Online activity recognition through kernel methods

Evgeni Pavlidis

September 15, 2014

Contents

Todo list

1 Introduction

1.1 Motivation

1.1.1 Online (active) learning of human activities

Use gaussian processes to learn new activities in real time

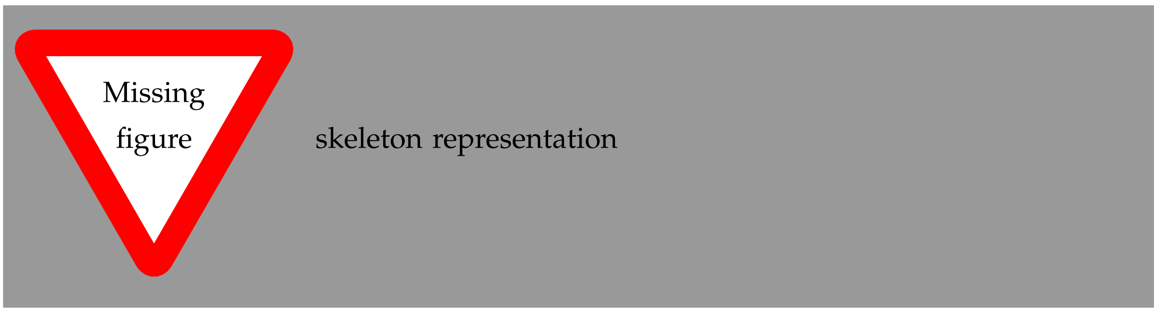
1.1.2 Evaluate Gaussian processes against different ml algorithms for activity recognition

Evaluate the performance of GPs in relation to the other solutions

1.2 Scope

1.2.1 Use only skeleton data

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



1.3 Notations and conventions

1.3.1 Matrices uppercase bold

1.3.2 Vectors lowercase bold

1.3.3 Constants lowercase

1.3.4 Parameters lowercase greek letters

1.3.5 GP notation

1.3.6 Expectation notation ¡. . . ¿

2 Concepts

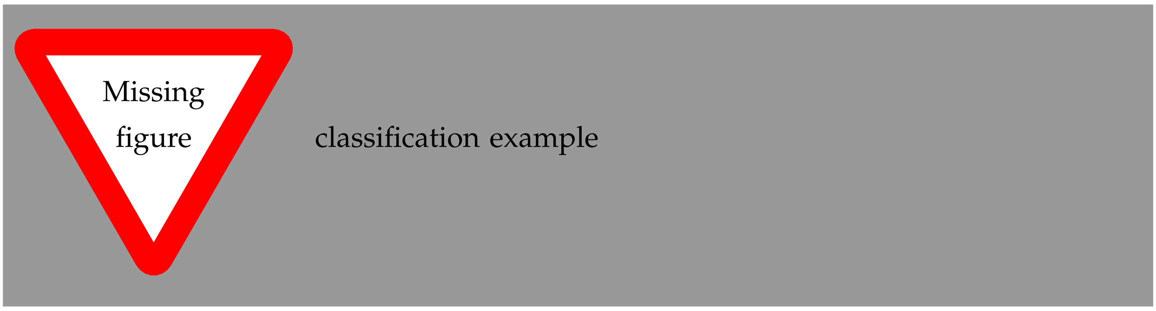
2.1 Machine Learning

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.

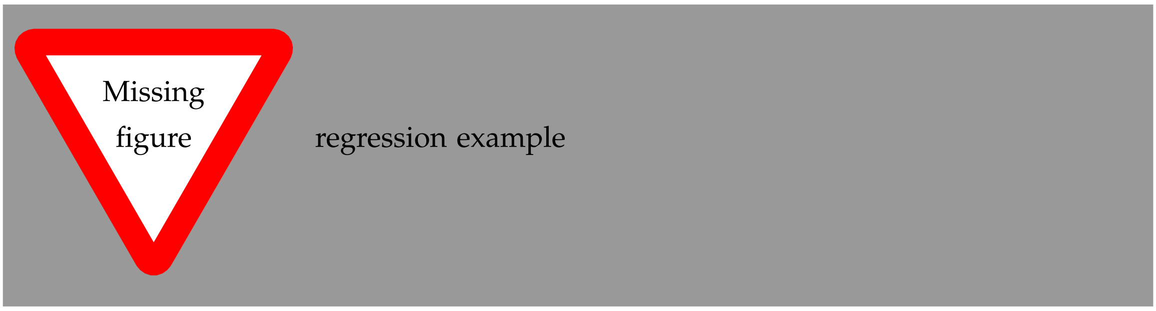
There are two distinct cases:

1. Classification

   Classification 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

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

2.1.2 Unsupervised learning

In unsupervised learning the algorithms tries to detect patterns in the unlabeled data. Pattern may be clusters of similar samples or a lower dimensional generative manifold. The last one is called Dimensionality Reduction. [2]

dimensionality reduction example

2.2 Generative vs. discriminative models

2.2.1 Generative

Generative models learn the underlying process which generates the data. Thus more data is needed to find an appropriate model. On the other side the model is very flexible and many attributes can be deduced later.

2.2.2 Discriminative

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

2.3 Gaussian Processes

A gaussian process can be seen as the bayesean posterior consisting of the product of the a (gaussian) functional prior and the observed samples.??? Another view is a kernelized regression with infinite parameters. [11]

A gaussian process is a non-parametric model and is governed by the hyperparameters of the used kernel. It can be seen as a gaussian distribution over functions.

2.3.1 Regression

nice GP regression example

2.3.2 Learning

GPs hyperparameter learning by variational optimization (data fit term + cov. regularizer)

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

2.3.3 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.

{

2.3.4 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. nice for Baysean
4. linear algebra operations (marginals and conditionals)

2.3.5 Disadvantages

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

2.3.6 Algorithms

1. Sparse GPs (IVM)
   a) IVM for multiple classes [12]

2.4 GP-LVM

The GP-LVM performs a non-linear dimensionality reduction from an observed space X to a latent space Y [8] 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 very strong latent variable model.

formulas etc.

elaborate GP-LVM

PCA

example GP-LVM, skeleton

2.4.1 PCA

Tipping and Bishop, Journal of the Royal Statistical Society (1999)

2.4.2 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 [9]. 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.

example BCGPLVM

2.4.3 Bayesian GP-LVM

An interesting approach for computing the likelihood of the latent variable mapping was proposed in [15]. 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.

explain ARD

2.4.4 Discriminative GP-LVM

Another improvement in the context of classification in latent space is the Discriminative GP-LVM [17]. 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 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.[17]

elaborate GDA

2.4.5 Subspace GP-LVM

2.4.6 Manifold Relevance Determination

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

explain MRD

2.4.7 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 [5]

3 Related work

3.1 Overview

3.1.1 a survey on vision based action recognition [10]

3.1.2 machine vision for human activities: a survey [16]

3.2 Histogram based approaches

3.2.1 Motion history image

3.2.2 Motion energy image

3.3 Sung et al. [13]

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

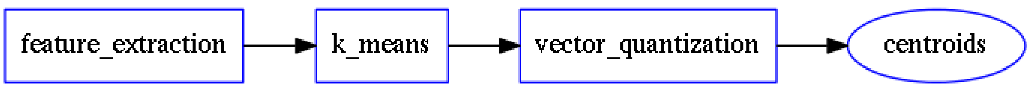
3.3.2 Naive classification: SVM

3.3.3 Maximum entropy markov model

Solved via max-flow/min-cut

3.4 RGB-D Camera-based Daily Living Activity Recognition [21]

3.4.1 Bag of Features



pipeline

See [1]

3.4.2 Features: Structural and Spatial motion

Feature capturing transition between two frames

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

3.4.3 Bag of Features approach (historgram of features)

3.4.4 Other: People identification (reidentification)

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

3.5.1 Learning both: activities and object detection/affordance

3.5.2 Using Markov Random Field and SVM for learing

3.6 Eigenjoints [20]

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

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

3.8 GPDM

In [18] the dynamics of the latent space is being modeled from time series data. In [19] 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 [4] where the classification is done through an SVM in the hyperparameter space. (only 2? features)

3.9 Dynamic time warping

3.10 See also

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

4 Analysis

4.1 Observations

- Person identification through covariance matrix of movement see. [7]
- 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.)
- Knowing which activity a human performs helps tremendously in classifying age and gender! (in the case that we do both)

4.2 Approaches

4.2.1 make the features invariant [14]

- view invariant (pos rel to torso)
- scale invariant (normalize length…) … time ?? invariant

4.2.2 Discriminative Sequence BCGPLVM

Use this to find the activity

1. DTW between walking and walking backwards very big …
2. not taking temporal dimension into account

4.2.3 Use the Joint Gate and Pose Manifold for age and gender detection

4.2.4 GPDM

1. approach to classify by hyperparameters not optimal

4.2.5 VarGPDS

1. very slow computation

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

1. GPDM can model the dynamics of the movement
2. has good properties (gaussian processes)
3. has intrinsic dim reduction
4. ?? shared GP-LVM to model different activities in the same latent manifold ??

4.3 Problems and solutions

4.3.1 limited sample data - probabilistic model + discriminative

Probabilistic (and generative ??) models are more accurate using fewer samples, because they model the probability directly ...

4.3.2 high dimensional - dim reduction(gp-lvm)

4.3.3 classification - BC GP-LVM + discriminative

4.3.4 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.

4.3.5 confidence is important !!!

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

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

4.4 Assumtpions

4.4.1 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.

4.4.2 Smooth skeleton transition !!!

4.4.3 Correctly labeled samples (no outliers)

4.5 Ideas

4.5.1 Presentation

1. make clear what is your contribution !!!
2. Black slides (important points)

4.5.2 Model

1. Take best three activites (uncertainty) with threshold
2. SPENCER: can help for (head tracking (bounding box), and pose estimation)
3. 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
4. Bhattacharyya distance
5. bag of features
   - no time dependency
   - no online capable because of k-means clustering
6. maximum entropy markov model
   - complex, performance not that good
7. GP-LVM
   - good to reduce the dimensionality
   - used in some papers
8. Learn a model of an activity and compare it with the help of a covariance function

4.5.3 Analogy LVM ¡-¿ marionettes

5 Approach

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 Discriminative Sequence Back-Constrained GP-LVM

In the paper "Discriminative Sequence Back-Constrained GP-LVM for MOCAP Based Action Recognition"[1] the authors propose a method for classifying MOCAP actions. 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.

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

[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.

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 [17] 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

5.2.1 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.

5.2.2 Improvements

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.

5.2.3 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.

5.3 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.

The Gaussian Process Regression Flow [6] can be used to model the trajectories 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.

5.3.1 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.

5.3.2 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.

5.3.3 Online learning

5.3.4 Novelty detection (anomaly detection)

In [6] 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.

5.3.5 Active learning

5.3.6 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.

5.3.7 Problems

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 ??

5.4 Implementation

5.4.1 Software

MATLAB - FGPLVM Dataset: CMU Motion capture dataset

- Emacs/Org-mode
- IPython

- SciPy/NumPy
- GPy
- mlpy

5.5 Bibliography

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] M. S. Kumar and R. Venkatesh Babu. Human gait recognition using depth camera: a covariance based approach. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, page 20. ACM, 2012.

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

[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] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

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

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