Articulated Tracking of Humans by Video Technology

Nico van der Aa[1,2], Coert van Gemeren[2], Remco Veltkamp[2], Lucas Noldus[1]

[1]Noldus Information Technology, Wageningen, The Netherlands. n.vanderaa@noldus.nl, l.noldus.noldus.nl

[2]Utrecht University, Utrecht, The Netherlands. c.j.vangemeren@uu.nl, r.c.veltkamp@uu.nl

Abstract

Measuring a person's behavior in any kind of domain by a computer system, starts with measuring and analyzing that person's movements. Video technology provides an unobtrusive way of capturing this information. In this paper, we will focus on articulated tracking, which is the field of estimating and tracking the body pose of a person. A pose consists of the joints and rigid parts of a person's skeleton at each frame. Alternatively, we use only the body orientation to estimate the direction at which a person is focused. We discuss the findings of our research on state-of-the-art methods in articulated tracking and body orientation estimation techniques with a distinct sensor setup to give researchers an idea of the challenges they might face for their application.

Introduction

To create tools for measuring human behavior in an automated way, the system must detect, track and analyze human motion. Camera technology provides a way to measure movements and activities in a scene unobtrusively. The field of capturing human motion is an important field within computer vision which is far from solved. A camera gives an array of pixels, including RGB colors for video cameras, or depth information for structured light cameras. The algorithm will provide details of the detected people as joint positions and orientations (e.g. of the torso). The main challenges are (1) to separate people in the foreground from the background by defining pixel regions which are unaffected by the endless amount of variability caused by clothing and background similarities, (2) to identify the limbs and cope with the symmetry of the kinematic structure of the human body, (3) to handle occlusions by objects, other people and self-occlusions, and (4) obtain these results in (near) real-time.

For articulated tracking, the literature divides these approaches in *model-based* (or *generative*) approaches versus *model-free* (or *discriminative*) approaches. Model-based approaches rely on an explicitly known parametric human model, and match the image observations to this predefined model. The kinematics of the model provide the basic restrictions for the human shape, such as the tree-structured kinematic constraints between adjacent body parts (e.g. torso-upper half-limb connection). In contrast, model-free approaches estimate a pose directly from observation, without using an accurate 3D model. They use the fact that the set of typical human poses is far smaller than the set of kinematically possible ones and train a model that directly recovers poses from observable image quantities. As our intention is to develop a tool to measure human motion in any kind of application, we restrict ourselves to model-based approaches. From literature, we select three model-based approaches based on a different sensor setup: (1) a single video camera; (2) a system of multiple calibrated cameras and (3) a depth sensor using structured light. The underlying idea is that the sensor and algorithm choice depends heavily on the application. As an example, a depth sensor has a range up to 8 meters and the infrared light principle will not work in an outdoor environment. In this paper we share our findings with the selected approaches.

For some applications, articulated tracking is not possible and also not needed. Think of a shopping center where many people walk around arbitrarily and we only want to know which direction they are facing or in which direction they are moving. Therefore, we have studied a state-of-the-art method for body orientation estimation.

Articulated tracking

In our research we consider three types of sensor setups to capture human motion: (1) a single static video camera, (2) a system of multiple static and calibrated video cameras, and (3) depth sensors using structured light. As we are interested in developing a general tool that is independent of the application, there are no constraints imposed on the pose to be detected. To keep it feasible, only a single person is assumed to be in the scene.

Monocular video camera

A single camera only provides a projection of the real world on a 2D image plane. The best strategy is to capture the projection of the skeleton on the camera view, which is the so-called 2D human pose. Pictorial structures is a model-based technique to estimate such 2D human poses. A pictorial structures model for a human being consists of a collection of body parts with connections between certain pairs of parts. It is a class of graphical models where the nodes of the graph represent body parts, and edges between parts encode pairwise geometric relationships. The appearance term (node) models the probability of a part being present at a particular location and orientation given the input image. A prior models the probability distribution over the pose, constraining the estimated pose to be plausible in terms of human articulation. To enable efficient inference, often two (unrealistic) assumptions are made: (i) the appearance of a part is assumed independent of its pose and that of the other parts; (ii) the prior over pose is a Gaussian with 'tree-structured' covariance. However, the main challenge for 2D pose estimation is the depth ambiguity, which occurs for example when an arm is behind the torso.

To estimate the 2D pose from videos, we analyzed the method of Ramanan *et al.* (2007), as it was tested on videos from various sources containing a wide range of activities and showing promising detection results at body part level accuracy. It employs pictorial structures in a tracking framework which includes a model-building phase and a detection phase. During the model-building phase the system selects a frame where the pose is distinctly present and all body parts are visible. This way it learns a model of the person who is to be tracked in the video sequence. In the detection phase, this model is applied to detect the current state of the pose in each frame of the video sequence. The main assumption of this method is that in both modules of the system, the scale of the person to be detected is known beforehand and it must remain approximately constant throughout the video. The system will not detect people wearing skirts, dresses or loose clothes, as body parts are modelled by rectangles. An example of the output is shown in Figure 1.

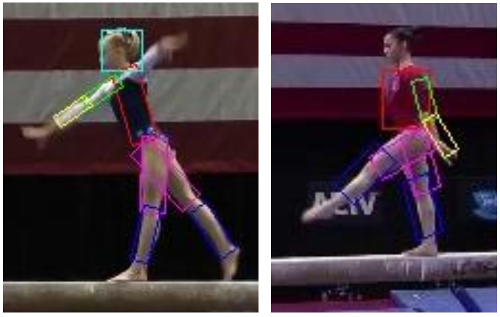

Figure 1. Examples of monocular pose estimation.

We analyzed the original method in terms of robustness and performance with respect to the type of input videos and found that, in contrast to what is stated in the original paper, the method is highly dependent on the set of input parameters, which do not translate across different videos. We also showed situations where one set of parameters can lead to different results within the same video, and we found that the motion model can eliminate the necessity to manually adjust these parameters when processing a single video. As a conclusion, the method under investigation does not solve the problem of occlusions as common by monocular pose estimation. The reader is referred to Ursu (2013) for more details.

Multiple static and calibrated cameras

Another way of handling occlusions is by including more camera views. Parts of the person that are occluded in one camera view, may still be visible in other views. If the cameras are static and calibrated, we know where they are with respect to each other and we can link their views to the real world.

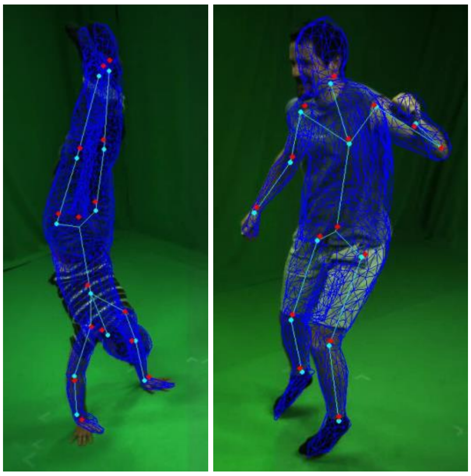

Figure 2. Examples of multi-view pose estimation. The white dots are our results and the red ones are the results of Gall *et al.* (2009)

We analyzed and extended the method presented by Gall *et al*. (2009), which is also a model-based pose estimation technique. The method estimates the pose by fitting a 3D model of the observed person to the images from the camera views. It accurately finds the skeleton pose and from this pose the mesh deformation is calculated by a three step algorithm. In the first two steps the motion of the skeleton and its shape are estimated. These steps consist of a combination of two optimization approaches: a local and a global optimization. For the local optimization a weighted least squares problem is solved to minimize the distance between the model and the silhouettes. The global optimization uses a combination of particle filters and simulated annealing to estimate the pose where the local optimization fails. Given the mesh model, camera calibration data and the camera video sequences, an accurate pose is estimated. The last step enhances the shape of the model to match the fine details of the observed person (like loose fitting clothes such as a robe or a skirt) in the images.

Our implementation of this method provides accurate results as shown in Figure 2, but fails in some cases for small body parts like hands and feet. Both the local and global optimization are based on silhouettes, so a proper background subtraction is required. The main challenge is the computational effort. Especially the global optimization step requires many iterations, causing the time taken to estimate the pose in each frame to be in the order of minutes. For a more detailed discussion on this method see Resodikromo (2012).

Depth sensor

With the introduction of structured light cameras like the Microsoft Kinect sensor, cheap alternative ways of capturing a scene have become available. Instead of 2D features or 3D matching, 3D depth information is now directly available. SDKs like NiTE (www.openni.org/files/nite/) include simple tools for body pose tracking. Although the applications are restricted to situations where the person is fully in the field of view and at most 8 meters away from the camera, NiTE provides a fast and relatively stable method for pose estimation. Similarly to the previous methods, the difficulties lie with the detection of the limbs as their appearance is the least distinctive and have the smallest dimensions, which result in wrong or lost detections of wrist and elbow joints.

In our research on 3D hand and finger tracking Koetzier (2014), we applied the 3Gear tool (www.threegear.com). Although it is necessary to put the camera within a meter from the hands, it is possible to combine the skeleton tracking with finger tracking. As the development of these SDKs will continue and successors of the depth cameras will follow rapidly, this provides a good starting point for applications within the requested constraints.

Body orientation

Instead of estimating the full body pose, which still faces many open issues, we can also restrict ourselves to finding the body orientation, which often is a sufficient cue for analyzing a person's activity including the focus of attention. Based on Chen and Odobez (2012), the body orientation of multiple human targets is estimated from a video sequence, captured by the view from a single moving camera. Accomplishing this goal requires a few stages, including human body detection and tracking. Additional computation, such as determining real-world 3D position coordinates of the targets, as well as its velocity and direction, can improve the results.

Estimating human body orientations can be formulated as a classification task with multiple classes of body angles. In the paper we discuss here, 8 angle classes are defined. We introduce a method that incorporates a set of different classifiers and cues, allowing us to be more flexible in choosing the classification methods, and to have the best results obtained from the combined response from several classifiers (committee). The input are video frames on which human tracking is performed using the method described in Choi *et al.* (2012), which is preferred over other tracking methods, as the input is allowed originate from a single moving camera. Besides, the method is able to provide estimates for the positions of the human targets in real world coordinates. This information is particularly useful to determine the velocity direction and magnitude, of the targets. Aside from the coordinates of the targets, the method also returns bounding boxes of the targets. From these bounding boxes Histogram of Gradient descriptors are extracted. These are then supplied to several pre-trained classifiers such as a Neural Network and a Support Vector Machine, which give probability estimates for each of the 8 angle classes.

In our implementation (see Ichim *et al.* (2014)) we add the face as an important cue to the overall orientation estimation, since it restricts the plausible angles. To include this information, face detection is performed on the bounding boxes. To maintain the consistency of the probabilistic framework, a uniform distribution, based on the presence or absence of a face, is generated. Other information is gained from velocity direction and magnitude, which is integrated in the framework by fitting a standard Gaussian distribution, centered on the velocity direction of an angle class in such a way that a relatively high velocity yields a high probability for the frontal direction, and a low probability for the other directions. A relatively low velocity yields the same probability for all directions. The response from all of the mentioned classifiers and additional cues are combined, and the final angle estimation the one with the highest probability. However, the final result is filtered using a sliding window to ensure temporal smoothness of the change in orientation over time. Compared to Chen and Odobez (2012), we were able to reduce the average error by more than 20 degrees and reduce the computation time by 400 times making this method near real-time. An example of the output is given in Figure 3.

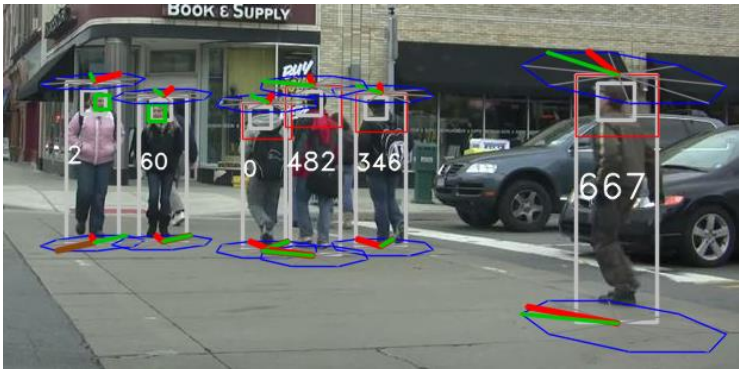

Figure 3 Example of the body orientation method. Top green denotes the head orientation, bottom green the body orientation. The red lines denote the temporal smoothened directions of the head and body, respectively.

Concluding remark

Human pose estimation from video technology remains a challenging field in computer vision due to occlusions in general, the dimensions of the different body parts and their variability in appearance. Obtaining human body orientation is feasible in real-time.

Acknowledgement

This research was supported by the FES project COMMIT/ and the NeurAS research project, which is part of IMDI-NeuroControl.

References

1. Chen, C., & Odobez, J. (2012). We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video. *IEEE Conference on Computer Vision and Pattern Recognition 2012.*

2. Choi, W., Pantofaru, C., & Savarese, S. (2012). A general framework for tracking multiple people from a moving camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(7).

3. Gall, J., Stoll, C., Aguiar, E. d., Theobalt, C., Rosenbahn, B., & Seidel, H.-P. (2009). Motion capture using joint skeleton tracking and surface estimation. *IEEE Conference on Computer Vision and Pattern Recognition.*

4. Ichim, M., Tan, R., Aa, N. van der, & Veltkamp, R. (2014). Human body orientation estimation using a committee based approach. *9th International Conference on Computer Vision Theory and Applications.* Lisbon.

5. Koetzier, M. (2014). *3D hand tracking using depth sensors.* Master thesis, Utrecht University, Information and Computing Science, Faculty of Science. Retrieved 22-6-2014, from < http://dspace.library.uu.nl/handle/1874/291559>

6. Ramanan, D., Forsyth, D., & Zisserman, A. (2007). Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(1), 65-81.

7. Resodikromo, J. (2012). *Marker-less 3D pose estimation.* Master thesis, Utrecht University, Information and Computing Sciences, Faculty of Sciences. Retrieved 22-6-2014, from <http://dspace.library.uu.nl/handle/1874/255396>

8. Ursu, E. (2013). *Pose Estimation in Video.* Master thesis, Utrecht University, Information and Computing Sciences, Faculty of Science. Retrieved 22-6-2014, from <http://dspace.library.uu.nl/handle/1874/282663>