FROM LEARNING TO UNDERSTANDING

USING MOTION INFORMATION TO RECOGNIZE AND RECONSTRUCT HUMAN ACTIONS FROM VIDEOS

Kuehne H., University of Karlsruhe (TH) – KIT, kuehne @ira.uka.de

Abstract:

Knowing what a person does and having the structural information to analyze and reproduce the ongoing motion can be seen as the one of the major research objectives in the field of video-based motion recognition. To allow both, the following work proposes a way to build up recognition and pose reconstruction as two parallel processing steps. The idea is to use the same input data, but to process it differently leading to two different outputs whose interpretation can be linked on different levels.

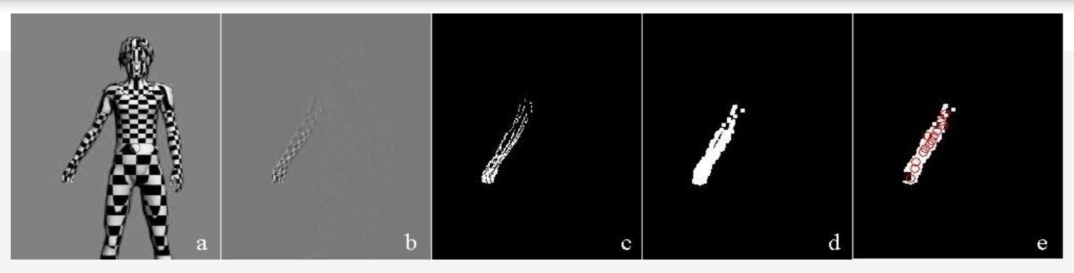
In this context, motion data gained from videos images can be seen as a good, abstract but also very general representation. As psycho-physiological experiments[1] show, can even a few moving points comprise enough information to recognize and reconstruct human actions, allowing even to analyze higher level characteristics like distinguishing between male and female. For recognition it is critical to find discriminative features that form a description which can be used as an input vector for machine learning algorithms. Here, histograms of sparse feature motion can be used for a very fast and robust learning and recognition of human actions. Reconstruction, however, is depending on information that can be used for model-based pose reconstruction. The output is a model configuration as well as the related joint angle trajectories that can be used for further analysis. Here, moving feature points can be used for optimizing the model configuration for the ongoing motion in a least-square model fitting approach, allowing to estimate the current pose from the preceding and successive motion. The combination of both aspects covers the knowledge of what goes on and how it is done. It allows to understand the action, to analyze and to compare it, as well as to interpret and even reproduce it.

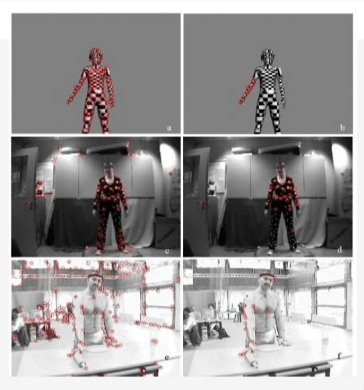
MOTION BASED FEATURE TRACKING

Extract motion information - track single moving points only

- Restrict the number of features to track to only the moving ones
- Detected regions with change of color, brightness or intensity

→ for intensity-based methods like Lucas-Kanade-feature tracking





Explanation of the algorithm:

a) Original image *I(t)*
b) Subtract *I(t)* and *I(t+1)* to obtain difference image I_{diff}
c) Binarize difference image I_{diff}
d) Dilate difference image → e) Result: Final mask image with new feature points

ACTION RECOGNITION

= „… classifying the captured motion as one of several types of actions…"[2] → find discriminative features for machine learning

Action recognition from sparse feature flow

a) Oriented histograms from sparse feature flow

Global histogram of overall motion directions
The weighted histogram for frame t is calculated from the motion vector of the feature points of images *I(t)* and *I(t+1)*.
→motion direction θ and motion intensity γ
Motion directions are weighted with norm value.
One bin of the histogram:
based on the motion angle γ.
k-th bin = sum of the intensity of all motion vectors with direction from $\left[\frac{1}{n}(k2\Pi),\frac{1}{n}((k+1)2\Pi)\right]$

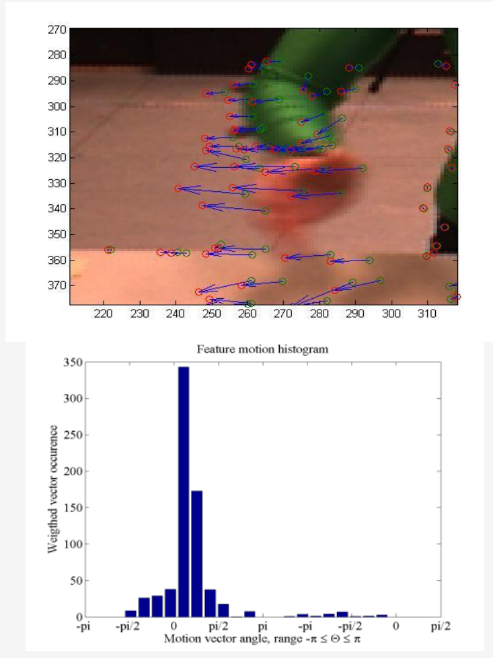


Fig.: Example for feature motion and bar plot of weighted motion histogram

b) Action units and grammar

- Complex tasks = concatenated action units

→the smallest entity, whose order can be changed

- The order is formulated by a grammar,

→ defines the action sequences = concatenation of action units

[Pouring water into a bowl] Rest position* - Take bowl - Take bottle - Pour - Put bottle back - Put bowl back - Rest position* (* = optional)

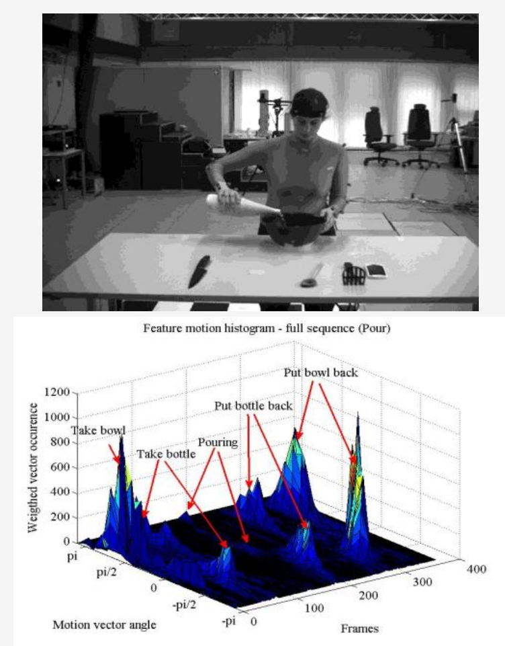
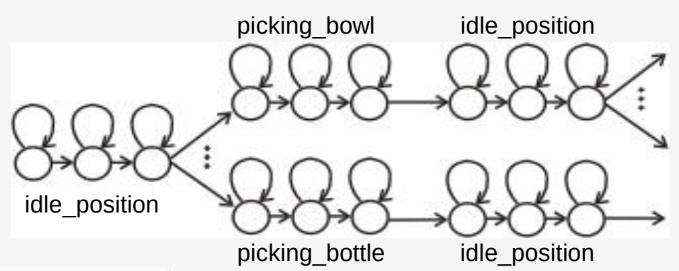


Fig: Example and feature motion histogram distribution for action sequence 'Pouring'

c) Recognition of action sequences

- low level modeling of action units using 4-state left-to-right HMMs
- action sequence = concatenation of action unit HMMs
- combined with a context free grammar,
- implicit automatic segmentation of the action sequences into action units



	5 sequences		10 sequences	
	HoOF	HoFF	HoOF	HoFF
Sequence recognition	100.0 %	100.0 %	100.0 %	100.0 %
Unit recognition with grammar	96.9 %	97.5 %	97.7 %	96.6 %
Unit recognition	89.7 %	89.0 %	90.4 %	78.4 %

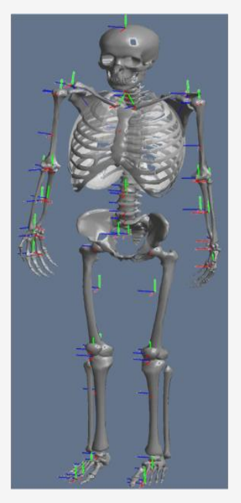
Tab.: Comparison of recognition performance of optical flow (HoOF) and feature flow (HoFF) based systems without and with action grammar for 5 sequences and 10 sequences

POSE RECONSTRUCTION

= "…Finding out how the individual limbs are configured in a given scene…" [2] → fit body model to the point cloud to estimate the current configuration

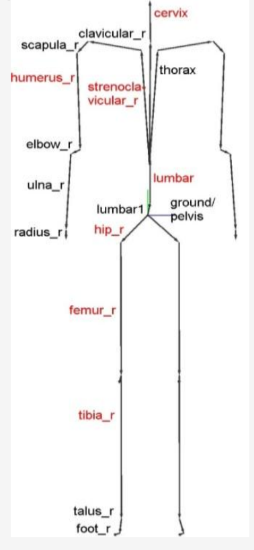
Least-square joint angle estimation from point clouds

- Inspired by motion capture → Consider stereo feature points as noisy, unlabed marker data
- Optimize body pose over:
  - least distance
  - least distance variation
- … over n frames



Human Body Model
•full definition of the human body:
maximum of 108 degrees of freedom
• following real human joint kinematic
•reduced model with
-35 body segments
-34 body joints
-44 degrees of freedom

Example with reduced degrees of freedom (DOFs)



Main Idea

Optimize the degrees of freedom of the model so, that the distance of the actual point set $x_0$and the result of the forward kinematics of the reconstructed pose x becomes minimal:

$$\min \sum_{x\in X} (d(x,x_0)\times weight(x_0))^2 \quad (1.1)$$

$$d(x,x_0)=\sqrt{(x_a-x_{0a})^2+(x_b-x_{0b})^2+(x_c-x_{0c})^2}$$

With X: the set of the result of the forward kinematics
$weight(x_0)$ the weight of the markers

Minimizing a large scale non linear function with bounds:

$$\min_{x\in\Re^n} f(x), l \le x \le u \quad (1.2)$$

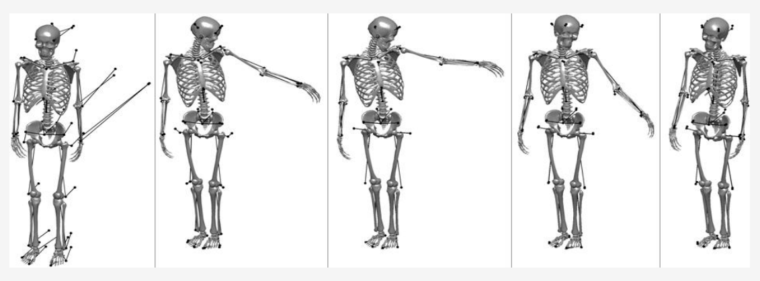


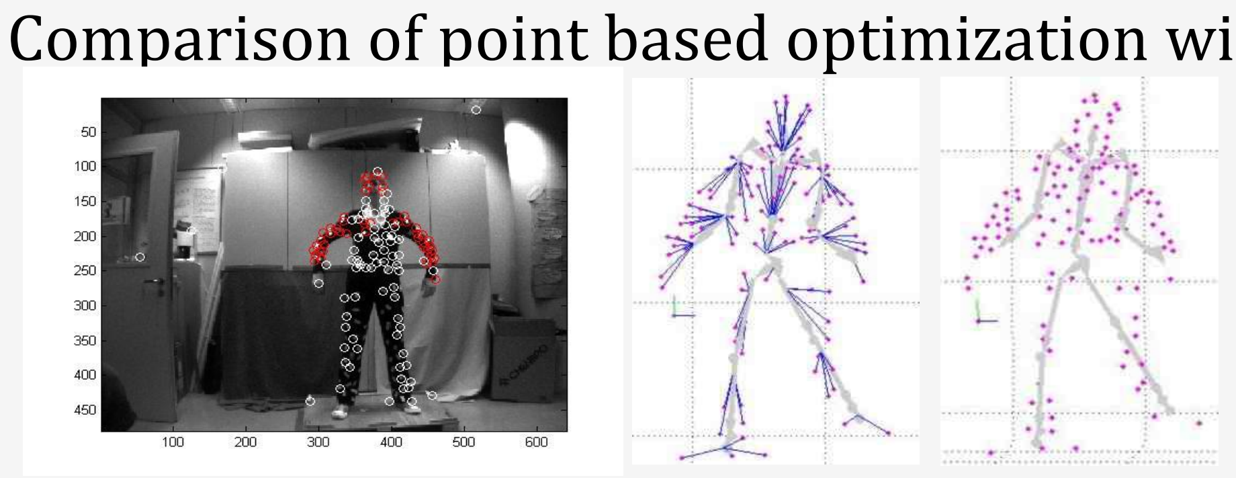
Fig: Reconstructed joint angles of a pointing gesture

Continuous integration vs. re-initialization

- First frame: distance-based optimization only
- N-th frame: optimization for known feature points + distance-based optimization for new features
- If overall error > thresh, reinitialize

Evaluation

Comparison of point based optimization with marker-based results



Pose reconstruction: Primary results

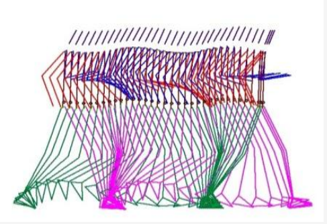
Coming up: evaluation with marker based pose reconstruction

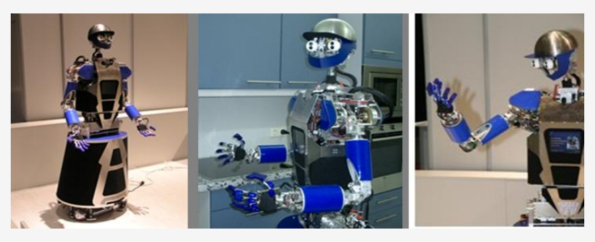
COMBINED INTERPRETATION

Result:

1) Estimated performed action + motion units
2) Joint angle trajectories for every action unit

→ Basis for further motion analysis e.g. for professional training or HCI
→ Basis for motion reproduction e.g. for humanoid robotics





Btw …. why using motion data? :

- abstract and general representation of action
- independent from many environmental conditions
- high level of anonymity

→Can be used for recognition and reconstruction

References

[1] Johansson, G., 1973. Visual perception of biological motion and a model for its analysis, Perception & Psychophysics, Vol. 14, No. 2, pp. 201 - 211.

[2] Moeslund, T.B., Granum, E., 2001. A survey of computer vision-based human motion capture, Computer Vision and Image Understanding, Vol. 81 , No. 3, March 2001, pp. 231 – 268.

Acknowledgements

This work was supported by the grant from the Ministry of Science, Research and the Arts of Baden-Württemberg.