ICCV 2011

Object Segmentation in Video: A Hierarchical Variational Approach for Turning Point Trajectories into Dense Regions

Peter Ochs and Thomas Brox
Computer Vision Group
University of Freiburg, Germany
{ochs, brox}@informatik.uni-freiburg.de

Abstract

Point trajectories have emerged as a powerful means to obtain high quality and fully unsupervised segmentation of objects in video shots. They can exploit the long term motion difference between objects, but they tend to be sparse due to computational reasons and the difficulty in estimating motion in homogeneous areas. In this paper we introduce a variational method to obtain dense segmentations from such sparse trajectory clusters. Information is propagated with a hierarchical, nonlinear diffusion process that runs in the continuous domain but takes superpixels into account. We show that this process raises the density from 3% to 100% and even increases the average precision of labels.

Figure 1. Object segmentation in videos based on point trajectories [7] yields sparse object labels. We propose a hierarchical variational model that propagates these labels preferably in homogenous regions. As a result we obtain a dense segmentation that reveals an even higher classification accuracy than the sparse labels and overcomes the over-segmentation issue of static image segmentation approaches [1].

1. Introduction

Current learning frameworks for visual recognition rely on manual annotation including manual segmentation of objects. Taking the best vision system solution to-date – the human brain – such annotation should not be necessary. Infants learn the visual appearance and shape of objects without being provided bounding boxes and segmentations by their parents. There is convincing evidence that infants obtain such object segmentations via motion cues [22, 17] and one could argue that computational vision systems should finally work in a similar way.

Motion analysis of point trajectories is a reasonably robust tool to extract object regions from video shots in a fully unsupervised manner, as recently demonstrated, e.g., in [20, 7]. However, these approaches have to struggle with the fact that motion estimation requires structures to match. In homogeneous areas of the image there are no such structures. This results in point trajectories to be sparse. Although the work in [7] is based on point trajectories derived from dense optical flow and the obtained trajectories could be made dense, trajectories in homogenous areas are less reliable and can hamper their clustering. Moreover, computational constraints require reducing the number of trajectories being analyzed. Spectral clustering of dense point trajectories would be far too slow. Hence, [7] yields trajectory clusters that look very appealing but are sparse – a severe problem, for instance, when aiming to learn shape priors or detectors from the segmentation.

In this paper, we present a variational method that creates dense segmentations from sparse clusters of point trajectories, as shown in Fig. 1. At first glance, this appears to be a simple interpolation problem. Our brain can easily fill the areas between the dots. However, a second glance reveals significant challenges. The point trajectories do not cover some of the most critical areas, especially near the object

boundaries. Those trajectories that exist near object boundaries are often assigned wrong labels because the underlying optical flow is imprecise in occlusion areas. Finally, in large homogeneous areas there is hardly any label information.

For a good label propagation, the key is to exploit color and edge information, which is complementary to the information used for generating the trajectory labels. Notably, segmentation based on color works the best in homogeneous areas, which are the problematic areas for motion based segmentation. We achieve this by spreading information depending on color homogeneity. To this end, we propose a hierarchical variational model, where we have continuous labeling functions on multiple levels. Each level corresponds to a superpixel partitioning at a certain granularity level. In contrast to a single-level model we have additional auxiliary functions at coarser levels which are optimized in a coupled diffusion process. To the best of our knowledge, this is the first continuous hierarchical model. The advantage of such an approach is that we obtain a structure-aware label propagation thanks to the superpixel hierarchy, while the final solution can be extracted at the finest level and is free of metrication errors or block artifacts known from discrete MRF models.

2. Related work

The problem we consider here is related to interactive segmentation, where the user draws a few scribbles into the image and the approach propagates these labels to the non-marked areas. Several techniques based on graph cuts [4], random walks [12], and intermediate settings [21] have been proposed. The latest techniques are built upon variational convex relaxation methods [23, 18, 14, 16], which avoid the discretization artifacts typical for classical MRFs defined on a graph. The variational technique we propose here in fact builds on the regularizer from [14].

None of these approaches consider a hierarchy. Moreover, labels from point trajectories differ from user scribbles in two ways. First, the labels are generated by an unsupervised approach and are likely to be erroneous, whereas interactive segmentation relies on the correctness of the user's input. This means, we do not have an interpolation but an approximation problem. For this reason, we do not follow the typical approach of estimating appearance statistics from the annotated areas combined with a typical region based segmentation. We rather formulate the problem as a label diffusion problem, where diffusion takes place on various hierarchy levels. Second, user annotation provides dense finite annotation areas, whereas trajectory labels constitute single points spread over the image. It is not immediate that a variational model acting on such infinitesimally small labels makes sense in the continuous limit. We show that our continuous model satisfies certain regularity properties and thus is a proper continuous formulation of the problem.

Our model is also related to image compression with anisotropic diffusion [11], where only a small set of pixel values is kept and the original image is sought to be restored by running a diffusion process on this sparse representation.

On the task of dense motion segmentation, there are many recent works that produce over-segmentations using superpixels, label propagation by optical flow, or other clustering methods [5, 13, 24, 15]. These over-segmentations do not provide object regions. User interactive video segmentation methods can avoid over-segmentation, but are no longer unsupervised [3, 19].

3. Single-level variational model

Given a video shot as input, we apply the clustering approach on point trajectories using the code from [7]. This yields a discrete sparse set of labels, which are consistent over the whole shot. Fig. 1 shows an example frame and the labels obtained with [7]. Approximately 3% of the pixels are labeled.

Let $\tilde{u} := (\tilde{u}_1, \dots, \tilde{u}_n) \colon \Omega \to \{0,1\}^n$, $n \in \mathbb{N}$ be a function indicating the n different point trajectory labels, i.e.,

$$\tilde{u}_i := \begin{cases} 1, & \text{if } x \in L_i \\ 0, & \text{else,} \end{cases} \tag{1}$$

where L_i is the set of coordinates occupied by a trajectory with label i and $\Omega \subset \mathbb{R}^2$ denotes the image domain. For simplicity, we focus on single images, although our model could be easily extended to compute a temporally consistent solution for all images of the video.

We seek a function $u := (u_1, \dots, u_n) \colon \Omega \to \{0,1\}^n$ that stays close to the given labels for points in $L := \bigcup_{i=1}^n L_i$. This is achieved by minimizing the energy

$$E_{\text{data}}(u) := \frac{1}{2} \int_\Omega c \sum_{i=1}^n (u_i - \tilde{u}_i)^2 dx, \tag{2}$$

where $c \colon \Omega \to \{0,1\}$ is the label indicator function, a characteristic function with value 1 on L and 0 else.

The above energy puts constraints only on points occupied by trajectories. On all other points, the minimizer can take any value. To force these points to take specific labels, we require a regularizer

$$E_{\text{reg}}(u) := \int_\Omega g \, \psi \left(\sum_{i=1}^n |\nabla_{\mathcal{B}\mathcal{V}} u_i|^2 \right) dx, \tag{3}$$

which is chosen such that it prefers compact regions with minimal perimeter and such that labels are propagated preferably in homogenous areas. The first is achieved by the regularized TV norm obtained with $\psi(s^2) := \sqrt{s^2 + \varepsilon^2}$

and $\varepsilon := 0.001$, the second by the diffusivity function $g\colon \Omega \to \mathbb{R}^+$

$$g(|\nabla I(x)|^2) := \frac{1}{\sqrt{|\nabla I(x)|^2 + \varepsilon^2}}. \tag{4}$$

Since u_i is binary-valued, it lies in the space of functions of bounded variation $\mathcal{BV}(\Omega)$, which includes functions with sharp discontinuities [2]. The ordinary gradient operator ∇ is not defined for such functions. Thus, we replace it by $\nabla_{\mathcal{BV}}$, which is the distributional derivative and is defined also for characteristic functions. In case of continuous functions $\nabla_{\mathcal{BV}} \equiv \nabla$.

Convex combination of the energies in (2) and (3) yields

$$\begin{aligned} E(u) := \frac{\alpha}{2} \int_\Omega c \sum_{i=1}^n (u_i - \tilde{u}_i)^2 \, dx \\ + (1-\alpha) \int_\Omega g\, \psi\left(\sum_{i=1}^n |\nabla_{\mathcal{BV}} u_i|^2 \right) dx \end{aligned} \tag{5}$$

s.t. $\sum_i u_i(x) = 1,\ \forall x$ with a model parameter $\alpha \in [0,1)$ that can be chosen depending on the credibility of the trajectory labels. In the limit $\alpha \to 1$, the minimizer will be an interpolation, otherwise it is an approximation that can correct erroneous labels.

3.1. Minimization

The functions u_i are binary-valued. For we can apply variational methods we must consider the relaxed problem, where we allow u_i to take values in the interval $[0,1]$. Such relaxations have been suggested for many similar problems [8, 18, 14]. Since both the energy and the relaxed feasible set are convex, we find the global minimizer of the relaxed problem. Except for the two-label setting in [8], projecting this minimizer back to the original feasible set generally does not ensure the global optimum but yields a good approximation.

For the relaxed problem, we may assume that u is continuous differentiable, so we may replace $\nabla_{\mathcal{BV}}$ by ∇. The Euler-Lagrange equations of the relaxed energy reads

$$\begin{aligned} 0 &= \alpha\, c \cdot (u_i - \tilde{u}_i) \\ &\quad - (1-\alpha)\, \mathrm{div} \left(g\, \psi' \left(\sum_{i=1}^n |\nabla u_i|^2 \right) \nabla u_i \right) \quad \forall i. \end{aligned} \tag{6}$$

We solve this nonlinear system with a fixed point iteration scheme, where the nonlinear factor $\psi'(s^2) = (s^2+\varepsilon^2)^{-\frac{1}{2}}$ is kept constant in each iteration. The resulting linear system is solved with successive over-relaxation (SOR). The constraint $\sum_i u_i(x) = 1,\ \forall x$ is enforced in each fixed point iteration by normalization as proposed in [9]. The obtained relaxed result is projected to $\{0,1\}^n$ via

$$u_i(x) = \begin{cases} 1, & \text{if } i = \mathrm{argmax}_i \{u_i | i = 1, ..., n\} \\ 0, & \text{else.} \end{cases} \tag{7}$$

3.2. Model consistency

It can be argued that c and $\tilde{u}$, as indicator functions, differ from 1 only on a discrete set, i.e., on a null-set. Following the theory of Lebesgue measures, the energy $E_{\text{data}} \equiv 0$. This problem can be avoided by a minor adaption of the model. Replace c by its C^∞-approximation $c_\varepsilon \in C^\infty(\Omega, [0,1])$ satisfying $c_\varepsilon \to c$ as $\varepsilon \to 0$ and the mass conservation $\int_{B_\varepsilon(x)} c_\varepsilon(x) dx = 1$, where $B_\varepsilon(x)$ denotes the ball with radius ε at x, e.g.,

$$c_\varepsilon(x) := \sum_{\tilde{x} \in L} \delta_\varepsilon(x - \tilde{x}). \tag{8}$$

Define $\tilde{u}_\varepsilon$ analogously as such an approximation to $\tilde{u}$. Note that the mappings c_ε and $\tilde{u}_\varepsilon$ are well-defined as mappings with range $[0,1]$ and $[0,1]^n$, respectively. The assumption of a discrete set of input labels assures non-overlapping ε-neighborhoods, and thus implies the well-definedness. Choosing $\varepsilon > 0$ small enough to satisfy this separation property for all $\tilde{x} \in L$ yields the consistent regularised energy $E_{\text{data},\varepsilon}$ by replacing c and $\tilde{u}$ by c_ε and $\tilde{u}_\varepsilon$ in (2). We omit the ε due to notational simplicity.

4. Multi-level variational model

The Euler-Lagrange equations in (6) can be interpreted as a nonlinear diffusion process subject to the constraint that the points in L should stay close to their original label $\tilde{u}$. These points serve as sources for spreading their respective label information. The antipoles that consume this label mass are other labels in the neighborhood. Depending on how much label mass is consumed by neighboring points and depending on α, a source point $x \in L$ can also change its label in the final solution u, but it still remains a source of the original label's mass.

Especially in homogeneous areas, the density of source points is low, i.e., the information must be propagated over large spatial distances, damped by noise and unimportant structures. To overcome this problem, we propose a hierarchical model. The finest level in this hierarchy corresponds to the single-level model we have introduced in the previous section. Additional levels make use of superpixels obtained with the approach from [1]. Fig. 2a illustrates the continuous hierarchy. Fig. 2b shows the corresponding discrete graph structure for helping readers who prefer to think in discrete terms.

Each level k, $k = 0, ..., K$, in our variational model represents a continuous function that is partitioned into M^k superpixels $\Omega_m^k, m = 1, ..., M^k$. For $k = 0$ we have the functions u^0 and I^0 as defined for the single-level model. For $k > 0$ we have the corresponding piecewise constant functions u^k and I^k, where $I^k(x) = \frac{1}{|\Omega_m^k|} \int_{\Omega_m^k} I^0(x') dx'$ takes the mean color of the corresponding superpixel Ω_m^k. The idea behind these additional auxiliary functions at

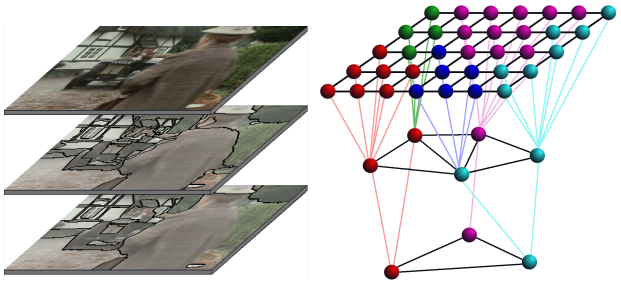

Figure 2. Illustration of the multi-level model. **Left: (a)** Continuous model, where each level is a continuous function. Coarser levels are piecewise constant according to their superpixel partitioning. **Right: (b)** Corresponding discrete graph structure in terms of pixels/superpixels showing the linkage between levels. Our model is in fact *not* a graph, as each level is continuous.

coarser levels is to define a label diffusion process that better adapts to the image structures at multiple scales.

We extend the single-level energy from (5) accordingly:

$$\begin{aligned} E(u) := \frac{\alpha}{2} \int_\Omega \rho\, c \sum_{i=1}^{n} (u_i^0 - \tilde{u}_i)^2 \, dx \\ + (1-\alpha) \sum_{k=0}^{K} \int_\Omega g^k \, \psi\left(\sum_{i=1}^{n} |\nabla_{\mathcal{BV}} u_i^k|^2 \right) dx \\ + (1-\alpha) \sum_{k=1}^{K} \int_\Omega g_l^k \, \psi\left(\sum_{i=1}^{n} |u_i^k - u_i^{k-1}|^2 \right) dx, \end{aligned} \tag{9}$$

where $u := (u_1^0, ..., u_n^0, u_1^1, ..., u_n^1, ..., u_1^K, ..., u_n^K)$ denotes the label function of the whole hierarchy. The first term is identical to the single-level model, except for an additional weighting function $\rho \colon \Omega \to \mathbb{R}$, which will be explained later. Label sources exist only in the finest level. They propagate their information to the coarser levels via the third term, which connects successive levels. The level diffusivity functions g_l^k have the same meaning as the spatial diffusivities g^k, but are defined based on the color distance between levels

$$g_l^k(x) := \frac{1}{\sqrt{|I^k(x) - I^{k-1}(x)|^2 + \varepsilon^2}}, \quad \varepsilon = 0.001 \tag{10}$$

rather than the image gradient $|\nabla I|^2$. The second term in (9) is a straightforward extension of the corresponding term in the single-level model.

What is the effect of the additional levels? The superpixels at coarser levels lead to regions with constant values I^k. Consequently, $\nabla I^k = 0$ within a superpixel, which leads to infinite diffusivities g^k. [1] In other words, within a superpixel, label information is propagated with infinite speed

[1] These diffusivities are only made finite for numerical reasons by means of the regularizing constant ε.

Figure 3. Difference between a discrete (top) and a continuous model (bottom). The discrete model shows block artifacts since its discretization error does not converge to 0 for finer grids sizes.

across the whole superpixel. Thanks to the connections between levels, this also affects points on the next finer level. Rather than traveling the long geodesic distance on the fine level hindered by noisy pixels and weak structures, information can take a shortcut via a coarser level where this noise has been removed.

The hierarchy comes with the great advantage that we need not choose the "correct" noise level, which just might not exist globally. Instead, we consider multiple levels from the superpixel hierarchy and integrate them all into our model. In theory, it is advantageous to have as many levels as possible. For computational reasons, however, it is wise to focus on a small number of levels. Our experiments indicate that three levels are already sufficient to benefit from the hierarchical model.

Since we formulated the hierarchy as a continuous, variational model rather than a common discrete, graphical model, we have the advantage that we do not suffer from discretization artifacts. This is shown in Fig. 3, where we compare our continuous model to an implementation based on the graph structure in Fig. 2b. Even though we have to discretize the Euler-Lagrange equations to finally implement the model on discrete pixel data, this discretization is *consistent*, i.e., the discretization error decreases as the image resolution increases. Moreover, a rotation of the grid does not change the outcome. These natural properties are missing in discrete models.

In (9) we have introduced a weighting function ρ that allows to give more weight to certain trajectories. The incentive is that the approach in [7] tends to produce wrong labels close to object boundaries due to inaccuracies of the optical flow in such areas. Hence, it makes sense to increase the influence of labels that are far away from object boundaries, whereas labels close to these boundaries should get less in-

Figure 4. Evolution of the label functions u_1^k on all three levels simultaneously. Intermediate states after 30, 300, 3000, and 30000 iterations are shown. For this visualization we did not use the cascadic multigrid strategy, which requires far fewer iterations to converge.

fluence. Since we do not yet know the object boundaries, the distance is approximated by the Euclidean distance to the superpixel boundaries $\partial\Omega_m$ at the coarsest level, which can be computed very efficiently [10]. Based on these distances we define

$$\rho(x) := \frac{dist(x, \partial\Omega_m)}{\frac{1}{|\Omega_m|}\sum_{x\in\Omega_m} dist(x, \partial\Omega_m)} \tag{11}$$

with $x \in \Omega_m$. This includes a normalization of the distance by the size and shape of the superpixel. In large homogenous regions, where optical flow estimation is most problematic, ρ increases slowly with the distance. In small superpixels, indicating textured areas, even points close to the boundary are assigned large weights.

4.1. Minimization

Minimization of the multi-level model is very similar to the single-level model. The Euler-Lagrange equations of (9) for the levels $k > 0$ read

$$\begin{aligned}\mathcal{D}_i^k := &- \operatorname{div}\left(g^k\,\psi'\left(\sum_{i=1}^n |\nabla u_i^k|^2\right)\nabla u_i^k\right)\\ &+ \left(g_l^k\,\psi'\left(\sum_{i=1}^n |u_i^k - u_i^{k-1}|^2\right)|u_i^k - u_i^{k-1}|\right)\\ &- \left(g_l^{k+1}\,\psi'\left(\sum_{i=1}^n |u_i^{k+1} - u_i^k|^2\right)|u_i^{k+1} - u_i^k|\right) = 0\end{aligned} \tag{12}$$

for all $i = 1, \dots, n$, stating a nonlinear system with variables on multiple levels. Using this term and the Neumann boundary conditions $u_i^{-1} = u_i^0$ and $u_i^{K+1} = u_i^K$ for all i, the Euler-Lagrange equations for $k = 0$ read

$$0 = \alpha\,\rho\,c\cdot\left(u_i^0 - \tilde{u}_i\right) + (1-\alpha)\,\mathcal{D}_i^0. \tag{13}$$

Again we apply fixed point iterations together with SOR. Intermediate states of the iterative method are shown in Fig. 4.

4.2. Implementation details

As labels within a superpixel are propagated with infinite speed, we can approximate the whole superpixel by a single constant value. Treating each superpixel as a single grid point, this leads to a considerable speedup as coarse levels consist of only few superpixels. In order to keep the advantages of the continuous model, the length of the interface between two superpixels and their diffusivity must be measured in a consistent manner using a properly discretized gradient operator. Such a discretization can be found, e.g., in [6, pp. 16].

At coarser levels, we further add non-local superpixel neighbors. These enable labels to cross obstacle regions that have a significantly different color, for instance the shadow in Fig. 7. We connect all superpixels that are separated by only one further superpixel. Since there is no direct interface that would define the amount of diffusion between two superpixels, we weight non-local diffusivities by the distance between the segments using a Gaussian function

$$w_{ab} := \exp\left(-\frac{dist^2(a,b)}{2\sigma^2}\right) \tag{14}$$

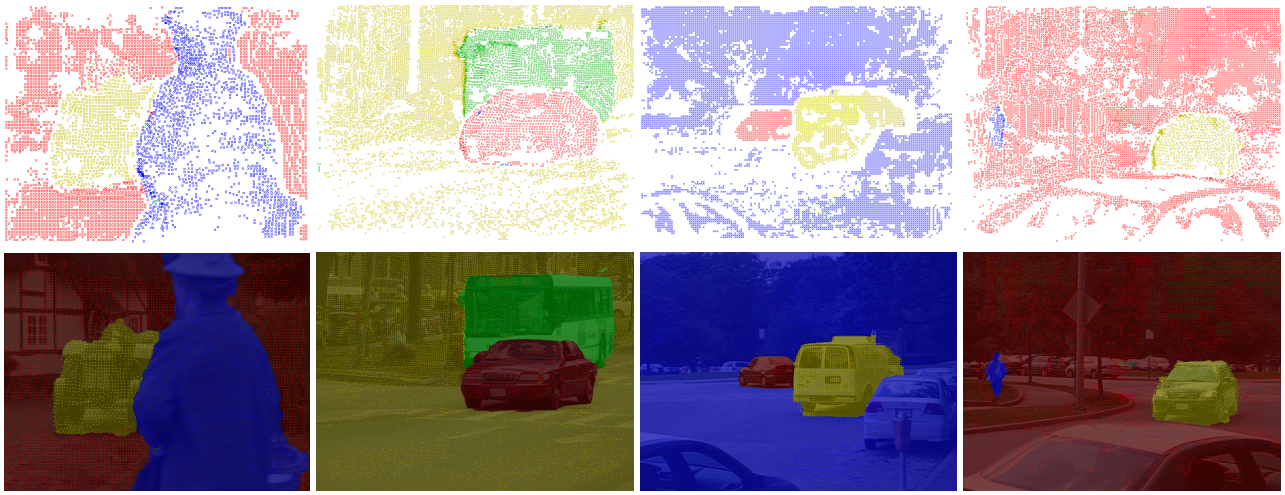

Figure 5. Some examples from the benchmark dataset in [7]. **Top row:** Sparse results from [7]. **Bottom row:** Dense results obtained with a 3-level model. Bright dots show the initial labels.

where a and b denote two superpixels and σ is a multiple of the image size.

To speed up convergence, we run a cascadic multigrid strategy on 4 grids. The coarser grids are obtained by downsampling with factor 0.5. Downsampling affects all functions at all levels including the superpixels. For the variational model at hand, this simple multigrid strategy leads to a significant speedup of factor 35. To allow others to replicate our experiments and to run the code on other sequences, we will provide executables of our implementation.

5. Experiments

For evaluation we use the benchmark introduced in [7]. It contains 204 annotated frames on 26 real world sequences. The evaluation tool coming with the benchmark allows to evaluate short term motion segmentation by running only on the first 10 or 50 frames as well as long term segmentation by running on the first 200 or all frames. Table 1 shows the density, the overall error (per pixel), and the average error (per ground truth region). We compare our variational model using different numbers of levels to the sparse result obtained from [7]. Although we raise the density to 100%, the errors do not increase compared to [7]. Even the contrary is true: the overall error drops from 6.68 to 5.33. This is quite surprising given that we assign labels in difficult locations in the image, such as occlusion areas. The numbers also show that the multi-level model always outperforms the single-level one. In most cases it is worth to use three levels. The results are slightly better than with two levels and the costs for the additional level are low. The good quantitative results are confirmed by the good visual quality of the results. Some examples are shown in Fig. 5.

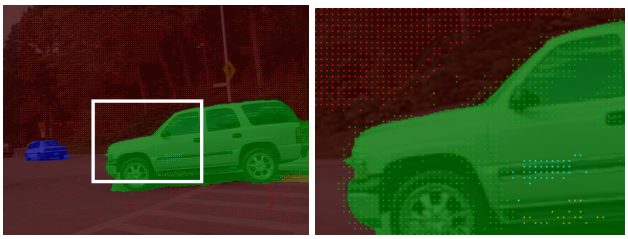

Figure 6. **Left:** Example from the dataset. **Right:** Detailed view. Brighter dots show the initial trajectory labels, background color the dense segmentation. Many erroneous initial labels are corrected by our dense segmentation.

In most cases, the regions agree with the true object regions.

Compared to the sparse trajectory labels, the classification accuracy increases due to the ability of the approach to correct erroneous labels. As shown in Fig. 6 quite large areas change their label. The amount of label correction also depends on the parameter $\alpha \in [0, 1)$ that controls the importance of the initial labels versus compact regions. One could think that the quality of the results is very sensitive to the choice of this parameter. In fact, Table 2 shows that this is not the case: as long as α is not too close to 0 or 1, the parameter can be chosen almost arbitrarily without affecting the average quality of the results.

Fig. 7 shows a qualitative comparison of the single-level and the multi-level model. In the multi-level model, information can spread more easily across larger areas. This is further supported by the non-local diffusivities we add at coarser levels, which allows to fill the court area between the legs despite the shadow and although there is not a single trajectory in this area. The information is propagated directly from other regions of the tennis court due to the

	Density	overall error	average error
First 10 frames			
3 level	100%	7.53	26.3
2 level	100%	7.40	26.7
single level	100%	7.87	26.0
superpixel voting	98.6%	7.90	25.7
sparse [7]	3.32%	7.67	25.4
First 50 frames			
3 level	100%	6.60	30.5
2 level	100%	6.62	30.6
single level	100%	6.94	30.6
superpixel voting	98.9%	7.26	30.6
sparse [7]	3.26%	6.91	32.5
First 200 frames			
3 level	100%	6.34	24.2
2 level	100%	6.14	24.2
single level	100%	6.37	24.2
superpixel voting	98.5%	6.49	24.3
sparse [7]	3.46%	6.21	31.6
All available frames			
3 level	100%	5.33	24.7
2 level	100%	5.35	25.2
single level	100%	5.37	24.5
superpixel voting	98.4%	5.72	24.3
sparse [7]	3.31%	6.68	27.7

Table 1. Evaluation on the dataset from [7] with $\alpha = 0.3$. Our variational approach raises the density to 100%. At the same time the error averaged over pixels (overall error) and over ground truth regions (average error) decreases. The hierarchical model performs better than the single-level model.

	Density	overall-error	average-error
All frames			
$\alpha = 0.9$	100%	5.33	24.7
$\alpha = 0.8$	100%	5.37	25.4
$\alpha = 0.5$	100%	5.32	24.5
$\alpha = 0.3$	100%	5.33	24.7
$\alpha = 0.1$	100%	5.55	26.3

Table 2. Robustness of the parameter α. As long as the initial labels are not completely ignored ($\alpha = 0$), the exact choice of α is not important. Numbers were obtained with the 3-level model.

same color.

As our multi-level approach uses superpixels, we also considered a naive voting approach that makes direct use of the same superpixels as in our two-level model. Each trajectory in a superpixel votes for its label and we assign the label with most votes. In case there is no trajectory in a superpixel or two labels get the same number of votes, the superpixel is not assigned any label. Tab. 1 shows that the variational approach clearly outperforms this naive voting method. Even the single-level model performs much better. This is because the finest level in the variational model ensures that we can find all object boundaries. In contrast, superpixel voting cannot recover from superpixels that erroneously cover two different objects. Moreover, the simple voting procedure ignores majorities in neighboring superpixels. Typical failure cases are shown in Fig. 8.

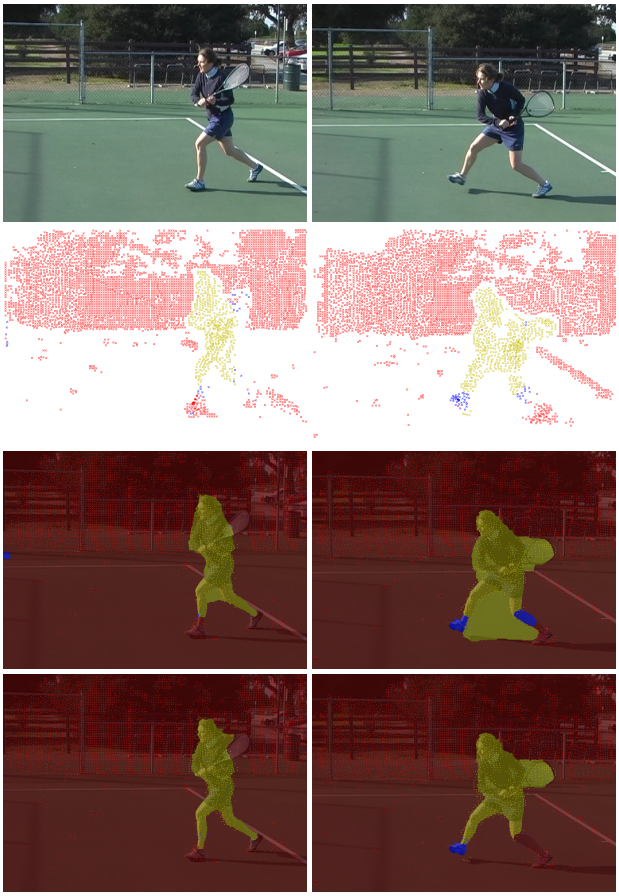

Figure 7. **Top row:** Two frames from the tennis sequence. **Second row:** Sparse labels [7]. **Third row:** Single-level model. Since there is no label information in the area between the legs and no direct connection to other parts of the tennis court, the single-level model can only interpolate the labels on the legs. **Bottom row:** Thanks to the better information flow inside superpixels and non-local diffusivities, the multi-level model can handle this hard case correctly. Remaining problems are due to incorrect motion clustering of the feet in [7] and must be approached there.

6. Conclusions

In this paper we have proposed a variational hierarchical model for generating dense segmentations from sparse sets of labels. The variational approach simultaneously op-

Figure 8. Segmentation by superpixel voting on the sample images shown in Fig. 7. Homogenous regions with few trajectories are dominated by false labels close to the object boundary.

timizes continuous functions on multiple superpixel levels and is, to the best of our knowledge, the first variational approach acting on multiple levels. We evaluated this approach on the Berkeley motion segmentation benchmark and showed that we obtain dense object segmentations with even higher accuracy than the original sparse input. We also showed that the multi-level model outperforms a single-level model as well as a voting approach based on superpixels. We believe this is another important step towards unsupervised object segmentation in video that ultimately may provide the technology enabling unsupervised learning.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:898–916, 2011. 1, 3

[2] H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*. SIAM, Philadelphia, 2006. 3

[3] X. Bai and G. Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. *ICCV*, 2007. 2

[4] Y. Boykov and G. Funka-Lea. Graph cuts and efficient n-d image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006. 2

[5] W. Brendel and S. Todorovic. Video object segmentation by tracking regions. *ICCV*, 2009. 2

[6] T. Brox. *From Pixels to Regions: Partial Differential Equations in Image Analysis*. PhD thesis, Saarland University, Germany, Apr. 2005. 5

[7] T. Brox and J. Malik. Object segmentation by long-term analysis of point trajectories. *ECCV*, 2010. 1, 2, 5, 6, 7

[8] T. Chan, S. Esedoḡlu, and M. Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal on Applied Mathematics*, 66(5):1632–1648, 2006. 3

[9] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. *ICML*, pages 272–279, New York, NY, USA, 2008. 3

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. TR2004-1963, Cornell University, Sept. 2004. 5

[11] I. Galić, J. Weickert, M. Welk, A. Bruhn, A. Belyaev, and H.-P. Seidel. Image compression with anisotropic diffusion. *Journal of Mathematical Imaging and Vision*, 31:255–269, 2008. 2

[12] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006. 2

[13] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. *CVPR*, 2010. 2

[14] J. Lellmann, F. Becker, and C. Schnörr. Convex optimization for multi-class image labeling with a novel family of total variation based regularizers. *ICCV*, 2009. 2, 3

[15] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. *CVPR*, 2011. 2

[16] C. Nieuwenhuis, B. Berkels, M. Rumpf, and D. Cremers. Interactive motion segmentation. *DAGM*, volume 6376 of *LNCS*, pages 483–492. Springer, 2010. 2

[17] Y. Ostrovsky, E. Meyers, S. Ganesh, U. Mathur, and P. Sinha. Visual parsing after recovery from blindness. *Psychological Science*, 20(12):1484–1491, 2009. 1

[18] T. Pock, A. Chambolle, D. Cremers, and H. Bischof. A convex relaxation approach for computing minimal partitions. *CVPR*, 2009. 2, 3

[19] B. L. Price, B. S. Morse, and S. Cohen. LIVEcut: learning-based interactive video segmentation by evaluation of multiple propagated cues. *ICCV*, 2009. 2

[20] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. *ICCV*, 2009. 1

[21] A. Sinop and L. Grady. A seeded image segmentation framework unifying graph cuts and random walkter which yields a new algorithm. *ICCV*, 2007. 2

[22] E. Spelke. Principles of object perception. *Cognitive Science*, 14:29–56, 1990. 1

[23] M. Unger, T. Pock, W. Trobin, D. Cremers, and H. Bischof. TVSeg - interactive total variation based image segmentation. *British Machine Vision Conference*, 2008. 2

[24] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Mulitple hypothesis video segmentation from superpixel flows. *ECCV*, LNCS. Springer, 2010. 2