# MOTION SEGMENTATION OF ARTICULATED STRUCTURES BY INTEGRATION OF VISUAL PERCEPTION CRITERIA

Hildegard Kuehne and Annika Woerner
*Institute for Anthropomatics, Karlsruhe Institute of Technology, Kaiserstr. 12, Karlsruhe, Germany*

Keywords: Motion segmentation, Articulated body tracking, Motion recognition.

Abstract: The correct segmentation of articulated motion is an important factor to extract and understand the functional structures of complex, articulated objects. Segmenting such body motion without additional appearance information is still a challenging task, because articulated objects as e.g. the human body are mainly based on fine, connected structures. The proposed approach combines consensus based motion segmentation with biological inspired visual perception criteria. This allows the grouping of sparse, dependent moving features points into several clusters, representing the rigid elements of an articulated structure. It is shown how geometric and time-based feature properties can be used to improve the result of motion segmentation in this context. We evaluated our algorithm on artificial as well as natural video sequences in order to segment the motion of human body elements. The results of the evaluation of parameter influences and also the practical evaluation show, that good motion segmentation can be achieved by this approach.

## 1 INTRODUCTION

The recovery of articulated structures from moving elements is one of the main abilities of the human perception system. In this context segmentation of articulated motion is an important factor to recognize complex motion and to understand the underlying functional structures.

Biological vision systems are able to understand complex structures from only few motion information. Just by seeing some moving features, we are able to understand the structure of the moving object and to recognize the perceived motion. One crucial step in this context is the correct grouping of motion information, in order to identify elements that are assumed to represent a rigid object, and to use this information to combine the motion elements for higher level recognition processes. An example for this has been given by Johansson's moving light displays (Johansson, 1973). The understanding of these biological mechanisms is still an open problem in neuroscience, but its importance for any vision system becomes increasingly clear and the work on this subject is still going on as can be seen e.g. at Giese and Poggio (Giese and Poggio, 2003).

Common motion segmentation algorithms are usually to unspecific to keep up with the abilities of biological vision systems. Many motion segmentation approaches are dealing with object tracking or scene understanding, so they are focused on the segmentation of compact, independent moving objects. When it comes to the handling of dependent motions of thin elements with only few data points as e.g. in gesture recognition, they will usually fail.

One step towards the segmentation of these structures could be the combination of well know consensus based motion segmentation with the constrains and connectivity rules of biological vision system. It is well known, that the visual perception usually follows a system of principles for the grouping of stationary and moving elements what has been described e.g. by Ullman (Ullman, 1983). Using these principles, we usually get a fast accurate guess about our environment.

In the here presented approach, a RANdom SAmple Consensus (RANSAC) algorithm is used to combine geometric criteria e.g. the affine projection of motion features with biological inspired constrains like center of mass distance, distance from main axis or motion vector distance to group sparse, dependent moving features to clusters, representing the rigid elements of articulated structures. It is shown how geometric and time-based feature properties can be used to improve the result of motion segmentation and help to overcome common problems in this context.



Kuehne H. and Woerner A. (2010).
MOTION SEGMENTATION OF ARTICULATED STRUCTURES BY INTEGRATION OF VISUAL PERCEPTION CRITERIA.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 54-59
DOI: 10.5220/0002829900540059

## 2 RELATED WORK

As motion segmentation is a broad field with applications in a lot of different contexts, we want to restrict the following overview to methods dealing with the clustering and grouping of feature points based on motion information.

A survey of common motion segmentation algorithms has been given by Tron and Vidal(Tron and Vidal, 2007). The main algorithms are explained and their performance is compared based on the results obtained with a benchmark set. The strengths and weaknesses of algorithms are also discussed here.

An example for RANSAC in context of motion segmentation is given by Yan and Pollefeys (Yan and Pollefeys, 2005), using RANSAC with priors to recover articulated structures. The presented algorithm is tested with a truck sequence with up to four depended moving segments. But motion segmentation by consensus can also be used to merge already segmented groups. Such an approach is proposed by Fraile et al. (Fraile et al., 2008). Here, a consensus method is used to merge feature groups tracked on video in order to analyze scenes from public transport surveillance cameras. Another reference is the approach presented by Pundlik and Birchfield (Pundlik and Birchfield, 2008) for motion segmentation at any speed. Here an incremental approach to motion segmentation is used to group feature points by a region-growing algorithm with an affine motion model.

## 3 MOTION SEGMENTATION BY CONSENSUS

One of the most popular applications of the RANSAC algorithm is probably the stitching of two or more overlapping images to a panoramic view. This is done by comparing a lot of different point correspondences in order to find the set that fits best into a projection to find the largest group of elements with the most uniform motion. This makes the algorithm very accurate with a high robustness against outliers. Translating this idea to the problem of articulated motion segmentation, we can assume more than one moving region which can be approximated by different projection matrices. For a video sequence with articulated body motion it is obvious that there is usually more than one motion projection. Given a set of 2D feature points $F^n = f_1^n, ..., f_k^n$ at frame $n$, the aim is to find all projections $P^n = P_1^n, ..., P_l^n$ that approximate the translations of the feature set from frame $n$ over the next $m$ frames.

It can be assumed that an articulated motion can be defined as a set of projections each determining a set of inliers, which is also called consensus set $CS$, so that the projection $P_i^n$ represents the projection of the points $f_{CS(i)}^n$ over the frames $n$ to $n+m$. As there is also no information about the number of expected projections, an iterative approach is chosen that does not need any prior knowledge about the number of regions but terminates when the largest regions are found. The iterative random sample consensus works as follows:

1. Estimate random minimal sample set *mss* from all given feature points $F^n$
2. Calculate the projection $P_{mss}^n$ from $f_{mss}^n$ over the next m frames
3. Apply the projection $P_{mss}^n$ to all feature points $F^n$
4. Calculate the error of every feature point defined by the error function $E(f^n)$ (see sec.5, equ.6). All features whose error is below the predefined threshold thresh are building the new consensus set $f_{CS}^n$
5. Calculated the overall cost of the consensus set by cost function $C(f_{CS}^n)$. (see sec.5, equ.9)
6. If the cost of the new consensus set is decreased or if the costs are the same and the size of the new consensus set has increased, update the final consensus set and its cost with the new one
7. Repeat the steps 1-6 until either all feature points had been assigned to a consensus set or the consensus set hasn't been updated for a predefined number of iterations or a predefined maximum number of iterations is reached

The final consensus set is assumed to be the best projection of the largest set of remaining feature points. So, the projection as well as the consensus set is defined as a new group and the features assigned to this group are removed from the feature set. This procedure is repeated until either the size of the last found consensus set or the number of remaining feature points becomes to small.

## 4 VISUAL PERCEPTION CRITERIA

Perceiving a group of moving features the biological perception systems usually depends a number of perceptual constrains, that help to group clusters of moving features. The following criteria are based on human interpretation of perception of rigid objects from 2D motion described by Ullman (Ullman, 1983). Assuming features are situated on one rigid element, they will probably follow one or more of follow criteria:

**Geometric Projection.** A feature point $f_a$ is rather located on the same rigid element as the random minimum sample set $f_{mss}$ if the symmetric reprojection error $e_p$ of $f_a$ of the projection $P_{mss}$ from $f_{mss}$ over all m frames is small:

$$e_p(f_a^n) = \frac{1}{m-1} \sum_{i=n}^{n+m-1} ((P_{mss}^i f_a^i) - f_a^{i+1})^2 + ((P_{mss}^i / f_a^{i+1}) - f_a^i)^2 \quad (1)$$

**Local Distance.** A feature point fa is rather located on the same rigid element if its distance d from the center of mass of the minimum sample set $M(f_{mss})$ over $m$ frames is small:

$$d(f_a^n, M(f_{mss}^n)) = \frac{1}{m} \sum_{i=n}^{n+m} \sqrt{(f_a^i - M(f_{mss}^i))^2} \quad (2)$$

**Motion Vector.** A feature point $f_a$ is rather located on the same rigid element if it has the same or a similar motion vector as the minimum sample set $f_{mss}$:

$$d_v(f_a^n) = \frac{1}{m-1} \sum_{i=n}^{n+m-1} (d(f_a^i, f_a^{i+1}) - d(f_m ss^i, f_m ss^{i+1}))^2 \quad (3)$$

**Axial Distance.** A feature point fa is rather located on the same rigid element if the distance da to the axis spanned by the minimum sample set $axis(f_{mss})$ is small:

$$d_a(f_a^n) = \frac{1}{m} \sum_{i}^{n+m} min(d(f_a(i), axis(f_{mss}^i))) \quad (4)$$

All these criteria are then integrated in the random sample consensus algorithm.

# 5 INTEGRATION OF PERCEPTION CRITERIA

The listed parameters are integrated in the RANSAC algorithm by using them as penalty factor for the overall error estimation. In a common RANSAC approach the error function (see sec.3, step 4) is based on the symmetric reprojection error, as has been described in equ.1. So, the common error function is defined by:

$$E_{org}(f_a^n) = e_p(f_a^n) \quad (5)$$

To integrate the predefined visual perception criteria, the related distances of the feature point to the actual minimum sample set are integrated in this function. To achieve this, all factors are weighted and added to the original error estimation. So the new consensus set error function is defined by:

$$E_{new}(f_a^n) = e_p(f_a^n) + w_d \cdot d(f_a^n, M(f_{mss}^n)) + w_v \cdot d_v(f_a^n) + w_a \cdot d_a(f_a^n) \quad (6)$$

Here, $w_d$, $w_v$ and $w_a$ represent the weighting factors for the local distance, motion vector and axial distance. The feature distance as well as the distance from the principal axis is normalized over the half image diagonal, whereas the motion vector distance is normalized from [0...1]. Additionally, the visual perception criteria is applied to the overall cost function of the consensus set (sec.3, step 4). Usually the cost function is based on the error function (equ.5) and is defined as:

$$C(f_{CS}^n) = \frac{1}{m} \sum_{i=n}^{n+m} C(f_{CS}^i) \quad (7)$$

where $n$ is the number of elements of the consensus set. The cost function for every element is defined as:

$$C(f_a^n) = \begin{cases} E_{org}(f_a^n), if E_{org}(f_a^n) < thresh \\ thresh, if E_{org}(f_a^n) \geq thresh \end{cases} \quad (8)$$

where *thresh* refers to the predefined threshold that has been used to select the consensus set (see sec.3, step 4). The visual perception criteria are integrated in the cost function by replacing the original error formulation by the new error function formulated in equ.6:

$$C(f_a^n) = \begin{cases} E_{new}(f_a^n), if E_{new}(f_a^n) < thresh \\ thresh, if E_{new}(f_a^n) \geq thresh \end{cases} \quad (9)$$

So both, the selection of the consensus set as well as the overall cost function are adapted and the influence of every criterium is controlled by the error function.

# 6 IMPLEMENTATION

The realization of the here presented approach has been done as follows: First, the feature points of a video sequences are detected and tracked by a motion-based feature tracking algorithm (Koehler and Woerner., 2008), which is mainly based on the pyramidal implementation of the KLT feature tracking method described by Bouget (Bouguet, 2002), following the 'good features to track' method of Shi and Tomasi (Shi and Tomasi, 1994).

For every frame $n$, the feature set is reduced to those changing continuously their position over the next $m$ frames to estimate a projection. Only, if the number of those features is larger than a predefined minimum, the RANSAC algorithm is applied.

The RANSAC implementation of the here presented approach is mainly based on the Matlab open source library by Marco Zuliani (Zuliani, 2008) and follows the description in section 3. The result for every frame is a set of groups representing the motion

segments for this frame. To avoid an over segmentation a maximum number of groups can be defined, so that only the largest groups are considered. This prevents the segmentation of groups with only few features that can also result from outliers or noise.

# 7 EVALUATION

The algorithm is evaluated on several video sequences with artificial and natural human body movements: an artificial rendered motion with a textured avatar lifting up his hands (Figure 1a), an artificial rendered motion with a walking avatar (Figure 1b) and a real human motion (Figure 1c) captured with a BumbleBee camera with 20fps and a resolution of 640x480px with duration of ca. 3 seconds. Each video sequence comprises ca. 60 frames. The features of the evaluated motion sequences are labeled by hand to get a ground truth for the clustering algorithm. For the hand labeling up to 10 clusters (head, body, left upper and lower arm, right upper and lower arm, left upper and lower leg, right upper and lower leg) are defined representing the significant rigid parts of the human body as shown in Figure 2. To evaluate the different perception criteria, we analyzed the correctness and specificity of the clustering of the labeled body segments.

## 7.1 Evaluation of Perception Criteria

The influence of the described perception criteria, local distance and mean motion as well as axial distance, on the clustering result is analyzed. Therefore, the feature points of the all video sequences are segmented on basis of a rotation-scaling-translation (RST) based-projection. For every frame, the feature motion over the last three frames has been considered. The segmentation results of every frame are compared to the ground truth and true positive and false positive rate is calculated. The true positive and false positive rate of the complete video sequence is calculated by the mean true positive and false positive rate over all frames.

To evaluate the clustering quality with regard to different weighting factors, $w_d$, $w_v$ and $w_a$, all combinations of weighting factors are tested for the values 0.0, 0.5, and 1.0 with increasing thresholds (0.1, 0.3, 0.5, 0.7, and 0.9). The best and the worst result, as well as the original RANSAC segmentation is shown in Figure 3. As can be seen the receiver operator characteristics (ROC) of segmentation with additional perception criterions vary to the original RANSAC segmentation. Best performance can be found for a weighting factor of $w_d = 1$, $w_v = 1$, $w_a = 0$

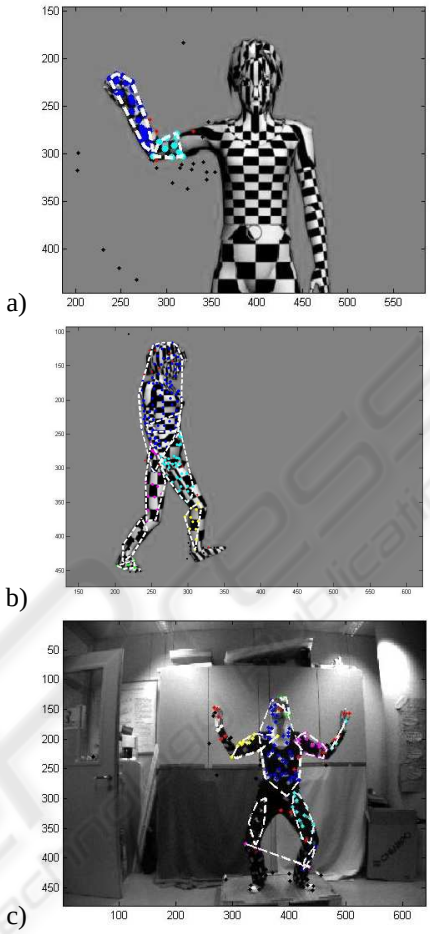


Figure 1: Results for the three different video sequences used for evaluation, a) and b) are artificial rendered waving and walking motions, c) is a video sequence with natural full body motion. The segmented regions are shown by different feature color.

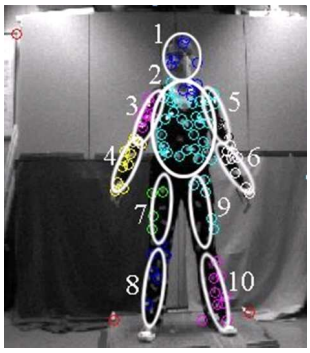


Figure 2: Ground truth for the evaluation of clustering and corresponding labeling of body segments: 1. head, 2. body, 3. upper right arm, 4. lower right arm, 5. upper left arm, 6. lower left arm, 7. upper right leg, 8. lower right leg, 9. upper left leg, 10. lower left leg.

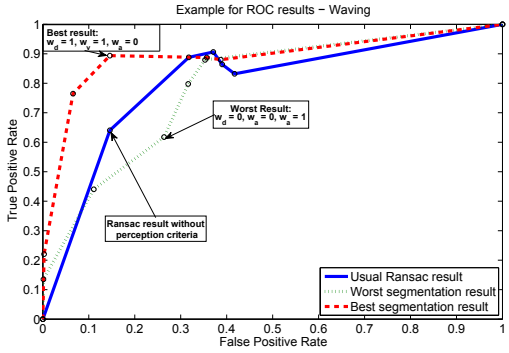


Figure 3: Best and worst ROC curves for RST based segmentation with different weights for artificial avatar motion 'Waving'. Best result has a weighting factor of $w_d = 1$, $w_v = 1$, $w_a = 0$ at $thresh = 0.3$, worst has a weighting factor of $w_d = 0$, $w_v = 0$, $w_a = 1$ at $thresh = 0.3$.

and $thresh = 0.3$ with a true positive rate of 0.8937 and a false positive rate of 0.1458. The samples including only the axial distances ($w_d = 0$, $w_v = 0$, $w_a = 1$, $thresh = 0.3$), are performing worse than usual RANSAC segmentation results. The best results of the true-positive and false-positive rate for the different criterions for the different video sequences are shown in Figure 4. The relation of true positive and false positive rate is usually better, the higher local distance $w_d$ and mean motion $w_v$ are weighted. They also show better performance when the weighting of the axial distance is low.

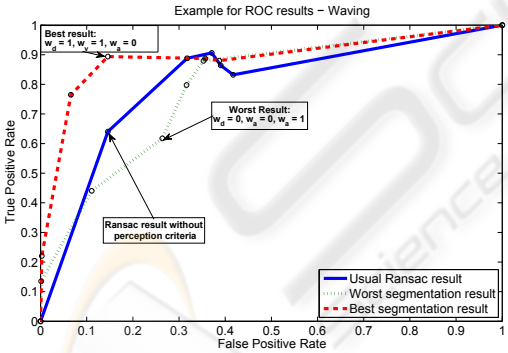


Figure 4: Comparison of best ROC results for RST based segmentation of different video motion sequences with the used weighting factors.

## 7.2 Evaluation of Segmentation over Three, Five and Ten Frames

As the segmentation is done over time, we have to consider the tradeoff between a long time period, which would be good to get reliable motion estimation and the problem that features tend to vanish because of occlusions etc. So, if the time period is chosen to long, it can happen that not enough features exist to reconstruct the motion. To analyze this tradeoff, we compared the best results of segmentations over three, five and ten frames. The feature points of the first video sequence are segmented with different weighting factors, and for every frame, the properties of the last three, five or ten frames are considered.

We can see that, comparing the best results of every segmentation (Figure 5), the true positive as well as the false positive rate decreases the more frames are used. Noteworthy is that the best result over ten frames, has been achieved without the integration of any additional weighting factors ($w_d = 0$, $w_v = 0$, $w_a = 0$, $thresh = 0.3$).

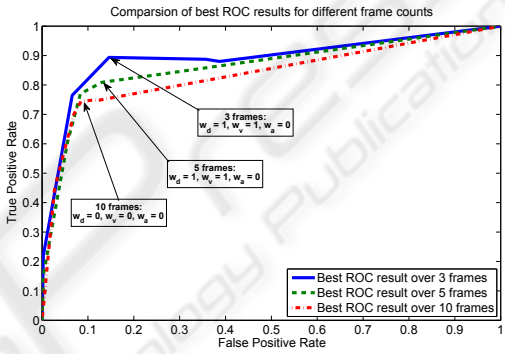


Figure 5: Comparison of best ROC result for segmentation of artificial avatar motion *Waving* over 3, 5 and 10 frames. True positive as well as the false positive rate of best results decrease with a higher number of frames used for segmentation.

## 7.3 Evaluation of RST and Homographic based Segmentation

To find elements which could be underlying by a rigid element, two different geometric projections can be used. From a geometrical point of view, it would be accurate to estimate the homographic projection of the feature motion, which needs at least 4 points to calculate the transformation. But from a perceptional point of view, also a Rotation-Translation-Scaling transformation, which only needs two points to be computed, can be assumed. This can be seen on a simple example of Johansson point light displays. Usually, the human perception system only needs one point at every joint to build up a human pose. This means that for the reconstruction and recognition of a rigid element, only two points are enough. So it is likely, that biological vision systems are recognizing information on the basis of RST transformations as well as on the basis of homographic projections.

To evaluate this characteristic, both projection criteria had been analyzed. To do this, the feature points

of the first video sequence are segmented on basis of a RST as well as on a homographic projection with different weighting factors.

Comparing the results which had been achieved with a RST projection with those of a homographic projection as can be seen in Figure 6, we can see that the segmentations on the basis of an RST projection has a much better relation of true positive and false positive rate than those on basis of a homographic projection. This could amongst others be caused by the fact that a RST projection is more robust against noise, because here, smaller variations don't have so much influence on the overall result as they would have considering a homographic projection.

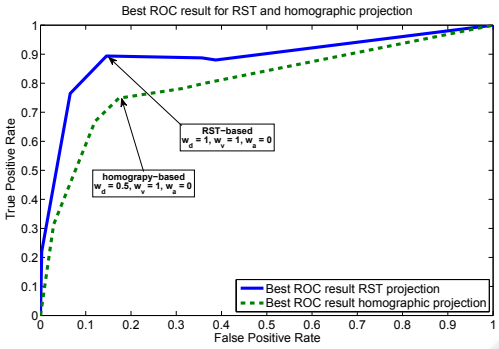


Figure 6: Comparison of best ROC results of RST based segmentation with homographic based segmentation for artificial avatar motion *Waving*.

## 8 CONCLUSIONS

We presented a motion segmentation approach that combines a consensus based motion segmentation algorithm with criteria from biological vision system in order to cluster sparse groups of feature points only by their motion information. It is show, that this combination has the potential to cluster also small, dependent moving features.

The results of the performance evaluation of parameter influences as well as the practical evaluation on artificial and real human motion video sequences show that good motion segmentation can be achieved by this approach

## REFERENCES

Bouguet, J. Y. (2002). Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm.

Fraile, R., Hogg, D., and Cohn, A. (2008). Motion segmentation by consensus. international conference on pattern recognition. *ICPR*.

Giese, M. and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4:179–192.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysis*, 14:201–211.

Koehler, H. and Woerner., A. (2008). Motion-based feature tracking for articulated motion analysis. In *IEEE Int. Conf. on Multimodal Interfaces (ICMI 2008), Workshop on Multimodal Interactions Analysis of Users a Controlled Environment.*, Chania, Greece.

Pundlik, S. J. and Birchfield, S. T. (2008). Real-time motion segmentation of sparse feature points at any speed. *IEEE Transactions on Systems, Man, and Cybernetics*, 38(3):731–742.

Shi, J. and Tomasi, C. (1994). Good features to track. In IEEE, editor, *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle.

Tron, R. and Vidal, R. (2007). A benchmark for the comparison of 3-d motion segmentation algorithms. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1–8.

Ullman, S. (1983). Computational studies in the interpretation of structure and motion: Summary and extension. *Human and Machine Vision*.

Yan, J. and Pollefeys, M. (2005). Articulated motion segmentation using ransac with priors. *ICCV Workshop on Dynamical Vision*.

Zuliani, M. (2008). Ransac toolbox for matlab. http://www.mathworks.com/matlabcentral/fileexchange/18555.