RAGDOLLS IN ACTION – ACTION RECOGNITION BY 3D POSE RECOVERY FROM MONOCULAR VIDEO

Verena Klinger and Michael Arens
FGAN-FOM
Gutleuthausstraße 1, 76275 Ettlingen, Germany
{klinger, arens}@fom.fgan.de

ABSTRACT

We present a novel approach to reconstruct and track articulated objects, specifically humans, in 3D from monocular videos for action recognition, by combining techniques from both image processing and 3D computer animation. The goal is to establish a system that is able to recognize basic actions (like walk, run) from frame to frame in a scene with more than one person. In a first step a feature-based detection algorithm extracts hints of body parts from input videos. Second, a virtual world is established in which the extracted 2D measurements are translated into 3D rays. A ragdoll - a model used to represent realistic human bodies in 3D animation - is attached to a bundle of rays and it adjusts itself with the help of physics simulation. Third, the ragdoll's pose is used to query a motion capture database in order to obtain predictions for the next time frame and to build the basis for action recognition.

KEYWORDS

pose reconstruction, people tracking, action analysis, monocular video, ragdoll, physics simulation

1. INTRODUCTION AND RELATED WORK

Reconstructing articulated objects from 2D video data in 3D is difficult, because many ambiguities have to be solved. The loss of information due to projecting the 3D world onto the 2D image plane makes it hard to find a plausible reconstruction in 3D. A lot of work has been done on the area of 3D pose reconstruction, both for multiple and monocular views.

Some approaches seek to recover the pose from monocular video data with use of silhouettes. Agarwal and Triggs (2006) present a learning based approach without an explicit body model. The pose is recovered by a direct non-linear regression against shape descriptor vectors which are extracted from silhouettes. Sminchisescu et al (2005) and Howe (2007) use motion capture data to synthesize training pairs to learn a mapping from images to poses. Zhao and Liu (2008) search solutions with a genetic algorithm in a PCA reduced solution space.

Body parts are used to recover the 3D pose in the following works. Barrón et al (2001) take manually selected landmarks of an observed person in an image to estimate anthropometry and pose, while Mori and Malik (2006) use pre-labeled example images to automatically extract joint positions. In Urtasun et al (2006) a body part tracker, initialized by hand, is used as basis for a generative tracking of the 3D pose in a latent variable space. A view-based body model is fitted into body part proposals in Sigal and Black (2006).

Delamarre and Faugeras (2001) use physical forces to drive the projection of their 3D model towards the extracted silhouettes in multiple views simultaneously. They need at least two calibrated cameras to realize the reconstruction. Vondrak et al (2008) use edge and silhouette information to track the 3D pose in one or more views. Physics simulation is employed to choose predictions which are physically plausible.

We think that an appropriate human body model and background knowledge have to be integrated into the reconstruction process. Therefore, we combine ideas from computer vision and computer animation in order to reconstruct and track 3D-poses from monocular video. Action recognition is based on the integration of motion capture data. We assume that the background in the video data is subject to rapid changes, as we

want to apply our system to outdoor settings with a moving camera. This assumption makes techniques based on background subtraction and silhouettes created from foreground segmentation impractical, therefore we use a feature based approach to detect and track humans in monocular videos in a first step. As a result we get consistent person tracks over time, which consist in a set of body parts.

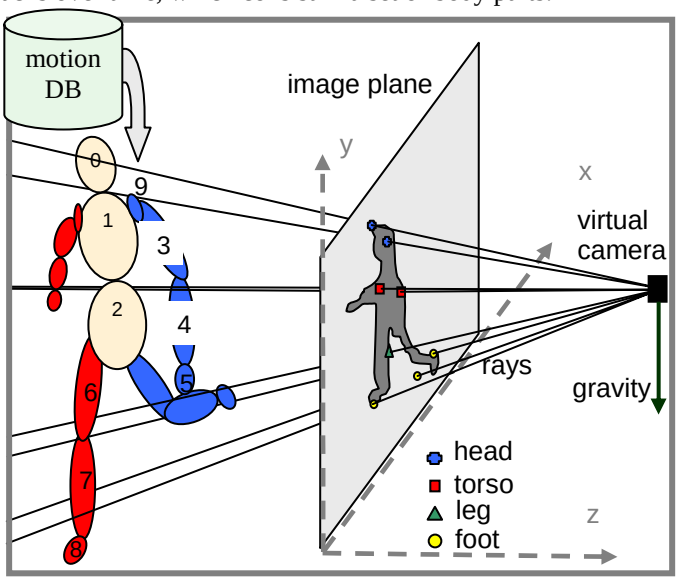


Figure 1. Overview of the system: rays are drawn from the virtual camera through the detected body parts. The ragdoll adapts to the ray bundle and the predictions from the data base. The ragdoll consists of rigid bodies that are labeled: head (0), spine (1), pelvis (2), left/right upper arm (3), left/right lower arm (4), left/right hand (5), left/right upper leg (6), left/right lower leg (7), left/right foot (8), left/right shoulder (9).

In step two, a virtual world is set up with a virtual camera, see figure 1 for an overview. The 2D-coordinates of body parts for each person hypothesis are transformed into 3D-rays having their origin in the camera center of the virtual world. These rays form a ray bundle for which a ragdoll modeling the kinematics of humans is initialized and attached to. We use the publicly available Bullet Physics Library (www.bulletphysics.com) to simulate the ragdolls' dynamics according to constraints, forces and collision events acting on them. In step three, predictions are extracted from a database holding real motion sequences stemming from motion capture data. This step also establishes a sound basis for action recognition.

We focus on the task of recognizing the correct action for each person in an image and not on an exact reconstruction of the pose. Right now action classes are restricted to *stand*, *walk*, *walk_run* and *run*. The reconstructed pose may differ from the one seen in the picture, especially in terms of left-right assignment. We also want to be able to reconstruct all people seen in a video.

In this paper we want to give insights into step two and three of our experimental system. The feature based tracking is based on the approach of Leibe et al (2008). They use an Implicit Shape Model (ISM) to detect instances of the object class of interest in images. Jüngling and Arens (2009) use this technique to detect people. They extend the ISM in the way that extracted features which build the code book during training are labeled with corresponding body parts to integrate body part classification into people detection. For our application this people detector is extended with a tracking-by-detection approach to get consistent tracks over time.

2. RAGDOLLS IN ACTION

Our ragdoll model consists of 17 rigid bodies connected by joints, with one, two, or three degrees of freedom (DOF). The joints are modeled as constraints limiting the DOF to ranges, for which human motion is plausible. The ragdoll's movement is controlled by physics simulation that translates forces and collision events acting on a ragdoll into linear and angular velocities, taking joint constraints into consideration. Figure 2 gives an overview of the reconstruction process described in the next paragraph.

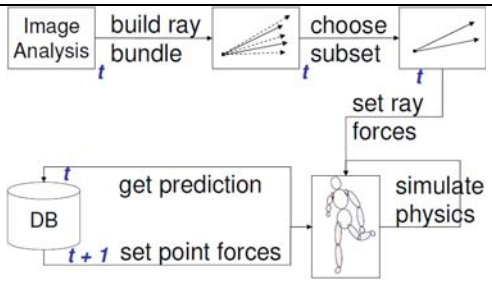


Figure 2. For frame t, body parts from the image analysis module build a ray bundle, from which a subset is chosen. Ray forces are set; a prediction from the database is fetched resulting in point forces. Ragdoll dynamics is simulated by the physics engine.

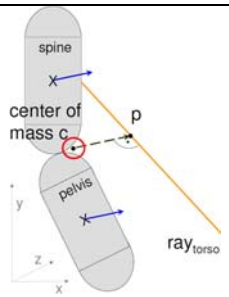


Figure 3. Ray force acts on the centers of involved body parts, directed from center of mass *c* to the point *p* of shortest distance.

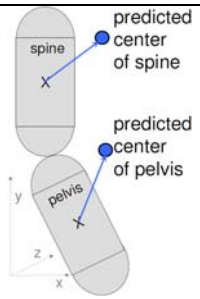


Figure 4. Point force acts on body part, directed to the predicted center.

The image analysis module produces a hypothesis for each detected person, tracked over time. The hypothesis consists of a set of image coordinates of body parts. There can be more than, e.g., two feet in such a set, while other parts like hands might not be found at all. A ray bundle is built for every hypothesis (compare figure 1). A subset of the ray bundle is chosen such that it contains one head, two feet, one torso, two legs etc. The choice is based on the distance from the corresponding ragdoll's body part to the ray. The closer the ray and the higher its reliability factor, the more likely that it joins the subset. Every ray in this subset results in a *ray force* that is directed from the center of mass *c* of involved body parts to the ray. Its length is proportional to the distance of *c* to point *p* on the ray, see figure 3. This force then acts on the centers of involved body parts. A ray labeled with *head* corresponds to the ragdoll's rigid body also labeled with *head*. The ray labeled with *torso* corresponds to the rigid bodies labeled with *pelvis* and *spine*; in this case the force is distributed on several body parts. Then the physics simulation translates forces into velocities and the ragdoll moves accordingly. After predefined number of simulation steps, the relative angles between the ragdoll's body parts are calculated to obtain a pose representation independent from global position and size and length of body parts. This *angle representation* is used to query the database (DB) that contains motion capture data. We use the CMU motion capture database (http://mocap.cs.cmu.edu) and recalculate it to meet the same angle representation used for our ragdolls. Every CMU motion sequence is labeled with the activity it represents, thus we take one sequence for *stand*, *walk*, *walk_run* and *run*, respectively. *walk_run* represents an action between walk and run comparable to slow jogging. These labels are also saved in our DB for each pose of a motion sequence. Now, querying the DB with an angle representation results in a set of nearest neighbors based on a distance measure on angles. A nearest neighbor represents an angle representation of a real pose in the DB and its action label. We take the angle representation with minimal distance to the query as prediction and calculate the predicted body part centers in 3D. With *point forces*, which act on each body part of the ragdoll and are directed to the predicted center of body parts, the prediction is integrated into the simulation, see figure 4. As all predictions carry a label, we get an implicit action classification for each frame. Right now, the predictions are uncorrelated from frame to frame, as we do not consider past predictions for the next choice. This results, too, in temporally independent action classification and action recognition.

3. RESULTS AND CONCLUSION

We tested our system on a short sequence of the PETS (2006) dataset (www.pets2006.net) that shows a girl running through the scene from left to the right, while two persons, the one occluding the other, enter the scene walking from right to left. Later, another person enters the scene walking from left to the right. The girl is represented by hypothesis 0, the next two persons cannot be separated by the image analysis module and therefore result in the single hypothesis 1, while the last person results in hypothesis 2.

Figure 5 illustrates the ragdolls in action. The screenshot of frame 57 shows all three hypotheses with the corresponding ragdolls. The other pictures depict the ragdolls for hypotheses 0 and 1 over four consecutive frames The black lines represent the subset of active rays in the ray bundle. The ragdolls move according to

measurements and predictions from frame to frame. If no measurements are available, as e.g. in frame 60 for hypothesis 0, then the prediction keeps the ragdoll in pose. Therefore the prediction compensates for measurement failures. It furthermore realizes action classification, rather implicitly, as it carries the action label of the original CMU motion.

The graphs in figure 6 show the assigned action classes for each hypothesis in a temporal order to give an impression of the continuity of action classification. Hypothesis 0 is labeled incorrectly for some frames after the initialization period, but then stabilizes at *run*. Separating the class *walk_run* and *walk* seems to be difficult for hypothesis 1. As body parts for this hypothesis are derived from two people, the estimation process is error prone. In contrast, the graph for hypothesis 2 tends to be smoother. Some jumps in the graphs can presumably be eliminated by applying an online filter of appropriate window size. The table in figure 6 shows how often an action class was chosen for each hypothesis. Note that for hypothesis 1 action recognition fails more often, because incoming body parts stem from two persons tracked as one hypothesis by the image analysis module.

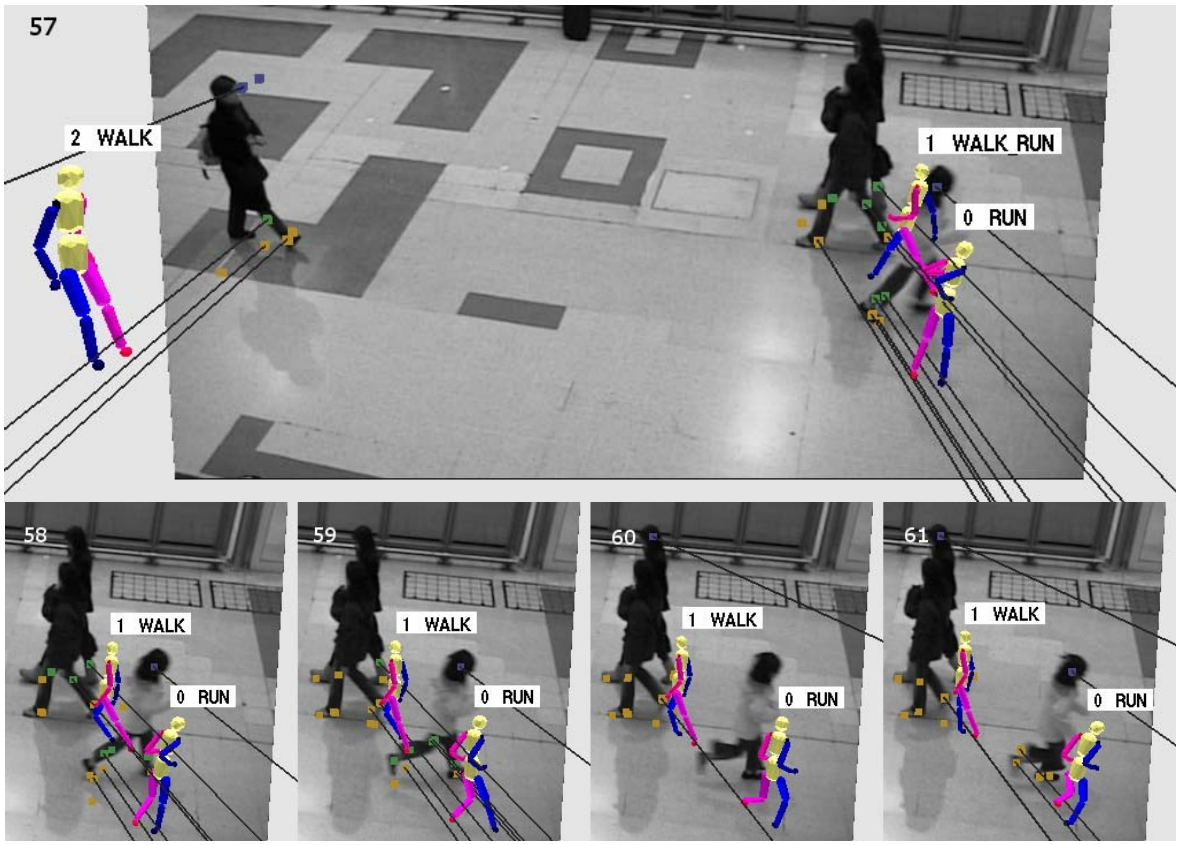
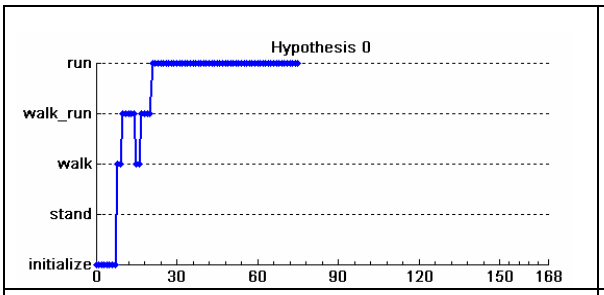
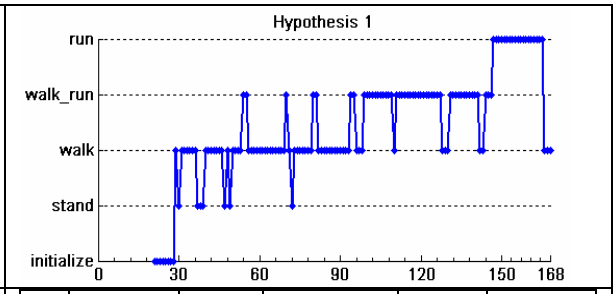


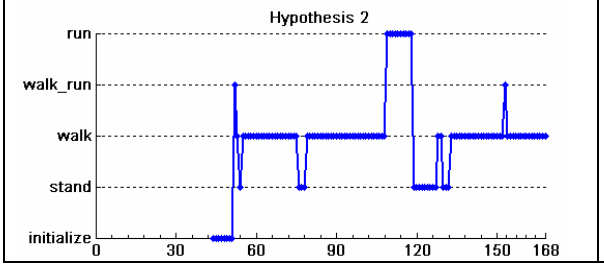
Figure 5. The screenshot of frame 57 depicts all three ragdolls with a label giving their id and current action. Note that only a subset of image features is used to build the ray bundle for each ragdoll. The smaller pictures below show ragdolls of hypotheses 0 and 1 for frames 58 – 61. In frame 60, there are no features at all for the girl, thus the ragdoll is kept in pose by the prediction only.

Using only one camera, we do have problems in placing the ragdolls correctly in 3D. Also, left and right get mixed up. The information on body parts is far from complete; therefore the reconstructed pose does not fit exactly the one in the image. Nevertheless we can recognize the performed action to some degree over time. Action recognition works when persons move from one side in the video to the other. Front and back movements are hard to distinguish, as the extracted body parts - and how we use them now - are not stable enough. The body parts are newly detected in every frame, as we do not use a tracking strategy here. Still, we reach up to ~80% of correctly labeled actions for correctly tracked persons without filtering the labels or underlying DB-predictions over time. For the biased case of hypothesis 1 where features of two persons are mixed, the rate of correct labels drops to ~46%.

We hope to get more stability by temporally filtering the predictions and by back-propagating information of the ragdolls' pose into the image analysis module.







id	stand	walk	walk_run	run	frames
	#	#	#	#	#
0	0	4	9	55	68
1	7	65	49	19	140
2	16	89	2	10	117
	%	%	%	%	
0	0	5.9	13.2	80.9	
1	5	46.4	35	13.6	
2	13.7	76.1	1.7	8.5	

Figure 6. The three graphs show the assigned action class per frame for each hypothesis in temporal order. The table gives the frequency of an action class for each hypothesis id as number of frames and as percentage. Colored regions depict the correct classes.

REFERENCES

A. Agarwal and B. Triggs, 2006. Recovering 3D Human Pose from Monocular Images. *Pattern Analysis & Machine Intelligence*, 28(1), pp. 44 – 58.

C. Barrón and I. A. Kakadiaris, 2001. Estimating Anthropometry and Pose from a Single Uncalibrated Image. In *Computer Vision and Image Understanding*, 81(3), pp. 269–284.

CMU Graphics Lab Motion Capture Database, http://mocap.cs.cmu.edu/

Q. Delamarre, O. Faugeras, 2001. 3D Articulated Models and Multiview Tracking with Physical Forces. In *Computer Vision and Image Understanding*. 81(3), pp. 328-357.

N. R. Howe, 2007. Silhouette Lookup for Monocular 3D Pose Tracking. In *Image and Vision Comp.*, 25(3), pp. 331–341.

K. Jüngling and M. Arens, 2009. Feature Based Person Detection Beyond the Visible Spectrum *Conference on Computer Vision and Pattern Recognition, Workshop OTCBVS.* To appear.

B. Leibe, A. Leonardis, and B. Schiele, 2008. Robust Object Detection with Interleaved Categorization and Segmentation. *International Journal of Computer Vision*, 77(1-3), pp.259–289.

G. Mori and J. Malik, 2006. Recovering 3D Human Body Configurations Using Shape Contexts. *Pattern Analysis & Machine Intelligence*, 28(7), pp.1052–1062.

PETS 2006 data sets, 2006. http://www.pets2006.net/

L. Sigal, M. Black, 2006. Predicting 3D People from 2D Pictures. *Conference on Articulated Motion and Deformable Objects*, pp. 185-195.

C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, 2005. Discriminative Density Propagation for 3D Human Motion Estimation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 390–397.

Bullet Physics Library, http://www.bulletphysics.com/Bullet/wordpress/

R. Urtasun, D. Fleet, and P. Fua, 2006. 3D People Tracking with Gaussian Process Dynamical Models. *Conference on Computer Vision and Pattern Recognition*, pp. 238-245.

M. Vondrak, L. Sigal, and O. C. Jenkins, 2008. Physical Simulation for Probabilistic Motion Tracking. *Conference on Computer Vision and Pattern Recognition*, pp. 1-8.

X. Zhao, Y. Liu, 2008. Generative Tracking of 3D Human Motion by Hierarchical Annealed Genetic Algorithm. In *Pattern Recognition.* 41(8), pp. 2470-2483.