Gesture Alignment Using Hidden Markov Models

Andrew Hershberger Salman Ahmad
{andrew.hershberger, saahmad}@cs.stanford.edu

Stanford University
CS228: Probabilistic Graphical Methods — Winter 2011

Abstract

Gesture recognition gaining interest in many domains. A common issue when learning from multiple training gestures is accounting for noise and different durations and speeds. This paper presents an algorithm that uses Hidden Markov Models to align training gestures and learn the hidden canonical gesture that they represent. The algorithm was used to align motion capture data from an Xbox Kinect. This paper present initial results along with a discussion of current limitations and paths for future work.

1 Introduction

Gesture recognition is becoming increasingly important in many fields from gaming to user interface design. Since it is difficult to manually encode gestures declaratively, there is a lot of interest in applying learning techniques to train a classifier that can recognize gestures from motion capture data.

A common problem with this approach is that the training examples are often different durations and different speeds. This paper provides an algorithm that aligns gestures based on the important actions that take place - e.g. the start of a wave or the midpoints in a jump. Additionally, the algorithm learns a canonical representation of the gesture that can be used for classification of new data.

The algorithm was evaluated using motion capture information from an Xbox Kinect. The raw RGBZ output was converted to *(x,y,z)* positions of 20 different control points on the human body as shown in Figure 1. An alignment hidden markov model was then used to align the different gestures.

The rest of this paper presents related work in this area, the graphical model used to encode the independencies of the data, a discussion of the algorithm, results, an analysis of current limitations, and logical areas of future work.

2 Related Work

Koller and Friedman [Koller and Friedman, 2009] provide an overview and analysis of different approaches to solving sequence labeling problems like gesture recognition. The hidden Markov model (HMM), a generative model, presents the challenge of modeling a distribution over the observed variables. By contrast, the maximum entropy Markov model (MEMM) and conditional random field (CRF) are discriminative approaches in which only the conditional distribution over the class labels must be modeled, thus avoiding the need to model the distribution over the observed variables directly. On the other hand, generative models may allow learning with less training data than would be required in the discriminative case.

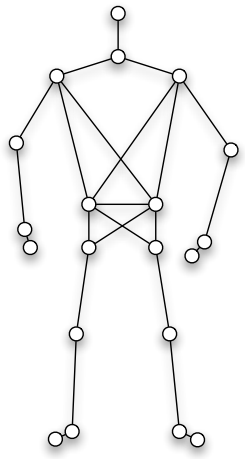

Figure 1: The location twenty control points that are tracked by the gesture alignment algorithm.

In one gesture recognition project, Wang, et al. [Wang *et al.*, 2006] employed a CRF variant called a hidden conditional random field (HCRF). Their approach, however, did not address the issue of dealing with varying length gestures or gestures that are compressed or expanded in time during various phases.

In a related study, Coates, et al. [Coates *et al.*, 2008] addressed the problem of learning an ideal pattern from multiple non-ideal demonstrations. They applied an iterative expectation maximization (EM) algorithm for aligning multiple input sequences while simultaneously learning the ideal target sequence. Their approach interleaved iterations of a Kalman smoother and dynamic time warping. The Kalman smoother [Murphy, 2002] was used to determine the means

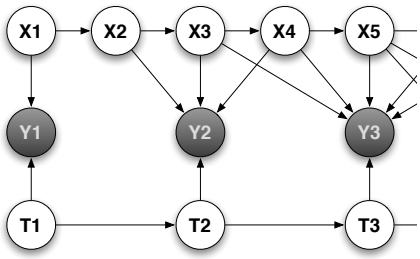

Figure 2: The graphical model using for gesture alignment. X represents the hidden, "canonical" gesture. Y represents the observed gesture from the training set. T represents an indexed mapping between the observed gesture to the canonical gesture.

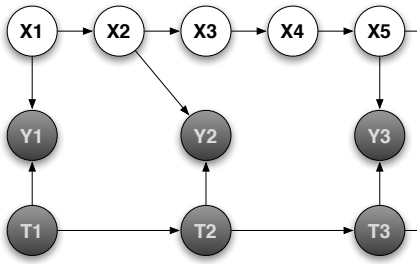

Figure 3: An example graphical model once T has been learned. In this case, the second and third frames from the observed gesture maps to the second and fifth frame of the canonical gesture.

and covariance matrices of the hidden target sequence. The dynamic time warping algorithm [Listgarten *et al.*, 2005] then determined the highest-likelihood mapping of the observed sequences onto the ideal. Eventually this iterative process converged producing tremendous results.

3 Graphical Model

images of the model before and after dynamic programming approach to DTW

It's an alignment HMM

4 Algorithm

ANDREW:

What we did: EM, DTW

Optimizing the algorithm when calculating q(:,:), tau;

Different smoothing

no prior knowledge of optimal trajectory

not using a bias function

Tried different allowed step sizes for d

5 Results

SALMAN:

Figures, writeup (we are geniuses).

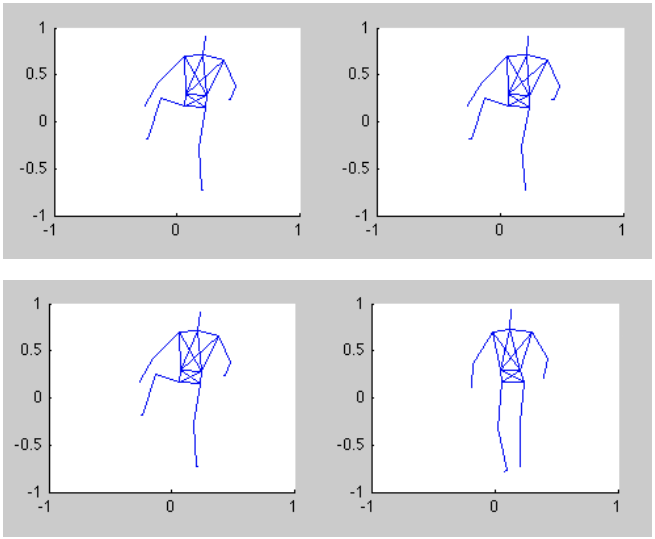

Figure 4: Motion capture data of a person performing a kick. Top: Data that has been aligned with our algorithm. Bottom: The Original, unaligned data. Both images were taken at the same time offset.

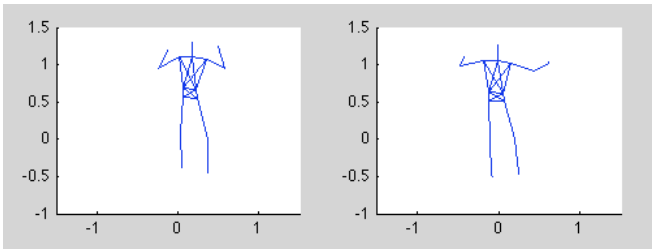

Figure 5: A failure case of our implementation. The algorithm does not have domain specific information about the dynamics of the real world, for example, gravity. The above data was taken from a person jumping. To align the data, the algorithm "freezes" the person in mid-air when this is obviously physically impossible.

6 Discussion and Future Work

ANDREW:

Why didn't it work very well? Smoothing made things worse.

Use different smoothing

Add prior knowledge of optimal trajectory: can incorporate effects of gravity - don't want things to hang in mid air.

Application to classification problem

Add more data (training data)

Add features to detect particular aspects of gestures.

Detect orientation differences

7 Conclusion

This paper presents a method to perform gesture alignment using a Hidden Markov Model. The algorithm was shown to be able to align certain gestures and learn a canonical gesture. The method was applied to motion capture data that was extracted from RGBZ images taken from an Xbox Kinect.

While the findings were some what promising it failed to work on a diverse set of gestures. There are obvious areas for future work. First, the model should incorporate our prior knowledge of the ideal gesture. For example, it would certainly help to encode that during a kick, one of the legs will be accelerating while the rest of the body stays still. Second, the algorithm should incorporate a dynamics model of the real world. This will allow the method to better deal with physical phenomenons like gravity.

References

[Coates *et al.*, 2008] Adam Coates, Pieter Abbeel, and Andrew Y. Ng. Learning for control from multiple demonstrations. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 144–151, New York, NY, USA, 2008. ACM.

[Koller and Friedman, 2009] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[Listgarten *et al.*, 2005] Jennifer Listgarten, Radford M. Neal, Sam T. Roweis, and Andrew Emili. Multiple alignment of continuous time series. In *in Advances in Neural Information Processing Systems*, pages 817–824. MIT Press, 2005.

[Murphy, 2002] Kevin Patrick Murphy. Dynamic bayesian networks: Representation, inference and learning, 2002.

[Wang *et al.*, 2006] Sy Bor Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1521 – 1527, 2006.