# Level-Set Person Segmentation and Tracking with Multi-Region Appearance Models and Top-Down Shape Information
## *Appendix*


Esther Horbert, Konstantinos Rematas, Bastian Leibe

UMIC Research Centre, RWTH Aachen University

{horbert,leibe}@umic.rwth-aachen.de


This appendix contains a detailed derivation of our segmentation and tracking framework [27] and lists typical values for important parameters.

## A. Derivation

Fig. 1 shows the results of our full model for the sequence WALKSTRAIGHT [26], illustrating the evolution of the three contours.

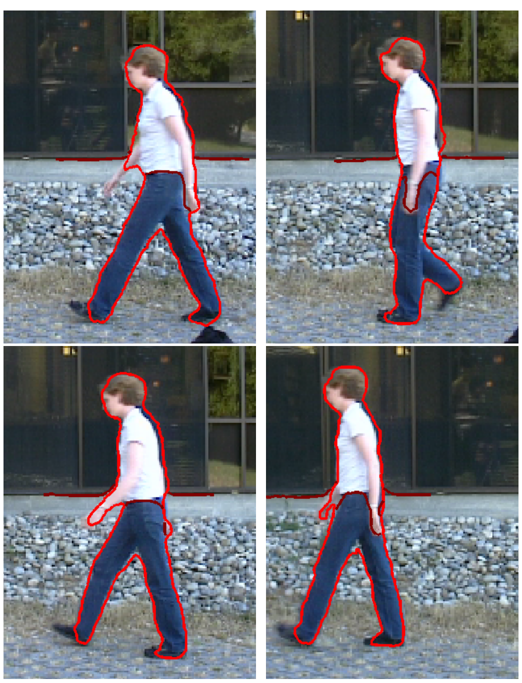

Figure 1. Results of our full model for the sequence WALKSTRAIGHT (125 frames, from [26]). This example nicely illustrates the evolution of the separating contours (dark red). The foreground contour $\mathbf{\Phi}_f$ is initialized with a horizontal line at 50% of the object frame height, the background contour $\mathbf{\Phi}_b$ with a line at 60%. However this is only the initialization and the lines can evolve to very different forms (as the person's contour does).

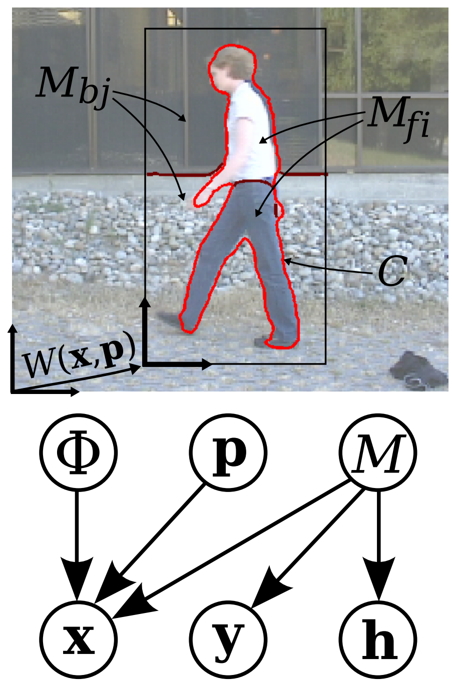


Figure 2. The generative model used in this approach, $M = \{M_{f1}, M_{f2}, M_{b1}, M_{b2}\}$.

| | |
|---|---|
| $\mathbf{x}$ | Pixel's coordinates inside reference frame |
| $\mathbf{y}$ | Pixel's color |
| $\mathbf{p}$ | Reference frame position |
| $\mathbf{h}$ | Shape model |
| $W(\mathbf{x},\mathbf{p})$ | Warp with parameters $\mathbf{p}$ |
| $M_{f1}, M_{f2}$ | Foreground regions |
| $M_{b1}, M_{b2}$ | Background regions |
| $P(\mathbf{y}\|M_k)$ | Appearance models |
| $\mathbf{\Phi}$ | Level set embedding function |
| $\{\mathbf{\Phi}_c, \mathbf{\Phi}_f, \mathbf{\Phi}_b\}$ | Embeddings for person and fore/background |
| $\mathbf{C}_k$ | Contour represented by the zero level set |
| $H_\epsilon(z)$ | Smoothed Heaviside step function |
| $\delta_\epsilon(z)$ | Smoothed Dirac delta function |

Table 1. Notation used in this paper

The joint distribution for one pixel given by the model in Fig. 2 is:

$$
\begin{aligned}
P(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i, \mathbf{\Phi}, \mathbf{p}, M) &= P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M)P(\mathbf{y}_i|M)P(\mathbf{h}_i|M)P(M)P(\mathbf{\Phi})P(\mathbf{p}) && (1)\\
& \text{(assumption: } \mathbf{y}, \mathbf{h} \text{ independent)} && (2)\\
P(\mathbf{x}_i, \mathbf{\Phi}, \mathbf{p}, M|\mathbf{y}_i, \mathbf{h}_i)P(\mathbf{y}_i)P(\mathbf{h}_i) &= P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M)P(\mathbf{y}_i|M)P(\mathbf{h}_i|M)P(M)P(\mathbf{\Phi})P(\mathbf{p}) && (3)\\
P(\mathbf{x}_i, \mathbf{\Phi}, \mathbf{p}, M|\mathbf{y}_i, \mathbf{h}_i) &= P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M)P(M|\mathbf{y}_i)P(M|\mathbf{h}_i)P(\mathbf{\Phi})P(\mathbf{p}) && (4)
\end{aligned}
$$

Marginalization over the models M yields the pixel-wise posterior probability of shape $\mathbf{\Phi}$ and location $\mathbf{p}$ given a pixel $\{\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i\}$:

$$
\sum_{k\in\{f1,f2,b1,b2\}} P(\mathbf{x}_i, \mathbf{\Phi}, \mathbf{p}, M_k|\mathbf{y}_i, \mathbf{h}_i) = P(\mathbf{x}_i, \mathbf{\Phi}, \mathbf{p}|\mathbf{y}_i, \mathbf{h}_i) = P(\mathbf{\Phi}, \mathbf{p}|\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i)P(\mathbf{x}_i) \quad (5)
$$

$$
= \sum_{k\in\{f1,f2,b1,b2\}} \Big\{P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M_k)\frac{P(\mathbf{y}_i|M_k)P(M_k)}{\sum_l P(\mathbf{y}_i|M_l)P(M_l)}P(M_k|\mathbf{h}_i)\Big\}P(\mathbf{\Phi})P(\mathbf{p}) \quad (6)
$$

It follows

$$
P(\mathbf{\Phi}, \mathbf{p}|\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i) = \frac{1}{P(\mathbf{x})}\sum_{k\in\{f1,f2,b1,b2\}} \Big\{P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M_k)\frac{P(\mathbf{y}_i|M_k)P(M_k)}{\sum_l P(\mathbf{y}|M_l)P(M_l)}P(M_k|\mathbf{h}_i)\Big\}P(\mathbf{\Phi})P(\mathbf{p}) \quad (7)
$$

We use a smoothed Heaviside step function $H_\epsilon$ to select the respective regions and a smoothed Dirac delta function $\delta_\epsilon$ to select the contours:

$$
\begin{aligned}
H_c &= H_\epsilon(\mathbf{\Phi}_c(\mathbf{x}_i)), \quad \tilde{H}_c = 1 - H_\epsilon(\mathbf{\Phi}_c(\mathbf{x}_i)) && (8)\\
H_f &= H_\epsilon(\mathbf{\Phi}_f(\mathbf{x}_i)), \quad \tilde{H}_f = 1 - H_\epsilon(\mathbf{\Phi}_f(\mathbf{x}_i)) && (9)\\
H_b &= H_\epsilon(\mathbf{\Phi}_b(\mathbf{x}_i)), \quad \tilde{H}_b = 1 - H_\epsilon(\mathbf{\Phi}_b(\mathbf{x}_i)) && (10)
\end{aligned}
$$

$$
P(M_k) = \frac{\eta_k}{\eta}, k \in \{f1, f2, b1, b2\}, \quad M_f = M_{f1} \cup M_{f2}, M_b = M_{b1} \cup M_{b2} \quad (11)
$$

The number of pixels in the four respective regions can be obtained as follows:

$$
N = \sum \eta_k,\ \ \eta_{f1} = \sum_{i=1}^{N} H_c H_f,\ \ \eta_{f2} = \sum_{i=1}^{N} H_c \tilde{H}_f, \quad (12)
$$

$$
\eta_{b1} = \sum_{i=1}^{N} \tilde{H}_c H_b,\ \ \eta_{b2} = \sum_{i=1}^{N} \tilde{H}_c \tilde{H}_b \quad (13)
$$

and thus the probability of pixel $x_i$ for each region:

$$P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},M_{f1})=\frac{H_cH_f}{\eta_{f1}},\quad P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},M_{f2})=\frac{H_c\tilde{H}_f}{\eta_{f2}}, \tag{14}$$

$$P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},M_{b1})=\frac{\tilde{H}_cH_b}{\eta_{b1}},\quad P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},M_{b2})=\frac{\tilde{H}_c\tilde{H}_b}{\eta_{b2}}. \tag{15}$$

Fusing the pixel-wise posteriors with a logarithmic opinion pool yields

$$P(\mathbf{\Phi},\mathbf{p}|\mathbf{x},\mathbf{y},\mathbf{h}) \tag{16}$$

$$=\prod_{i=1}^{N}\sum_{k\in\{f1,f2,b1,b2\}}\Big[P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},M_k)\frac{P(\mathbf{y}_i|M_k)P(M_k)}{\sum_l P(\mathbf{y}_i|M_l)P(M_l)}P(M_k|\mathbf{h}_i)\Big]P(\mathbf{\Phi})P(\mathbf{p}) \tag{17}$$

$$=\prod_{i=1}^{N}\Big[\frac{H_cH_f}{\eta_{f1}}\frac{P(\mathbf{y}_i|M_{f1})P(M_{f1})}{\sum_l P(\mathbf{y}_i|M_l)P(M_l)}P(M_{f1}|\mathbf{h}_i)+\frac{H_c\tilde{H}_f}{\eta_{f2}}\frac{P(\mathbf{y}_i|M_{f2})P(M_{f2})}{\sum_l P(\mathbf{y}_i|M_l)P(M_l)}P(M_{f2}|\mathbf{h}_i) \tag{18}$$

$$+\frac{\tilde{H}_cH_b}{\eta_{b1}}\frac{P(\mathbf{y}_i|M_{b1})P(M_{b1})}{\sum_l P(\mathbf{y}_i|M_l)P(M_l)}P(M_{b1}|\mathbf{h}_i)+\frac{\tilde{H}_c\tilde{H}_b}{\eta_{b2}}\frac{P(\mathbf{y}_i|M_{b2})P(M_{b2})}{\sum_l P(\mathbf{y}_i|M_l)P(M_l)}P(M_{b2}|\mathbf{h}_i)\Big]P(\mathbf{\Phi})P(\mathbf{p}) \tag{19}$$

$$=\prod_{i=1}^{N}\Big[H_cH_fP_{f1}+H_c\tilde{H}_fP_{f2}+\tilde{H}_cH_bP_{b1}+\tilde{H}_c\tilde{H}_bP_{b2}\Big]P(\mathbf{\Phi})P(\mathbf{p}) \tag{20}$$

$$=\prod_{i=1}^{N}P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},\mathbf{y}_i,\mathbf{h}_i)P(\mathbf{\Phi})P(\mathbf{p}) \tag{21}$$

where

$$P_k=\frac{P(\mathbf{y}_i|M_k)P(M_k)P(M_k|\mathbf{h}_i)}{\eta_k\sum_l P(\mathbf{y}_i|M_l)P(M_l)}=\frac{P(\mathbf{y}_i|M_k)P(M_k|\mathbf{h}_i)}{\sum_l \eta_l P(\mathbf{y}_i|M_l)},\qquad k\in\{f1,f2,b1,b2\} \tag{22}$$

$$P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},\mathbf{y}_i,\mathbf{h}_i)=H_cH_fP_{f1}+H_c\tilde{H}_fP_{f2}+\tilde{H}_cH_bP_{b1}+\tilde{H}_c\tilde{H}_bP_{b2}. \tag{23}$$

$P(\mathbf{y}_i|M_k)$ is computed from the appearance models, *i.e.* color histograms.

Eq. (21) contains $P(M_k|h)$ for four regions, but the detector only provides probabilities for two regions: foreground and background. However, (21) selects a model for each region by use of $H$, $H_f$ and $H_b$ respectively. We set

$$P(M_f|h)=P(M_{f1}|h)+P(M_{f2}|h)=H_fP(M_f|h)+\tilde{H}_fP(M_f|h) \tag{24}$$

$$P(M_b|h)=P(M_{b1}|h)+P(M_{b2}|h)=H_bP(M_b|h)+\tilde{H}_bP(M_b|h). \tag{25}$$

Thus it holds:

$$\text{Either } P(M_f|h)=P(M_{f1}|h) \text{ or } P(M_f|h)=P(M_{f2}|h) \text{ and} \tag{26}$$

$$\text{either } P(M_b|h)=P(M_{b1}|h) \text{ or } P(M_b|h)=P(M_{b2}|h). \tag{27}$$

This means in practice $P(M_{f1}|h)$ and $P(M_{f1}|h)$ are both set to $P(M_f|h)$ (background accordingly), which is possible because for each pixel one of the subregions is selected.

We now specify $P(\mathbf{\Phi})$ as the internal energy of the level set embedding function(s). It contains a geometric prior that rewards a signed distance function: the gradient of the level set function is a normal distribution with mean 1. It thus makes the level set embedding function numerically stable without the need for periodic re-initializations [20]. The second term (as in [10]) describes the length of the contour and rewards a smoother contour. This is very useful for cluttered scenes, where pixels with foreground or background appearance can form very small regions, which can easily result in a very uneven contour.

$$P(\mathbf{\Phi})=\prod_{i=1}^{N}\Big[\frac{1}{\sigma\sqrt{2\pi}}\exp\Big(-\frac{(|\nabla\mathbf{\Phi}(\mathbf{x}_i)|-1)^2}{2\sigma^2}\Big)\ \exp\Big(-\lambda|\nabla H_\epsilon(\mathbf{\Phi})|\Big)\Big] \tag{28}$$

where $\sigma$ and $\lambda$ are the weights of the priors.
Maximizing the posterior is equivalent to minimizing its negative logarithm:

$$\begin{aligned}\mathcal{E}(\mathbf{\Phi}) =& -\log(P(\mathbf{\Phi},\mathbf{p}|\mathbf{x},\mathbf{y},\mathbf{h})) \\ \propto& -\left(\sum_{i=1}^{N}\left\{\log(P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},\mathbf{y}_i,\mathbf{h}_i)) - \frac{(|\nabla\mathbf{\Phi}|-1)^2}{2\sigma^2} - \lambda|\nabla H_\epsilon(\mathbf{\Phi})|\right\} + N\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log(P(\mathbf{p}))\right) \end{aligned} \tag{29}$$

### A.1. Derivation of the Segmentation Framework

For segmentation we optimize (29) w.r.t $\mathbf{\Phi}$, so the last two terms can be dropped, the rest is then differentiated by calculus of variation:

$$\begin{aligned}\frac{\partial\mathbf{\Phi}_c}{\partial t} &= -\frac{\partial\mathcal{E}(\mathbf{\Phi}_c)}{\partial\mathbf{\Phi}_c} = \frac{\delta H_f P_{f1} + \delta\tilde{H}_f P_{f2} - \delta H_b P_{b1} - \delta\tilde{H}_b P_{b2}}{P(\mathbf{x}|\mathbf{\Phi},\mathbf{p},\mathbf{y},\mathbf{h})} - \frac{1}{\sigma^2}\left[\nabla^2(\mathbf{\Phi}_c) - \operatorname{div}\left(\frac{\nabla\mathbf{\Phi}_c}{|\nabla\mathbf{\Phi}_c|}\right)\right] \\ &\quad + \lambda\delta_\epsilon(\mathbf{\Phi}_c)\operatorname{div}\left(\frac{\nabla\mathbf{\Phi}_c}{|\nabla\mathbf{\Phi}_c|}\right)\end{aligned} \tag{30}$$

$$\begin{aligned}\frac{\partial\mathbf{\Phi}_f}{\partial t} &= -\frac{\partial\mathcal{E}(\mathbf{\Phi}_f)}{\partial\mathbf{\Phi}_f} = \frac{H_c\delta_f P_{f1} - H_c\delta_f P_{f2} + \tilde{H}_c H_b P_{b1} + \tilde{H}_c\tilde{H}_b P_{b2}}{P(\mathbf{x}|\mathbf{\Phi},\mathbf{p},\mathbf{y},\mathbf{h})} - \frac{1}{\sigma^2}\left[\nabla^2(\mathbf{\Phi}_f) - \operatorname{div}\left(\frac{\nabla\mathbf{\Phi}_f}{|\nabla\mathbf{\Phi}_f|}\right)\right] \\ &\quad + \lambda\delta_\epsilon(\mathbf{\Phi}_f)\operatorname{div}\left(\frac{\nabla\mathbf{\Phi}_f}{|\nabla\mathbf{\Phi}_f|}\right)\end{aligned} \tag{31}$$

$$\approx^{(\mathbf{\Phi}_f : H \overset{!}{>} 0)} \frac{\delta_f(P_{f1} - P_{f2})}{P(\mathbf{x}|\mathbf{\Phi},\mathbf{p},\mathbf{y},\mathbf{h})} - \frac{1}{\sigma^2}\left[\nabla^2(\mathbf{\Phi}_f) - \operatorname{div}\left(\frac{\nabla\mathbf{\Phi}_f}{|\nabla\mathbf{\Phi}_f|}\right)\right] + \lambda\delta_\epsilon(\mathbf{\Phi}_f)\operatorname{div}\left(\frac{\nabla\mathbf{\Phi}_f}{|\nabla\mathbf{\Phi}_f|}\right) \tag{32}$$

$$\frac{\partial\mathbf{\Phi}_b}{\partial t} = -\frac{\partial\mathcal{E}(\mathbf{\Phi}_b)}{\partial\mathbf{\Phi}_b} \quad \text{accordingly.}$$

We evolve the two additional level set functions interleaved with the original level set function $\mathbf{\Phi}_c$. In this way, the four appearance models are optimized at the same time, which leads to more robust and accurate segmentation results.

In our implementation we use $\sigma^2 = 50$, $\lambda = 2$, $\tau = 2$, $\epsilon = 6$ and

$$H_\epsilon(x) = \begin{cases} 0 & \text{if } x < -\epsilon \\ \frac{x}{2\epsilon} + \frac{1}{2\pi}\sin\left(\frac{\pi x}{\epsilon}\right) + \frac{1}{2} & \text{if } |x| < \epsilon \\ 1 & \text{if } x > \epsilon \end{cases} \tag{33}$$

$$\delta_\epsilon(x) = \begin{cases} \frac{1}{2\epsilon}\left(1 + \cos\left(\frac{\pi x}{\epsilon}\right)\right) & \text{if } |x| < \epsilon \\ 0 & \text{else} \end{cases} \tag{34}$$

### A.2. Derivation of the Tracking Framework

In preparation for differentiation w.r.t $\mathbf{p}$ some terms in (29) can be dropped:

$$\mathcal{E}(\mathbf{\Phi}) \propto -\left(\sum_{i=1}^{N}\{\log(P(\mathbf{x}_i|\mathbf{\Phi},\mathbf{p},\mathbf{y}_i,\mathbf{h}_i))\} + \log(P(\mathbf{p})) + const.\right) \tag{35}$$

Now the warp $\mathbf{W}(\mathbf{x}_i, \Delta\mathbf{p})$ is introduced into (35), *i.e.* pixels $\mathbf{x}_i$ are warped with parameters $\mathbf{p}$. $P(\mathbf{p})$ is dropped for the moment, this is handled with drift correction, as in [3]:

$$\mathcal{E}(\mathbf{\Phi}) \quad \propto \quad -\sum_{i=1}^{N}\log\left\{P(\mathbf{W}(\mathbf{x}_i,\Delta\mathbf{p})|\mathbf{\Phi},\mathbf{p},\mathbf{y}_i,\mathbf{h}_i)\right\} \tag{36}$$

We maximize w.r.t $\mathbf{p}$:

$$\mathbf{p} = \quad \arg\max_{\mathbf{p}}\left\{\sum_{i=1}^{N}\log P(\mathbf{W}(\mathbf{x}_i,\Delta\mathbf{p})|\mathbf{\Phi},\mathbf{p},\mathbf{y}_i,\mathbf{h}_i)\right\} \tag{37}$$

We use a second order Newton optimization scheme as in [4]: With the short-hand notation $P(...) = P(\mathbf{W}(\mathbf{x}_i, \Delta\mathbf{p})|\boldsymbol{\Phi}, \mathbf{p}, \mathbf{y}_i, \mathbf{h}_i)$:

$$\Delta\mathbf{p} = \Big[\sum_{i=1}^{N} \frac{\left(\frac{\partial P(...)}{\partial \mathbf{p}}\right)^2}{P(...)}\Big]^{-1} \sum_{i=1}^{N} \frac{\partial P(...)}{\partial \mathbf{p}} \tag{38}$$

where

$$\begin{aligned}\frac{\partial P(...)}{\partial \mathbf{p}} &= (\mathbf{J}_c H_f + H_c \mathbf{J}_f) P_{f1} + (\mathbf{J}_c \tilde{H}_f - H_c \mathbf{J}_f) P_{f2} - \mathbf{J}_c H_b P_{b1} - \mathbf{J}_c \tilde{H}_b P_{b2} \\ &= \mathbf{J}_c (H_f P_{f1} + \tilde{H}_f P_{f2} - H_b P_{b1} - \tilde{H}_b P_{b2}) + \mathbf{J}_f (H_c P_{f1} - H_c P_{f2})\end{aligned} \tag{39}$$

with

$$\mathbf{J}_c = \frac{\partial H_c}{\partial \boldsymbol{\Phi}_c} \frac{\partial \boldsymbol{\Phi}_c}{\partial \mathbf{x}} \frac{\partial \mathbf{W}}{\partial \Delta\mathbf{p}} = \delta_\epsilon(\boldsymbol{\Phi}_c(\mathbf{x}_i)) \nabla \boldsymbol{\Phi}_c(\mathbf{x}_i) \frac{\partial \mathbf{W}}{\partial \Delta\mathbf{p}}, \tag{40}$$

$$\mathbf{J}_f = \frac{\partial H_f}{\partial \boldsymbol{\Phi}_f} \frac{\partial \boldsymbol{\Phi}_f}{\partial \mathbf{x}} \frac{\partial \mathbf{W}}{\partial \Delta\mathbf{p}} = \delta_\epsilon(\boldsymbol{\Phi}_f(\mathbf{x}_i)) \nabla \boldsymbol{\Phi}_f(\mathbf{x}_i) \frac{\partial \mathbf{W}}{\partial \Delta\mathbf{p}} \tag{41}$$

We assume that the background does not move with the foreground, thus $H_\epsilon(\Phi_b(\mathbf{x}_i))$ is constant w.r.t the derivation $\frac{\partial}{\partial \mathbf{p}}$. Eq. (39) illustrates that both the person's contour and the division line of the foreground contribute to the warp, whereas the division line of the background does not contribute: $\mathbf{J}_c$ and $\mathbf{J}_f$ contain the factor $\delta_\epsilon(\boldsymbol{\Phi}_k)$, the Dirac delta function of the respective level set function, which is only greater than zero in a narrow band around the contour (with width $\epsilon$).
For parameters $\mathbf{p} = (s, t_x, t_y)^{\mathrm{T}}$ with scale and translation the warp is:

$$\mathbf{W}(\mathbf{x}; \mathbf{p}) = \begin{pmatrix} 1+s & 0 & t_x \\ 0 & 1+s & t_y \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} (1+s)\cdot x + t_x \\ (1+s)\cdot y + t_y \end{pmatrix} \tag{42}$$

The term $\frac{\partial \mathbf{W}}{\partial \mathbf{p}}$ is the Jacobian of the warp and with $\mathbf{W}(\mathbf{x}; \mathbf{p}) = (W_x(\mathbf{x}; \mathbf{p}), W_y(\mathbf{x}; \mathbf{p}))^{\mathrm{T}}$:

$$\frac{\partial \mathbf{W}}{\partial \mathbf{p}} = \begin{pmatrix} x & 1 & 0 \\ y & 0 & 1 \end{pmatrix} \tag{43}$$

In our implementation the rigid registration typically converges after 10 to 40 iterations.

## References


[3] C. Bibby and I. Reid. Robust Real-Time Visual Tracking using Pixel-Wise Posteriors. *ECCV*, 2008.

[4] C. Bibby and I. Reid. Real-time Tracking of Multiple Occluding Objects using Level Sets. *CVPR*, 2010.

[10] D. Cremers, M. Rousson, and R. Deriche. A Review of Statistical Approaches to Level Set Segmentation Integrating Color, Texture, Motion and Shape. *IJCV*, 72:195-215, 2007.

[20] C. Li, C. Xu, C. Gui, and M. Fox. Level Set Evolution without Re-initialization: A New Variational Formulation. *CVPR*, 2005.

[26] H. Sidenbladh and M. J. Black. Learning the statistics of people in images and video. *IJCV*, 54(1-3):183-209, 2003.

[27] E. Horbert, K. Rematas and B. Leibe. Level-Set Person Segmentation and Tracking with Multi-Region Appearance Models and Top-Down Shape Information. *ICCV*, 2011.