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 retrieves planar models of clothed, non-rigidly moving human limbs from given video input in a fully autonomous manner. This is achieved by combining and extending state-of-the-art algorithms for tracking, clustering and probabilistic image segmentation, allowing them to self-adapt to the processed footage. Application of our system to video sequences of humans yields precise limb templates, which trace well the true body structure of the captured performer.

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 Salient region detection

To get an idea of 'where to look' for the captured human body in a given video input sequence, our system performs fully autonomous two-phase salient region detection. In phase one, a sparse, yet precise description of the sought-after foreground shape is found by employing frame-differencing techniques.

For the frame displaying the highest activity potential (the *reference frame* , exemplary depicted in fig. 1), phase two fleshes out this initial foreground estimate using the advanced *graph cut* -based pixel classification scheme of [1]. The resulting, high-quality *foreground proposal map* is depicted in fig. 2 (green layer).

Via simple image processing techniques, an additional *static background proposal map* is set up, comprising exclusively image regions that show no motion throughout the complete input sequence (indicated in fig. 2, yellow layer).

Eventually, the salient region detection process comes up with a *non-static background proposal map* (fig. 2, red layer). This map is defined on all image regions which are truly part of the background, yet are occasionally obscured by the foreground subject.

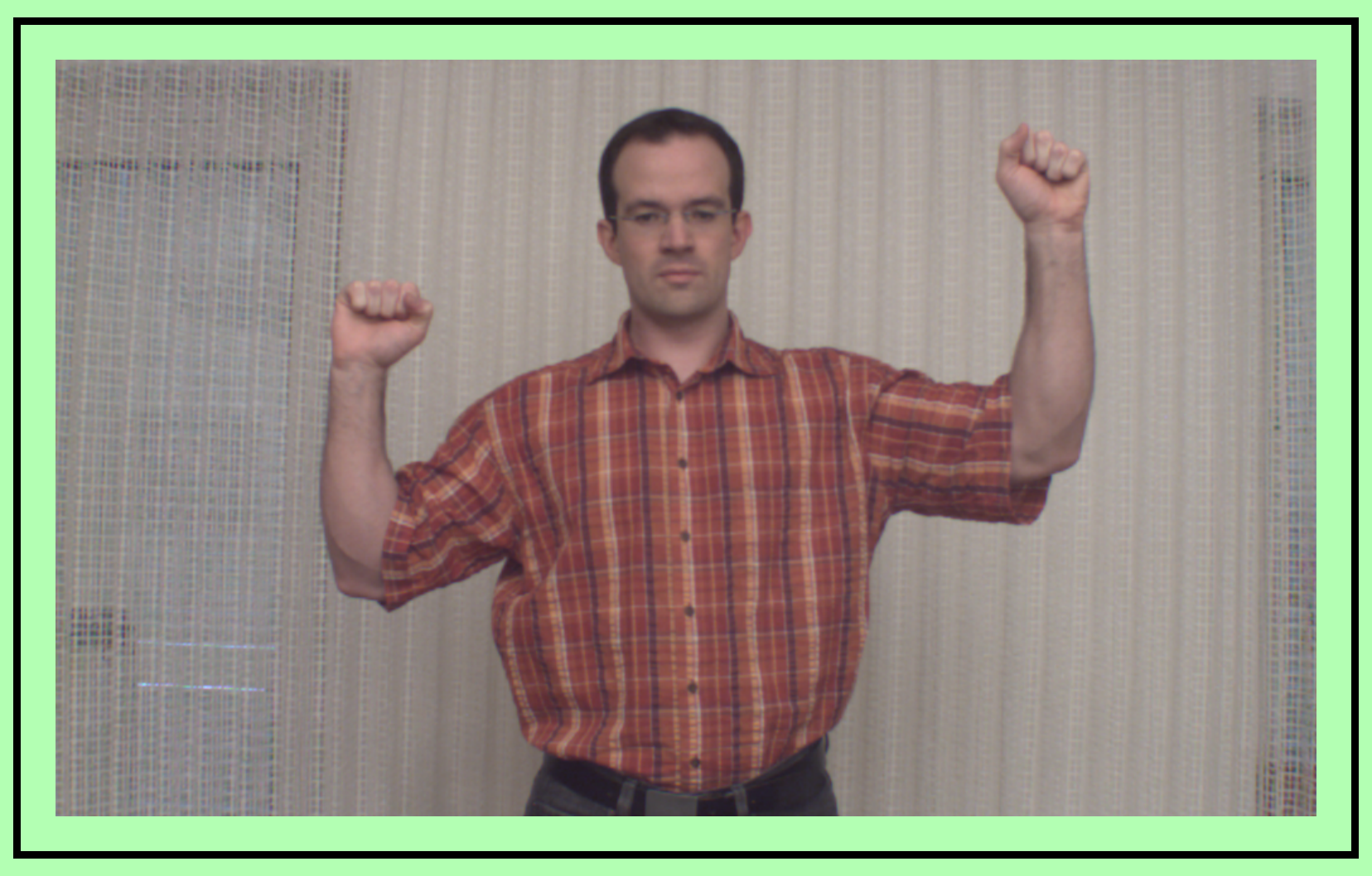
Figure 1: Reference frame

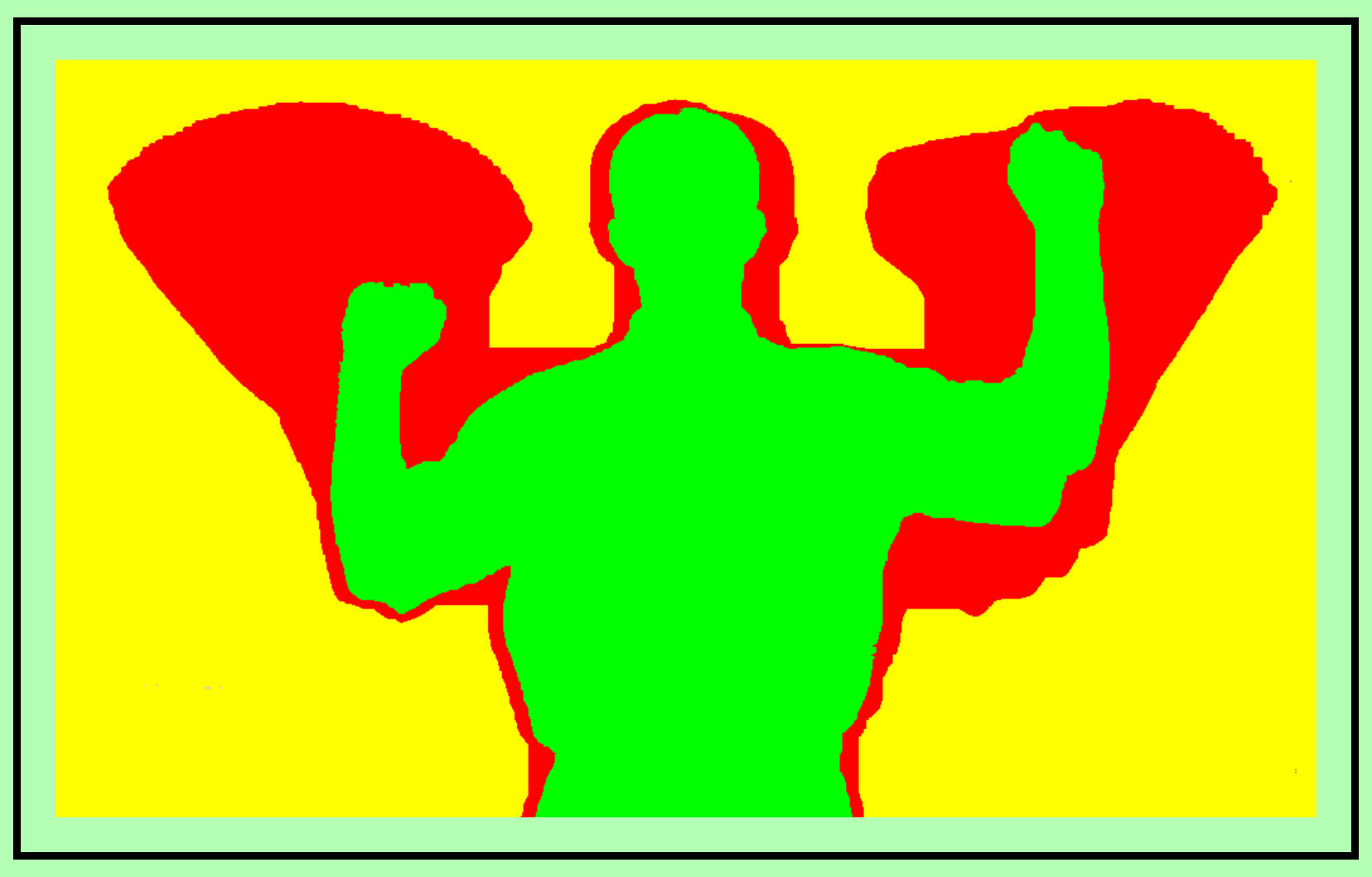
Figure 2: Proposal maps

3 Feature placement and tracking

Based on the results of salient region detection, body motion is followed via *sparse feature tracking* . Features to be tracked are plain image patches. Using the above proposal maps as a guidance, foreground features are distributed relatively dense on the interesting foreground shape, while the static background is only sparsely sampled. The non-static background areas are kept a void space; features placed there would show unpredictable tracking behavior.

The resulting feature placement is shown in fig. 3, foreground features are indicated by yellow dots, background features are shown as blue dots.

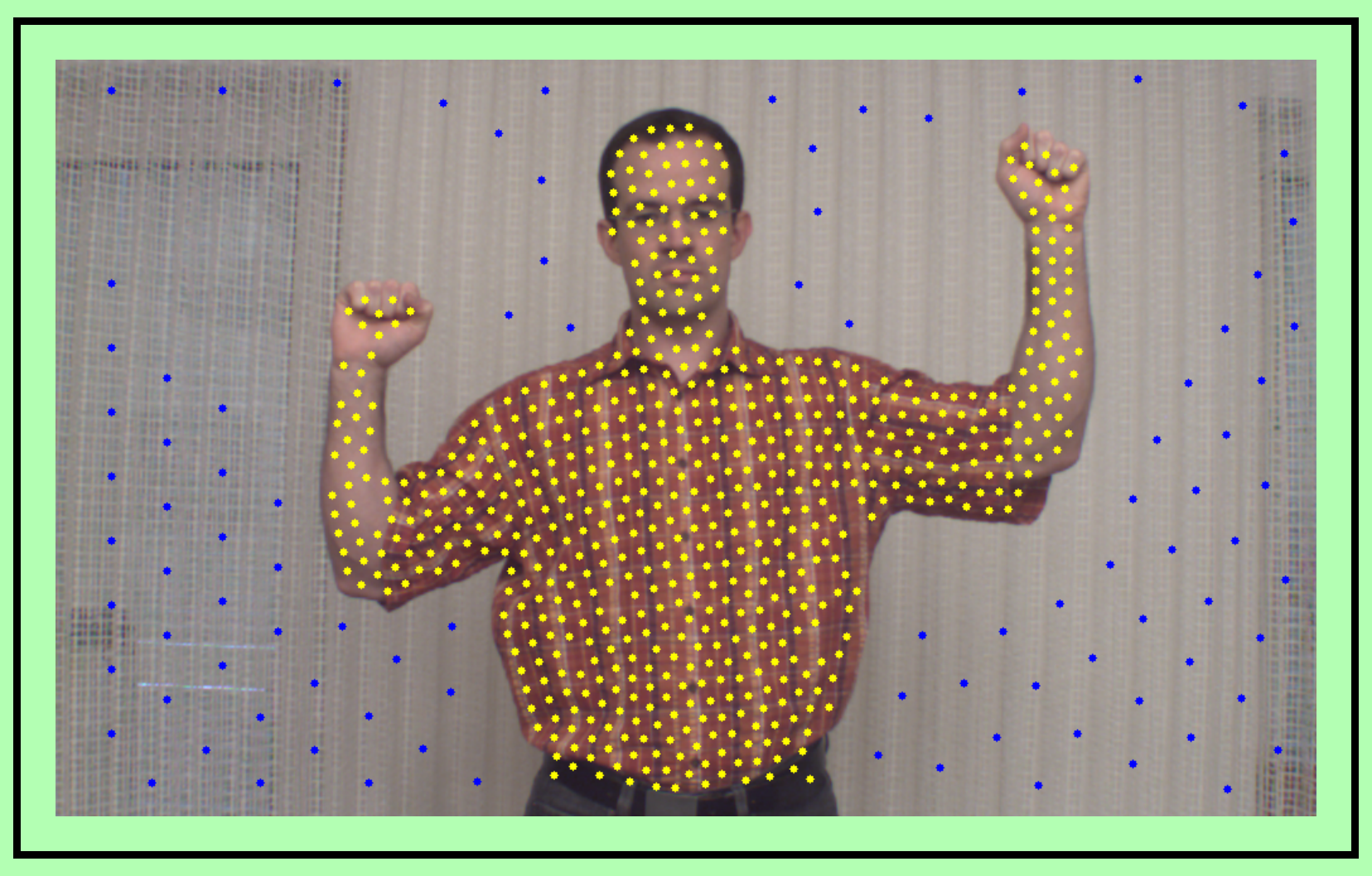
Figure 3: Feature initialization

Tracking the resulting large feature set over all input frames requires a fast and robust tracking framework. For our purposes, the method of [2] turned out appropriate; the feature trajectories resulting from the tracking process are shown in fig. 4.

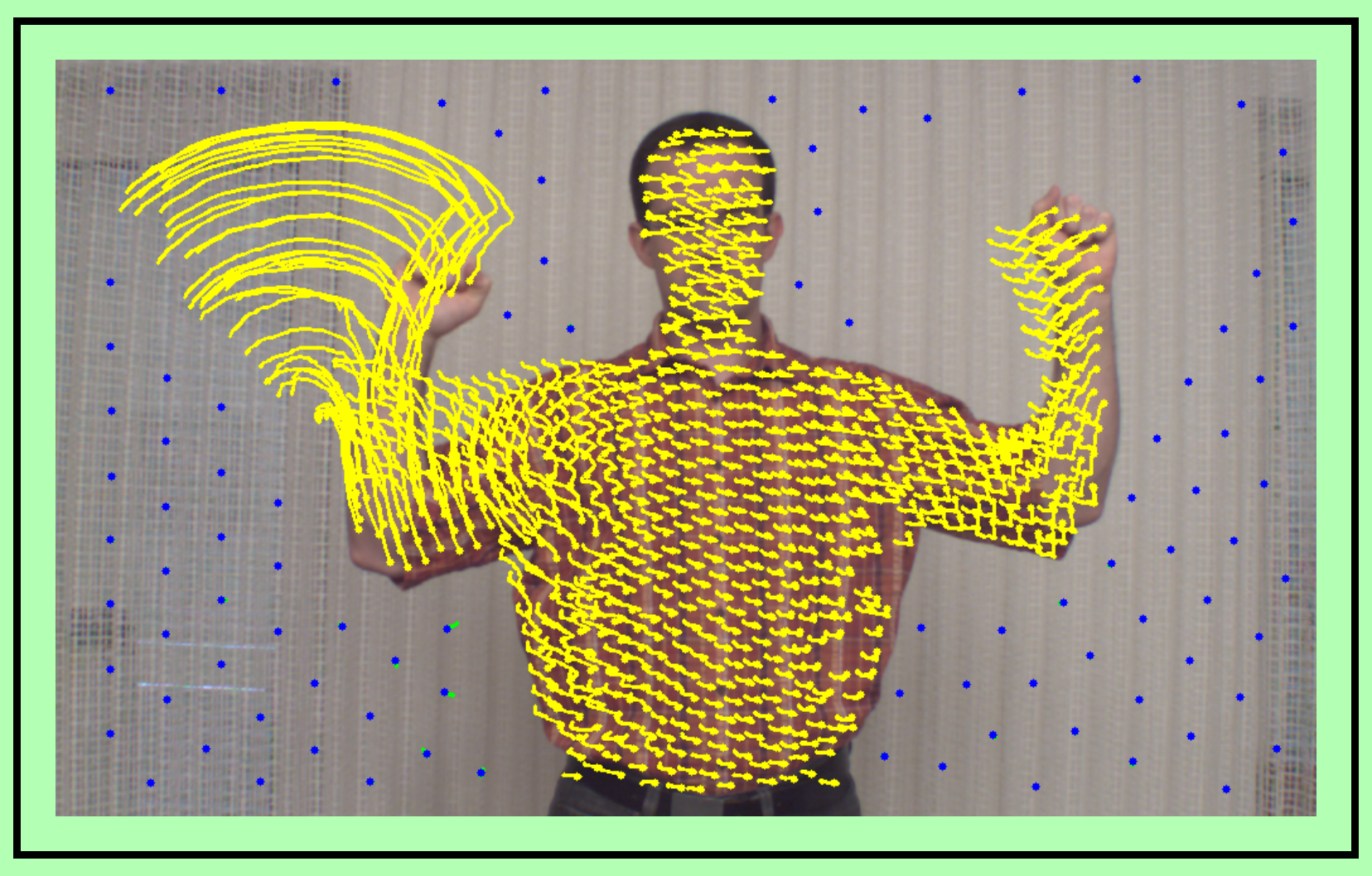
Figure 4: Tracking results, using the scheme of [2]

4 Coherent motion segmentation

As human limbs move as monolithic entities, limb retrieval can be adressed by finding coherently moving feature clusters. For this, we employ *spectral clustering* techniques: endowing the *normalized cut* mechanism of [3] with the *self-tuning capabilities* proposed by [4], our segmentation framework reliably segments the single body parts (*limb proposals*) without human intervention (see fig. 5).

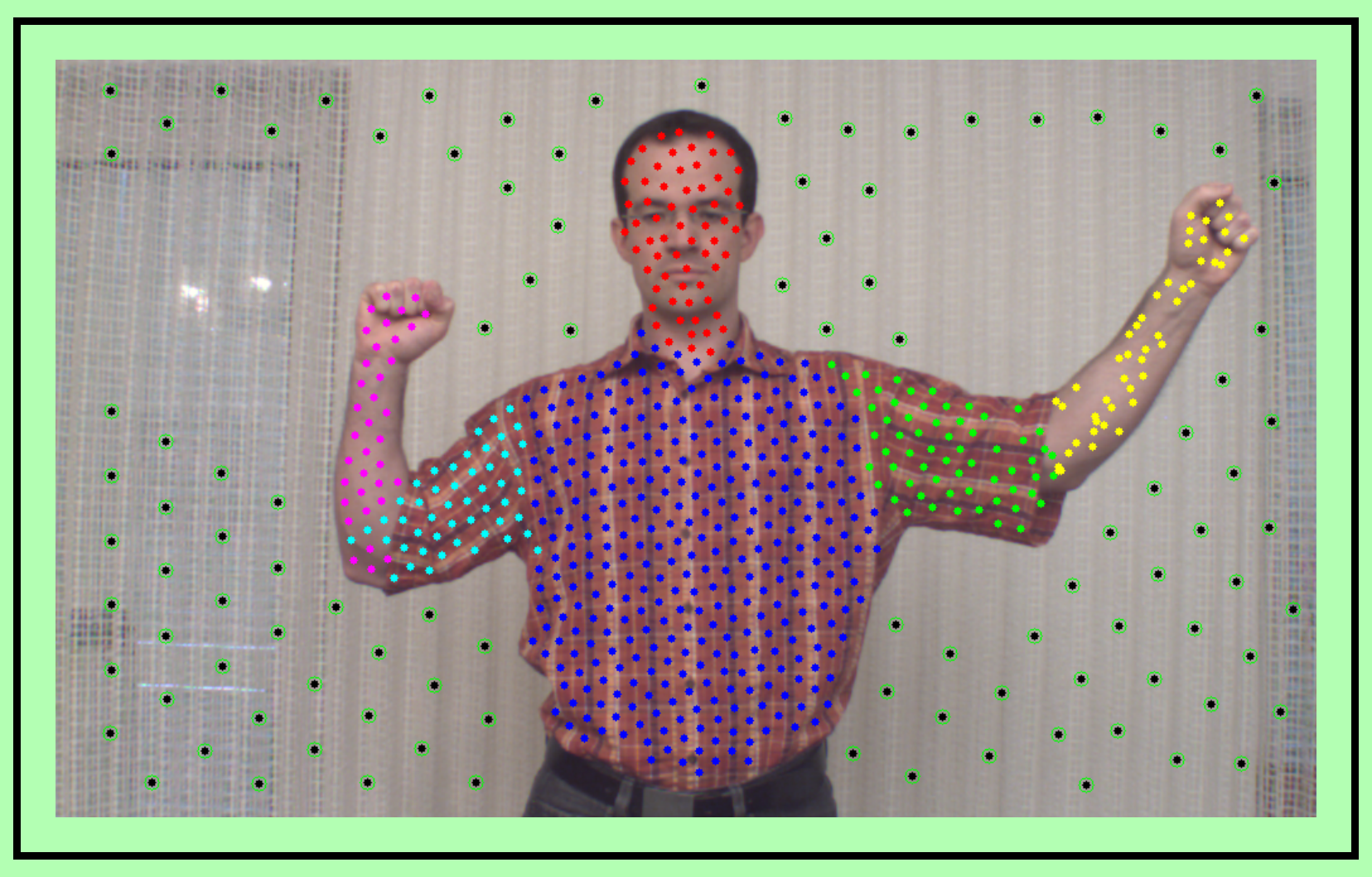
Figure 5: Results of self-tuning spectral clustering

5 Skeleton construction

Given the limb proposals, it is straightforward to infer the kinematic structure of the captured body. Possible joint candidates are allocated via techniques proposed by [5], assuming the sought-after skeleton to be tree-like. Results of the skeleton extraction stage are depicted in fig. 6; identified joints are indicated by green dots, skeleton 'bones' are shown in white.

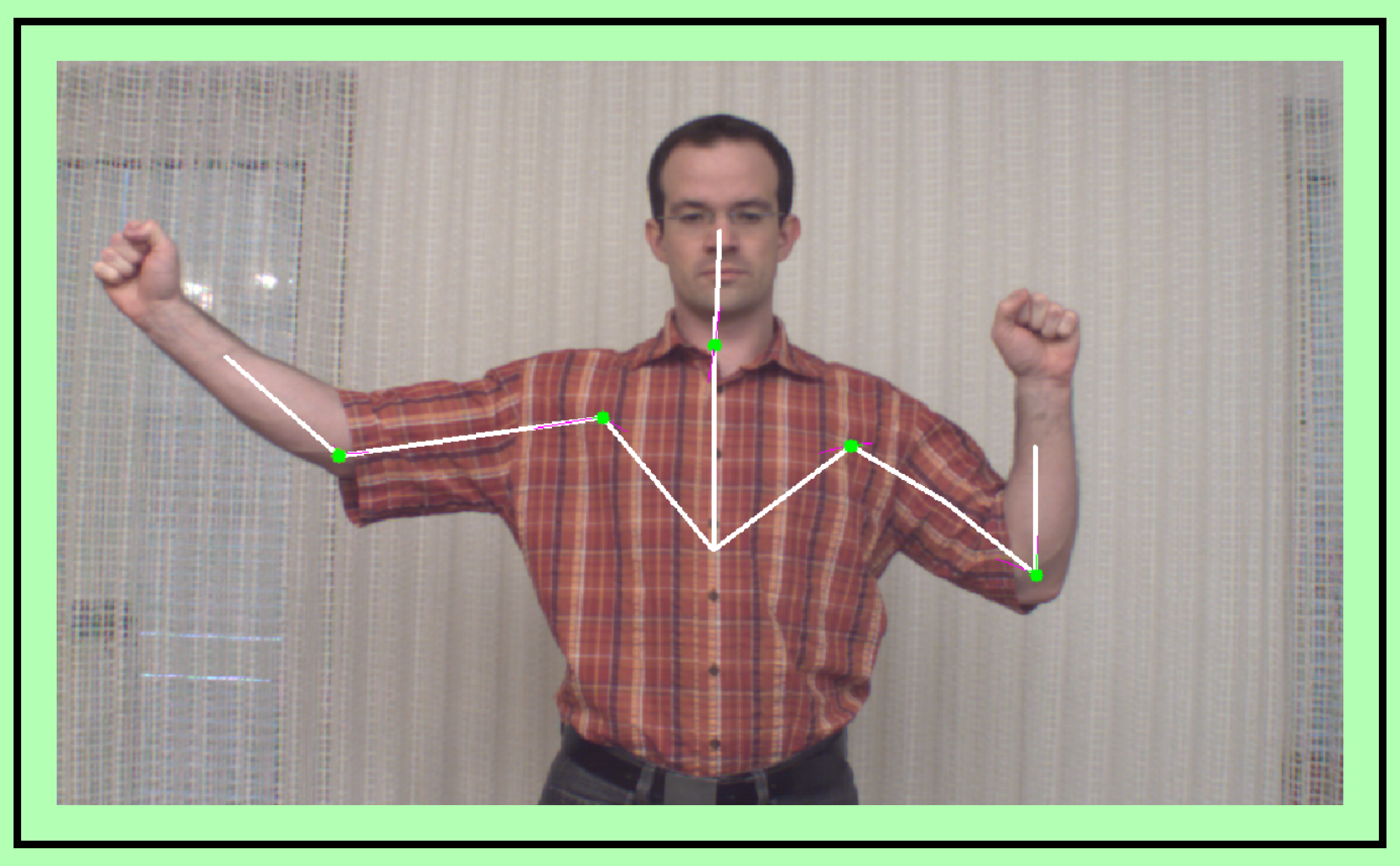
Figure 6: Skeleton retrieval

6 Refining limb models

Obviously, the sparse limb proposals give only a rough impression of the true limb shape. To arrive at more detailed *limb models* , fleshing out these sparse representations becomes inevitable. To do so, we employ, similar to [5], a *Markov Random Field* (MRF)-based pixel labeling scheme. The limb proposals herein serve as seed inputs, and are used to build combined motion, shape and appearance model of the sought-after body parts. These models constitute the basis for the class conditional probability of the underlying MRF, thereby enforcing the created limb templates to stay in accord with image observations. The MRF prior probability is chosen equal to the Potts model (cf. [6]), favoring spatial coherence of the refined body parts. Eventually, the MAP probability estimate of the MRF yields the required, densified limb models. Performing MAP-MRF estimation can be done in a variety of ways [7], we here modify the approach presented by [5]. Running our refinement scheme on several selected training sequences gives the results depicted in fig. 7.

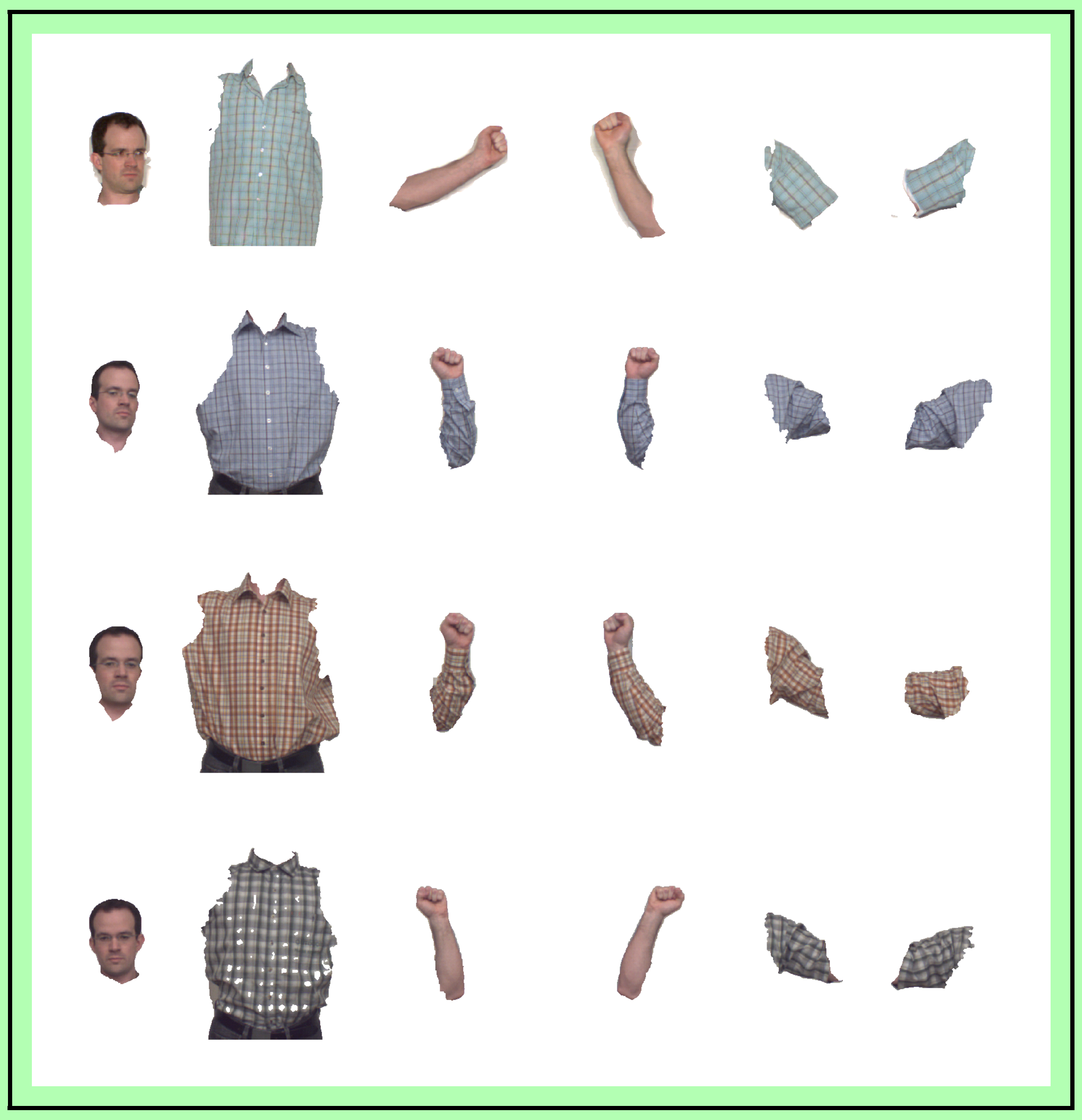
Figure 7: Results of limb refinement

7 Discussion

Observing the precision of the obtained limb models, we plan to combine them via kinematic constraints encoded in the skeleton graph; this yields full-fledged upper human body models which can in turn be matched against previously unseen input images to perform posture analysis. Beyond that, we would like our system to learn conceptual models of the human body. Therefore, it will be necessary to combine information from several exemplar models into a more generic *meta model* . This new model has to selectively emphasize characteristics common to all human individuals. Such characteristics might include similar face or hand color, head shape or body joint limits. Using the meta models, performance of posture estimation in novel situations is deemed to significantly increase.

References

[1] Y. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d-images. In *Proceedings of the Intern. Conf. on Computer Vision*, volume 1, pages 105–112, 2001.

[2] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams: a factorization method - detection and tracking of point features. Technical report, Carnegie Mellon University, 1991.

[3] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[4] L. Z. Manor and P. Perona. Self-tuning spectral clustering. In *18th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1601–1608, 2004.

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

[6] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 648, 1998.

[7] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, 2008.