Thomas Walther
Rolf P. Würtz

Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany
[thomas.walther,rolf.wuertz]@neuroinformatik.rub.de
http://www.neuroinformatik.ruhr-uni-bochum.de/

SPP 1183

Learning to look at humans - what are the parts of a moving body?

Abstract

We present a system that can segment articulated, non-rigid motion without a priori knowledge of the number of clusters present in the analyzed scenario. We combine existing algorithms for tracking and extend clustering techniques by a self-tuning heuristic. Application to video sequences of humans shows good segmentation into limbs.

1 Motivation

Despite considerable effort creating an artificial system capable of analyzing human body pose and motion is still an open challenge. Such a *pose estimation system* (PE-system) would enable machines to communicate with their users in a more natural way (body language interpretation) or to survey activities of individuals to anticipate their intentions (traffic/security). Existing PE-Systems are by far no match for the human brain when it comes to the task of motion and pose estimation, let alone behavior interpretation. These systems are narrowly tuned to their field of application and work with relatively inflexible, pre-defined models of human shape and motion. In our project, we attempt to create a PE-System which initially has no idea of its environment and the humans inhabiting it. Instead, it should gather knowledge during its lifetime and build up its own environmental and human model, like the human visual cortex does at some time in its development. For this, we combine state-of-the-art computer vision techniques and biologically inspired principles like (controlled) self-organization and machine learning.

2 Feature tracking

Tracking human motion with sparse features provides an acceptable tradeoff between accuracy and computational effort. Using the sparse tracking technique described in [1] together with the feature initialization techniques of [2] we are able to produce feasible tracking of features through all frames of our image sequences. Since tracking quality is strongly dependent on the degree of saliency of the tracked features, we used easily trackable clothing in all of our training experiments with real data (figure 1).
When clustering feature trajectories produced by the feature tracking stage, our system relies on the technique of self-tuning spectral clustering, proposed by [3]. Nevertheless, several improvements in order to use their technique for motion segmentation had to be made which are described below.

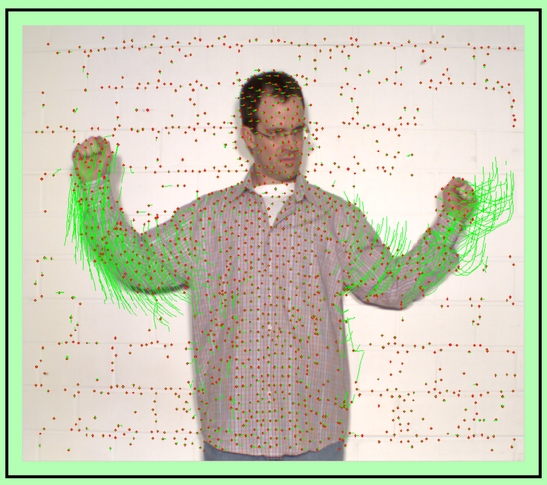

Figure 1: Feature tracking via the KLT-Algorithm [1], [4]

3 Trajectory distance measure

Since trajectories represent a spatio-temporal sequence of image coordinates, changing the distance measure proposed by [3] is inevitable. We use a distance measure inspired by [5]:

$$a_{ij} = \alpha \sum_{t=A}^{B} (\Delta_{ij}^t - \overline{\Delta_{ij}})^2 + \beta \sum_{t=A}^{B-1} \left\| \mathbf{v}_i^t - \mathbf{v}_j^t \right\|^2 \qquad (1)$$

with $\Delta_{ij}^t = \left\| \mathbf{x}_i^t - \mathbf{x}_j^t \right\|$, $\mathbf{v}^t{}_i = \mathbf{x}_i{}^t - \mathbf{x}_i{}^{t-1}$, and $\overline{\Delta_{ij}}$ is the mean of Δ_{ij} over frames $[A \dots B]$.

4 Heuristic scale selection

The clustering quality of the spectral clustering method proposed by [3] strongly depends on a local kernel parameter, σ_i. This paramter is found using a heuristic that turned out to be insufficient for our purposes. Therefore, we adopted a novel heuristic well-adapted to our demands:

1. Let **E** be a trajectory data source with N elements, consider element e_i in frame t.
2. Define the Delaunay-neighborhood of e_i as $N_i = [n_1 \dots n_{k_i}]$, consisting of those k_i data elements which are neighbors of e_i in a Delaunay-Triangulation in $\mathbf{R}^2$ constructed over the positions of all elements in **E** at time t.
3. For each element e_i build up a sorted list $L_i = [v_1, \dots v_{k_i}]$ with $a_{iv_1} \leq a_{iv_2} \leq \dots a_{iv_{k_i}}$ and $v_1 \dots v_{k_i} \in N_i$, where a_{ij} is the Euclidean distance between data element e_i and e_j at time t.
4. When element e_i moves in an arbitrary way, it is highly likely that its M closest Delaunay-neighbors $L_i[1 \dots M]$ will move coherently with e_i. In our experiments, we used $M = 3$.
5. Each element e_i owns a sorted list T_i of trajectory distances between itself and all other elements of **E**: $T_i = [a_{i1}, \dots, a_{iN}]$ with $a_{i1} \leq a_{i2} \dots \leq a_{in}$. a_{ij} is identical to (1).
6. The first P entries of T_k with $k \in L_i[1 \dots M] \cup \{i\}$ will, with high likelihood, represent distances from e_i to elements of **E** being in the spatial vicinity of e_i and simultaneously moving coherently with e_i. Let $\bar{\sigma}_k$ be the mean of those P distances for each data element e_k. $P = 3$ in all our experiments.
7. For each data element e_i sum up all $\bar{\sigma}_k$ with $k \in L_i[1 \dots M] \cup \{i\}$. Let this sum be σ_i.
8. Experimentally, this σ_i-value turned out to be too small, so we decided to artificially enlarge it by coupling it to the total number of Delaunay-neighbors of i: $\sigma_i = \sigma_i \cdot \frac{K}{M}$.

This heuristic approach can surely be rendered into producing wrong results by using strongly corrupted input data or constructing pathological situations. Nevertheless, it yields very good results in practice, as can be seen in figure 2 and figure 3.

5 Iterative clustering methodology

Our system combines recursive clustering adopted from [6] with self-tuning spectral clustering from [3]. Clustering starts with the data set **E** and an empty list of limb clusters. The Θ-score, as given in (2) is a good measure for clustering quality in every iteration.

$$\Theta = 1 - \frac{1}{K} \left(\frac{\sum_{j=1}^{K} \sum_{i=1}^{N} \frac{\vec{y}_{ji}^2}{\max_p \vec{y}_{pi}^2}}{N} - 1 \right). \qquad (2)$$

We could identify three different types of outcomes for subdividing a cluster **C** into subclusters $\mathbf{C}_1$ and $\mathbf{C}_2$ during iterative clustering:

Case 1: $\Theta < 0.975$: invalid clustering, **C** is added to the list of limb clusters.

Case 2: $\Theta \in [0.975 \dots 0.995[$: imperfect clustering, try to boost clustering quality by identifying one outlier in $\mathbf{C}_1$ and $\mathbf{C}_2$. Those are temporarily removed from **C**, subdivision is again attempted, and the outliers are then reassigned to the subclusters they were in before if the new subdivision is successful.

Case 3: $\Theta > 0.995$: valid clustering, continue recursive splitting on both $\mathbf{C}_1$ and $\mathbf{C}_2$.

6 Results and discussion

Limb segmentation results for some highly non-rigid motion examples are shown in figure 2, where three clusters were identified, and in figure 3, where the motion is more complex and led to the identification of seven clusters. The results are quite precise, nearly all generated segmentations are very close to the actual limb structure. It can be stated that in the case of well-behaved input data (separate limbs showing strong motion, tracker data smooth and continuous) our system finds correct segmentations of the complete articulated body structure.

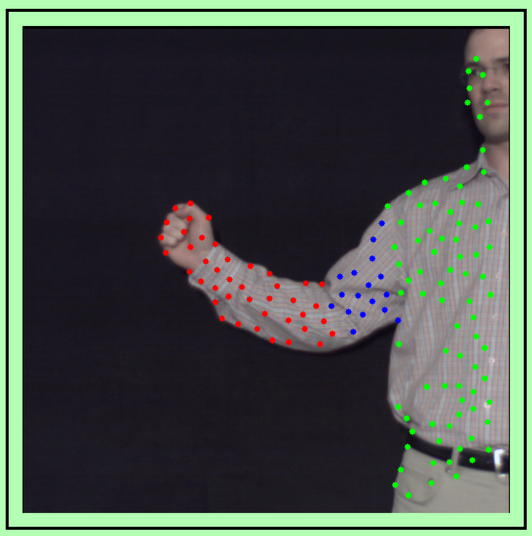

Figure 2: A simple sequence, which is correctly segmented into body, forearm and upper arm.

Figure 3: A more complex sequence including structured background. Clustering selects both forearms and upper arms, body, head, and background.

Figure 4: Segmented sequence, automatically estimated skeleton overlaid

References

[1] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.

[2] Jianbo Shi and Carlo Tomasi. Good features to track. In *Proceedings of CVPR'94*, Seattle, June 1994.

[3] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems, NIPS*, volume 17, 2004.

[4] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of IJCAI81*, pages 674–679, 1981.

[5] Nils Krahnstoever. *Articulated Models from Video*. PhD thesis, Pennsylvania State University, 2003.

[6] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. PAMI*, 22:888–905, 2000.