Human Pose Tracking

Benjamin Braithwaite (cpg608)
Shuhab Hussain (rtc525)

October 22, 2013

1 Introduction

Human motion tracking is useful in a variety of areas: surveillance, human-computer interaction, the film industry and video games. The performance and efficiency of tracking methods has increased rapidly in the last few years, but there are still limitations regarding difficult poses to be resolved. Current state-of-the-art methods are implemented in the OpenNI and Microsoft Kinect tracking frameworks but these have limitations regarding certain difficult human poses, for example when the head is not directly above the neck as shown in Figure 1.

A recent paper [Ganapathi et al., 2012] describes a different solution which is supposedly not only more robust when tracking difficult poses but also substantially faster and more accurate than Microsoft Kinect's methods described in [Shotton et al., 2011]. The new method is based on a Ray-Constrained ICP model which defines a likelihood on single depth measurements for certain poses. The optimal pose is found by accelerated gradient descent constrained by physical and motion models. We have implemented this new method for human pose tracking and compared it with OpenNI on varied poses. [Insert something regarding results here..]

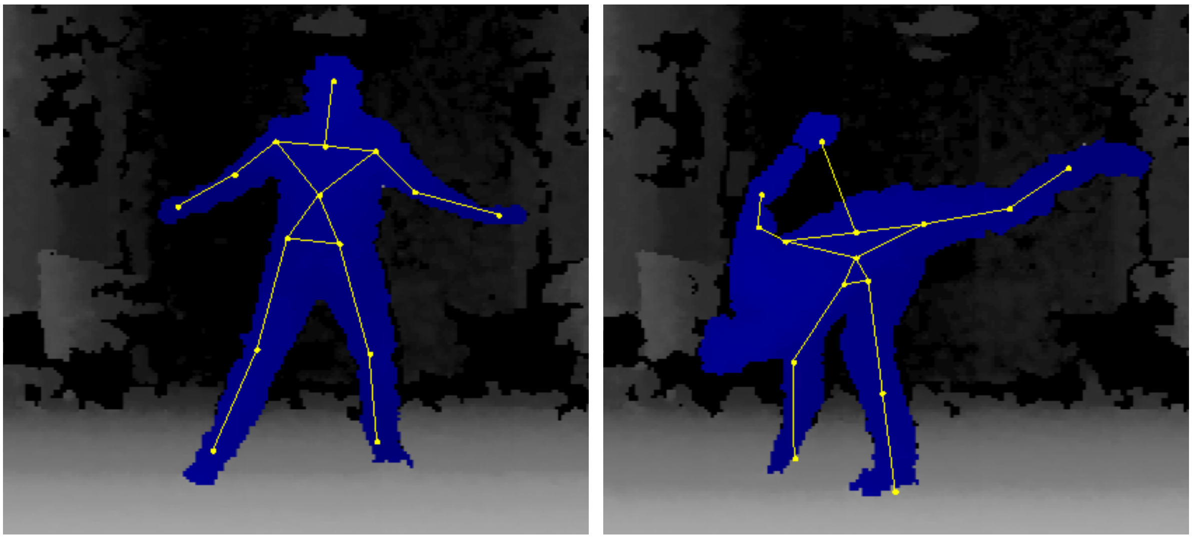

Figure 1: A good and a bad example are shown using the OpenNI tracker. The bad example shows for instance the left hand is mistaken for the head.

2 Literature Study

2.1 Introduction

Traditionally motion tracking has been done using motion capture systems involving electromagnetic markers placed on key parts of the body. This method has two disadvantages: it is expensive and intrusive. By using a marker-less vision-based system we can capture a depth image that we then use to estimate a pose. Using vision-based systems we can not only expand the areas where pose estimation can occur, such as surveillance and human-computer interaction, but also make it cheaper and easier.

We provide an overview of the different methods and techniques used to do such vision-based pose tracking that are currently being used. Each approach to the problem has many different components, which are grouped into various categories. First, a method is used to acquire depth images. Then, as described in [Poppe, 2007], there are two approaches: model-based and model-free. In the model-based approach we construct a number of mathematical models which describe constraints or likelihoods on the space of poses. In accordance with these, we then use an estimation method to calculate best-matching poses based on the depth images. Model-free methods are not bound to these mathematical models, and are instead learning- or example-based.

2.2 Acquisition of depth images

With the advent of depth capturing cameras in recent years, the ability to acquire depth images for use in motion capturing has been simplified since we don't need to have multiple camera sources, and thus also more affordable. The Kinect camera from Microsoft uses a technique called *structured light*, which projects an IR grid over the scene and measures the deformation of this grid due to the scene. In [Shotton et al., 2011] they employ a Kinect camera that uses this technique for their data acquisition.

Another technique using light called *time-of-flight* (ToF) measures the delay in light after projecting it to the scene. Based on this it can accurately and quite quickly construct a depth image of the scene. In the paper [Ganapathi et al., 2010] they use a ToF camera for their depth images.

2.3 Model-based methods

The purpose of model-based methods is to describe constraints or likelihoods on poses.

2.3.1 Body models

Shape models describe how a pose is translated to a body surface given a number of parameters. For example in [Ganapathi et al., 2012], each joint has an associated radius that defines the form of the connecting capsules. Another type of body model introduces constraints on possible poses, for instance the allowed length of links or angles (degrees of freedom) between joints.

2.3.2 Measurement models

A measurement model describes the likelihood of a given depth measurement given a pose. As such it requires a shape model, described above. In the estimation step we can then use this likelihood to find the optimal pose given a depth image. There are two main approaches:

ICP

The Iterative Closest Point (ICP) algorithm defines a function that maps single depth measurements to the nearest point on the body surface for a given pose. The likelihood of a measurement is higher the closer it is to the body surface. This method was used efficiently in [Grest et al., 2005], though they assume a fixed skeletal size and starting position.

Ray Casting

The Ray Casting model is slightly different. The likelihood of a depth measurement is based on how close it is to the depth the given pose would theoretically have. This is a more accurate model, but harder to optimize. In [Ganapathi et al., 2012], a combination of ICP and Ray Casting is used, of which they have an efficient optimization scheme.

2.3.3 Transition models

Transition models are used in the case of having a series of depth images taken closely after each other. They introduce constraints on how poses are allowed to evolve over time. We generally limit the change in joint positions and length of links according to some parameters to enforce smooth movement.

2.4 Estimation

Estimation is used to find a pose that minimizes the error between observation and the pose projected to the body shape model. It is also possible to use a learning-based projection function or example set instead of a shape model. Estimation methods are divided into top-down and bottom-up.

2.4.1 Top-down

A top-down approach tries to match various pose estimates with the observation, and pick the best one according to the models. This can be done using gradient ascent on the pose likelihood starting from an initial pose, as in [Ganapathi et al., 2012]. There are two disadvantages of this method. The first is the initialization of the first pose before any subsequent pose estimation can be done. The second is the computational cost associated with estimating the pose of the whole body.

2.4.2 Bottom-up

A bottom-up method estimates the individual body parts and then assembles the human body from those, taking into account the body constraints. This method requires different body part detectors to minimize false matches, since there can be many parts of the depth image that can look limb-like. The Microsoft Research paper [Shotton et al., 2011] uses a bottom-up method based on classifying each depth measurement with a Randomized Decision Forest.

2.5 Model-free methods

Methods that don't use a body model instead build a direct relation between depth images and poses. There are two main approaches.

2.5.1 Learning-based

Machine learning algorithms can be trained for this purpose, using training data consisting of depth images and their respective poses. An example, though based on intensity images, is [Agarwal and Triggs, 2006]. Given an image, they calculate local gradient histograms, followed non negative matrix factorization, to learn a set of bases that correspond to local body features, and estimate pose by direct regression.

2.5.2 Example-based

Example-based methods also use a database of depth images and their respective poses, but instead of learning the relation, perform a similarity search on the input image to retrieve a number of candidate poses, and interpolate these to return a single pose estimate. This kind of method is used by [Mori and Malik, 2006], who employ shape contexts to encode the edges of objects.

2.6 Results

We will briefly present mention some of the recent results in pose tracking. [Shotton et al., 2011] were the first to use a Kinect camera to produce depth images. They have compared their results with the then state-of-the-art method used by [Ganapathi et al., 2010]. Their algorithm outperformed Ganapathi's and was also 10 times faster. The precision for different body parts can be seen in figure 2. One of the shortcomings of Shottons method is its failure to detect uncommon poses such as hand stands or cart wheels. Since the method is learning-based it needs to be trained with these poses.

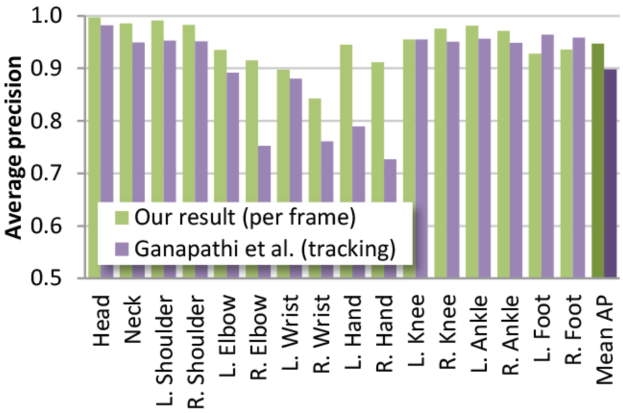

Figure 2: Comparison results from [Shotton et al., 2011]

[Ganapathi et al., 2012] used a different approach for motion capturing than in their 2010 paper and compared it to their earlier results. This method doesn't use any example set or learning data, and is supposedly able to detect poses in various difficult situations, such as cart wheels. It is also overall more accurate and 16 times faster than [Shotton et al., 2011]. The results can be seen in figure 3. They have managed to outperform their old method in tracking accuracy for every body part.

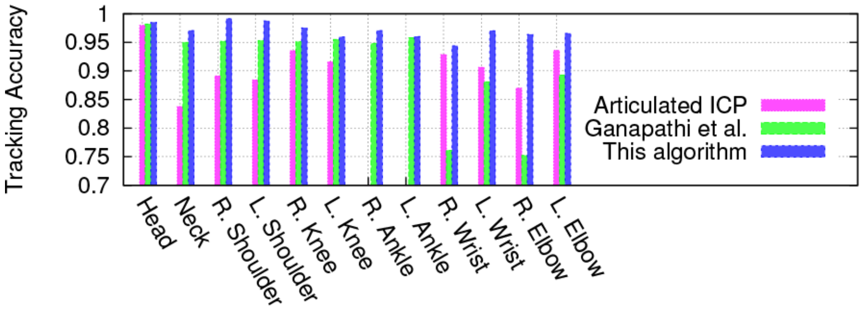

Figure 3: Results from [Ganapathi et al., 2012]

References

[Agarwal and Triggs, 2006] Agarwal, A. and Triggs, B. (2006). A local basis representation for estimating human pose from cluttered images.

[Ganapathi et al., 2010] Ganapathi, V., Plagemann, C., Koller, D., and Thrun, S. (2010). Real time motion capture using a single time-of-flight camera.

[Ganapathi et al., 2012] Ganapathi, V., Plagemann, C., Koller, D., and Thrun, S. (2012). Real-time human pose tracking from range data.

[Grest et al., 2005] Grest, D., Woetzel, J., and Koch, R. (2005). Nonlinear body pose estimation from depth images.

[Mori and Malik, 2006] Mori, G. and Malik, J. (2006). Recovering 3d human body configurations using shape contexts.

[Poppe, 2007] Poppe, R. (2007). Vision-based human motion analysis: An overview.

[Shotton et al., 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images.