining criteria result in different estimated hierarchies (Section 6.2.3). Let us assume that each neighbor-joining criterion represents one user who performed the organization. Our task is to assign a figure-of-merit to each estimated hierarchy so that we can rank the possible choices. While estimating this goodness measure (or figure-of-merit) for organizational structures is straight-forward in the supervised scenario [1], it is not clear how the quality of the database organization can be judged in the absence of labels. Lack of labels, or prototypical examples, or metadata makes this task quite hard. One also has to remember that the number of groups may not be known *a priori*, and different end-users might want the database to be divided up into different number of groups. Assuming availability of labels (or metadata) for large video collections is impractical in practice, so this is a key practical issue. We propose a solution to this problem by computing a performance measure on the estimated hierarchy (Section 6.2.4). The main insight in our solution is to note that, for a good organizational hierarchy, *the cophenetic distances* computed during the agglomeration *should obey the input pair-wise distance relationships* estimated from pair-wise action distances [2] [Rohlf and Fisher, 1968].

[1] One can easily evaluate the goodness of the organization based on the labels of the data samples and the groups they end up in - assuming the availability of labels or metadata.

[2] In our work, the pair-wise action distances are measured using our robust compressed domain action similarity approach.

6.1.2 Problem Statement

Given a set of videos and a user-defined *space-time scale* of actions, we would like the system to: *(a) automatically* and *efficiently* organize the videos into a hierarchy based on action similarity; *(b)* estimate clusters; *(c)* compute a performance measure on the estimated hierarchy (even without access to the labels); and *(d)* select one representative exemplar for each cluster.

6.1.3 Related Work

Any practical system that records and stores digital video is likely to employ video compression such as H.263+ [Cote *et al.*, 1998] or H.264 [Wiegand *et al.*, 2003]. It has long been recognized that some of the video processing for compression can be reused in video analysis or transcoding; this has been an area of active research (see for example [Chang, 1995; Wee *et al.*, 2002]) in the last decade or so. For large databases of videos, techniques that operate directly on compressed domain features are more suitable since they offer a significant speed-up in processing time.

Typically, example based query systems operate by assigning a similarity (or dissimilarity) score to each target video based on the example video [Shechtman and Irani, 2005; Yeo *et al.*, 2006; Yeo *et al.*, 2008]. These scores are not always metrics, so one needs to find ways to convert these scores into meaningful metrics in order to perform unsupervised grouping.

There has been much prior work in human action recognition; an excellent review of such methods has been presented by Aggarwal and Cai [Aggarwal and Cai, 1997]. We are interested in approaches that work on video without relying on capturing or labeling body landmark points (see [Yilmaz and Shah, 2005; Parameswaran and Chellappa, 2005] for recent examples of the latter approach). Efros *et al.* [Efros *et al.*, 2003] require the extraction of a stabilized image sequence before using a rectified optical flow based normalized correlation measure for measuring similarity. Our motion similarity measure uses some key ideas from their work, while choosing to operate in compressed domain instead

of the pixel domain. Shechtman and Irani [Shechtman and Irani, 2005] exhaustively test motion-consistency between small space-time (ST) image intensity patches to compute a correlation measure between a query video and a test video. While their method is highly computationally intensive, they are able to detect multiple actions (similar or different) in the test video and also perform localization in both space and time. Laptev and Lindeberg [Laptev and Lindeberg, 2003] adopted a local feature based approach. Schüldt *et al.* [Schüldt *et al.*, 2004] propose an approach based on local ST features [Laptev and Lindeberg, 2003] in which Support Vector Machines (SVM) are used to classify actions in a large database of action videos that they collected. Dollar *et al.* [Dollar *et al.*, 2005] adopt a similar approach, but introduce a different spatio-temporal feature detector which they claim can find more feature points. Ke *et al.* [Ke *et al.*, 2005] also use an image intensity based approach, but apply boosting to ST volumetric features computed from image intensity. Since this approach does not output any measure of motion similarity or dissimilarity, despite its speed, it is not well suited for *unsupervised* organization of large action databases. Since these three methods ([Schüldt *et al.*, 2004; Ke *et al.*, 2005; Dollar *et al.*, 2005]) only accounts for motion implicitly through the use of image intensity, it is also not clear how appearance-invariant these methods really are.

We review the prior work in action recognition and/or video retrieval in the compressed domain, while noting that methods that simply perform classification or detection without computing a similarity or dissimilarity measure are not well-suited for building *unsupervised* organizational hierarchies across large databases. Dimitrova et al. [Dimitrova and Golshani, 1994] assume that motion vectors are coarse approximations of optical flow but unlike our approach, they estimate object trajectories explicitly using motion vectors. Chang et al. assume that objects can be segmented and tracked easily in order to compute features [Chang *et al.*, 1998]. Some approaches segment a single video into shots and organize neighboring shots into a hierarchy for browsing the video but they do not build action based hierarchies across a large collection of videos [Yeung and Liu, 1995; Ngo *et al.*, 2002]. Sahouria et. al. compute principal components of the motion vectors as

low-dimensional representations of videos to classify sports videos into different classes of sports [Sahouria and Zakhor, 1999]. While this approach might work well for scene analysis, recognizing or categorizing articulated motions (such as human motions) may be difficult using this approach. Ozer *et al.* [Ozer *et al.*, 2000] applied Principal Component Analysis (PCA) on motion vectors from *segmented* body parts for dimensionality reduction before classification. They require that the sequences must have a fixed number of frames and be *temporally aligned*. Babu *et al.* [Babu *et al.*, 2002] trained a Hidden Markov Model (HMM) to classify each action, where the emission is a codeword based on the histogram of motion vector components of the *whole* frame. In later work [Babu and Ramakrishnan, 2003], they extracted Motion History Image (MHI) and Motion Flow History (MFH) [Davis and Bobick, 1997] from compressed domain features, before computing global measures for classification.

Since we would like to build a system that can organize large collections of activity videos, one of the key requirements is the ability to quickly localize and recognize actions. In our previous work [Yeo *et al.*, 2006; Yeo *et al.*, 2008], we used the motion vector information to compute motion similarity between a query video and a target video with a similarity measure that takes into account differences in both orientation and magnitude of motion vectors. Shechtman et al.'s approach for estimating action similarity [Shechtman and Irani, 2005] is computationally complex compared to our method and may be unsuitable for use in organizing large video databases. Babu et al. [Babu *et al.*, 2002] use codewords based on the histogram of motion vector components of the whole frame; this approach requires some preprocessing of the motion vector components into appropriate feature representation, followed by trained a Hidden Markov Model (HMM) for classification. Our approach to computing motion similarity is significantly simpler in comparison, and hence it results in better overall efficiency for estimating all pairwise distances across data points.

6.2 Proposed Approach

Let $\Phi_X \doteq \{X^p\}_{p=1}^P$ be the given set of videos, where $P \in \mathbb{Z}_+$ is the cardinality of the set, and let $\tilde{N} \times \tilde{M} \times \tilde{T}$ be the user-specified space-time scale of interest. Each video X^p has an action label $y^p \in \{1, .., K\}$, where K is the number of actions in the collection. Assume that X^p is a video with T^p frames, with each frame containing $N^p \times M^p$ macroblocks. We assume that an *action* induces a motion field that can be observed as a spatio-temporal pattern; let $\vec{V}^p$ be the spatio-temporal pattern (motion field) associated with video X^p. Furthermore, $\vec{V}^p_{n,m}(i) = [V^{p,u}_{n,m}(i) \quad V^{p,v}_{n,m}(i)]$ denotes the motion vector at location (n, m) in frame i of X^p. We assume that similar actions will induce similar motion fields - i.e., $y^p = y^q \iff D(X^p, X^q) < \gamma$ for some acceptance threshold γ, where $D(\cdot)$ is the distance metric defined between the videos based on their motion fields (defined in Section 6.2.2). We will use $(\boldsymbol{u})_+$ as a shorthand for $\max(0, \boldsymbol{u})$.

Figure 6.2 shows the flow of our algorithm for organizing the videos (Φ_X) with minimal user input. Each of the steps shown in the figure is explained in the following sections. For an extensive discussion on the intuition behind the steps involved in computation of action similarity, please refer to [Yeo *et al.*, 2008].

1. Compute pair-wise action distances between videos using chosen similarity measure $\rho(\cdot, \cdot)$ on coarse optical flows estimated from given videos.

2. Perform hierarchical agglomerative clustering on videos using the computed pair-wise distances $D_{\text{SIM}}(\cdot, \cdot)$ and an appropriate neighbor-joining criterion.

3. Stop agglomerative process either when the remaining clusters are too far apart to be merged (distance criterion) or when number of clusters reaches a user-defined limit (number criterion).

4. Choose a representative exemplar for each cluster.

5. Evaluate the clustering performance using the pair-wise distance matrix D_{SIM} and the cophenetic distances D_{COPH} derived from the hierarchy.

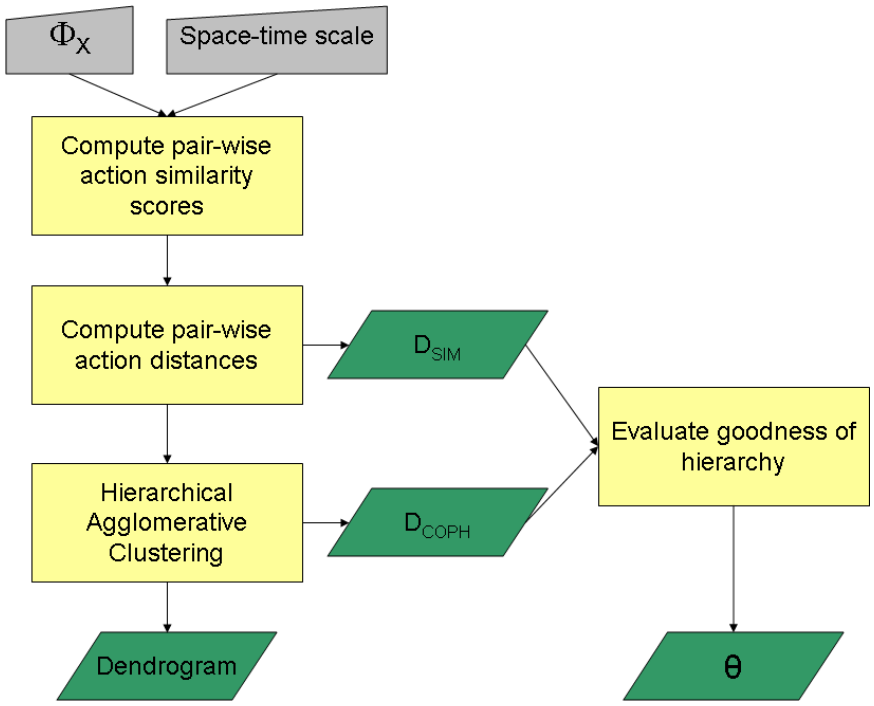

Figure 6.2: Data flow for our proposed approach: Given a set of videos Φ_X and a user-defined space-time scale for actions, we compute pair-wise action similarity scores between all pairs of videos, and then convert them to symmetric action distances, D_{SIM}. We use D_{SIM} in hierarchical agglomerative clustering to produce a dendrogram, which is a binary hierarchical tree representing the videos, and the pair-wise cophenetic distances D_{COPH}, which are distances computed from the constructed dendrogram. The cophenetic correlation coefficient, Θ, is the correlation coefficient between D_{SIM} and D_{COPH}, and can be used to evaluate the goodness of the hierarchy.

6.2.1 Computation of efficient pair-wise action similarity scores

In order to compute non-symmetric pair-wise action similarity scores between X^{test} and X^{query}, we carry out the following steps [Yeo *et al.*, 2008] as illustrated in Figure 6.3:

1. Obtain the motion field estimate, $\vec{V}$, for a video X from its compressed-domain motion vectors, keeping only the reliable estimates as indicated by a confidence map computed from DCT AC coefficients [Coimbra and Davies, 2005]. Motion vectors have been found to be a coarse but reasonable estimate of the motion field, and using them allows our approach to be computationally efficient.

2. At a particular macroblock location (n, m) of the test video, compute the frame-to-frame motion similarity measure, $\tilde{S}_{n,m}(i, j)$, between the i^{th} test video frame and the j^{th}th query video frame (cropped to $\tilde{N} \times \tilde{M}$ macroblocks). In our experiments, we used two methods to compute $\tilde{S}_{n,m}(i, j)$: *Normalized Correlation between Non-negative motion Channels (NCNC)*, and *Non-Zero Motion block Similarity (NZMS)* (discussion follows).

3. To enforce temporal consistency of the similarity between X^{test} and X^{query}, we convolve $\tilde{S}_{n,m}(i, j)$ with a smoothing kernel $H_\alpha \in \mathbb{R}^{T \times T}$. The resultant aggregated similarity matrix is $S_{n,m}(i, j) = (\tilde{S}_{n,m} \star H_\alpha)(i, j)$[3]. α is a parameter that allows us to control how tolerant we are to different action rates [Yeo *et al.*, 2008].

4. After repeating the above two steps over space and time (see Figure 6.4), we compute a confidence score which tells us how likely the action in the query video is occurring at the (n, m) macroblock of test video frame i by taking the maximum of the aggregated similarity matrix over a space-time window:

$$C(n, m, i) = \max_{\substack{\max(i-\frac{T}{2},0) \le k \le \min(i+\frac{T}{2}, T^{\text{test}}-1) \\ 0 \le j \le \tilde{T}-1}} S_{n,m}(k, j) \tag{6.1}$$

[3]Note that the convolution is performed separately for each (n, m), and is only over the (i, j) frame indices.

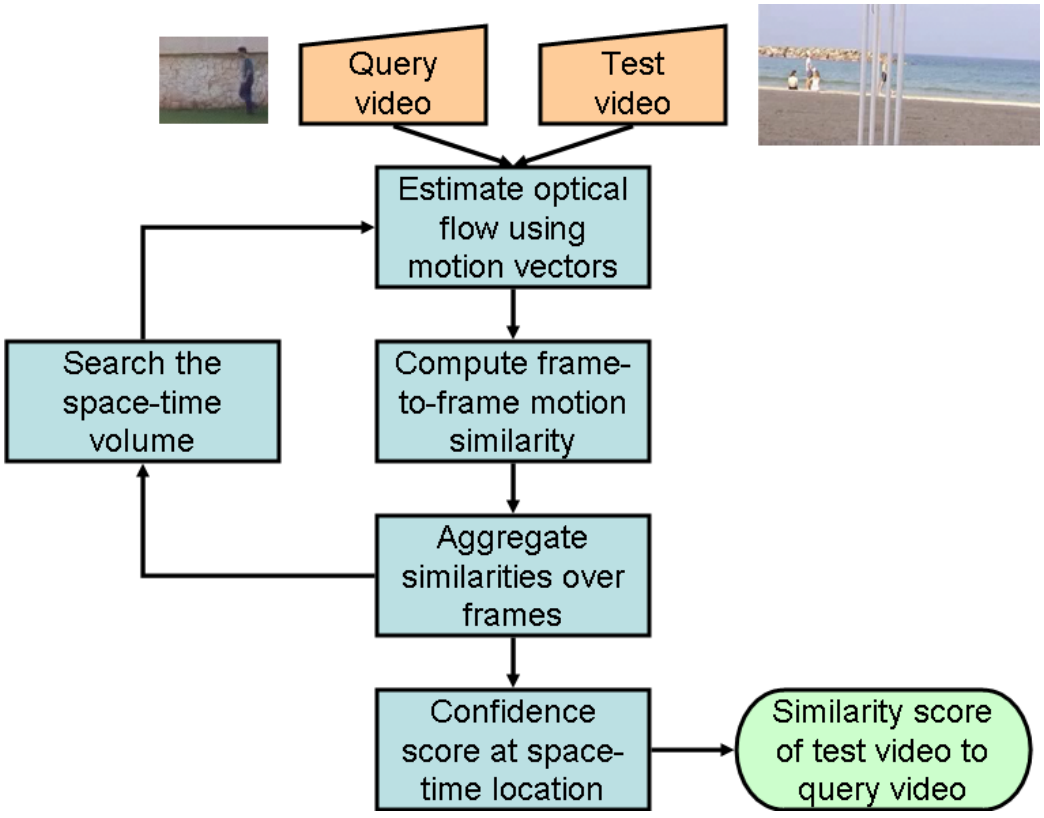

Figure 6.3: Flow chart of action recognition and localization method: Optical flow in the query and test videos are first estimated from motion vector information. Next, frame-to-frame motion similarity is computed between all frames of the query and test videos. The motion similarities are then aggregated over a series of frames to enforce temporal consistency. To localize, these steps are repeated over all possible space-time locations. If an overall similarity score between the query and test videos is desired, a final step is performed with the confidence scores.

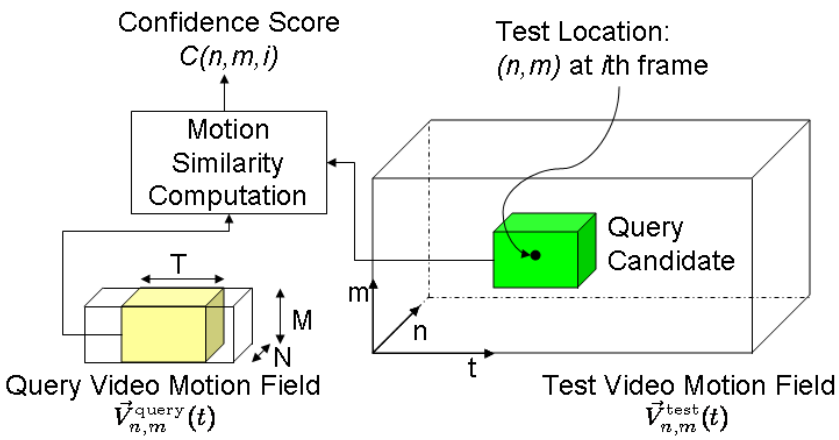

Figure 6.4: Illustration of space-time localization and similarity score computation: The query video space-time patch is shifted over the entire space-time volume of the input video, and the similarity, $C(n, m, i)$ is computed for each space-time location.

5. Compute the similarity, $\rho(X^{\text{test}}, X^{\text{query}})$, of the test video to the query video by:

$$\rho(X^{\text{test}}, X^{\text{query}}) = \frac{\sum_{i=0}^{T^{\text{test}}-1} \eta(i) \left(\max_{n,m} C(n, m, i)\right)}{\sum_{i=0}^{T^{\text{test}}-1} \eta(i)} \tag{6.2}$$

where $\eta(i)$ is an indicator function which returns one if at least T frames in the $2T$-length temporal neighborhood centered at frame i have significant motion and returns zero otherwise. A frame is asserted to have significant motion if at least δ proportion of the macroblocks have reliable motion vectors of magnitude greater than ϵ.

In our experiments, we used $\alpha = 2.0$, $\tilde{N} = \tilde{M} = 6$, $T = 17$, $\tilde{T} = 2T + 1 = 35$, $\delta = \frac{1}{30}$ and $\epsilon = 0.5$ pixels/frame.

Let us elaborate on the methods for computing $\tilde{S}_{n,m}(i, j)$:

6.2.1.1 Normalized Correlation between Non-negative motion Channels

To compute Normalized Correlation between Non-negative motion Channels (NCNC), each $\vec{V}_{n,m}$ is first split into non-negative motion channels (e.g. left, right, up and down) [Efros *et al.*, 2003; Yeo *et al.*, 2008]. An $\tilde{N} \times \tilde{M}$ patch of these motion channels with top-left corner at (n, m) is stacked into a single vector $\vec{U}_{n,m} \in \mathbb{R}^{4\tilde{N}\tilde{M}}$. $\tilde{S}^{\text{NCNC}}_{n,m}(i, j)$ is then computed as

follows:

$$\tilde{S}^{\text{NCNC}}_{n,m}(i,j) = \frac{\langle \vec{U}^{\text{test}}_{n,m}(i), \vec{U}^{\text{query}}(j) \rangle}{\|\vec{U}^{\text{test}}_{n,m}(i)\| \|\vec{U}^{\text{query}}(j)\|} \tag{6.3}$$

6.2.1.2 Non-Zero Motion block Similarity (NZMS)

$\tilde{S}^{\text{NZMS}}_{n,m}(i,j)$ is computed as follows [Yeo *et al.*, 2008]:

$$\tilde{S}^{\text{NZMS}}_{n,m}(i,j) = \frac{1}{Z_{n,m}(i,j)} \sum_{k=0}^{\tilde{N}-1} \sum_{l=0}^{\tilde{M}-1} f(\vec{V}^{\text{test}}_{k+n,l+m}(i), \vec{V}^{\text{query}}_{k,l}(j)) \tag{6.4}$$

$$f(\vec{V_1}, \vec{V_2}) = \begin{cases} \frac{\left(\langle \vec{V_1}, \vec{V_2} \rangle\right)_+}{\max\left(\|\vec{V_2}\|^2, \|\vec{V_1}\|^2\right)} & \text{if } \|\vec{V_1}\| > 0 \text{ and } \|\vec{V_2}\| > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6.5}$$

The normalizing factor, $Z_{n,m}(i,j)$, in (6.4) is:

$$Z_{n,m}(i,j) = \sum_{k=0}^{\tilde{N}-1} \sum_{l=0}^{\tilde{M}-1} \mathbb{I}\Big[\|\vec{V}^{\text{test}}_{k+n,l+m}(i)\| > 0 \text{ or } \|\vec{V}^{\text{query}}_{k,l}(j)\| > 0\Big] \tag{6.6}$$

6.2.2 Computation of pair-wise action distances

Using the similarity scores computed from Section 6.2.1, we compute the pair-wise symmetric action distances for videos X^p and X^q as follows:

$$D_{\text{SIM}}(X^p, X^q) = \frac{1}{\max\left(\frac{1}{2}\left(\rho(X^p, X^q) + \rho(X^q, X^p)\right), \beta\right)} \tag{6.7}$$

where β represents the smallest value of $\rho(.,.)$ admissible. In our experiments, we choose $\beta = 0.01$.

6.2.3 Hierarchical agglomerative clustering of actions

We apply hierarchical agglomerative clustering (HAC) [Webb, 1999] to construct a binary tree (also called *dendrogram*) containing all the elements of Φ_X as leaf nodes. Divisive methods (e.g. K-means, K-medoids) for constructing dendrogram are usually sensitive to initialization [Webb, 1999]. To address this sensitivity with divisive methods, typically one needs to perform many randomly initialized trials in order to obtain a good clustering solution, thus resulting in loss of computational efficiency. In contrast, HAC constructs the dendrogram in a sequential and *deterministic* fashion using a neighbor-joining (also called linkage) criterion. We use four different linkage criteria in our experiments: *Single linkage*, *Complete linkage*, *Average linkage* and *Ward linkage* [Hair *et al.*, 1995].

6.2.3.1 Linkage Criteria

Single linkage method uses minimum distance between the clusters as the merging criterion, where distance between clusters is defined as the distance between closest pair of elements(one element drawn from each cluster) [Hair *et al.*, 1995]. Pairs consisting of one element from each cluster are used in the calculation. The first cluster is formed by merging the two groups with the shortest distance. Then the next smallest distance is found between all of the clusters. The two clusters corresponding to the smallest distance are then merged. The merging process for complete linkage method is similar to single linkage, but the merging criterion is different: the distance between clusters is defined as the distance between most distant pair of elements(one element drawn from each cluster) [Hair *et al.*, 1995]. The merging process for average linkage method is similar to single or complete linkage, but the merging criterion is the average distance between all pairs, where one element of the pair comes from each cluster [Hair *et al.*, 1995]. The distance between two clusters in Ward's linkage method is defined as the incremental sum of the squares between two clusters [Hair *et al.*, 1995].

The user defines a stopping condition for the agglomeration, L^{STOP}, which is the farthest allowable merging distance between clusters. L^{STOP} is used to cut the dendrogram at an

appropriate level and obtain the clusters. After computing the matrix of pair-wise action distances $D_{\text{SIM}} \in \mathbb{R}^{P \times P}$ as described in Section 6.2.2, we apply HAC to obtain the hierarchy. The *cophenetic distance* between videos X^p and X^q, $D_{\text{COPH}}(X^p, X^q)$, computed in the HAC procedure, is their linkage distance when first merged into the same cluster [Webb, 1999].

Algorithm 1: Hierarchical Agglomerative Clustering

Input: D_{SIM}, K^{STOP} OR L^{STOP}
Output: clustering tree formed, K^{FINAL}
Initialize: $t = 0$, $K^0 = P$ (i.e. each video X^p is one cluster) , compute $L^0(\cdot, \cdot)$;
repeat
  Find $(k, l) = \arg\min_{i,j} L^t(i, j)$, $L^t_{\text{MERGE}} = L^t(k, l)$;
  if $(K^t < K^{\text{STOP}})$ *AND* $(L^t_{\text{MERGE}} \leq L^{\text{STOP}})$ **then**
  end
  break;
  Merge cluster k and l;
  Increment t, $K^t = K^{t-1} - 1$;
  Update $L^t(\cdot, \cdot)$ for the new clustering;
until $(K^t > K^{\text{STOP}})$ *OR* $(L^t_{\text{MERGE}} < L^{\text{STOP}})$;
$K^{\text{FINAL}} = K^t$;

6.2.4 Measuring the goodness of the estimated hierarchy

Different choices in clustering parameters, such as distance metric or linkage criteria, lead to different hierarchies (dendrograms). For a good hierarchy, *the cophenetic distances*, D_{COPH}, *should obey the input pair-wise distance relationships* specified by D_{SIM} [Rohlf and Fisher, 1968]. By measuring how well D_{COPH} satisfies pair-wise distance relationships specified by D_{SIM}, we can estimate the goodness of the clustering performance. The *Cophenetic Correlation Coefficient*, $\Theta \in [0, 1]$, for a dendrogram is defined as the correlation coefficient between D_{COPH} obtained from the dendrogram, and D_{SIM} used to construct the dendrogram [S.Farris, 1969]. Thus Θ is a measure of *how faithfully the dendrogram represents the dissimilarities* among videos in the given set Φ_X; its magnitude should be close to 1 for a high-quality solution. Θ is useful in comparing alternative dendrograms obtained by using

different neighbor joining strategies.

6.3 Experimental Results and Discussion

We use a publicly available[4] comprehensive dataset compiled by [Schüldt *et al.*, 2004] to perform our evaluations. This dataset consists of different actions (boxing, handclapping, handwaving, running, jogging and walking) performed by 25 different people over 4 different environments (outdoors [d1], outdoors with scale variations [d2], outdoors with different clothes [d3] and indoors [d4]). Since the two similarity measures we used are not designed for scale-varying actions, we considered only the three non-scale-varying environments.

6.3.1 Action Classification Performance

In order for the proposed organization scheme to be robust to various nuisance factors such as action variations across people or appearance variations, we must make sure that the action matching component must be robust first. We have performed classification experiments to verify the goodness of the compressed domain action recognition algorithm used in our approach. Our results demonstrate that our approach is robust to noise in motion vector estimates (Section 6.3.1.1), person-to-person variations in the rate of actions (Section 6.3.1.2), as well as significant background clutter in videos (Section 6.3.1.3).

To evaluate performance within each environment, we perform a leave-one-out full-fold cross-validation, i.e. to classify each video in the dataset, we use the remaining videos that are not of the same human subject as the training set. This will improve the statistical significance of our results given the limited number of videos in the dataset. To perform classification, we simply use Nearest Neighbor Classification (NNC) by evaluating the video action similarity score with each of the videos in the training set.

The action classification confusion matrix for our algorithm when using NZMS is shown in Table 6.1, while that using NCNC [Efros *et al.*, 2003] is shown in Table 6.2. Each

[4] http://www.nada.kth.se/cvap/actions/

Table 6.1: Confusion matrix using NZMS

	Box	Hc	Hw	Run	Jog	Walk
Boxing	0.86	0.07	0.05	0.00	0.00	0.01
Hand**c**lapping	0.03	0.89	0.08	0.00	0.00	0.00
Hand**w**aving	0.00	0.04	0.96	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.79	0.21	0.00
Jogging	0.00	0.00	0.00	0.01	0.97	0.01
Walking	0.00	0.00	0.00	0.00	0.07	0.93

Table 6.2: Confusion matrix using normalized correlation (NCNC)

	Box	Hc	Hw	Run	Jog	Walk
Boxing	0.86	0.00	0.01	0.00	0.00	0.12
Hand**c**lapping	0.43	0.32	0.24	0.00	0.00	0.00
Hand**w**aving	0.01	0.01	0.97	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.97	0.03	0.00
Jogging	0.00	0.00	0.00	0.21	0.79	0.00
Walking	0.00	0.00	0.00	0.00	0.61	0.39

entry of the matrix gives the fraction of videos of the action corresponding to its row that were classified as an action corresponding to the column. Our overall percentage of correct classification is 90%. As a comparison against state of the art methods that work in the pixel domain, we note here for reference that Schüldt *et al.* [Schüldt *et al.*, 2004], Dollar *et al.* [Dollar *et al.*, 2005] and Ke *et al.* [Ke *et al.*, 2005] report classification accuracies of 72%, 81% and 63% respectively on the same dataset. While the methodology and classification methods used in these works differ, our results compare very favorably on the same benchmark dataset, even though we use compressed domain features and a very simple classifier.

Using NZMS, most of the confusion is between "Running" and "Jogging", with a significant proportion of "Jogging" videos being erroneously classified as "Running". Looking

Table 6.3: Classification performance with and without thresholding confidence map

Method	With thresholding	Without thresholding
NZMS	90.0%	81.2%
NCNC	71.7%	72.5%

at the actual videos visually, we found it hard to distinguish between some "Running" and "Jogging" actions. In fact, there are certain cases where the speed of one subject in a "Jogging" video is faster than the speed of another subject in a "Running" video.

6.3.1.1 Robustness to Noisy Estimates of Motion Vectors

Table 6.3 shows the effects of the reliability of motion vectors on action classification performance using our proposed approach. By removing noisy estimates of the optical flow, we are able to achieve a 10% gain in classification performance. Since motion vectors are generally noisy, these results demonstrate that our method of comparing actions is robust to the imperfections in motion vector estimates.

6.3.1.2 Robustness to Variations in Action Rates

To understand the effect of α on classification, we ran an experiment using NZMS with varying values of α. Table 6.4 shows the results of this experiment. We see that the classification performance is relatively stable over a range of α. This shows that the aggregation step described in Section 6.2.1 is critical for action classification, and provides robustness against person-to-person variability in rates at which people perform various actions. Our results demonstrate that our approach is robust to person-to-person variations in the rate of actions.

Table 6.4: Classification performance with varying α

α	Classification performance
1.0	88.2%
2.0	90.0%
3.0	91.0%
4.0	90.8%
No aggregation	62.5%

6.3.1.3 Robustness to Background Clutter

Figure 6.5 shows an example (the "beach" test sequence and walking query sequence from Shechtman and Irani [Shechtman and Irani, 2005]) which we used as a test case to evaluate the robustness of our action similarity measure in the presence of significant background clutter, with respect to the state of art ([Shechtman and Irani, 2005])). In the test sequence, there are both static background clutter, such as people sitting and standing on the beach, and dynamic background clutter, such as sea waves and a fluttering umbrella. Since both these methods (our method and [Shechtman and Irani, 2005]) can also localize actions within space-time, they are quite robust to background clutter, and give a more accurate estimate of the action similarity based on embedded actions.

6.3.1.4 Computational costs of pairwise similarity computation

On a Pentium-4 2.6 GHz machine with 1 GB of RAM, it took just under 11 seconds to process a test video of 368×184 pixels with 835 frames on a query video that is of 80×64 pixels with 23 frames. We extrapolated the timing reported in [Shechtman and Irani, 2005] to this case; it would have taken about 11 hours. If their multi-grid search was adopted, it would still have taken about 22 minutes. Our method is able to perform the localization, albeit with a coarser spatial resolution, up to *3 orders of magnitude faster*. On the database compiled in [Schüldt *et al.*, 2004], each video has a spatial resolution of 160×120 pixels, and has an average of about 480 frames. For each environment, we would need to perform 22500 cross-comparisons across the dataset. Yet, each run took an average of about 8 hours. In

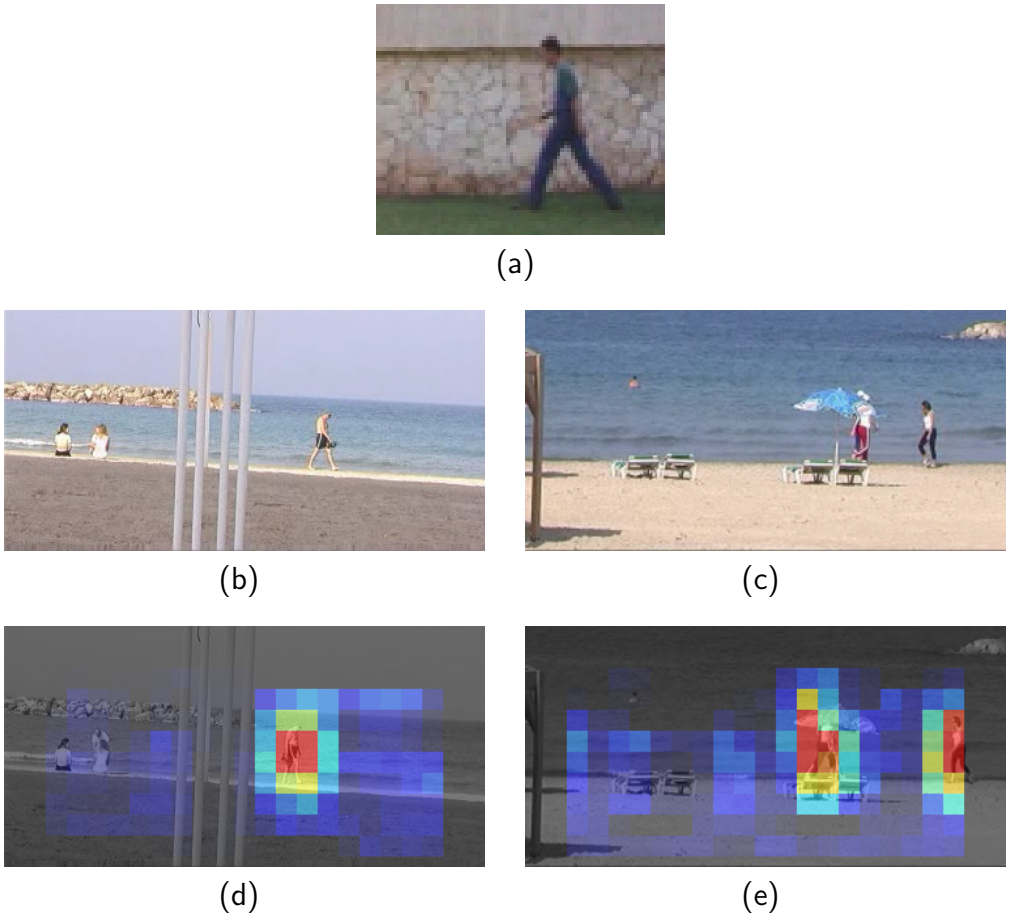

Figure 6.5: Evaluating the robustness to background clutter: The false color in (d) and (e) denotes detection responses, with blue and red indicating a low and high response respectively. (a) A frame from the query video, (b) An input video frame with one person walking, (c) An input video frame with two people walking, (d) Detection of one person walking, (e) Detection of two people walking. Our method's ability to localize allows us to only compute the action similarity at the precise space-time location where target action is situated, hence making our approach robust to background clutter.

contrast, [Shechtman and Irani, 2005] would have taken an extrapolated run time of 3 years. These results demonstrate that our approach is well-suited for organizing large databases of action videos.

6.3.2 Estimating Organizational Structure

From each action video, we create a query video by cropping out a space-time volume in an automatic fashion. Since automatic determination of space-time scale is very hard, we let the user specify the size of an approximate space-time bounding box, $\tilde{N} \times \tilde{M}$ macroblocks by $\tilde{T}$ frames, for the entire collection of videos[5]. The system then looks in each action video for a $\tilde{M} \times \tilde{N} \times \tilde{T}$ space-time volume that contains the most number of significant motion vectors, where $\vec{V}$ is significant if $\|\vec{V}\| > \epsilon$ (as defined in Section 6.2.1).

We adopt two different criteria for evaluating the performance of our organization scheme. The first is based on the ability of the hierarchy to infer meaningful exemplars from the dataset and the second is based on the *F-score* [Larsen and Aone, 1999] used in information-retrieval literature.

6.3.2.1 Inferring action exemplars

In each cluster, an exemplar is defined as the element that has the minimum pair-wise distance with respect to all the other elements in the cluster. *A meaningful hierarchy would organize the videos in such a way that exemplars from each cluster would represent a distinct action from the dataset.* In Figure 6.1, we show the estimated action hierarchy constructed using NCNC action similarity measure with Ward linkage neighbor-joining criterion. Notice that the actions such as running, walking and jogging were grouped separately compared to actions such as boxing, handwaving or handclapping. Intuitively, this fits well with what a human operator would do given the same task. Among the 4 linkage criteria we used, we found qualitatively that the combination of NCNC and Ward linkage gives the best

[5]This implicitly constrains the system to consider actions of approximately similar space-time scale.

Table 6.5: F_1 scores for different environments and clustering methods

Environment	**NZMS / HAC Ward**	**NCNC / HAC Ward**	**NZMS / K-medoids**	**NCNC / K-medoids**
d1	0.7384	0.8496	0.8089	0.7514
d3	0.7220	0.6659	0.7122	0.6480
d4	0.7774	0.7614	0.7515	0.6601

inference for exemplars of actions in the database.

6.3.2.2 Evaluating retrieval performance

As discussed in Section 1.1, one of the key utilities of a well-organized database is that it allows efficient retrieval. We adopt *Balanced F-score* [Larsen and Aone, 1999] from information-retrieval literature for evaluating the goodness of hierarchies estimated in our approach. Treating the action video X^p (with label y^p and in cluster $C^p \in \{1, \cdots, K\}$) as a query video, we define the following:

1. N_1^p is the number of videos in cluster C^p with label y^p,
2. N_2^p is the number of videos in cluster C^p,
3. N_3^p is the number of videos in Φ_X with label y^p.

For the query video X^p, we compute *precision* as $Pr^p = N_1^p/N_2^p$ and its *recall* as $Rc^p = N_1^p/N_3^p$. The *Balanced F-score* [Larsen and Aone, 1999], F_1^p, for this query is the harmonic mean of its *precision* and *recall*: $F_1^p = \frac{2 \cdot Pr^p \cdot Rc^p}{Pr^p + Rc^p}$. We average F_1^p to get $F_1 = \frac{\sum_{p=1}^{P} F_1^p}{P}$. Since the labels in our dataset are for six actions, for the purpose of making comparisons, we only consider F_1 using a value of L^{STOP} such that the number of estimated clusters is 6. We also compute Θ as described in Section 6.2.4. Figure 6.6 shows the variation of F_1 with Θ for different neighbor joining criteria and action environments. The correlation coefficient between F_1 and Θ is 0.77 for NZMS and 0.73 for NCNC respectively - suggesting that Θ can be used to predict clustering performance across various linkage criteria even

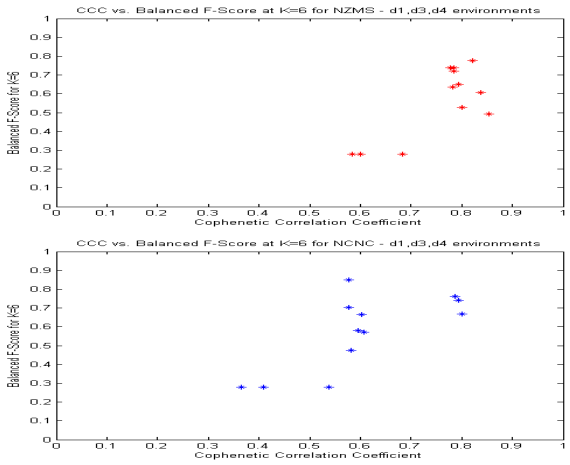

Figure 6.6: Plots (top: NZMS, bottom: NCNC) showing positive correlation between Cophenetic Correlation Coefficient (Θ) and the Balanced F-score (F_1), suggesting that the goodness of hierarchy correlates well with clustering performance.

in the absence of labels. We also compare F_1 scores of our proposed approach with those of a baseline clustering scheme, K-medoids [Webb, 1999], with $K = 6$. We run K-medoids with 200 different random initializations and pick the best F_1 score over all the runs. Due to space constraints, we show only results for HAC using Ward linkage and K-medoids in Table 6.5. It is clear from the results that HAC almost always gives favorable clustering results, without any initialization issues while efficiently producing a useful hierarchy.

6.4 Conclusions

We have demonstrated an efficient unsupervised approach for organizing large collections of videos into a meaningful hierarchy based on the similarity of *actions* embedded in the videos [Ahammad *et al.*, 2007]. By using a fast compressed domain action similarity measure, our method can efficiently operate on large databases of activity videos. Our results show that our method is robust to noise in estimates of motion vectors, significant background clutter and variations in appearance as well as rates of actions across people.

Based on the evidence of high correlation between Θ and F_1 for a given D_{SIM}, we

conjecture that *the unsupervised hierarchical solution for actions that has high* Θ *would also be a solution with high* F_1, thus hinting at good clustering and efficient retrieval performance. Clearly, the figure-of-merit we proposed may not be the only or the most optimal measure for measuring the goodness of estimated hierarchy. This is a direction we plan to investigate further. While compressed domain features are efficient to compute and use, the next logical step would be to extend this framework to include features from raw video and investigate other clustering criteria in the future. Moving forward, it will be an exciting next step to extend this framework to a general system that organizes multimedia databases in a content-based manner.

The estimated organizational hierarchy facilitates quick navigation of the database. Using this hierarchy, we have shown how to select representative videos (exemplars) from a dataset. Our method does not assume any *a priori* knowledge of the number of groups in the database, so if the user decides that the number of clusters are different from a previous hypothesis, it is easy to accommodate such a user request (without any re-computation) by simply allowing the end-user to cut the estimated hierarchy at the appropriate level. Once the organizational hierarchy is estimated, the database can be quickly indexed by assigning a unique action tag to each cluster[6]. These derived action tags can then be combined with other features (such as color, texture etc.) to build more complex queries or to develop organizational principles for managing video databases. Using this unsupervised estimate of the organizational hierarchy, one could potentially use active feedback to create a very effective organization system that uses a human in the loop.

[6]User can easily label a cluster simply by identifying the cluster exemplar and propagating the label downwards in the hierarchy.

Chapter 7

Future Directions

The joint pattern alignment algorithm (JPA) requires no correspondence solving stage, since we attempt to align the entire template without choosing any landmark points. This can lead to problems in the cases where the structure of interest is partially occluded. To handle partial occlusions better, it might be worth considering a hybrid approach that combines JPA and land-mark based methods (such as shape-context or geometric blur). The convergence of the JPA algorithm can get very slow as the number of examples increases. The convergence can also potentially slow down significantly as the dimensionality of the data samples increases (due to the inherent increase in the number of associated transformation parameters). Efficient extensions to JPA, where increasing number of samples or high-dimensionality of the signals can be handled easily, are worth investigating - since both these issues arise in many practical applications. It is worth investigating if there are better nonparametric entropy estimators for real-valued, multi-variate signals that could be used in JPA routine. In our work on geometric alignment, we have only used affine space of transformations. Extending JPA to nonrigid transformations can really improve the goodness of the alignment (albeit with an associated cost in increased number of parameters). These improvements will have a clear impact on the goodness of final solution from the alignment process.

We have shown how the JPA framework can be applied to a variety of applications

such as high-throughput gene expression atlas generation (Chapter 3), denoising random-field bias from magnetic resonance images (Chapter 4) and characterizing evoked neural responses of subjects (Chapter 5). Our characterization of spatial gene expression in Drosophila imaginal discs is not exhaustive (mainly due to resource constraints related to data collection). Applying our pipeline to more genes could be potentially informative on a number of levels. Further analysis of genes known to play a role in the patterning and development of imaginal discs, and the quantification of the precise extent of spatial expression of these genes may provide a more detailed view of the roles of and interactions between these important genes. Our semi-supervised high-throughput JPA framework could also be applicable to construct high-throughput data driven atlases of spatial expression for other biological or neurobiological substrates. In other biological datasets where shape prior is a strong cue (such as Mouse atlases), a similar approach to ours can be taken to construct data driven gene expression atlases. Another natural direction to extend our approach is to apply this procedure to three-dimensional datasets such as image stacks from confocal microscopy studies of in situ stained tissues (such as Drosophila embryos) to construct data-driven gene expression atlases. In our MRI bias removal work, we approximated the RF bias using sine-cosine bases. This was a convenient choice, if not physically appropriate. One direction to extend our bias removal work would be in investigating bias fields that are physically motivated based on the device physics. Using the nonparametrically characterized ERP signals from single trial experiments, one could potentially build a classification scheme that could convert the cortical signals into discrete symbols (a task that is of interest in the domain of brain-computer interfaces (BCI)). Our proposed approach to estimating the ERP signal can be seen as a denoising step in the context of BCI. In terms of further applications, learning compact representations of human actions (for example, using 4-dimensional space-time data from a tele-immersive set-up) would be very interesting. It could potentially lay the ground work for annotating human actions in tele-immersive environments.

To extend or improve our work on organizing large databases of action videos, we

would like to investigate better kernels for estimating pair-wise action distances. It is also of interest to investigate better statistical measures for choosing the best hierarchy in an unsupervised manner. Once a large database of actions is organized using our approach, one could combine the derived action tags with other features (color, texture etc.) for better content-based query system for videos. Our approach relied on an action similarity measure that is not scale-invariant by design. We would like to investigate scale and view invariant methods for action similarity, and incorporate multimodal cues (such as signals from body mounted sensors or audio sensors) into action analysis. Robust methods for action analysis and organization of activity data could pave the way for interactive action-initiated camera networks in future.

Bibliography

[Adams *et al.*, 2000] MD. Adams, SE. Celniker, RA. Holt, CA. Evans, JD. Gocayne, PG. Amanatides, SE. Scherer, PW. Li, RA. Hoskins, RF. Galle, RA. George, SE. Lewis, S. Richards, M. Ashburner, SN. Henderson, GG. Sutton, JR. Wortman, MD. Yandell, Q. Zhang, LX. Chen, RC. Brandon, YH. Rogers, RG. Blazej, M. Champe, BD. Pfeiffer, KH. Wan, C. Doyle, EG. Baxter, G. Helt, CR. Nelson, GL. Gabor, JF. Abril, A. Agbayani, HJ. An, C. Andrews-Pfa nnkoch, D. Baldwin, RM. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, EM. Beasley, KY. Beeson, PV. Benos, BP. Berman, D. Bhandari, S. Bolshako v, D. Borkova, MR. Botchan, J. Bouck, P. Brokstein, P. Brottier, KC. Burtis, DA. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, JM. Cherr y, S. Cawley, C. Dahlke, LB. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, AD. Mays, I. Dew, SM. Dietz, K. Dodson, LE. Doup, M. Downes, S. Dugan-Rocha, BC. Dunkov, P. Dunn, KJ. Durbin, CC. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, AE. Gabrielian, NS. Garg, WM. Gelbart, K. Glasser, A. Glodek, F. Gong, JH. Gorrell, Z. Gu, P. Guan, M. Harris, NL. Harris, D. Harvey, TJ. Heiman, JR. Hernandez an d J. Houck, D. Hostin, KA. Houston, TJ. Howland, MH. Wei, C. Ibegwam, M. Jalali, F. Kalush, GH. Karpen, Z. Ke, JA. Kennison, KA. Ketchum, BE. Kimmel, CD. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, AA. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Ma ttei, TC. McIntosh, MP. McLeod, D. McPherson, G. Merkulov, NV. Milshina, C. Mobarry, J. Morris, A. Moshrefi, SM. Mount, M. Moy, B. Murphy, L. Murphy, DM. Muzny, DL. Nelson, DR. Nelson, KA. Nelson, K. Nixon, DR. Nusskern, JM. Pacleb, M. Palazzolo, GS. Pittman, S. Pan, J. Pollard, V. Puri, MG. Reese, K. Reinert, K. Remington, RD. Saunders, F. Scheeler, H. Shen, BC. Shue, I. Siden-Kiamos, M. Simpson, MP. Skupski, T. Smith, E. Spier, AC. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, AH. Wang, X. Wang, ZY. Wang, DA. Wass arman, GM. Weinstock, J. Weissenbach, SM. Williams, . WoodageT, KC. Worley, D. Wu, S. Yang, QA. Yao, J. Ye, RF. Yeh, JS. Zaveri, M. Zhan, G . Zhang, Q. Zhao, L. Zheng, XH. Zheng, FN. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, HO. Smith, RA. Gibbs, EW. Myers, GM. Rubin, and JC. Venter. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–95, 2000.

[Aggarwal and Cai, 1997] J. Aggarwal and Q. Cai. Human motion analysis: a review. In *Proceedings of IEEE Nonrigid and Articulated Motion Workshop*, pages 90–102, 1997.

[Ahammad *et al.*, 2005] Parvez Ahammad, Cyrus Harmon, Ann Hammonds, Shankar Sastry, and Gerald Rubin. Joint nonparametric alignment for analyzing spatial gene expression patterns of drosophila imaginal discs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[Ahammad *et al.*, 2006a] Parvez Ahammad, Ruzena Bajcsy, and Shankar Sastry. A framework for characterization and comparison of event related neuronal activity. Technical Report UCB/EECS-2006-128, EECS Department, University of California, Berkeley, October 2006.

[Ahammad *et al.*, 2006b] Parvez Ahammad, Ruzena Bajcsy, Gregory Simpson, and Shankar Sastry. Tools for characterization and classification of brain activity via Magnetoencephalography. In *Cold Spring Harbor Laboratory Conference on Engineering Principles in Biological Systems*, 2006.

[Ahammad *et al.*, 2007] Parvez Ahammad, Chuohao Yeo, Kannan Ramchandran, and S. Shankar Sastry. Unsupervised discovery of action hierarchies in large collections of activity videos. In *Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2007.

[Alizadeh *et al.*, 2000] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[Arbeitman *et al.*, 2002] M. N. Arbeitman, E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White. Gene expression during the life cycle of drosophila melanogaster. *Science*, 297:2270–2275, 2002.

[Arnone and Davidson, 1997] M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124:1851–1864, 1997.

[Arthur, 2006] Wallace Arthur. D'Arcy Thompson and the theory of transformations. *Nature Reviews Genetics*, 7:401–406, May 2006.

[Babu and Ramakrishnan, 2003] R. Venkatesh Babu and K. R. Ramakrishnan. Compressed domain human motion recognition using motion history information. In *Proc. IEEE International Conference on Image Processing*, pages 321–324, Barcelona, Spain, September 2003.

[Babu *et al.*, 2002] R. Venkatesh Babu, B. Anantharaman, K.R. Ramakrishnan, and S.H. Srinivasan. Compressed domain action classification using HMM. *Pattern Recognition Letters*, 23(10):1203–1213, August 2002.

[Bajcsy and Kovacic, 1989] R. Bajcsy and S. Kovacic. Multiresolution elastic matching. *Computer Vision, Graphics, Image Processing*, 46:1–21, 1989.

[Baldock *et al.*, 2003] RA. Baldock, JB. Bard, A. Burger, N. Burton, J. Christiansen, G. Feng, B. Hill, D. Houghton, M. Kaufman, J. Rao, J. Sharpe, A. Ross, P. Stevenson, S. Venkataraman, A. Waterhouse, Y. Yang, and DR. Davidson. Emap and emage: a framework for understanding spatially organized data. *Neuroinformatics*, 1(4):309–25, 2003.

[Beirlant *et al.*, 1997] J. Beirlant, E. Dudewicz, L. Gyorfi, and E. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6:17–39, 1997.

[Bell and Sejnowski, 1995] AJ Bell and TJ Sejnowski. An information-maximization approach to blind source separation and deconvolution. *Neural Computation*, 7:11291159, 1995.

[Belongie *et al.*, 2002] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[Belouchrani *et al.*, 1993] A. Belouchrani, K. Abed-Meraim, J-F. Cardoso, and E. Moulines. Second-order blind separation of correlated sources. In *Proceedings of International Conference on Digital Signal Processing*, pages 346–351, Nicosia, Cyprus, 1993.

[Berg, 2005] Alexander C. Berg. *Shape Matching and Object Recognition*. PhD thesis, University of California, Berkeley, 2005.

[Berman *et al.*, 2002] BP. Berman, Y. Nibu, BD. Pfeiffer, P. Tomancak, SE. Celniker, M. Levine, GM. Rubin, and MB. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A*, 99(2):757–62, 2002.

[Beverley and Ponsonby, 2003] C. Beverley and D. Ponsonby. *The anatomy of insects and spiders*. Chronicle Books, 2003.

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[Bra,] Brain web project.

[Butler *et al.*, 2003] Miranda J. Butler, Thomas L. Jacobsen, Donna M. Cain, Michael G. Jarman, Michael Hubank, J. Robert S. Whittle, Roger Phillips, and Amanda Simcox. Discovery of genes with highly restricted expression patterns in the Drosophila wing disc using DNA oligonucleotide microarrays. *Development*, 130:659–670, 2003.

[Cao *et al.*, 2000] J Cao, N Murata, S Amari, A Cichocki, T Takeda, H Endo, and N Harada. Single-trial magnetoencephalographic data decomposition and localization based on independent component analysis approach. *IEICE Transactions on Fund. Electr.*, 9:17571766, 2000.

[Chang *et al.*, 1998] Shih-Fu Chang, W. Chen, H.J. Meng, H. Sundaram, and Di Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):602 – 615, Sept. 1998.

[Chang, 1995] Shih-Fu Chang. Compressed-domain techniques for image/video indexing and manipulation. In *Proc. IEEE International Conference on Image Processing*, pages 314–317, 1995.

[Christensen *et al.*, 1996] G.E. Christensen, R.D. Rabbitt, and M.I. Miller. Deformable templates using large deformation kinematics. *IEEE Transactions on Image Processing*, 5(10):1435–1447, October 1996.

[Chui and Rangarajan, 2000] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proceedings of IEEE CVPR*, pages 44–51, 2000.

[Cohen, 1993] S. M. Cohen. The development of drosophila melanogaster. volume II, page 747841. Cold Spring Harbor Laboratory Press, 1993.

[Coimbra and Davies, 2005] Miguel T. Coimbra and Mike Davies. Approximating optical flow within the MPEG-2 compressed domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):103–107, 2005.

[Collignon *et al.*, 1995] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated multi-modality image registration based on information theory. *Information Processing in Medical Imaging*, pages 263–274, 1995.

[Collins *et al.*, 1998] D.L. Collins, A.P. Zijdenbos, J.G. Kollokian, N.J. Sled, C.J. Kabani, C.J. Holmes, and A.C. Evans. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17:463–468, 1998.

[Cote *et al.*, 1998] G. Cote, B. Erol, M. Gallant, and F. Kossentini. H. 263+: video coding at low bit rates. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(7):849866, 1998.

[Davis and Bobick, 1997] James Davis and Aaron Bobick. The representation and recognition of action using temporal templates. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–934, 1997.

[Dimitrova and Golshani, 1994] N. Dimitrova and F. Golshani. Rx for semantic video database retrieval. In *MULTIMEDIA '94: Proceedings of the second ACM international conference on Multimedia*, pages 219–226, New York, NY, USA, 1994. ACM Press.

[Dollar *et al.*, 2005] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop onVisual Surveillance and Performance Evaluation of Tracking and Surveillance*, page 6572, 2005.

[Efros *et al.*, 2003] Alexei Efros, Alexander Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Proc. IEEE International Conference on Computer Vision*, Nice, France, October 2003.

[Eugene *et al.*, 1979] O. Eugene, A. Yund, and J. W. Fristrom. *Tissue Culture Association Manual 5*. 1979.

[Fan, 2003] Ayres Fan. A variational approach to MR bias correction. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 2003.

[Fischler and Elschlager, 1973] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22:67–92, 1973.

[Fowlkes *et al.*, 2005] C. Fowlkes, C. Luengo Hendriks, S. Kernen, M. Biggin, D. Knowles, D. Sudar, and J. Malik. Registering drosophila embryos at cellular resolution to build a quantitative 3d atlas of gene expression patterns and morphology. In *CSB Workshop on BioImage Data Minning and Informatics*, 2005.

[Grenander and Miller, 1994] Ulf Grenander and Michael I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(4):549–603, 1994.

[Grenander, 1993] U. Grenander. *General Pattern Theory: A Mathematical Study of Regular Structures*. Oxford University Press Inc., New York, NY, 1993.

[Hair *et al.*, 1995] Joseph Hair, Rolph Anderson, Ronald Tatham, and William Black. *Multivariate Data Analysis*. Prentice Hall, New York, NY, 4 edition, 1995.

[Hamalainen *et al.*, 1993] M. Hamalainen, R. Hari, R. Ilmoniemi, J. Knuutila, and O.V. Lounasmaa. Magnetoencephalography theory, instrumentation, and applications to noninvasive studies of signal processing in the human brain. *Reviews of Modern Physics*, 65:413–497, 1993.

[Handy, 2004] Todd C. Handy. *Event-Related Potentials : A Methods Handbook*. The MIT Press, 2004.

[Harmon *et al.*, 2007] Cyrus L. Harmon, Parvez Ahammad, Ann Hammonds, Richard Weiszmann, Susan E. Celniker, S. Shankar Sastry, , and Gerald M. Rubin. Comparative analysis of spatial patterns of gene expression in drosophila melanogaster imaginal discs. In *Research in Computational Molecular Biology*, volume 4453/2007 of *Lecture Notes in Computer Science*, pages 533–547. Springer Berlin / Heidelberg, 2007.

[Harmon, 2007] Cyrus Harmon. *Spatial Patterns of Gene Expression in Drosophila melanogaster Imaginal Discs*. PhD thesis, University of California at Berkeley, 2007.

[Held, 2002] Lewis I. Held, Jr. *Imaginal discs: the genetic and cellular logic of pattern formation*. 2002.

[Hyvrinen and Oja, 1997] Aapo Hyvrinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, October 1997.

[Jaskowski and Verleger, 1999] P. Jaskowski and R. Verleger. Amplitudes and latencies of single-trial ERP's estimated by amaximum-likelihood method. *Biomedical Engineering, IEEE Transactions on*, 46(8):987–993, 1999.

[Jung *et al.*, 1999] T-P Jung, S Makeig, M Westerfeld, J Townsend, E Courchesne, and TJ Sejnowski. Independent component analysis of single-trial event-related potentials. In J-F Cardoso, Ch Jutten, and Ph Loubaton, editors, *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation: ICA99*, pages 173–178, Aussois, France, 1999.

[Ke *et al.*, 2005] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Tenth IEEE International Conference onComputer Vision (ICCV) 2005*, volume 1, 2005.

[Kirkpatrick *et al.*, 1983] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[Klebes *et al.*, 2002] A. Klebes, B. Biehs, F. Cifuentes, and TB. Kornberg. Expression profiling of Drosophila imaginal discs. *Genome Biology*, 3(8):Research0038.1–Research0038.16, 2002.

[Knuth *et al.*, 2006] Kevin H. Knuth, Ankoor S. Shah, Wilson Truccolo, Mingzhou Ding, Steven L. Bressler, , and Charles E. Schroeder. Differentially Variable Component Analysis: Identifying Multiple Evoked Components Using Trial-to-Trial Variability. *Journal of Neurophysiology*, 95(5):3257–3276, Feb 2006.

[Kumar *et al.*, 2002] Sudhir Kumar, Karthik Jayaraman, Sethuraman Panchanathan, Rajalakshmi Gurunathan, Ana Marti-Subirana, and Stuart J. Newfeld. BEST: a novel computational approach for comparing gene expression patterns from early stages of Drosophila melanogaster development. *Genetics*, 162:2037–2047, 2002.

[Laptev and Lindeberg, 2003] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *Proc. IEEE International Conference on Computer Vision*, Nice, France, October 2003.

[Larsen and Aone, 1999] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *KDD '99: Proceedings of the fifth ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 16–22, New York, NY, USA, 1999. ACM Press.

[Learned-Miller and Ahammad, 2005] Erik G. Learned-Miller and Parvez Ahammad. Joint MRI bias removal using entropy minimization across images. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

[Lein *et al.*, 2007] Ed S. Lein, Michael J. Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F. Boe, Mark S. Boguski, Kevin S. Brockway, Emi J. Byrnes, Lin Chen, Li Chen, Tsuey-Ming Chen, Mei C. Chin, Jimmy Chong, Brian E. Crook, Aneta Czaplinska, Chinh N. Dang, Suvro Datta, Nick R. Dee, Aimee L. Desaki, Tsega Desta, Ellen Diep, Tim A. Dolbeare, Matthew J. Donelan, Hong-Wei Dong, Jennifer G. Dougherty, Ben J. Duncan, Amanda J. Ebbert, Gregor Eichele, Lili K. Estin, Casey Faber, Benjamin A. Facer, Rick Fields, Shanna R. Fischer, Tim P. Fliss, Cliff Frensley, Sabrina N. Gates, Katie J. Glattfelder, Kevin R. Halverson, Matthew R. Hart, John G. Hohmann, Maureen P. Howell, Darren P. Jeung, Rebecca A. Johnson, Patrick T. Karr, Reena Kawal, Jolene M. Kidney, Rachel H. Knapik, Chihchau L. Kuan, James H. Lake, Annabel R. Laramee, Kirk D. Larsen, Christopher Lau, Tracy A. Lemon, Agnes J. Liang, Ying Liu, Lon T. Luong, Jesse Michaels, Judith J. Morgan, Rebecca J. Morgan, Marty T. Mortrud, Nerick F. Mosqueda, Lydia L. Ng, Randy Ng, Geralyn J. Orta, Caroline C. Overly, Tu H. Pak, Sheana E. Parry, Sayan D. Pathak, Owen C. Pearson, Ralph B. Puchalski, Zackery L. Riley, Hannah R. Rockett, Stephen A. Rowland, Joshua J. Royall, Marcos J. Ruiz, Nadia R. Sarno, Katherine Schaffnit, Nadiya V. Shapovalova, Taz Sivisay, Clifford R. Slaughterbeck, Simon C. Smith, Kimberly A. Smith, Bryan I. Smith, Andy J. Sodt, Nick N. Stewart, Kenda-Ruth Stumpf, Susan M. Sunkin, Madhavi Sutram, Angelene Tam, Carey D. Teemer, Christina Thaller, Carol L. Thompson, Lee R. Varnam, Axel Visel, Ray M. Whitlock, Paul E. Wohnoutka, Crissa K. Wolkey, Victoria Y. Wong, Matthew Wood, Murat B. Yaylaoglu, Rob C. Young, Brian L. Youngstrom, Xu F. Yuan, Bin Zhang, Theresa A. Zwingman, and Allan R. Jones. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445:168–176, 2007.

[Lipshutz *et al.*, 1999] RJ. Lipshutz, SP. Fodor, TR. Gingeras, and DJ. Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–4, 1999.

[Makeig *et al.*, 1997] S Makeig, T-P Jung, A Bell, D Ghahremani, and TJ Sejnowski TJ. Blind separation of auditory event-related brain responses in independent components. *Proc Natl Acad Sci USA*, 94:10979–10984, 1997.

[Makeig *et al.*, 2002] S Makeig, M Westerfield, T-P Jung, S Enghoff, J Townsend, E Courchesne, and TJ Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295:690–694, 2002.

[Miller *et al.*, 2000] Erik G. Miller, Nicholas Matsakis, and Paul A. Viola. Learning from one example through shared densities on transforms. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[Miller, 2002] Erik G. Miller. *Learning from One Example in Machine Vision by Sharing Probability Densities*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 2002.

[Ngo *et al.*, 2002] C.W. Ngo, T.C. Pong, and H.J. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *IEEE Transactions on Multimedia*, 4(4):446–458, 2002.

[Nishimura, 1996] Dwight Nishimura. *Principles of Magnetic Resonance Imaging*. Stanford University, Palo Alto, CA, 1996.

[Ozer *et al.*, 2000] Burak Ozer, Wayne Wolf, and Ali N. Akansu. Human activity detection in MPEG sequences. In *Proc. IEEE Workshop on Human Motion*, pages 61–66, Austin, USA, December 2000.

[Parameswaran and Chellappa, 2005] V. Parameswaran and R. Chellappa. Human action-recognition using mutual invariants. *Computer Vision and Image Understanding*, 98(2):294–324, 2005.

[Peng and Myers, 2004] Hanchuan Peng and Eugene W. Myers. Comparing In situ mRNA expression patterns of Drosophila embryos. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, 2004.

[Pennec, 2006a] Xavier Pennec. Intrinsic statistics on Reimannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, July 2006.

[Pennec, 2006b] Xavier Pennec. *Statistical Computing on Manifolds for Computational Anatomy*. Habilitation à diriger des recherches, Université Nice Sophia-Antipolis, December 2006.

[Polak, 1997] Elijah Polak. *Optimization: algorithms and consistent approximations*. Springer-Verlag New York, Inc., New York, NY, USA, 1997.

[Quiroga and Garcia, 2003] R.Q. Quiroga and H. Garcia. Single-trial event-related potentials with wavelet denoising. *Clinical Neurophysiology*, 114(2):376–390, 2003.

[Rohlf and Fisher, 1968] James F. Rohlf and David R. Fisher. Tests for hierarchical structure in random data sets. *Systematic Zoology*, 17(4):407–412, Dec 1968.

[Sahouria and Zakhor, 1999] E. Sahouria and A. Zakhor. Content analysis of video using principal components. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(8):1290–1298, December 1999.

[Schüldt *et al.*, 2004] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proc. International Conference on Pattern Recognition*, pages 32–36, Cambridge, UK, August 2004.

[S.Farris, 1969] James S.Farris. On the cophenetic correlation coefficient. *Systematic Zoology*, 18(3):279–285, Sep 1969.

[Shechtman and Irani, 2005] Eli Shechtman and Michal Irani. Space-time behavior based correlation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 405–412, San Diego, USA, June 2005.

[Snodgrass, 1935] R. E. Snodgrass. *Principles of Insect Morphology*. Mcgraw-Hill, 1935.

[Tang *et al.*, 2002] AC Tang, BA Pearlmutter, NA Malaszenko, and DB Phung. Independent components of magnetoencephalography: single-trial response onset times. *Neuroimage*, 17:1773–1789, 2002.

[Thompson, 1942] D. W. Thompson. *On Growth and Form*. Cambridge University Press, Cambridge, UK, 1942.

[Tomancak *et al.*, 2002] Pavel Tomancak, Amy Beaton, Richard Weiszmann, Elaine Kwan, ShengQiang Shu, Suzanna E Lewis, Stephen Richards, Michael Ashburner, Volker Hartenstein, and Gerald M Rubin. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 3(12):1–14, 2002.

[Truccolo *et al.*, 2003] W. Truccolo, K.H. Knuth, A. Shah, S.L. Bressler, C.E. Schroeder, and M. Ding. Estimation of single-trial multicomponent ERPs: Differentially variable component analysis (dVCA). *Biological Cybernetics*, 89(6):426–438, 2003.

[Vasicek, 1976] O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society Series B*, 31:632–636, 1976.

[Vasudevan *et al.*, 2008] Ramanarayan Vasudevan, Parvez Ahammad, Shankar Sastry, and Ruzena Bajcsy. Single trial multicomponent ERP estimation via nonparametric entropy minimization. In *Computational and Systems Neuroscience Meeting (CoSyne)*, 2008.

[Vedaldi and Soatto, 2007] Andrea Vedaldi and Stefano Soatto. A complexity-distortion approach to joint pattern alignment. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1425–1432. MIT Press, Cambridge, MA, 2007.

[Vigario *et al.*, 2000] R. Vigario, J. Sarela, V. Jousmiki, M. Hamalainen, and E. Oja. Independent component approach to the analysis of eeg and meg recordings. *IEEE Transactions on Biomedical Engineering*, 47(5):589–593, May 2000.

[Viola and Wells, 1997] P. A. Viola and W. M. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[Viola, 1995] Paul A. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1995.

[Webb, 1999] Andrew Webb. *Statistical Pattern Recognition*. Oxford: Oxford University Press, 1999.

[Wee *et al.*, 2002] Susie Wee, Bo Shen, and John Apostolopoulos. Compressed-domain video processing. Technical Report HPL-2002-282, Hewlett-Packard, 2002.

[Wells *et al.*, 1996] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. Jolesz. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15:429–442, 1996.

[Wiegand *et al.*, 2003] T. Wiegand, G. Sullivan, G. Bjntegaard, and A. Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560576, 2003.

[Yeo *et al.*, 2006] Chuohao Yeo, Parvez Ahammad, Kannan Ramchandran, and S. Shankar Sastry. Compressed domain real-time action recognition. In *Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2006.

[Yeo *et al.*, 2008] Chuohao Yeo, Parvez Ahammad, Kannan Ramchandran, and Shankar Sastry. High speed action recognition and localization in compressed domain videos. *IEEE Transactions on Circuits and Systems for Video Technology: special issue on Video Surveillance: 2008*, 2008.

[Yeung and Liu, 1995] Minerva M. Yeung and Bede Liu. Efficient matching and clustering of video shots. In *IEEE International Conference on Image Processing*, volume 1, pages 338–341, 1995.

[Yilmaz and Shah, 2005] A. Yilmaz and M. Shah. Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras. In *Proceedings of Tenth IEEE International Conference on Computer Vision (ICCV), 2005.*, volume 1, 2005.

[Yuille *et al.*, 1992] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.