Localised Mixture Models in Region-based Tracking

Christian Schmaltz[1], Bodo Rosenhahn[2], Thomas Brox[3], and Joachim Weickert[1]

[1] Mathematical Image Analysis Group, Faculty of Mathematics and Computer Science, Building E1 1 Saarland University, 66041 Saarbrücken, Germany
{schmaltz,weickert}@mia.uni-saarland.de
[2] Leibniz Universität Hannover 30167 Hannover, Germany
rosenhahn@tnt.uni-hannover.de
[3] University of California, Berkeley Berkeley, CA, 94720, USA
brox@eecs.berkeley.edu

Abstract. An important problem in many computer vision tasks is the separation of an object from its background. One common strategy is to estimate appearance models of the object and background region. However, if the appearance is spatially varying, simple homogeneous models are often inaccurate. Gaussian mixture models can take multimodal distributions into account, yet they still neglect the positional information. In this paper, we propose localised mixture models (LMMs) and evaluate this idea in the scope of model-based tracking by automatically partitioning the fore- and background into several subregions. In contrast to background subtraction methods, this approach also allows for moving backgrounds. Experiments with a rigid object and the HumanEva-II benchmark show that tracking is remarkably stabilised by the new model.

1 Introduction

In many image processing tasks such as object segmentation or tracking, it is necessary to distinguish between the region of interest (foreground) and its background. Common approaches, such as MRFs or active contours build appearance models of both regions with their parameters being learnt either from a-priori data or from the images [1–3]. Various types of features can be used to build the appearance model. Most common are brightness and colour, but any dense feature set such as texture descriptors [4] or motion [5] can be part of the model.

Apart from the considered features, the statistical model of the region is of large interest. In simple cases, one assumes a Gaussian distribution in each region. However, since usually object regions change their appearance locally, such a Gaussian model is too inaccurate. A typical example is the black and white stripes of a zebra, which leads to a Gaussian distribution with a grayish mean that does neither describe the black nor the white part very well. In order to deal with such cases, Gaussian mixture models or kernel density models have been

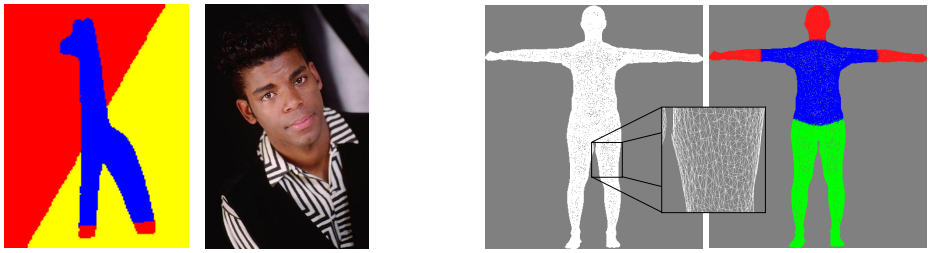


Fig. 1. Left: Illustrative examples of situations where object (to be further specified by a shape prior) and background tregion are not well modelled by identically distributed pixels. In (a), red points are more likely in the background. Thus, the hooves of the giraffe will not be classified correctly. In (b), the dark hair and parts of the body are more likely to belong to the background. Localised distributions can model these cases more accurately. **Right**: Object model used by the tracker in one of our experiments (c) and decomposition of the object model into three different components (d), as proposed by the automatic splitting algorithm from [6]. There are 22 joint angles in the model, resulting in a total of 28 parameters that must be estimated.

proposed. These models are much more general, yet still impose the assumption of identically distributed pixels in each region, i.e., they ignore positional information. The left part of Fig. 1 shows two examples where this is insufficient.

In contrast, a model which is sensitive for the location in the image was proposed in [7]. The region statistics are estimated for each point separately, thereby considering only information from the local neighbourhood. Consequently, the distribution varies smoothly within a region. A similar local statistical model was used in [8]. A drawback of this model is that it blurs across discontinuities inside the region. As the support of the neighbourhood needs to be sufficiently large to reliably estimate the parameters of local distributions, this blurring can be quite significant. This is especially true when using local kernel density models, which require more data than a local Gaussian model.

The basic idea in the present paper is to segment the regions into subregions inside which a statistical model can be estimated. Similar to the above local region statistics, the distribution model integrates positional information. The support for estimating the distribution parameters is usually much larger as it considers all pixels from the subregion, though. Splitting the background into subregions and employing a kernel density estimator in each of those allows for a very precise region model relying on enough data for parameter estimation.

Related to this concept are Gaussian mixture models in the context of background subtraction. Here, the mixture parameters are not estimated in a spatial neighbourhood but from data along the temporal axis. This leads to models which include very accurate positional information [9]. In [10], an overview of several possible background models ranging from very simple to complex models is given. The learned statistics from such models can also be combined with a

conventional spatially global model as proposed in [11]. For background subtraction, however, the parameters are learned in advance, i.e., a background image or images with little motion and without the object must be available. Such limitations are not present in our approach. In fact, our experiments show that background subtraction and the proposed localised mixture model (LMM) are in some sense complementary and can be combined to improve results in tracking. Also note that, in contrast to image labeling approaches that also split the background into different regions such as [12], no learning step is necessary.

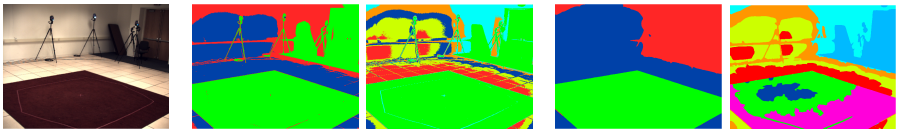
A general problem that arises when making statistical models more and more precise is the increasing amount of local optima in corresponding cost functions. In Fig. 1 there is actually no reason to put the red hooves to the giraffe region or the black hair to the person. A shape prior and/or a close initialisation of the contour is required to properly define the object segmentation problem. For this reason we focus in this paper on the field of model based tracking, where both a shape model and a good initial separation into foreground and background can be derived from the previous frame. In particular, we evaluated the model in silhouette-based 3-D pose tracking, where pose and deformation parameters of a 3-D object model are estimated such that the image is optimally split into object and background [13, 6]. The model is generally applicable to any other contour-based tracking method as well. Another possible field of application is semi-supervised segmentation, where the user can incrementally improve the segmentation by manually specifying some parts of the image as foreground or background [1]. This can resolve above ambiguities as well.

Our paper is organised as follows: We first review the pose tracking approach used for evaluation. We then explain the localised mixture model (LMM) in Section 3. While the basic approach only works with static background images, we remove this restriction later in a more general approach. After presentation of our experimental data in Section 4, the paper is concluded in Section 5.

2 Foreground-Background Separation in Region-Based Pose Tracking

In this paper, we focus on tracking an articulated free-form surface consisting of rigid parts interconnected by predefined joints. The state vector χ consists of the global pose parameters (3-D shift and rotation) as well as n joint angles, similar to [14]. The surface model is divided into l different (not necessarily connected) components $M_i, i = 1, \ldots, l$, as illustrated in Fig. 1. The components are chosen such that each component has a uniform appearance but its appearance differs from other components. Such a model has been proposed in [6]. There are many more tracking approaches than the one presented here. We refer to the surveys [15, 16] for an overview.

Given an initial pose, the primary goal is to adapt the state vector such that the projections of the object parts lead to maximally homogeneous regions in the image. This is stated by the following cost function which is sought to be minimised in each frame:



(a) (b) (c) (d) (e)

Fig. 2. Example of a background segmentation. **From left to right:** (a) Background image. (b,c) K-means clustering with three and six clusters. (d,e) Level set segmentation with two different parameter settings.

$$E(\chi) = -\sum_{i=0}^{l} \int_{\Omega} \Big(v_i(\chi, x) P_{i,\chi}(x) \log p_{i,\chi}(x) \Big) \mathrm{d}x, \tag{1}$$

where Ω denotes the image domain. The appearance of each component i and of the background ($i = 0$) is modelled by a probability density function (PDF) $p_i, i \in 0, \ldots, l$. The PDFs of the object parts are modelled as kernel densities, whereas we will use the LMM for modelling the background as explained in later sections.

$P_{i,\chi}$ is the indicator function for the projection of the i-th component M_i, i.e. $P_{i,\chi}(x)$ is 1 if and only if a part of the object with pose χ is projected to the image point x. In order to take occlusion into account, $v_i(\chi, x) : \mathbb{R}^{6+n} \times \Omega \mapsto \{0, 1\}$ is a visibility function that is 1 if and only if the i-th object part is not occluded by another part of the object in the given pose. Visibility can be computed efficiently using openGL.

The cost function can be minimised locally by a modified gradient descent. The PDFs are evaluated at silhouette points x_i of each projected model components. These points x_i are then moved along the normal direction of the projected object, either towards or away from the components, depending on which of the regions' PDF fits better at that particular point. The point motion is transferred to the corresponding change of the state vector by using a point-based pose estimation algorithm as described, e.g., in [7].

3 Localised Mixture Models

In the above approach, the object region is very accurately described by the object model, which is split into various parts that are similar in their appearance. Hence, the local change of appearance within the object region is taken well into account. The background region, however, consists of a single appearance model and positional changes of this appearance are so far neglected.

Consider a red-haired person standing on a red carpet which is facing the camera. Then, only a very small part of the person is red, compared to a large part of the background. As a larger percentage of pixels lying outside the person

are red, red pixels will be classified to belong to the outside regions. Thus, the hair will be considered as not being part of the object, which deteriorates tracking. This happens despite the fact that the carpet is far away from the hair.

The idea to circumvent this problem is to separate the background into multiple subregions each of which is modelled by its own PDF. This can be regarded as a mixture of PDFs, yet the mixture components exploit the positional information telling where the separate mixture components are to be applied.

3.1 Case I: Static Background Image Available

If a static background image is available, segmenting the background is quite simple. In contrast to the top-level task of object-background separation, the regions need not necessarily correspond to objects in the scene. Hence, virtually any multi-region segmentation technique can be applied for this. We tested a very simple one, the K-means algorithm [17, 18], and a more sophisticated level set based segmentation, which considers multiple scales and includes a smoothness prior on the contour [19]. In the K-means algorithm the number of clusters is fixed, whereas the level set approach optimises the number of regions by a homogeneity criterion, which is steered by a tuning parameter. Thus, the number of subregions can vary.

Fig. 2 compares the segmentation output of these two methods for two different parameter settings. The results with the level set method are much smoother due to the boundary length constraint. In contrast, the regions computed with K-means have more fuzzy boundaries. This can be disadvantageous, particularly when the localisation of the model is not precise due to a moving background as considered in the next section.

After splitting the background image into subregions, a localised PDF can be assembled from PDFs estimated in each subregion j. Let $L(x, y)$ denote the labelling obtained by the segmentation, we obtain the density

$$p(x, y, s) = p_{L(x,y)}(s), \tag{2}$$

where s is any feature used for tracking. It makes most sense to use the same density model for the subregions as used in the segmentation method. In case of K-means this means that we have a Gaussian distribution with fixed variance:

$$p_j^{\text{kmeans}}(s) \propto \exp\left(-\frac{(s-\mu_j)^2}{2}\right), \tag{3}$$

where μ_j is the cluster centre of cluster j. The level set based segmentation method is build upon a kernel density estimator

$$p_j^{\text{levelset}}(s) = K_\sigma * \frac{\sum_{(x,y)\in\Omega_j} \delta(s, I(x, y))}{|\Omega_j|} \tag{4}$$

where δ is the Dirac delta distribution and K_σ is a Gaussian kernel with standard deviation σ. Here, we use $\sigma = \sqrt{30}$. The PDF in (2) can simply be plugged into the energy in (1). Note that this PDF needs to be estimated only once for the background image and then stays fixed, whereas the PDFs of the object parts are reestimated in each frame to account for the changing appearance.

3.2 Case II: Potentially Varying Background

For some scenarios, generating a static background image is not possible. In outdoor scenarios, for example, the background usually changes due to moving plants or people passing by. Even inside buildings, the lighting conditions – and thus the background – typically vary. Furthermore, the background could vary due to camera motion. In fact, varying backgrounds can appear in many applications and render background subtraction methods impossible.

In general, the background changes only slowly between two consecutive frames. This can be exploited to extend the described approach to non-static images or to images where the object is already present. Excluding the current object region from the image domain the remainder of the image can be segmented as before. This is shown in Fig. 5. To further deal with slow changes in the background, the segmentation can also be recomputed in each new frame. This takes changes in the localisation or in the statistics into account.

A subtle difficulty appearing in this case is that there may be parts of the background not available in the density model because these areas were occluded by the object in the previous frame. When reestimating the pose parameters of the object model, the previously occluded part can appear and needs some treatment. In such a case we choose the nearest available neighbour and use the probability density of the corresponding subregion. That is, if Ω_j is the jth subregion as computed by the segmentation step, the local mixture density is:

$$p(x, y, s) = p_{j^*}(x, y) \qquad \text{with} \qquad j^* = \operatorname*{argmin}_j \left(\operatorname{dist}((x, y), \Omega_j)\right). \tag{5}$$

4 Experiments

We evaluated the described region statistics on sequence S4 of the HumanEva-II benchmark [20]. For this sequence, a total of four views as well as static background images are available. Thus, this sequence allows us to compare the variant that uses a static background image to the version without the need for such an image. The sequence shows a man walking in a circle for approximately 370 frames, followed by a jogging part from frame 370 to 780, and finally a "balancing" part until frame 1200. Ground truth marker data is available for this sequence and tracking errors can be evaluated via an online interface provided by Brown University. Note that the ground truth data between frame 299 and 334 is not available, thus this part is ignored in the evaluation. In the figures, we plotted a linear interpolation between frame 298 and 335.

Table 1 shows some statistics over tracking results with different models. The first line in the table shows an experiment in which background subtraction was used to find approximate silhouettes of the person to be tracked. These silhouette images are used as additional features, i.e. in addition to the three channels of the CIELAB colour space, for computing the PDFs of the different regions. This approach corresponds to the one in [6]. Results are improved when using the

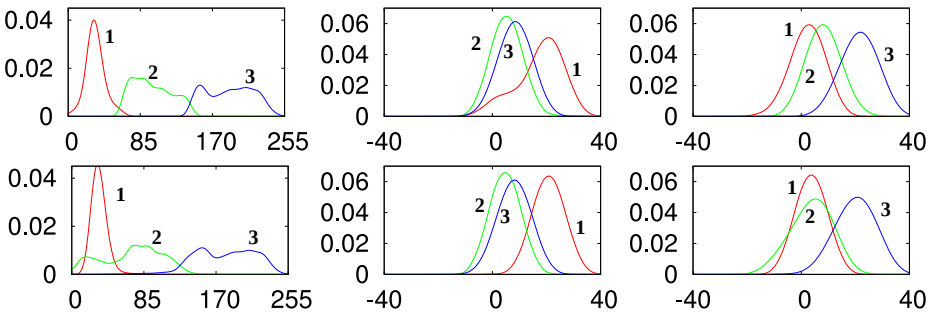


Fig. 3. PDFs estimated for the CIELAB colour channels of the subregions shown in Fig. 5. Each colour corresponds to one region. **From left to right**: Lightness channel, A channel and B channel. **Top**: Estimated PDFs when using the level-set-based segmentation. **Bottom**: Estimated PDFs when computing the subregions using K-means. Due to the smoothness term, the region boundaries are smoother resulting in PDFs that are separated less clearly when using the level-set-based method than with the K-means algorithm. Nevertheless, the level set approach performed better in the quantitative evaluation.

LMM based on level set segmentation. This can be seen by comparing the first and third line of the table. The best results are achieved when using both the silhouette images as well as the LMM (fifth line). The level set based LMM yields slightly better results than K-means clustering. See Fig. 4 for a tracking curve illustrating the error per frame for the best of these experiments

Fig. 5 shows segmentation results without using the background image, hence dropping the assumption of a static background. Fig. 3 visualises the estimated PDFs for each channel in each subregion. Aside from some misclassified pixels close to the occluded area (due to tracking inaccuracies, and due to the fact that a human cannot be perfectly modelled by a kinematic chain), the background is split into reasonable subparts and yields a good LMM. Tracking is almost as good as the combination with background subtraction, as indicated by the lower part of Table 1, without requiring a strictly static background any more. The same setting with a global Parzen model fails completely, as depicted in Fig. 4, since fore- and background are too similar for tracking at some places.

In order to verify the true applicability of the LMM in the presence of non-static backgrounds, we tracked a tea box in a monocular sequence with a partially moving background. Neither ground truth nor background images are available for this sequence, making background subtraction impossible. As expected, the LMM can handle the moving background very well. When using only the Parzen model for the background, a 90° rotation of the tea box is missed by the tracker as shown in the left part of the lower row in Fig. 6.

If we add Gaussian noise with standard deviation 10, the Parzen model completely fails (right part in lower row of Fig. 6) while tracking still works when using the LMM.

Table 1. Comparison of different tracking versions for sequence S4 of the HumanEva-II benchmark as reported by the automatic evaluation script. Each line shows the model used for the background region, if images of the backgrounds were used, the average tracking error in millimetre, its variance and its maximum, as well as the total time used for tracking all 1200 frames.

Model	BG image	Avg error	Variance	Max.	Time
Parzen model + BG subtraction	yes	46.16	276.81	104.0	4h 31m
LMM (K-means)	yes	49.63	473.90	114.2	4h 34m
LMM (level set segmentation)	yes	42.18	157.31	93.6	4h 22m
BG subtraction + LMM (K-means)	yes	42.96	178.19	92.6	4h 27m
BG subtraction + LMM (LS segm.)	yes	41.64	153.94	83.8	4h 29m
Parzen model	no	451.11	24059.41	728.4	5h 12m
LMM (K-means)	no	52.64	588.66	162.7	9h 19m
LMM (level set segmentation)	no	49.94	168.61	111.2	19h 9m

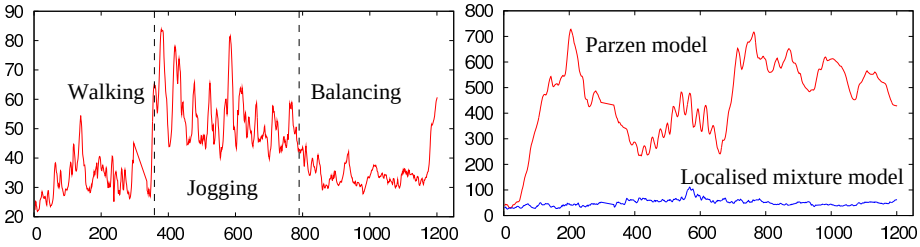


Fig. 4. Tracking error per frame of some tracking results of sequence S4 from the HumanEva-II dataset, automatically evaluated. **Left**: LMM where background subtraction information is supplemented as an additional feature channel. This plot corresponds to the fifth line in Table 1. **Right**: Global kernel density estimator (red) and LMM (blue). Here, we did not use the background images or any information derived from them. These plots correspond to the last (blue) and third last (red) line of Table 1.

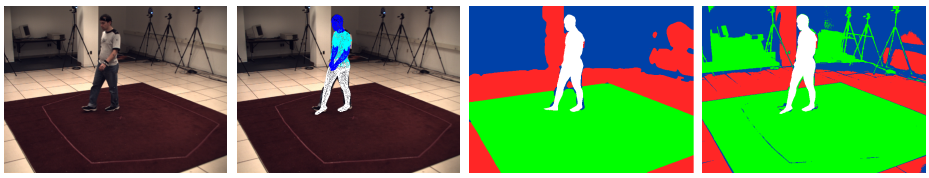


Fig. 5. Segmentation results for frame 42 as seen from camera 3. **Leftmost**: Input image of frame 42 the HumanEva-II sequence S4. **Left**: Object model projected into the image. The different colours indicate the different model component. **Right**: Segmentation with level-set-based segmentation and using K-means with 3 regions. The white part is the area occluded by the tracked object, i.e. the area removed from the segmentation process. Every other colour denotes a region. Although no information from the background image was used, segmentation results still look good.

Fig. 6. Experiment with varying background. **Upper row**: Model of the tea box to be tracked, input image with initialisation in first frame, and tracking results for frame 50, 150 and 180. **Lower row**: Input image (frame 90), result when using LMM, result with Parzen model, and results with Gaussian noise with LMM and the Parzen model.

5 Summary

We have presented a localised mixture model that splits the region whose appearance should be estimated into distinct subregions. The appearance of the region is then modelled by a mixture of densities, each applied in its local vicinity. For the partitioning step, we tested a fast K-means clustering as well as a multi-region segmentation algorithm based on level sets. We demonstrated the relevance of such a localised mixture model by quantitative experiments in model based tracking using the HumanEva-II benchmark. Results clearly improved when using this new model. Moreover, the approach is also applicable when a static background image is missing. In such cases tracking is only successful with the localised mixture model. We believe that such localised models can also be very beneficial in other object segmentation tasks, where low-level cues are combined with a-priori information, such as semi-supervised segmentation, or combined object recognition and segmentation.

References

1. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics **23**(3) (2004) 309–314
2. Criminisi, A., Cross, G., Blake, A., Kolmogorov, V.: Bilayer segmentation of live video. In: Proc. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, IEEE Computer Society (2006) 53–60
3. Paragios, N., Deriche, R.: Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision. Journal of Visual Communication and Image Representation **13**(1/2) (2002) 249–268

4. Sifakis, E., Garcia, C., Tziritas, G.: Bayesian level sets for image segmentation. Journal of Visual Communication and Image Representation **13**(1/2) (March 2002) 44–64
5. Cremers, D., Soatto, S.: Motion competition: A variational approach to piecewise parametric motion segmentation. International Journal of Computer Vision **62**(3) (May 2005) 249–265
6. Schmaltz, C., Rosenhahn, B., Brox, T., Weickert, J., Wietzke, L., Sommer, G.: Dealing with self-occlusion in region based motion capture by means of internal regions. In Perales, F.J., Fisher, R.B., eds.: Articulated Motion and Deformable Objects. Volume 5098 of Lecture Notes in Computer Science., Springer (July 2008) 102–111
7. Rosenhahn, B., Brox, T., Weickert, J.: Three-dimensional shape knowledge for joint image segmentation and pose tracking. International Journal of Computer Vision **73**(3) (July 2007) 243–262
8. Morya, B., Ardon, R., Thiran, J.P.: Variational segmentation using fuzzy region competition and local non-parametric probability density functions. In: Proc. Eleventh International Conference on Computer Vision, IEEE Computer Society Press (October 2007)
9. Grimson, W., Stauffer, C., Romano, R., Lee, L.: Using adaptive tracking to classify and monitor activities in a site. In: Proc. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, IEEE Computer Society Press (1998) 22–29
10. Pless, R., Larson, J., Siebers, S., Westover, B.: Evaluation of local models of dynamic backgrounds. In: Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2003) 73–78
11. Sun, J., Zhang, W., Tang, X., Shum, H.Y.: Background cut. In Leonardis, A., Bischof, H., Pinz, A., eds.: Computer Vision – ECCV 2006, Part II. Volume 3952 of Lecture Notes in Computer Science., Berlin, Springer (May 2006) 628–641
12. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: Proc. Twelfth International Conference on Computer Vision, IEEE Computer Society Press (2008)
13. Dambreville, S., Sandhu, R., Yezzi, A., Tannenbaum, A.: Robust 3D pose estimation and efficient 2D region-based segmentation from a 3D shape prior. In Forsyth, D., Torr, P., Zisserman, A., eds.: Computer Vision – ECCV 2008, Part II. Volume 5303 of Lecture Notes in Computer Science., Berlin, Springer (2008) 169–182
14. Bregler, C., Malik, J., Pullen, K.: Twist based acquisition and tracking of animal and human kinematics. International Journal of Computer Vision **56**(3) (January 2004) 179–194
15. Gavrila, D.M.: The visual analysis of human movement: a survey. Computer Vision and Image Understanding **73**(1) (January 1999) 82–98
16. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding **108**(1-2) (October 2007) 4–18
17. Elkan, C.: Using the triangle inequality to accelerate k-Means. In: Proceedings of the Twentieth International Conference on Machine Learning, AAAI Press (2003) 147–153
18. Gehler, P.: Mpikmeans (2007) http://mloss.org/software/view/48/.
19. Brox, T., Weickert, J.: Level set segmentation with multiple regions. IEEE Transactions on Image Processing **15**(10) (October 2006) 3213–3218
20. Sigal, L., Black, M.J.: HumanEva: Synchronized video and motion capture dataset for evaluation of articulated motion. Technical Report CS-06-08, Department of Computer Science, Brown University (September 2006)