Analyzing gait using a time-of-flight camera

Rasmus R. Jensen, Rasmus R. Paulsen, and Rasmus Larsen

Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
{raje, rrp, rl}@imm.dtu.dk, www.imm.dtu.dk

Abstract. An algorithm is created, which performs human gait analysis using 3-dimensional data from a *Time-of-flight* camera. For each frame in a sequence the camera supplies cartesian coordinates in space for every pixel in the frame. By using an articulated model the subject pose is estimated in the depth map in each frame. The pose estimation is based on likelihood, gradient, smoothness and a shape prior used to solve a Markov random field. Based on the pose estimates, and the prior that movement is locally smooth, a sequential model is created, and a gait analysis is done on this model. The output data are: Speed, Cadence (steps per minute), Step length, Stride length (stride being two consecutive steps also known as a gait cycle), and Range of motion (angles of joints). The created system produces good output data of the described output parameters and requires no user interaction.

-Keywords: *Time-of-flight camera, Markov random fields, gait analysis, computer vision*

1 Introduction

Recognizing and analyzing human movement in computer vision can be used for different purposes such as biomechanics, biometrics and motion capture. In biomechanics it helps us understand how the human body functions, and if something is not right it can be used to correct this.

Top athletes have used high speed cameras to analyze their movement either to improve on technique or to help recover from an injury. Using several high speed cameras, bluescreens and marker suits an advanced model of movement can be created, which can then be analyzed. This optimal setup is however complex and expensive, a luxury which is not widely available. Several approaches aim to make simpler tracking of movement.

Using several cameras but without bluescreens nor markers [10] creates a visual hull in space from silhouettes by solving a spacial Markov random field using graph cuts and then fitting a model to this hull.

Based on a large database [8] is able to find a pose estimate in sublinear time relative to the database size. This algorithm uses subsets of features to find the nearest match in parameter space.

An Earlier study uses the *Time-of-flight* (*TOF*) camera to estimate pose using key feature points in combination with a an articulated model to solve problems with ambiguous feature detection, self penetration and joint constraints [12].

To minimize expenses and time spend on multi camera setups, bluescreens, markersuits, initializing algorithms, annotating etc. this article aims to deliver a cheap alternative that analyzes gait.

Using the *Posecut* algorithm [4] on output from a *TOF* camera with no restrictions on neither background nor clothing a system is presented that can deliver a gait analysis with a simple setup and no user interaction. The project object is to broaden the range of patients benefiting from an algorithmic gait analysis.

2 Introduction to the algorithm finding the pose

This section will give a brief overview of the algorithm used to solve the problem of finding the pose of the subject. To do a gait analysis the pose has to be estimated in a sequence of frames. This is done using the *Posecut* algorithm on the depth stream provided by a *TOF* camera [1]. The *Posecut* algorithm uses 4 terms to define an energy minimization problem and find the pose of the subject as well as segmenting between subject and background:

Likelihood term: This term is based on statistics of the background. It is based on a probability function of a given pixel being labeled background.

Smoothness prior: This is a prior based on the general assumption that data is smooth. Neighbouring pixels are expected to have the same label with higher probability than having different labels.

Gradient term: Neighbouring pixels with different labels are expected to have depth values that differs from one another. If the values are very similar but the labels different, this is penalized by this term.

Shape prior: Trying to find the pose of a human, a human shape is used as a prior.

2.1 Random fields

A frame in the sequence is considered to be a random field. A random field consists of a set of discrete random variables $\{X_1, X_2, \ldots, X_n\}$ defined on the index set I. In this set each variable X_i takes a value x_i from the label set $L = \{L_1, L_2, \ldots, L_k\}$ presenting all possible labels. All values of x_i, $\forall i \in I$ are represented by the vector $\mathbf{x}$ which is the configuration of the random field and takes values from the label set L^n. In the following the labeling is a binary problem, where $L = \{\text{subject}, \text{background}\}$.

A neighbourhood system to X_i is defined as $N = \{N_i | i \in I\}$ for which it holds that $i \notin N_i$ and $i \in N_j \Leftrightarrow j \in N_i$. A random field is said to be a Markov field, if it satisfies the positivity property:

$$P(\mathbf{x}) > 0 \qquad \forall \mathbf{x} \in L^n \tag{1}$$

And the Markovian Property:

$$P(x_i|\{x_j : j \in I - \{i\}\}) = P(x_i|\{x_j : j \in N_i\}) \tag{2}$$

Or in other words any configuration of $\mathbf{x}$ has higher probability than 0 and the probability of x_i given the index set $I - \{i\}$ is the same as the probability given the neighbourhood of i.

2.2 The likelihood function

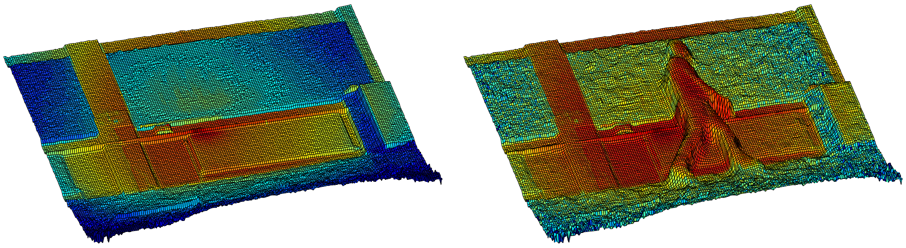
The likelihood energy is based on the negative log likelihood and for the background distribution defined as:

$$\Phi(\mathbf{D}|x_i = \text{background}) = -\log p(\mathbf{D}|x_i) \tag{3}$$

Using the Gibbs measure without the normalization constant this energy becomes:

$$\Phi(\mathbf{D}|x_i = \text{background}) = \frac{(\mathbf{D} - \mu_{\text{background,i}})^2}{\sigma^2_{\text{background,i}}} \tag{4}$$

With no distribution defined for pixels belonging to the subject, the subject likelihood function is set to the mean of the background likelihood function. To estimate a stable background a variety of methods is available. A well known method, models each pixel as a mixture of Gaussians and is also able to update these estimates on the fly [9]. In our method a simpler approach proved sufficient. The background estimation is done by computing the median value at each pixel over a number of frames. Figure 1(a) shows the background depth model, while Figure 1(b) shows a frame with the subject.



(a) Background model

(b) Subject depth image

Fig. 1. Depth images of background and subject, both images are rotated to emphasize the spatial properties.

2.3 The smoothness prior

This term states that generally neighbours have the same label with higher probability, or in other words that data are not totally random. The generalized Potts model where $j \in N_i$ is given by:

$$\psi(x_i, x_j) = \begin{cases} K_{ij} & x_i \neq x_j \\ 0 & x_i = x_j \end{cases} \tag{5}$$

This term penalizes neighbours having different labels. In the case of segmenting between background and subject, the problem becomes binary and is referred to as the Ising model [3]. The parameter K_{ij} determines the smoothness in the resulting labeling.

2.4 The contrast term

It is expected that two adjacent pixels with the same label have similar camera distances, which implies that adjacent pixels with different labels have different distances. By decreasing the cost of neighbouring pixels with different labels exponentially with an increase in difference in intensity, this term favours neighbouring pixels with similar distance to have the same label. This function is defined as:

$$\gamma(i,j) = \lambda \exp\left(\frac{-g^2(i,j)}{2\sigma^2_{background,i}}\right) \tag{6}$$

Where $g^2(i,j)$ is the gradient in the depth map and approximated using convolution with gradient filters. The parameter λ controls the cost of the contrast term, and the contribution to the energy minimization problem becomes:

$$\Phi(\mathbf{D}|x_i, x_j) = \begin{cases} \gamma(i,j) & x_i \neq x_j \\ 0 & x_i = x_j \end{cases} \tag{7}$$

2.5 The shape prior

To ensure that the segmentation is human like and wanting to estimate a human pose, a human shape model consisting of ellipses is used as a prior. The model is based on measures from a large Bulgarian population study [7], and the model is simplified such that it has no arms, and the only restriction to the model is that it cannot overstretch the knee joints. The hip joint is simplified such that the hip is connected in one point as studies shows that a 2D model can produce good results in gait analysis [2]. Pixels near the shape model in a frame are more likely to be labeled subject, while pixels far from the shape are more likely to be background.

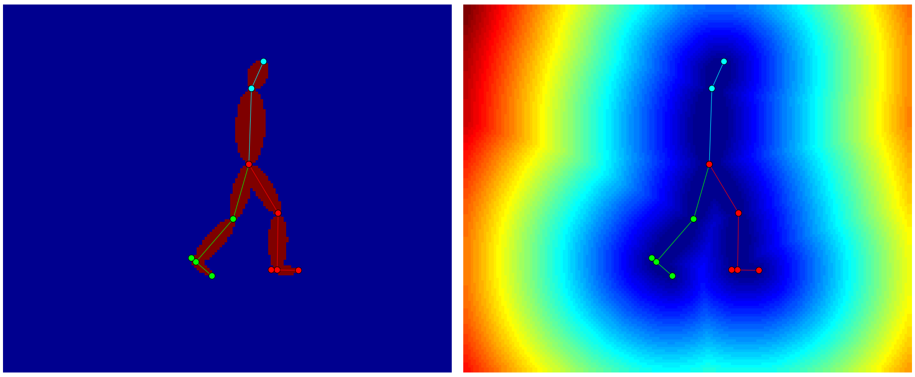
The cost function for the shape prior is defined as:

$$\Phi(x_i|\mathbf{\Theta}) = -\log(p(x_i|\mathbf{\Theta})) \tag{8}$$

Where $\mathbf{\Theta}$ contains the pose parameters of the shape model being position, height and joint angles. The probability $p(x_i|\mathbf{\Theta})$ of labeling subject or background is defined as follows:

$$p(x_i = \text{subject}|\mathbf{\Theta}) = 1 - p(x_i = \text{background}|\mathbf{\Theta}) = \frac{1}{1 + \exp(\mu * (\text{dist}(i, \mathbf{\Theta}) - d_r))} \tag{9}$$

The function $\text{dist}(i, \mathbf{\Theta})$ is the distance from pixel i to the shape defined by $\mathbf{\Theta}$, d_r is the width of the shape, and μ is the magnitude of the penalty given to points outside the shape. To calculate the distance for all pixels to the model, the shape model is rasterized and the distance found using the *Signed Euclidian Distance Transform* (*SEDT*) [11]. Figure 2 shows the rasterized model and the distances calculated using the *SEDT*.



(a) Rasterized model

(b) Distance map

Fig. 2. Raster model and the corresponding distance map.

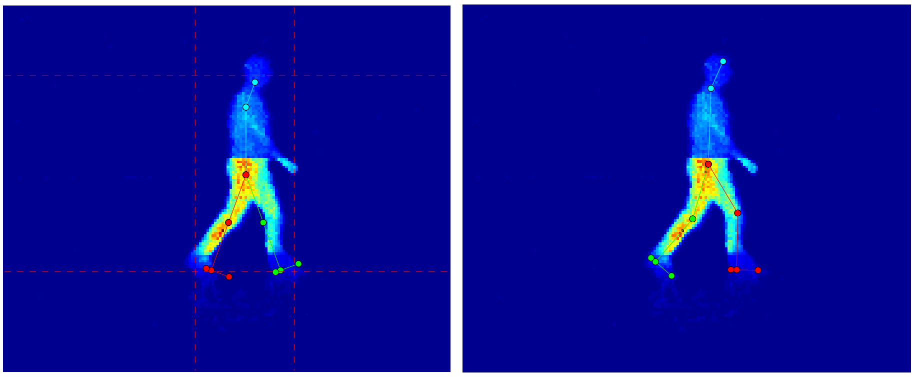
2.6 Energy minimization

Combining the four energy terms a cost function for the pose and segmentation becomes:

$$\Psi(\mathbf{x}, \mathbf{\Theta}) = \sum_{i \in V} \left(\Phi(\mathbf{D}|x_i) + \Phi(x_i|\mathbf{\Theta}) + \sum_{j \in N_i} (\psi(x_i, x_j) + \Phi(\mathbf{D}|x_i, x_j)) \right) \tag{10}$$

This Markov random field is solved using *Graph Cuts* [5], and the pose is optimized in each frame using the pose from the previous frame as initialization. To find an initial frame and a pose, the frame that differs the most from the background is chosen based on the background log likelihood function. As a rough guess on where the subject is in this frame, the log likelihood is summed first along the rows and then along the columns. These two sum vectors are used to guess the first and last rows and columns that contains the subject (Fig 3(a)). From the initial guess the pose is optimized according to the energy problem by searching locally. Figure 3(b) shows the optimized pose, notice that the legs change place during the optimization. This is done based on the depth image such that the closest leg is also closest in the depth image (green is the right side in the model), which solves an ambiguity problem in silhouettes.

The pose in the remaining frames is found using the previous frame as an initial guess and then optimizing this. This generally works very well, but problems sometimes arise when the legs pass each other as feet or knees of one leg tend to get stuck on the wrong side of the other leg. This entanglement is avoided by not allowing crossed legs as an initial guess and instead using straight legs close together.



(a) Initial guess

(b) Optimized pose

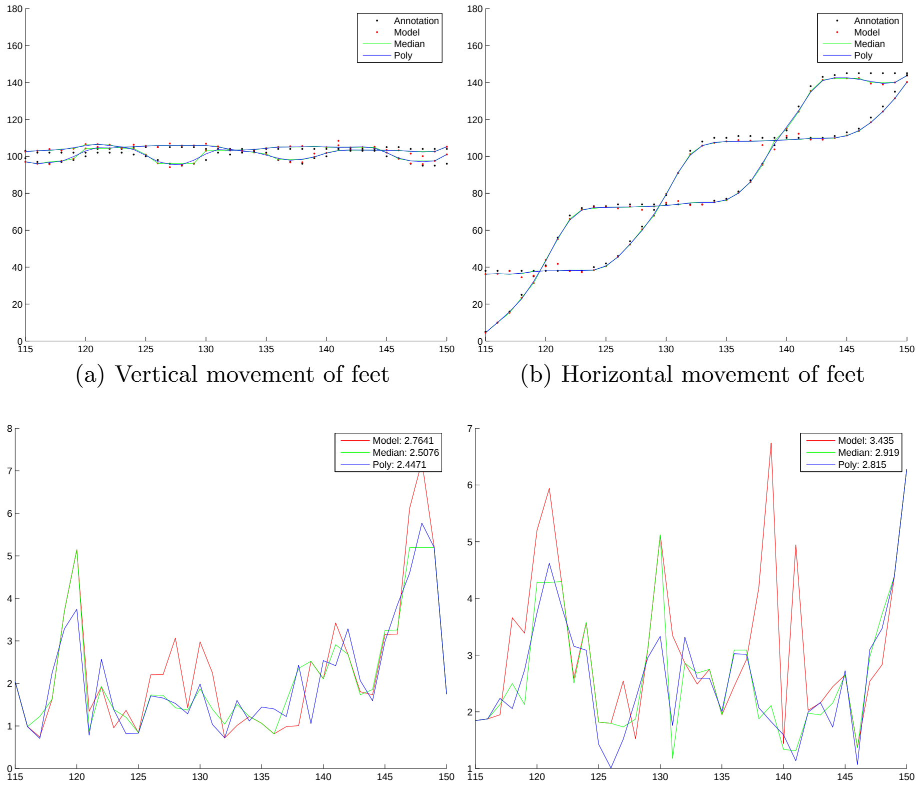
Fig. 3. Initialization of the algorithm.

3 Analyzing the gait

From the markerless tracking a sequential model is created. To ensure local smoothness in the movement before the analysis is carried out a little postprocessing is done.

3.1 Post processing

The movement of the model is expected to be locally smooth, and the influence of a few outliers is minimized by using a local median filter on the sequences of point and then locally fitting polynomials to the filtered points. As a measure of ground truth the foot joints of the subject has been annotated in the sequence to give a standard deviation in pixels of the foot joint movement. Figure 4 shows the movement of the feet compared to the annotated points and the resulting error. The figure shows that the curve fitting of the points gives an improvement on the accuracy of the model, resulting in a standard deviation of only a few pixels. If the depth detection used to decide which leg is left and which is right fails in a frame, comparing the body points to the fitted curve can be used to detect and correct the incorrect left right detection.



(a) Vertical movement of feet

(b) Horizontal movement of feet

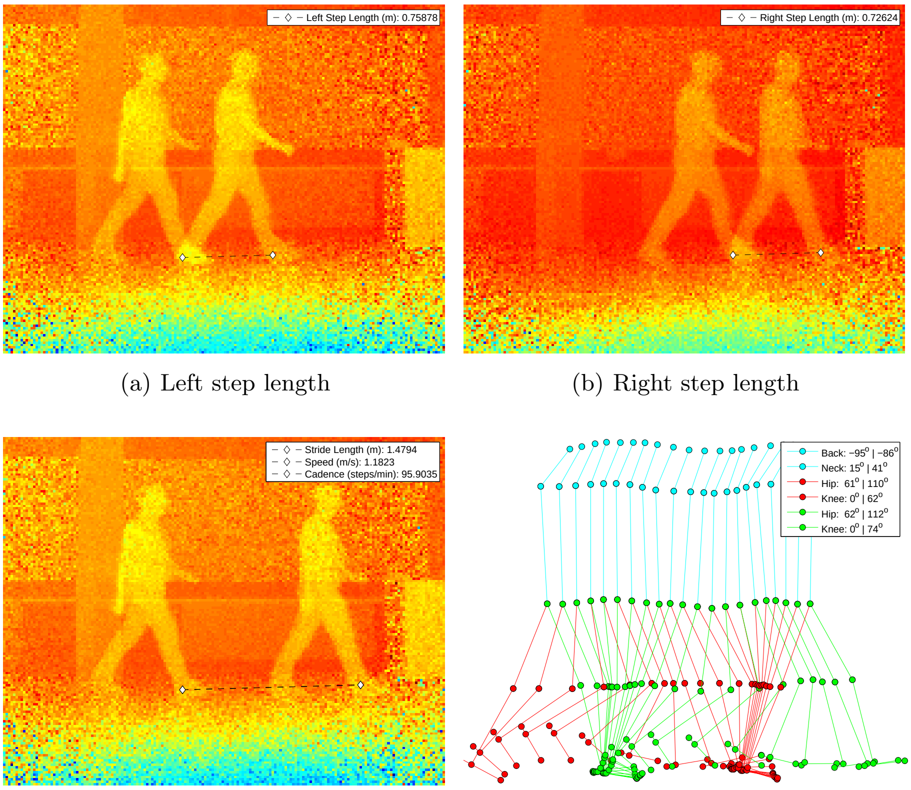
(c) Error of right foot

(d) Error of left foot

Fig. 4. 4(a) shows the vertical movement of the feet for annotated points, points from the pose estimate, and for curve fittings. 4(b) shows the points for the horizontal movement. 4(c) shows the pixelwise error for the right foot for each frame and the standard deviation for each fitting. 4(b) shows the same but for the left foot.

3.2 Output parameters

With the pose estimated in every frame the gait can now be analyzed. To find the steps during gait, the frames where the distance between the feet has a local maximum are used. Combining this with information about which foot is leading, the foot that is taking a step can be found. From the provided Cartesian coordinates in space and a timestamp for each frame the step length (Fig. 5(a) and 5(b)), stride length, speed and cadence (Fig. 5(c)) are found. The found parameters are close to the average found in a small group of subjects aging 17 to 31 [6], even though based only on very few steps and therefore expected to have some variance, this is an indication of correctness. The range of motion is found as the clockwise angle from the x-axis in positive direction for the inner limbs (femurs and torso) and the clockwise change compared to the inner limbs for the outer joints (ankles and head). Figure 5(d) shows the angles and the model pose throughout the sequence.



(a) Left step length

(b) Right step length

(c) Stride length, speed and cadence

(d) Range of motion

Fig. 5. Analysis output.

4 Conclusion

A system is created that autonomously produces a simple gait analysis. Because a depth map is used to perform the tracking rather than an intensity map, there are no requirements to the background nor to the subjects clothes. No reference system is needed as the camera provides a such. Compared to manual annotation in each frame the error is very little. For further analysis on gait the system could easily be adapted to work on a subject walking on a treadmill. The adaption would be that there is no longer a movement in space (it is the treadmill conveyor belt moving) hence speed and stride lengths should be calculated using step lengths. With the treadmill adaption averages could be found of the different outputs as well as standard deviations.

Currently the system uses a 2 dimensional model and to optimize precision in the joint angles the subject should move in an angle perpendicular to the camera. While the distances calculated depends little on the angle of movement the joint angles have a higher dependency. The dependence of the joint angles could be minimized using a 3 dimensional model. It does however still seem reasonable that the best results would come from movement perpendicular to the camera, whether using a 3 dimensional model or not.

The camera used is the SR-3000 at a framerate of about 18 Fps, which is on the low end in tracking movement, which is why a better precision could be obtained with a higher framerate. Due to the fact that movement from one frame to the next will be relatively shorter, the processing time would not be augmented greatly, bearing in mind that the pose from the previous frame is used as an initialization for the next.

References

1. Mesa. Website, 2008. www.mesa-imaging.ch.
2. E. B. Alkjaer, T. Simonsen and P. Dygre-Poulsen. Comparison of inverse dynamics calculated by two- and three-dimensional models during walking. *2001 Gait and Posture*, pages 73–77, 2001.
3. Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.
4. M. Bray, P. Kohli, and P.H.S. Torr. Posecut: simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. *Computer Vision-ECCV 2006. 9th European Conference on Computer Vision. Proceedings (Lecture Notes in Computer Science Vol.3952)*, pages 642–55, 2006.
5. V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.
6. Mark D. Latt, Hylton B. Menz, Victor S. Fung, and Stephen R. Lord. Walking speed, cadence and step length are selected to optimize the stability of head and pelvis accelerations. *Experimental Brain Research*, 184(2):201–209, 2008.
7. Gergana Stefanova Nikolova and Yuli Emilov Toshev. Estimation of male and female body segment parameters of the bulgarian population using a 16-segmental mathematical model. *Journal of Biomechanics*, 40(16):3700–3707, 2007.

8. G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 750–757 vol.2, 2003.
9. C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, 2:246–252 Vol. 2, 1999.
10. Chengkai Wan, Baozong Yuan, and Zhenjiang Miao. Markerless human body motion capture using Markov random field and dynamic graph cuts. *Visual Computer*, 24(5):373–380, 2008.
11. Q.-Z. Ye. The signed Euclidean distance transform and its applications. *[1988 Proceedings] 9th International Conference on Pattern Recognition*, pages 495–499 vol.1, 1988.
12. Youding Zhu, B. Dariush, and K. Fujimura. Controlled human pose estimation from depth image streams. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 1–8, 2008.