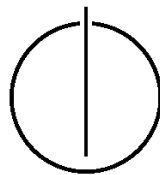# FAKULTÄT FÜR INFORMATIK

DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatik

Online Activity Recognition from skeletal features with Gaussian Processes

Evgeni Pavlidis

FAKULTÄT FÜR INFORMATIK
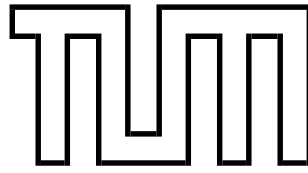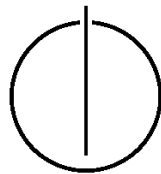
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatik

Online Activity Recognition from skeletal features with Gaussian Processes

Echtzeit Aktivitätserkennung duch Skelettmerkmale mittels Gauss'sche Prozesse

Author:	Evgeni Pavlidis
Supervisor:	Prof. Dr. Daniel Cremers
Advisor:	Dr. Rudolph Triebel
Date:	September 25, 2014

Abstract

English abstract

Zusammenfassung

German Abstract

Ich versichere hiermit, dass ich diese Masterarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den September 25, 2014   Evgeni Pavlidis

Contents

Todo list

1 Introduction

1.1 Motivation

Activity recognition is a big research field in machine learning and robotics. Being able to infer what human actors are doing helps in many practical robotic tasks. An example is doing short-time prediction for collision-avoidance. Moreover in social robotics it is crucial to know what humans are doing when reasoning about the current state of the robot's environment.

To do activity recognition the human pose has to be inferred for each frame. With the advent and further development of RGBD (color and depth) sensors it is now possible to perform skeletal tracking of persons. This allows us to decouple pose estimation and activity learning, which make the problem a bit easier.

For real robotic tasks it is very important that the activity recognition is online i.e. runs in real-time.

1.2 Problem statement

make a distinction between action and activity

1.2.1 Online (active) learning of human activities

Use gaussian processes to learn new activities in real time
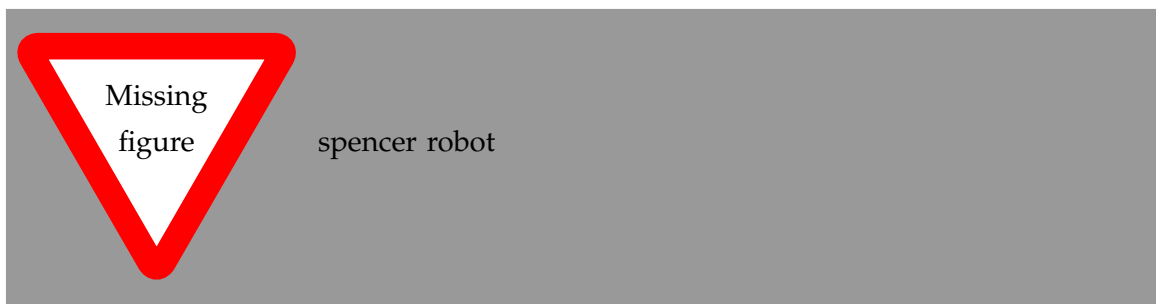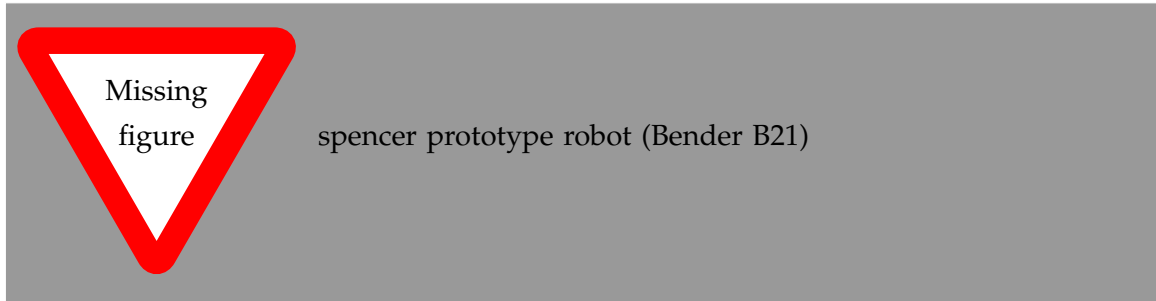
1.2.2 Evaluate Gaussian processes against different ml algorithms for activity recognition

Evaluate the performance of GPs in relation to the other solutions

1.2.3 Scope

1.2.4 The SPENCER project

Missing figure

spencer prototype robot (Bender B21)

Missing figure

spencer robot

With the modern technologies (ms Kinect SDK, primesense ...) it is possible to decouple skeleton tracking and learning

Missing figure

skeleton representation
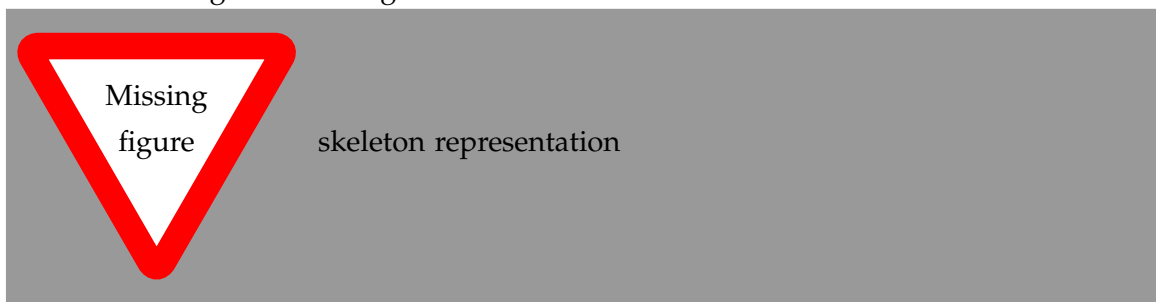
1.3 Notations and conventions

- Matrices uppercase bold

- Vectors lowercase bold

- Constants lowercase

- Parameters lowercase greek letters

- GP notation

- Expectation notation ¡...¿

1.4 Structure of the thesis

Introduction The first chapter introduced the topic of this work. The motivation and the scope is explained.

Basic Concepts This chapters summarizes some basic concepts and models used in our approach. For the most part Gaussian Process Regression and Gaussian Process - Latent Variable Models are being explained.

Related Work The third chapter gives an overview of methods used in similar approaches and then analyses strength and weaknesses of these methods in regards to online activity recognition.

Approach The fourth chapter presents three approaches to real time activity recognition and their implementations.

Evaluation In the fifth chapter the two approaches are being discussed. The first one is an implementation of "Discriminative Sequence Back-constrained {GP}-{LVM} for {MOCAP} based Action Recognition}" [**?**]. The second one being a novel approach which is learns a motion flow field in the latent space through a Gaussian Process Regression.

Results and Outlook The last chapter summarizes the results of the two approaches and gives a brief outlook of future improvements.

2 Basic Concepts

This chapter introduces some basic concepts needed to understand the proposed approaches. First an high-level overview is given on machine learning and its terminology. After that the gaussian distribution is presented in its univariate and multivariate variants. Following is an explanation of Gaussian Processes, their different interpretations and properties. Last the Gaussian Process - Latent Variable Model is being introduced along with some extensions for learning a backward mapping and optimizing it for discrimination in the case of multiple-classes.
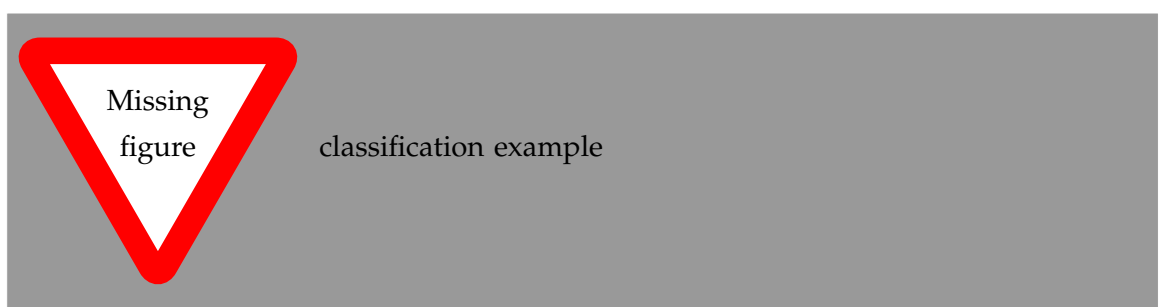
2.1 Machine Learning

2.1.1 Supervised learning

Supervised learning is the task of classification or regression when the data is being labeled. The algorithm then takes the labeled samples (and maybe some confidence values) and infers the model parameters (or hyperparameters) accordingly.

There are two distinct cases:

1. Classification


classification example

Classification the task of learning which category a sample belongs to. A prominent example is Spam filtering. By taking a large number of emails which are labeled either as spam or as ham (regular email), the algorithm deduces a model which can classify unknown samples into these two categories.

2. Regression

Missing figure   regression example
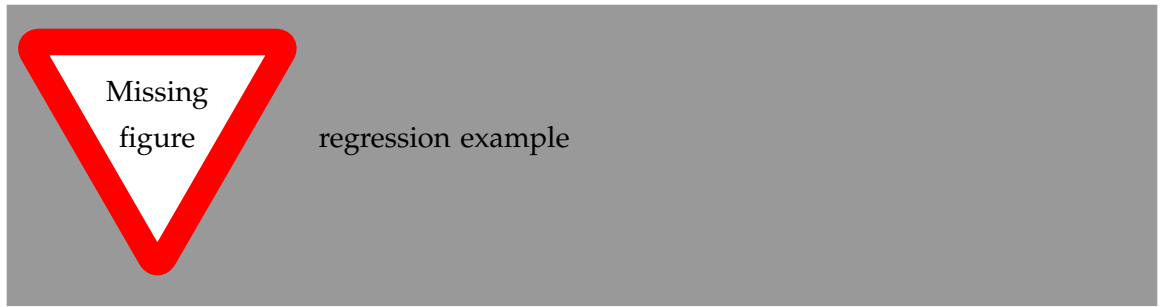
Regression is a terminus in machine learning and means function approximation. Here the domain of the samples label is continuous.

In most cases we search for a good model that explains the data we have. Parametric models, for example, try to learn the ... When searching for an appropriate model it is also important that we try to capture the underlying relationship without compromising the generalization property, which is the ability of the model to correctly predict unseen samples. The case that an algorithm learns the relationship of the data that is used to train the model (training data) but cannot predict new samples is called overfitting. The opposite is called underfitting.

Very often the parameter search is done by maximizing the probability of the data given the model parameters.

$$\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \frac{p(\boldsymbol{\theta}|\boldsymbol{X}) * p(\boldsymbol{X})}{p(\boldsymbol{\theta})}$$

where θ are the model parameters and X is the data.

2.1.2 Unsupervised learning

In contrast to supervised learning in unsupervised learning we have no labeled data i.e. there is no supervisor giving each sample a category (classification) or a value (regression). In this case we can only derive properties of the generation process. Therefore we try to detect patterns in the unlabeled data. These pattern may be clusters of similar samples or a lower dimensional generative manifold from which the samples are generated. The last one is called Dimensionality Reduction which will be also a subject in this work. [**?**]

Missing figure   dimensionality reduction example
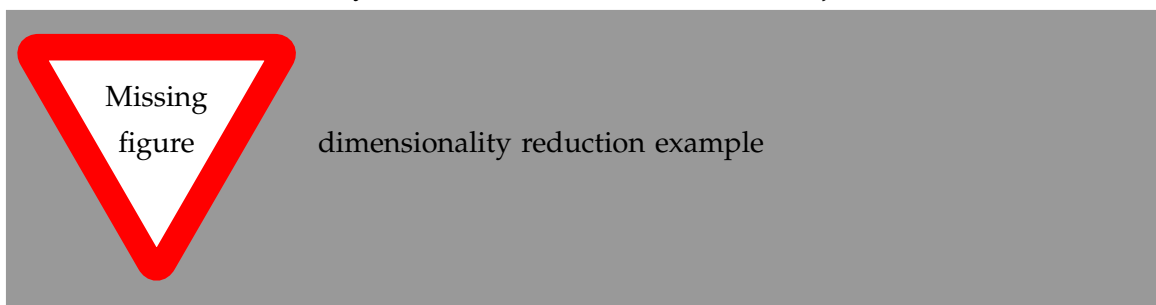
2.1.3 Generative models

Generative methods model the underlying process which generates the data. In Bayesian terms we model the likelihood and the. Thus more data is needed to find an appropriate model. On the other side the model is very flexible and many attributes have a natural interpretation. An example of this is

generative model example

2.1.4 Discriminative models

A discriminative model is only concerned with modeling the actual posterior. This way fewer samples are needed to find the model parameters but by not taking the prior into account the model becomes more generative and is susceptible to overfitting.

2.1.5 Online learning

Algorithms which can be gradually optimized towards a good solution using streaming batches of samples are considered to do online learning. In contrast to online learning online recognition means that the algorithm works in real-time and fast recognition is possible.

2.1.6 Active learning

Very often the bottleneck of powerful supervised learning techniques is that they rely on correctly labeled data. Since labeling has to be performed by a human it is very difficult and costly to label large amount of data. By identifying more important samples by their entropy, thus information ability of selecting a good model, it is possible to achieve good results with fewer samples.

Letting the algorithm select such samples and query only their labels from a human, who is now actively participating in the learning loop, is called active learning.

Active learning is in practice an convenient way to acquire new informative samples without letting someone go over a huge amount of data to label.

2.2 The gaussian distribution

2.2.1 Univariate gaussian distribution

In the one dimensional case the gaussian distribution is well known and understood. Moreover many processes in nature can be modeled with this distribution and for this reason it is also called the Normal distribution. The probability of an event is very high on a certain "point" (its meain value μ) and it drops quickly on each side with the standard deviation σ.

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x-\mu}{2\sigma^2}}$$

One disadvantage of this distribution which we can see from the above formula is that it can model only one hypothesis. This is also the case for the gaussian distributions of multiple (multivariat gaussian distribution) and infinite (gaussian process) dimensions.

2.2.2 Multivariate gaussian distribution

The multivariat gaussian distribution is the generalization of the gaussian distribution in higher dimensions.

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

The two parameters of the distribution are:

mean $\boldsymbol{\mu} = E[x]$ Representing the most probable vector

covariance $\boldsymbol{\Sigma}$ Representing the mutual variance for each pair of the elements of the random vector: $\boldsymbol{\Sigma}_{ij} = Cov[x_i, x_j]$

The exponent is mahalanobis distance, which measures the distance of a point to the ellipsoid defined by the covariance matrix.

cite

2.2.3 Properties of gaussian distributions

Write about total probability and such

Aside for being an appropriate model for many processes occurring in nature, gaussian distributions are also very nice to work with. The marginal and conditional of two gaussian distributions are also gaussian.

One reason GPs are straightforward and work is the math behind them. It is just linear algebra operations.

Linear maps for gaussian distributions:

Product of two multivariate gaussian distributions:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_x}, \boldsymbol{\Sigma_x}) \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu_y}, \boldsymbol{\Sigma_y}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_x})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_y})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_y})}$$

Marginal of a multivariate gaussian:

Conditional of a multivariate gaussian:

cite Gaussian Winter School slides Philipp Hennig

2.3 Gaussian Processes

Consider the multivariate gaussian distribution above. If we want to model the distribution of discrete function defined over a finite interval, we can treat each element of the vector $\boldsymbol{x}$ as an point of the function. Thus we can we can view the multivariate gaussian distribution as a probability function over the function space. Letting the dimensionality

d go to infinity (the distance between each point goes to zero) we can model continuous functions. ... ueberleitung GPs ...

In this case the mean is a point in the function space, thus a function $E[\boldsymbol{x}] = f(x)$. And because of the fact that we now have infinite dimensions the covariance can be seen as an *"infinite* matrix/"*, thus a function of two elements: $Cov(x, y)$.

Thus it can be seen as a gaussian distribution over the function space. [**?**]

ssen~gaussian2006~

A gaussian process can be also seen as the bayesean posterior consisting of the product of the (gaussian) functional prior and the observed samples.??? Another view is a kernelized regression with infinite parameters. [**?**]

A gaussian process is a non-parametric model and is governed by the hyperparameters of the used kernel. This also means that the model is less prune to overfitting which is an important property as it not needed to perform cross validation.

2.3.1 ARD kernel

2.3.2 Regression

nice GP regression example
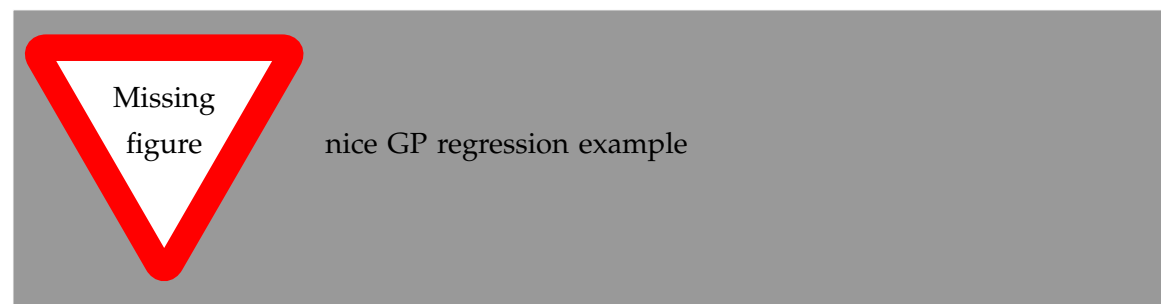
2.3.3 Learning

In the case of a GP the learning phase is different than in parametric models, where the model parameters are inferred from the data. GPs hyperparameter learning by variational optimization (data fit term + cov. regularizer)

In contrast to parametric models gaussian processes are less prune to overfitting because of the covariance regularizer term.

$$E(\theta) = \frac{1}{2} \log(K) - \frac{y^T K^{-1} y}{2}$$

2.3.4 Classification

Classifying with GPs is a little more involved, because of the discriminative function and the fact that the likelihood explain problems of GP classification right} is not a Gaussian. For this reason different models exist which try to approximate this likelihood.

2.3.5 Advantages

1. non parametric Because the model is not parametric it does not suffer from

2. probabilistic The hyperparameters can be interpreted. The lenghtscale controls how much neighboring points contribute to the covariance of the function.

3. generative

4. nice for Bayesian

5. linear algebra operations (marginals and conditionals)

2.3.6 Disadvantages

1. Unimodal

2. susceptible to outliers The student-t distribution is robust against outliers but is much harder to deal with.

3. high computational complexity $\mathcal{O}(n^3)$

   There are sparse GP methods which approximate the model. An example are the informative vector machines (IVM) which reduce the complexity to $\mathcal{O}(d^2 n)$ where d is a subset of the samples selected by maximum entropy. [?]

4. memory heavy A GP saves all data points...

2.3.7 Algorithms

1. Sparse GPs (IVM)

   a) IVM for multiple classes [?]

2.4 GP-LVM

The GP-LVM performs a non-linear dimensionality reduction from an observed space X to a latent space Y [?] It does this by maximizing the likelihood

$$p(Y|X) = p(Y|f)p(f|X)$$

using a gaussian prior for the mapping f. Technically it a GP-LVM is a product of Gaussian Processes which model a regression of the mapping from observed space to one latent dimension. The model learns a (non-linear) mapping from latent space to observed space. This means also that if we want to compute the latent position of a new observed sample we have to compute the .... Using a linear kernel the model generalizes to PCA. By using a non linear kernel a non-linear mapping is inferred making it a non-linear latent variable model.

formulas etc.

elaborate GP-LVM PCA

Missing figure

example GP-LVM, skeleton
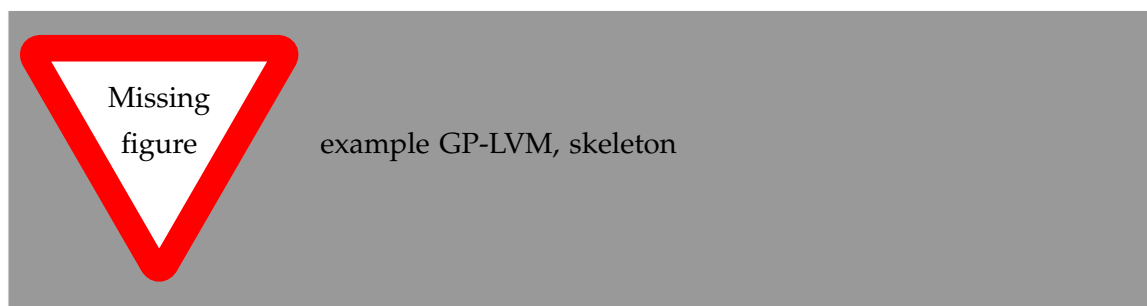
Analogy LVM ¡-¿ marionettes

2.4.1 Kernels

2.4.2 Principal Components Analysis

The Principal Components Analysis is an method which.

This method has an Tipping and Bishop, Journal of the Royal Statistical Society (1999)

2.4.3 Back-constraints GP-LVM

One problem with this model is that it does not preserve local distances in the latent space.This is because it tries to explain the data by moving distant samples from the observed space also far apart in the latent space. This problem is addressed by Lawrence et al. in the back-constrained GP-LVM [?]. A mapping $g_i(y_i) = x_i$ is introduced which constrains the points in latent space to be more near if they are also near in the observed space. Instead of optimizing directly on X the back-constrained GP-LVM optimizes using the mapping instead.

Having this back-constraints also gives us a mapping from observed space to latent space which can be used to project a new sample into the latent space without costly maximum likelihood estimates.

Missing figure

example BCGPLVM
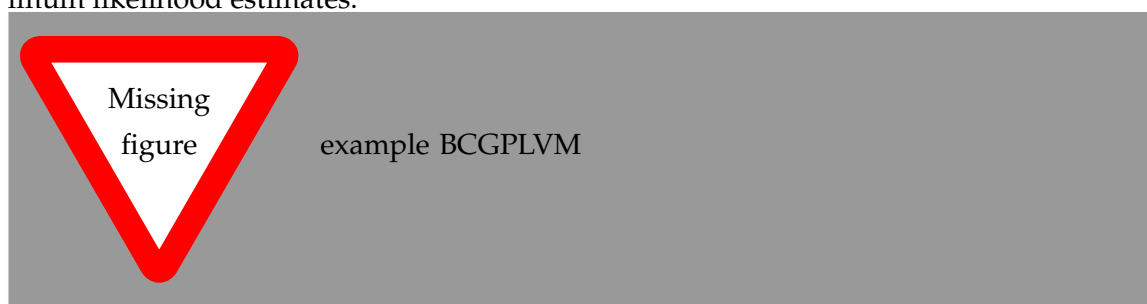
2.4.4 Discriminative GP-LVM

Another improvement in the context of classification in latent space is the Discriminative GP-LVM [?]. Using the GDA a prior is being enforced on the LVM which ensures that samples from one class are more clustered and different classes are more separated in the

elaborate GDA

latent space. This is done by maximizing the between-class separability and minimizing the within-class variability while optimizing the log likelihood of the GP-LVM.[**?**]

2.4.5 Other variants

1. Bayesian GP-LVM An interesting approach for computing the likelihood of the latent variable mapping was proposed in [**?**]. By using a variational method it becomes possible to marginalize over X. Doing so the mapping can be learned together with an ARD kernel. This way the dimensionality of the manifold can be learned from the data.

2. Subspace GP-LVM

3. Manifold Relevance Determination Combining the Subspace GP-LVM with the variational approach and the ARD kernel it is possible to learn the manifold .[**?**]

> explain
> ARD

> explain
> MRD

2.4.6 Advantages

2.4.7 Disadvantages

1. The objective function is non-convex This in fact is the biggest problem as it limits its use on real world data, because for more complex data and (latent manifold structures) there will likely be many local minima. For this reason it is crucial to choose a good initialization. Examples are PCA, Local Linear Embedding or Isomap.

2.4.8 GP-LVM for human motion

As the space of human motion is high-dimensional (spatio-temporal) dimensionality reduction is crucial for a number of models dealing with human motion (e.g. [**?**]). The GP-LVM preserve the distances in the mapping and are therefore suitable to model human motion with high noise of the poses see Urtasun DGPLVM Newest addition is [**?**]

2.5 Dynamic Time Warping

The Dynamic Time Warping is an algorithm which tries to find a minimal warping path between two sequences. Since we are not interested in the path itself but in the cost of the minimal path we define the DTW as a mapping from two time series to an real value. The sequences can be of arbitrary length.

The recursive definitions – excluding some corner cases – reveals the workings of this method.

$$\text{dtw}_{x,y}(i,j) = \text{dist}(x_i, y_j) + \min(\text{dtw}_{x,y}(i-1,j), \text{dtw}_{x,y}(i,j-1), \text{dtw}_{x,y}(i-1,j-1))$$

Where $\text{dist}(x,y)$ is a distance function which tells how close two points are. The DTW can be computed with dynamic programming and has a runtime complexity of $\mathcal{O}(nm)$ where n, m are the lengths of the two sequences.

It is closely relates to the *Longest Common Subsequence* where, instead of minimizing the total warping cost we maximize a common subsequence which is contained in both sequences.

We consider DTW to be a distance which is not entirely correct as the triangle inequality does not hold. Nevertheless it gives us a notion of how similar two time series are and since it is non-negative ($d(x, y) >= 0$), symmetric ($d(x, y) = d(y, x)$) and respects the identity property ($d(x, x) = 0$) it can be used to define a meaningful kernel. [?]

3 Related work

This chapter will introduce some models and their corresponding algorithms for activity recognition. An emphasis is led on methods which work with skeleton data. In the last part a short analysis is done on these methods and some observations are discussed.

3.1 Overview

Activity recognition is a difficult task as we have to make sure our algorithm will discriminate between different classes – activities – but also will leave room for inner class variations. These variations are the result of different persons performing activities differently. A simple example is walking, where different person has a different walking style – also called gait. Also different environments will result in actions to be performed slightly differently. [**?**]

There are many methods which learn from videos and try to explain. This approach is very flexible but also has several drawbacks. One of which is that it is very hard to achieve scale and view-invariance. Furthermore inferring the human pose is very difficult and ambiguous.

For these reasons we will consider only data with pose information in this thesis.

3.1.1 machine vision for human activities: a survey [**?**]

Generative models such as HMM Discriminative models such as CRF

3.2 Histogram based approaches

3.2.1 Motion history image

3.2.2 Motion energy image

3.3 Dynamic time warping

3.4 A class of space-varying parametric motion fields for human activity recognition

3.5 Methods using skeleton features

3.5.1 Gaussian Mixture Based HMM for Human DailyActivity Recognition Using 3D Skeleton Features

3.5.2 Sung et al. [?]

1. Features: Skeleton data + HOG features of RGBD image and depth image

2. Naive classification: SVM

3. Maximum entropy markov model Solved via max-flow/min-cut

3.5.3 RGB-D Camera-based Daily Living Activity Recognition [?]

pipeline
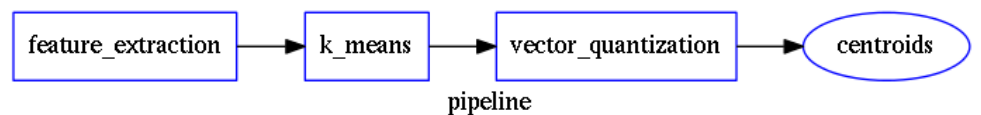
1. Bag of Features
   See [1]

2. Features: Structural and Spatial motion Feature capturing transition between two frames

3. Bag of Features approach (historgram of features)

4. Other: People identification (reidentification)

3.5.4 Learning Human Activities and Object Affordances from RGB-D Videos

1. Learning both: activities and object detection/affordance

2. Using Markov Random Field and SVM for learing

---

[1]RGB-D Camera-based Daily Living Activity Recognition - Chenyang Zhang, Student Member, IEEE and Yingli Tian, Senior Member, IEEE

3.5.5 Eigenjoints [?]

3.5.6 Gaussian Process - Latent Conditional Random Field (GP-L CFR)

[?] use GP-LVM to reduce dimensionality of human motion. (earlier approach was Gibbs sampling)

3.5.7 GPDM

In [?] the dynamics of the latent space is being modeled from time series data. In [?] this model is being used to model human motion by applying a GP-LVM to the high-dimensional mocap data and simultaneously learning the dynamic transition in the latent space:

$$x_{t_{k+1}} = f(x_k)$$

$f(x)$ is being modeled by a gaussian process.

This model was applied for activity recognition in [?] where the classification is done through an SVM in the hyperparameter space. (only 2? features)

3.5.8 Joint Gait Pose Manifold

The JGPM models the activity and the gait in an common latent space. This way several samples from different persons are modeled with the addition with the gait and do not corrupt the class model.

3.6 Analysis

Skeleton features are sufficient but other features can be useful:

- hand

- head pose recognition

- situation awareness ...

3.6.1 Observations

- Activities are represented by the dynamics of the poses, and thus we try to capture this dynamic model

- Difference between activity and action Activities are composed of actions

- Context information can tremendously help in classification of activities (e.g. object detection and human anticipation)

- Skeleton data is sufficient for classification ([?]_????) and also robust to changes in appearance (most state-of-the-art methods work with visual features) and also unobtrusive and sensible data doesn't need to be stored (like face features etc.)

- hierarchical learning: Some methods learn the actions that a activity is composed of. This practice is also very common in HMM models as they model discrete states and their temporal dependencies

- DTW is a good measure but has several drawbacks, such as in cyclic activities where some motions can be repeated several times

3.6.2 Approaches

1. Discriminative Sequence BCGPLVM Use this to find the activity

   a) DTW between walking and walking backwards very big ...

   b) not taking temporal dimension into account

2. GPDM

   a) approach to classify by hyperparameters not optimal

3. VarGPDS

   a) very slow computation

4. Classify by dynamics of the skeleton (this should bring good classification)

   a) GPDM can model the dynamics of the movement

   b) has good properties (gaussian processes)

   c) has intrinsic dim reduction

   d) ?? shared GP-LVM to model different activities in the same latent manifold ??

3.6.3 Problems and solutions

1. limited sample data - probabilistic model + discriminative Probabilistic (and generative ??) models are more accurate using fewer samples, because they model the probability directly ...

2. high dimensional - dim reduction(gp-lvm)

3. classification - BC GP-LVM + discriminative

4. time series data - GPDM An can be modeled as a sequence of consecutive poses. Hence a dynamical model. By using a dynamical model classification becomes more discriminative.

5. confidence is important !!! Using a probabilistic model (especially gaussian processes) we also get a confidence which in turn can be used for active learning

6. high dim. noise =¿ GP-LVM is very robust because of the nature of optimization (distance is preserved instead of locality)

3.6.4 Assumtpions

1. Skeleton tracking is correct and stable For the algorithm we assume that the skeleton extraction from RGBD data works as expected. This is far from the truth with current skeleton tracking algorithms but we also get confidences of the poses. This way we can prune a large number of incorrect poses and because we model the dynamics and do not compare poses this is not a big problem.

2. Smooth skeleton transition !!!

3. Correctly labeled samples (no outliers)

3.6.5 Ideas

1. Use hand and/or head features
   a) Head direction is important
   b) Hand structure is very important for most tasks
   c) Object interrelation ???
   d) Use HOG for hand features only

2. bag of features
   - no time dependency
   - no online capable because of k-means clustering

4 Approach

In this chapter two main approaches are presented to do online activity recognition. The first is an extended version of the **Discriminative Sequence Back-constrained GP-LVM** method. Two improvements are proposed. The first one is using also the velocities of the poses for the dimensionality reduction. The second one is a more discriminative Kernel alignment by using the Mahalanobis distance for the DTW. The second approach is a novel method which models the motion flow field inside latent space to capture the dynamics of each activity, named **GP-Latent Motion Flow**. Along with this model we propose an *energy minimization* approach for online activity recognition which also tries to classify several activities inside a time interval.

4.1 Feature extraction

We get the joint positions and the angles between them in the camera frame defined by the used depth camera (.e.g Microsoft's Kinect). We want these data in the frame of the skeleton. When extracting features we have to make sure that we have scale and view invariant features of the skeleton.

One way to achieve scale invariance is to normalize all link lengths in respect to the torso link. To make the pose view invariant we have to define a local skeleton frame which captures the skeletons *orientation* in the world coordinate system.

local skeleton frame
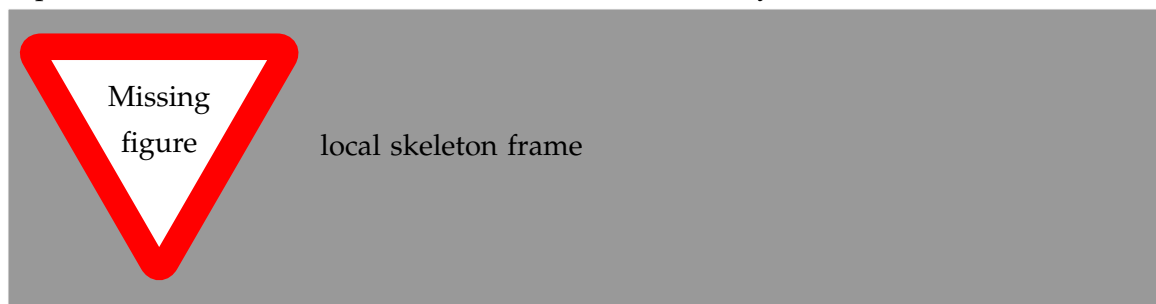
Another way to achieve scale and view invariance is to not consider the 3D points of the joints all together but instead to take only the angles of the joints as features.

An interesting approach is used in [**?**], which is to define a polar coordinate frame for each joint and use the two angles as features. This way we also reduce the observation space.

As discussed in `Related Work` many methods also make the extracted features time invariant (e.g. Eigenjoints). However since we want to model the dynamics we do not use time invariant features.

4.2 Dynamic time warping with mahalanobis distance

This idea was also proposed in the context of handwritten signature verification. Learning Mahalanobis distance for DTW based online signature verification maybe use fisher discriminant analysis as in the disr. seqBCGPLVM

4.3 Discriminative Sequence Back-Constrained GP-LVM

In the paper "Discriminative Sequence Back-Constrained GP-LVM for MOCAP Based Action Recognition"[?] the authors propose a method for classifying MOCAP actions.
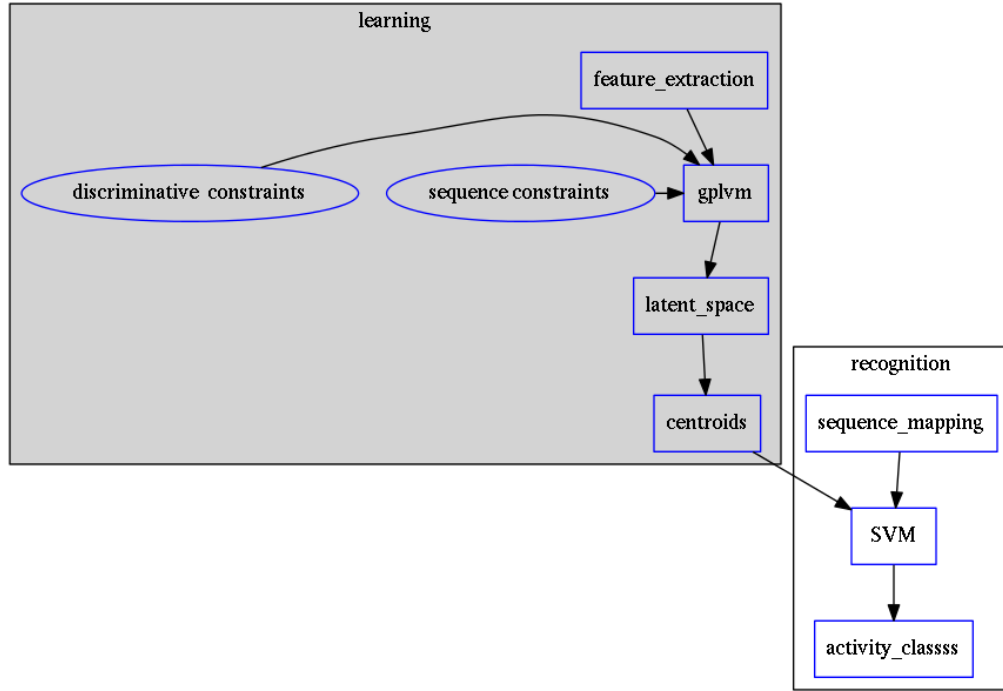
Sequence back-constrained GP-LVM pipeline ... CITATION

By using a similarity feature for the sequences in the observed space and constraining the optimization to preserve this measure the local distances between the sequences are transferred into the latent space. This has two advantages. First of all the sequences have a meaningful clustering in the latent space. Second by also learning the back-constraint it is possible to calculate the centroid of a sequence in the latent space directly without maximizing a likelihood. This in turn is being used to do real-time classification for actions. The mapping is defined as a linear combination of the DTW distance between every other sequence. For every latent dimension q we have:

$$g_q(Y_s) = \sum_{m=1}^{S} a_{mq} k(Y_s, Y_m)$$

where the similarity measure is $k(Y_s, Y_m) = \gamma e^{DTW(Y_s, Y_m)}$. This measure is to be preserved in the latent spaces.

$$g_q(Y_s) = \mu_{sq} = \frac{1}{L_s} \sum_{n \in J_s} x_{nq}$$

This constraints are being enforced in the optimization by adding Lagrangians to the objective function.

Missing figure

example of discriminative and back-constrained latent space
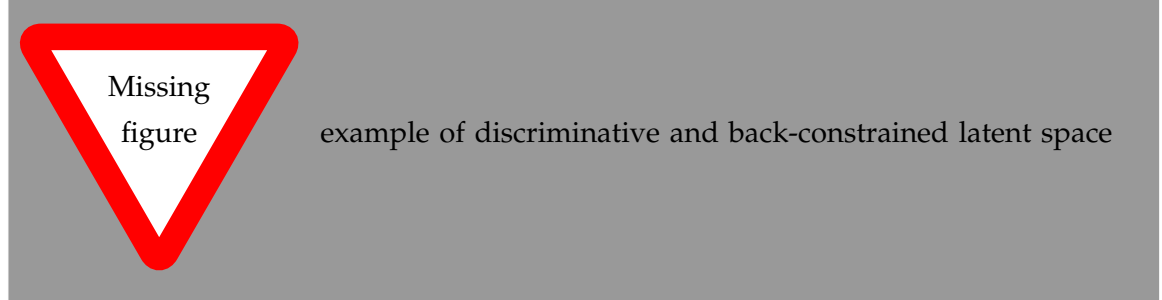
Furthermore, by applying the Discriminative GP-LVM we ensure that poses of different activities are separated from each other and poses from similar activities are located closer together. This ensures that the centroid of an activity is more informative and thus discriminative. The Discriminative GP-LVM works by also maximizing the between class variance and minimizing the in-class similarity [?] Also by applying the Discriminative GP-LVM the clustering of similar actions and the distances of different actions is enhanced which allows for a better classification. Recognition is being done by applying the mapping above to the new sequence and using a SVM in the latent space.

expain D GP-LVM properly

4.3.1 Advantages

Recognition can be done in real time by using the learned back constrained. The centroid in the latent space is being calculated for the whole sequence and classified by the SVM.

4.3.2 Shortcomings

As the optimization for GP-LVM is determined by the above similarity measure and the discriminative criterion online optimization is very difficult. It is thus highly likely that performing a gradient online optimization will be stuck in an local minimum.

Also one problem with the real-time recognition is that determining when a activity has ended/begun is very difficult. Also as we do not know how long a sequence is we have to calculate the centroid for several time frames.

4.3.3 Extensions:

1. Learn poselets (pose and velocities) to capture dynamics The GP-LVM learns a mapping for each pose but does not consider velocities and accelerations. If we take a pose along with its first and second moments (let us call them poselets) as the high-dimensional space we allow for the temporal displacements to be also modeled. The latent space represents the poselet and the DTW kernel in the constraint captures also the motion of the activity.

2. Use mahalanobis for the DTW

## 4.4 GP- Latent Motion Flow Field (based on the gp regreesion flow)

Many models which use GP-LVM to reduce the high dimensional space into fewer dimension. These approaches make the problem more feasible but the problem remains how to do classification for time-series data. Human motions are mostly characterized by the dynamics of the model (temporal dimension). So we have to compare trajectories in the latent space. One idea is to use GPRF as classification can be done using second order dynamics which should give better results. Going further the activity itself is characterized by the first and second moments of the trajectory function. By explicitly modeling the velocity of the trajectory we can take changes in the joint movement into account.

The Gaussian Process Regression Flow [**?**] can be used to model the trajectories in the latent space.

The GP-LMF method is inspired by this model. The difference being that in the case of activity recognition we do not know the starting position and also the trajectories can have significantly different lengths. For this reason it is very difficult to normalize with respect to the time dimension. Nevertheless, resulting from the properties of Gaussian Process regression, we have also a dense mean flow field and dense variances. This allows us perform efficient and robust online recognition in the latent space.

This model is attractive for two reasons. First real-time classification of incomplete trajectories is possible. Second it is possible to do online learning by simply adding the new class as a new flow field to the pool of GPs. It is very difficult to adjust the other models for online learning, because of the problem that we can get stuck in a local minimum when optimizing the parameters of the GP.

The idea is to learn a motion field in the latent space for each activity. This can be achieved by learning the velocity function of the latent point just like in the GPRF model presented above. With the difference that we do not use the spatio-temporal domain but learn only the flow in the latent space. The reason being that we do not have starting and ending positions for each activity and also the lengths can be variable. On top of that we also want to recognize an activity which is being interrupted by another activity, so we can't fix the lengths of the trajectories.

Missing figure   example of several flow fields inside latent space
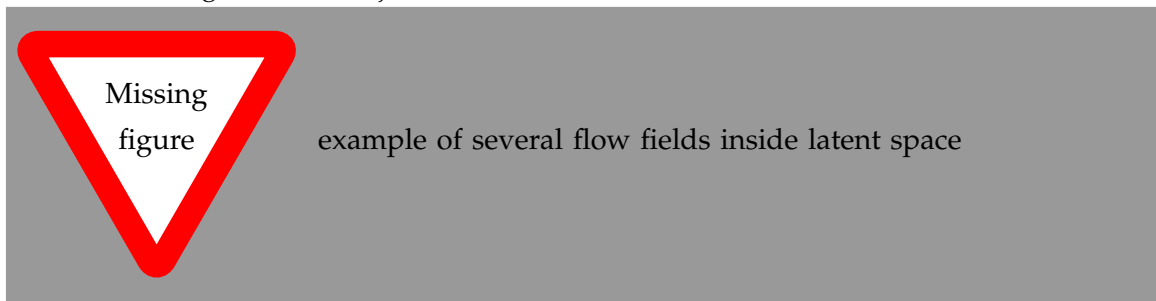
Each activity has its own flow field. Recognition and prediction is done by calculating the energy of the currently moving point with each different field. The field with the

minimum energy represents the most probable activity as the point follows more closely its "current" of motion.

Variances in the speed of performing an activity can be modeled by giving the point in the latent space a mass which can be adjusted in real time. When a point has greater mass then it needs more energy to be propagated along the flow field (the overall activity is slower) and vice versa.

An advantage of this method is that activities with repetitive motions, such as walking or running, can be learned without using periodic kernels or other means to model them explicitly. Repetitive motions can be seen as just multiple samples of the same motion which define the flow field.
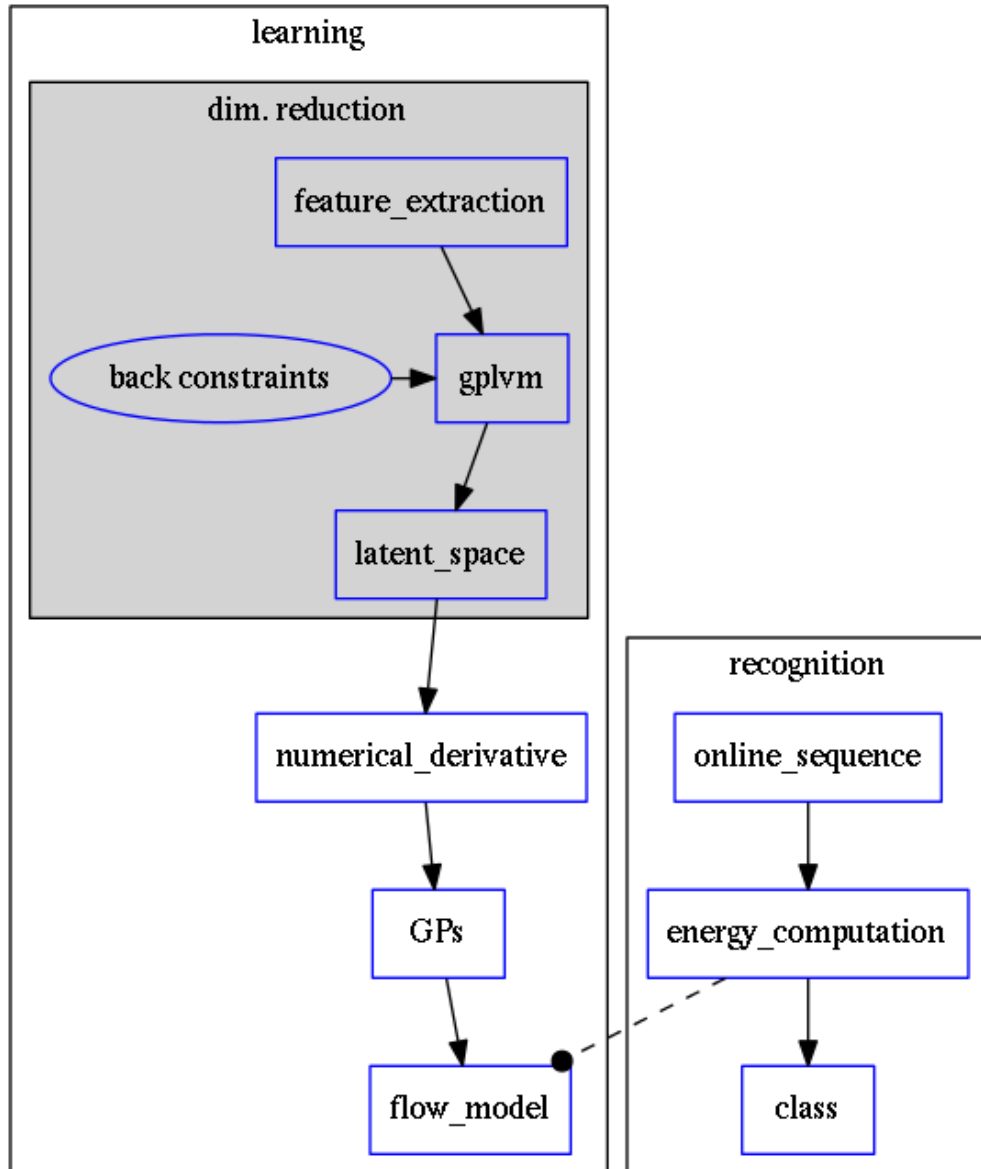
Gaussian Process - Latent Motion Flow

4.4.1 Learning the flow field

We deploy GP for learning the flow field which gives us several advantages.

Missing figure — latent space (several samples of one activity) with flow field
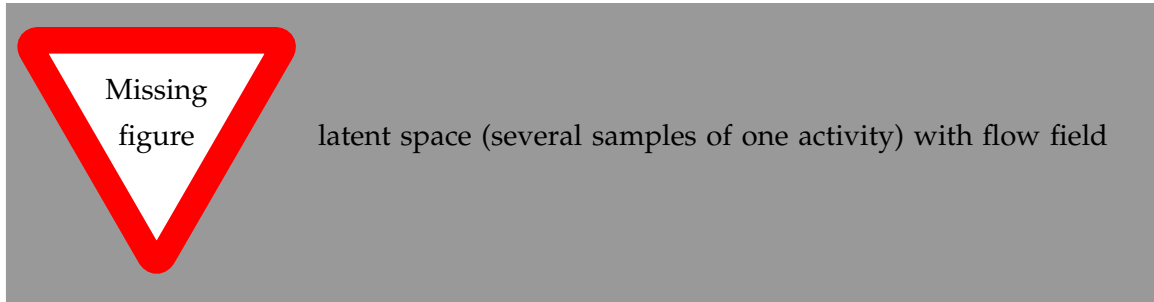
1. Effects of the hyperparameters

   Changing the *lengthscale* defines how much each point is contributing to the regression process. It can be interpreted as a smoothness factor which governs how strong the interpolation of the flow field is performed on the latent points.

   Changing the signal variance controls how much

Missing figure — effect of hyperparameters on the resulting flow field
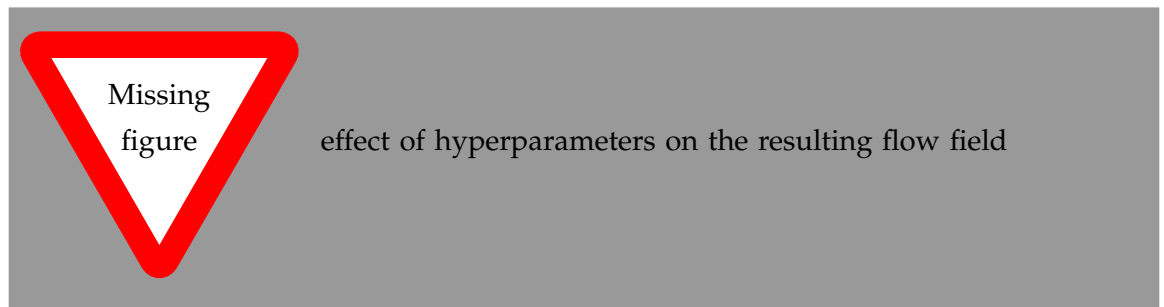
4.4.2 Interpretation

The proposed model has a natural interpretation. A point represents a pose in latent space and an activity is a trajectory in time inside the same space. With the flow field we learn the motion tendencies for each pose. When performing recognition we for each separate flow the energy needed to traverse the flow. If we consider that the point has a mass we can model the speed at which activities are being done. This way we can recognize when a point leaves an activity, which represents a *motion current*, and passes over to some other activity.

4.4.3 Advantages

1. Recognition The current activity is being mapped into the latent space. Through the learned back-constrained. The recognition is being performed solely in the latent space. By propagating the current position by each flow field we can calculate the next possible pose. By comparing the similarity considering the variances we have a measure of how well the current activity resamples each flow field e.g. learned activity.

2. Prediction If we have detected the activity predicting is simply a matter of propagating the pose through the flow field by taking the mean of the GP.

3. Online learning

4. Natural interpretation

5. Novelty detection (anomaly detection) In [**?**] the authors present the ability of the GPRF model for anomaly detection. This approach is also suitable for finding new classes as the above energy value can be used to recognize novel activities. The reasoning is that if we cannot find an flow field with a small energy the activity has to be unobserved.

6. Active learning

7. Multiple Hypothesis Prediction Since we have a GP representing our flow field we can predict future point positions with the mean value. Moreover also having informative variances we can sample several possible trajectories. This can be accomplished using an particle filter. Hence we can have multi-hypothesis predictions along with their probabilities.

4.4.4 Problems

1. Dimensionality reduction Performing a non-linear dimensionality reduction is no easy task. Testing was done with only two dimensions as it easier to visualize the latent space and the resulting flow fields. A latent space with higher dimension will naturally make the reduction more robust and the field will have a more natural interpretation. . . .

2. Stable class mean flow field When learning a stable flow field from several samples the field can degenerate with the inclusion of strong variable paths. Therefore it is important to ensure that the algorithm learns stable paths. This can be achieved by sampling uniform random sampling from all samples of the same activity.

active learning - problem ??

4.4.5 Learning the motion flow field

One problem we encounter by learning the motion flow field from several samples is complexity of the Gaussian Process. There are two solutions for this. The first one is to use a sparse GP model. The second one is to sample points from all samples and use only those that are most suitable for the regression. If we take IVM as the sparse GP model both approaches can be seen as equivalent as the IVM will automatically take the most informative samples.

4.4.6 Recognition

Energy minimization:

$$E_{t_1,t_n} = \sum_{i \in \mathcal{T}, j \in \mathcal{T}} E_{i,j} + regularizer$$

where the regularizer ensures that we do not change flow fields often. . .

5 Evaluation

5.1 Datasets

5.1.1 Cornell Activity Dataset

Active learning using Gaussian Processes. We will use the "Cornell Activity Datasets (CAD-60 & CAD-120)"[1] to learn and evaluate the performance of an implementation of Gaussian Processes.

The data set s consist of an sequence of frames which include:

- Image data

- RGBD data

- Skeleton information: (joint position and orientation)

- annotated meta information (e.g. activity)

5.2 Software

MATLAB - FGPLVM Dataset: CMU Motion capture dataset

- Emacs/Org-mode

- IPython

- SciPy/NumPy

- GPy

- mlpy

5.3 Discriminative Sequence back-constrained GP-LVM

As the idea was to implement the algorithm in an language that can be easily integrated into the ROS infrastructure we implemented the model in Python. Unfortunately we were not able to perform an appropriate dimensionality reduction. We believe that the many constraints on the optimization and the highly different data is very hard to optimize. For this reason we choose to implement a new model basing on motion flow fields.

---

[1]Human Activity Detection from RGBD Images, Jaeyong Sung, Colin Ponce, Bart Selman, Ashutosh Saxena. In AAAI workshop on Pattern, Activity and Intent Recognition (PAIR), 2011.

5.4 Mahalanobis DTW

5.5 Gaussian Process - Latent Motion Field

6 Conclusions and Outlook

6.1 Summary

6.1.1 Dimensionality reduction for all activities is very difficult (also with extra constraints)

6.1.2 Dynamics is a good measure for classification of human activities

6.1.3 Contributions

1. Advantages and Disadvantages of dimensionality reduction with GP-LVM for human motions

2. Improvement of the DTW measure with the mahalanobis distance ????????

3. A novel approach for activity recognition (prediction??)

6.2 Outlook

6.2.1 Energy minimization evaluation

6.2.2 Active activity learning !!!

6.2.3 Automatic segmentation of activities

List of Figures

Bibliography

[1] Discriminative sequence back-constrained GP-LVM for MOCAP based action recognition:. pages 87–96. SciTePress - Science and and Technology Publications, 2013.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, August 2006.

[3] Andreas Damianou, Carl Ek, Michalis Titsias, and Neil Lawrence. Manifold relevance determination. *arXiv preprint arXiv:1206.4610*, 2012.

[4] Hamed Jamalifar, Vahid Ghadakchi, and Shohreh Kasaei. 3d human action recognition using gaussian processes dynamical models. In *Telecommunications (IST), 2012 Sixth International Symposium on*, pages 1179–1183. IEEE, 2012.

[5] Yun Jiang and Ashutosh Saxena. Modeling high-dimensional humans for activity anticipation using gaussian process latent crfs. In *Robotics: Science and Systems, RSS*, 2014.

[6] Kihwan Kim, Dongryeol Lee, and Irfan Essa. Gaussian process regression flow for analysis of motion trajectories. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1164–1171. IEEE, 2011.

[7] Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.

[8] Neil Lawrence, Matthias Seeger, and Ralf Herbrich. Fast sparse gaussian process methods: The informative vector machine. *Advances in neural information processing systems*, pages 625–632, 2003.

[9] Neil D. Lawrence and Joaquin Quiñonero-Candela. Local distance preservation in the GP-LVM through back constraints. In *Proceedings of the 23rd international conference on Machine learning*, pages 513–520. ACM, 2006.

[10] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, June 2010.

[11] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. University Press Group Limited, January 2006.

[12] Matthias Seeger and Michael I. Jordan. Sparse gaussian process classification with multiple classes. Technical report, Citeseer, 2004.

[13] Hiroshi Shimodaira, Ken-ichi Noma, Mitsuru Nakai, and Shigeki Sagayama. Dynamic time-alignment kernel in support vector machine. In *NIPS*, volume 14, pages 921–928, 2001.

[14] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Unstructured human activity detection from rgbd images. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 842–849. IEEE, 2012.

[15] Ilias Theodorakopoulos, Dimitris Kastaniotis, George Economou, and Spiros Fotopoulos. Pose-based human action recognition via sparse representation in dissimilarity space. *Journal of Visual Communication and Image Representation*, 25(1):12–23, January 2014.

[16] Michalis Titsias and Neil Lawrence. Bayesian gaussian process latent variable model. 2010.

[17] Pavan Turaga, Rama Chellappa, Venkatramana S. Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, 2008.

[18] Raquel Urtasun and Trevor Darrell. Discriminative gaussian process latent variable model for classification. In *Proceedings of the 24th international conference on Machine learning*, pages 927–934. ACM, 2007.

[19] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models. In *NIPS*, volume 18, page 3, 2005.

[20] J.M. Wang, D.J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, February 2008.

[21] Xiaodong Yang and YingLi Tian. Effective 3d action recognition using EigenJoints. *Journal of Visual Communication and Image Representation*, 2013.

[22] Chenyang Zhang and Yingli Tian. RGB-d camera-based daily living activity recognition. *Journal of Computer Vision and Image Processing*, 2(4):12, 2012.