

# Objektsegmentierung im Video: Ein hierarchischer, variationaler Ansatz, um Punkttrajektorien auf dichte Regionen zu erweitern

Hauptseminar Recent Advances in Computer Vision

Steffen Fuchs

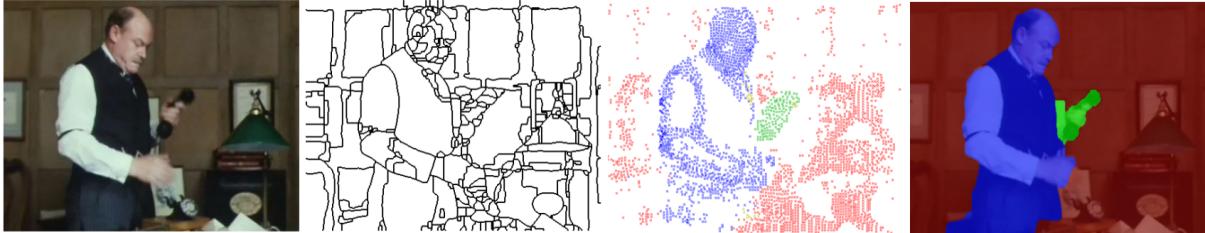


Abb. 1. von links nach rechts: original Bild, statische Segmentierung, Segmentierung auf Punkttrajektorien, hier vorgestelltes Verfahren. Klassische, statische Bildsegmentierungsverfahren wie [1] führen häufig zu Übersegmentierungen, während Objektsegmentierungen, die ausschließlich Punkttrajektorien verwenden [5], nur unvollständige Labelinformationen liefern. Die Autoren P. Ochs und T. Brox stellen ein hierarchisches, variationales Model vor, das Labels vorzugsweise in homogene Bereiche propagiert. Sie versuchen somit, die Vorteile beider Verfahren zu vereinen. Dadurch erhält man eine dichte Segmentierung, die darüber hinaus auch eine höhere Genauigkeit als die ursprünglichen Labels aufweist

**Kurzbeschreibung**—Punkttrajektorien zeichneten sich in der Vergangenheit als leistungsstarkes Hilfsmittel ab, um unbeaufsichtigt qualitativ hochwertige Segmentierungen auf Videosequenzen durchzuführen. Sie können zwar die Langzeit-Bewegungsunterschiede zu ihrem Vorteil nutzen, liefern auf Grund ihres Rechenaufwandes und der Schwierigkeiten in homogenen Regionen in der Regel aber nur spärliche Informationsdichte. Die Autoren P. Ochs und T. Brox stellen mit ihrer Arbeit eine variationale Methode vor, um aus den Gruppierungen dieser grob verteilten Trajektorien eine dichte Segmentierung zu gewinnen. Die Information wird dabei durch einen hierarchischen, nicht-linearen Diffusionsprozesses propagiert, welcher zwar im kontinuierlichen Bereich arbeitet, jedoch Superpixel mit berücksichtigt. Es wird gezeigt, dass dieser Prozess nicht nur die Informationsdichte von 3% auf 100% erhöht, sondern auch die durchschnittliche Genauigkeit der Labels verbessert.



## 1 EINLEITUNG

Lernansätze die gegenwärtig Verwendung in der visuellen Erkennung finden, hängen sehr von der manuellen Markierung und Segmentierung von Objekten ab. Betrachtet man das bis heute beste visuelle Erkennungssystem - das menschliche Gehirn - wird klar, dass solche manuellen Hilfestellungen nicht notwendig sind. Säuglinge erlernen die visuellen Formen und Eigenschaften von Objekten auch ohne dass ihnen ihre Eltern Bounding Boxes darum legen oder eine Segmentierung zur Verfügung stellen. Es gibt überzeugende Beweise dafür, dass Säuglinge diese Art von Objektsegmentierung durch den Einsatz Bewegung durchführen [20, 16]. Man könnte damit argumentieren, dass rechnergestützte, visuelle Systeme sich letztendlich immer weiter dem menschlichen Sehvermögen annähern sollten.

Die Bewegungsanalyse von Punkttrajektorien ist ein praktisches und robustes Werkzeug, um in Videosequenzen die Regionen von Objekten ohne menschliches Zutun automatisch bestimmen und extrahieren zu können. [19, 5] zeigte dies erst kürzlich. Diese Ansätze verlangen jedoch danach, dass für die Bewegungsschätzung immer auch genügend Strukturen in den Bildern vorhanden sind, um entsprechende Übereinstimmungen finden zu können. In homogenen Gebieten gibt es diese Strukturen aber nicht, was dazu führt, dass die resultierenden Punkttrajektorien nur spärlich vorhanden sind. In der Arbeit von [5] werden die Punkttrajektorien zwar aus dem dichten, optischen Flussfeld berechnet und resultierenden Trajektorien würden somit ebenfalls für das gesamte Bild zur Verfügung stehen, jedoch sind diese in homogene Regionen weniger zuverlässig und können das Clustern behindern. Zudem verlangt die nur eingeschränkt zur Verfügung stehende Rechenkraft nach der Reduzierung der Trajektorien, die analysiert

werden müssen. Das Clustern von dichten Punkttrajektorien würde viel zu lange dauern.

Die Autoren stellen in ihrem Artikel eine Methode basierend auf der Variationsrechnung vor, die aus wenigen Clustern von Punkttrajektorien eine dichte Segmentierung erstellt (Abbildung 1). Auf dem ersten Blick mag dies nach einem simplen Interpolationsproblem aussehen, da unser Verstand ganz einfach die Lücken zwischen den Punkten füllen kann. Bei genauerer Betrachtung werden jedoch einige Schwierigkeiten deutlich. So werden zum Beispiel einige der kritischen Bereiche überhaupt nicht von den Trajektorien abgedeckt, ganz besonders sind davon die Grenzen der Objekte betroffen. Den Trajektorien, die an den Grenzen jedoch vorhanden sind, wurden in den meisten Fällen falsche Labels zugewiesen, da der zugrunde liegende optische Fluss gerade bei Okklusion ungenau wird. Des Weiteren ist besonders in homogenen Bereichen auf Grund mangelnder Struktur nahezu keine Information über die entsprechenden Labels vorhanden.

Der Schlüssel, um die bestehenden Labels zuverlässig propagieren zu können, liegt darin, sich die Farb- und Kanteninformationen zunutze zu machen, was sich sehr gut mit der Trajektorienbestimmung ergänzt. Bemerkenswert ist, dass eine Segmentierung basierend auf Farbdaten gerade in homogenen Bereichen am besten arbeitet, eben dort wo die Probleme der bewegungsbasierten Segmentierung liegen. Dies wird erreicht, indem die Information in Abhängigkeit der Farbhomogenität verbreitet wird.

Am Ende der Arbeit wird ein hierarchischer variationaler Ansatz vorgestellt, mit kontinuierlicher Labelfunktion auf mehreren Ebenen. Jede Ebene entspricht einer Unterteilung in Superpixeln mit einer speziellen Grobkörnigkeit. Im Vergleich zu einem Modell mit nur einer Ebene stehen Hilfsfunktionen auf den größeren Ebenen zur Verfügung, die mittels eines verbindenden Diffusionsprozesses optimiert

werden.

Der Vorteil dieses Verfahrens liegt darin, dass das Propagieren der Labels, dank des hierarchischen Ansatzes, die Struktur auf unterschiedlichen Skalen berücksichtigt, während Metrisierungsfehler und Blockartefakte vermieden werden können. Diese treten besonders bei diskreten Markov Random Fields (MRF) Modellen auf.

## 2 VERWANDTE ARBEITEN

Die hier behandelte Problematik ist sehr verwandt mit der interaktiven Segmentierung, bei der der Benutzer selbst zunächst einfache Markierungen im Bild vornimmt und anschließend propagiert das entsprechend gewählte Verfahren diese Labels zu den restlichen Stellen. Mehrere existierende Techniken basieren auf Graph Cuts [3] oder Random Walks [11]. Die aktuellsten Techniken bauen auf konvex relaxierte Variationsrechnungen [21, 17, 13, 15] auf, die es, im Gegensatz zu den klassischen, auf einem Graph definierten MRF, vermeiden, Diskretisierungsartefakte zu erzeugen. Die hier vorgestellte variationale Methode baut auf die Regularisierung von [13] auf.

Keines der genannten Verfahren zieht einen hierarchischen Ansatz in Betracht. Vielmehr unterscheiden sich die Labels der Punkttrajektorien von denen, die durch die Benutzer eingezeichnet werden, in zwei Dingen: Zum einen sind erstere unbeaufsichtigt generiert und damit mit großer Wahrscheinlichkeit fehlerbehaftet, während bei letzteren stets von der Richtigkeit der Benutzereingabe ausgegangen wird. Dies bedeutet, dass man nicht von einem Interpolationsproblem ausgeht, sondern ein Approximationsproblem vorliegt. Zum anderen erhält man durch den Benutzer eine dichte und endliche Vorauswahl, während die Labels der Trajektorien sich nur aus einzelnen Punkten zusammen setzen und über das gesamte Bild verteilt sind.

Das hier vorgestellte Modell ist ebenfalls verwandt mit dem Bildkompressionsverfahren durch anistrophe Diffusion [10]. Dieses Verfahren erhält nur eine kleine Menge der Pixelwerte der ursprünglichen Bildpunkte und versucht die übrigen mittels Diffusionsprozess wieder herzustellen.

Des Weiteren gibt es zahlreiche, aktuelle Arbeiten zur dichten Bewegungssegmentierung, die eine Übersegmentierung durch den Einsatz von Superpixeln, Labelpropagation mittels optischen Fluss oder anderen Clusterverfahren erzielen [4, 12, 22, 14]. Diese liefern jedoch nicht die tatsächlichen Objektrektionen. Einige interaktive Videosegmentierungen versuchen dies zu vermeiden [2, 18], jedoch verlangen diese Verfahren wiederum die Eingabe eines Benutzers und laufen somit nicht unbeaufsichtigt ab.

## 3 BERECHNUNG VON PUNKTTRAJEKTORIEN

Als Grundlage der Trajektorienberechnung dient das von [6] vorgestellte Verfahren Large Displacement Optical Flow (LDOF). Dies wird verwendet, um ein optisches Flussfeld  $\mathbf{w} = (u, v)^T$  zwischen zwei aufeinanderfolgende Bildern einer Videosequenz zu berechnen.

Zunächst wird auf dem ersten Frame des Videos eine Menge von Punkten initialisiert, deren Bewegung auf den darauf folgenden Frames verfolgt werden soll (Abbildung 2). Theoretisch könnte man jedes einzelne Pixel eines Bildes versuchen zu verfolgen, jedoch würde dies zu einem ungemeinen Anstieg der erforderlichen Rechenkraft führen, zum anderen sind besonders Punkte, die in Bereichen ohne Struktur liegen nur sehr schwer zu verfolgen. Daher werden die Punkte entfernt, die keine Struktur besitzen, bzw. deren zweiter Eigenwert  $\lambda_2$  des Strukturtensors

$$J_\rho = K_\rho \sum_{k=1}^3 (\nabla I_k)(\nabla I_k)^T \quad (1)$$

einen verhältnismäßig kleinen Wert annimmt. Dabei entspricht  $K_\rho$  einem Gauß-FILTER mit Standardabweichung  $\rho = 1$  und  $\nabla I_k$  dem Gradienten im Farbkanal  $k$ .

Jeder dieser Punkte kann nun über den zuvor bestimmten optischen Fluss auf das nächste Frame

$$(x_{t+1}, y_{t+1})^T = (x_t, y_t)^T + (u_t(x_t, y_t), v_t(x_t, y_t))^T \quad (2)$$

verfolgt werden. Da der optische Fluss Subpixel-genau arbeitet, ergeben sich für  $x$  und  $y$  üblicherweise Koordinatenwerte zwischen dem Gitter. Eine bilinear Interpolation liefert wieder die entsprechenden Punkte auf dem Gitter.

Um die Richtigkeit der Trajektorien zu gewährleisten, muss zuverlässig erkannt werden, wann ein Punkt verdeckt wird und nicht mehr weiter verfolgt werden kann. Ansonsten würde ein Punkt die Bewegung von zwei unterschiedlichen Objekten beschreiben. Um einen Punkt auf Okklusion zu testen, wird dazu der Vorwärts- und Rückwärtsfluss überprüft. Diese sollten im nicht verdeckten Fall genau entgegen gerichtet sein:  $u_t(x_t, y_t) = -\hat{u}_t(x_t + u_t, y_t + v_t)$  und  $v_t(x_t, y_t) = -\hat{v}_t(x_t + u_t, y_t + v_t)$ , wobei  $\hat{\mathbf{w}}_t := (\hat{u}_t, \hat{v}_t)^T$  dem optischen Fluss von Frame  $t + 1$  nach  $t$  entspricht. Falls diese Bedingung nicht erfüllt sein sollte, dann wurde der Punkt in  $t + 1$  entweder verdeckt oder der Fluss wurde nicht richtig bestimmt. Da es jedoch immer zu kleinen Berechnungsfehlern im optischen Fluss kommen kann, wird für die Überprüfung eine kleine Toleranz zugelassen, die sich linear zur Bewegungsgeschwindigkeit verhält:

$$|\mathbf{w} + \hat{\mathbf{w}}|^2 < 0.01(|\mathbf{w}|^2 + |\hat{\mathbf{w}}|^2) + 0.5 \quad (3)$$

Des Weiteren werden Punkte an Bewegungskanten nicht weiter verfolgt, da die genaue Schätzung der Grenzen durch den optischen Fluss immer leicht variiert. Dies kann zu einem ähnlichen Effekt wie bei der Okklusion führen, bei dem ein Punkt plötzlich auf die andere Seite der Grenze fällt und somit die Bewegung von zwei unterschiedlichen Objekten beschreibt. Um dies zu verhindern, wird eine zusätzliche Bedingung für die korrekte Verfolgung eingeführt:

$$|\nabla u|^2 + |\nabla v|^2 > 0.01|\mathbf{w}|^2 + 0.002 \quad (4)$$

Letztendlich wird noch versucht, die leeren Bereiche des Bildes, die durch Okklusion entstanden sind, wieder zu füllen. Daher werden in jedem neuen Frame neue Punkte auf die gleiche Weise wie im ersten Bild initialisiert.

### 3.1 Segmentierung von Punkttrajektorien

Nachdem nun die Punkttrajektorien zur Verfügung stehen, wird das Objektsegmentierungsverfahren von [6] zum Clustern verwendet. Wie in Abbildung 2 zu sehen ist, können diese Trajektorien sich über sehr lange Zeiträume erstrecken, können zeitweise verdeckt werden oder zu jeder Zeit neue dazukommen. Würden nun nur die Trajektorien ausgewählt werden, die die gesamte Videosequenz abdecken, würde das Set am Ende sehr klein oder sogar ganz leer sein. Daher wird zum Vergleich eine paarweise Affinität zwischen allen Trajektorien definiert, die mindestens ein gemeinsames Frame besitzen. Diese Affinitäten definieren dann einen Graphen, auf dem anschließend ein Spectral Clustering Verfahren angewendet werden kann. Durch diese Transitivität können selbst Trajektorien, die niemals ein gemeinsames Frame besitzen, dem selben Cluster zugeordnet werden.

Schließlich gilt es noch zu beachten, dass es zu Situationen kommen kann, in denen man zwei Objekte mit gleicher Bewegung nicht von einander unterscheiden kann. Die tatsächliche Information liegt daher nicht in der gemeinsamen Bewegung, sondern im Bewegungsunterschied. Sobald also zum Beispiel eine Person ihre Bewegungsrichtung ändert und diese sich nicht mehr mit der anderen Person gleich, besitzt man eine genug Daten, um zu wissen, dass diese beiden Regionen im Bild nicht zusammen gehören (Abbildung 3).

Daraus kann nun eine Distanzmetrik zwischen zwei Trajektorien  $A$  und  $B$  definierte werden

$$d^2(A, B) = \max_t d_t^2(A, B), \quad (5)$$

an der Stelle, an der der Bewegungsunterschied zwischen zwei Punkten am größten ist. Die Distanz zwischen zwei Punkten zum Zeitpunkt  $t$  ist definiert als:

$$d_t^2(A, B) = d_{sp}(A, B) \frac{(u_t^A - u_t^B)^2 + (v_t^A - v_t^B)^2}{5\sigma_t^2} \quad (6)$$

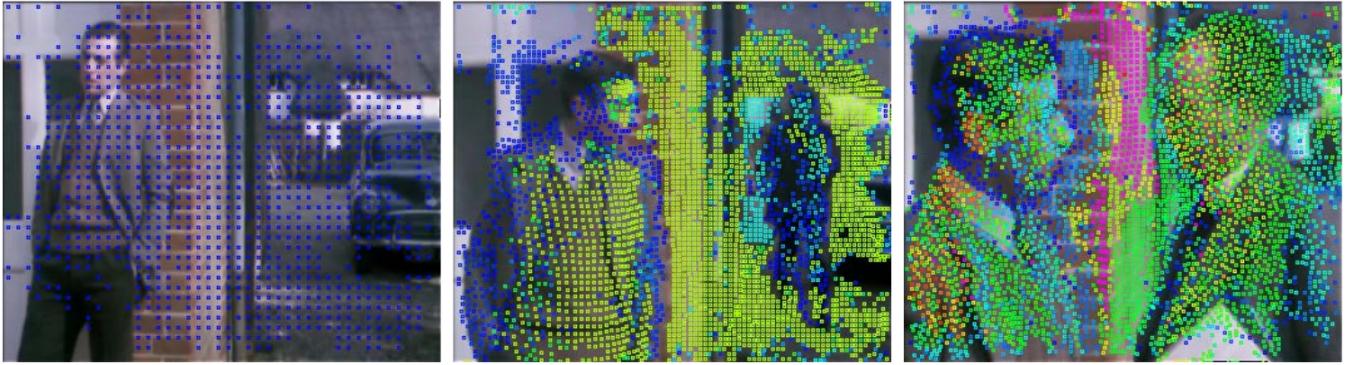


Abb. 2. Von links nach rechts: Das erste Frame ursprünglichen Punkt. Sie wurden nur dort initialisiert, wo auch genügend Struktur vorhanden ist. Frames 21 und 400 zeigen das Tracking der Punkte. Die Farbe gibt die Dauer des Trackings an (angefangen mit blau als jüngste, über grün, gelb, rot und magenta als älteste Punkte).

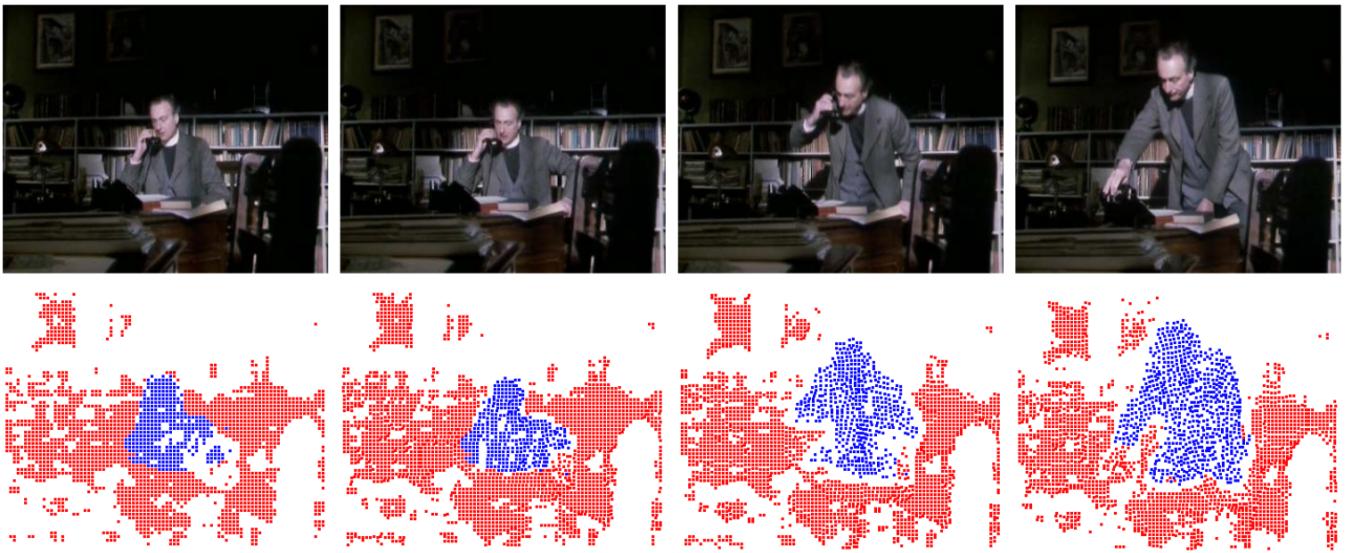


Abb. 3. Frames 0, 30, 50, 80 eines Teils aus dem Film *Miss Marple: Murder at the vicarage*. Bis zu Frame 30 gibt es kaum Bewegung, da die Person sitzt. Die meiste Information wird geliefert, sobald sie aufsteht. Mit Hilfe der Langezeit-Verfolgung, ist diese Information auch im ersten Frame verfügbar.

Dabei bezeichnet  $d_{sp}(A, B)$  die mittlere, euklidische Distanz von A und B in einem gemeinsamen Zeitfenster. Diese räumliche Gewichtung weist nahen Punkten eine größere Bedeutung zu.  $u_t := x_{t+5} - x_t$  und  $v_t := y_{t+5} - y_t$  bezeichnet die gemittelte Bewegung eines Punktes über fünf Frames, was zusätzliche Genauigkeit bietet.  $\sigma_t$  schließlich fügt eine Normalisierung der Distanz hinzu

$$\sigma_t = \min_{a \in \{A, B\}} \sum_{t'=1}^5 \sigma(x_{t+t'}^a, y_{t+t'}^a, t+t'), \quad (7)$$

wobei  $\sigma : \mathbb{R}^3 \rightarrow \mathbb{R}$  die Varianz des lokalen optischen Flusses darstellt. Diese Normalisierung soll sicherstellen, dass schnelle und langsame Bewegungen gleichermaßen behandelt werden.

Schließlich werden diese Distanzen mittels

$$w(A, B) = \exp(-\lambda d^2(A, B)), \quad \lambda = 0.1 \quad (8)$$

in Affinitäten umgewandelt, die dann eine  $n \times n$  Matrix  $W$  für die gesamte Sequenz bilden, wobei  $n$  die Anzahl der Trajektorien ist.

Auf diese Affinitätsmatrix können nun gewöhnliche Clusterverfahren angewendet werden. Die Autoren von [5] stellen jedoch auch eine angepasste Variante des Spectral Clusterings vor, dass um einen räumlichen Regularisator erweitert wurde, mit dem Ziel, Übersegmentierungen zu vermeiden.

#### 4 EINSTUFIGES, VARIATIONALES MODEL

Wie in den vorangegangenen Sektionen beschrieben stehen nun die erwähnten Punktrajektorien und Labelinformation zur Verfügung. Diese seien beschrieben als Labelfunktion  $\tilde{u} := (\tilde{u}_1, \dots, \tilde{u}_n) : \Omega \rightarrow \{0, 1\}^n, n \in \mathbb{N}$ , die  $n$  verschiedene Labels repräsentiert, wobei

$$\tilde{u}_i := \begin{cases} 1, & \text{if } x \in L_i \\ 0, & \text{else} \end{cases} \quad (9)$$

und  $L_i$  das Set von Koordinaten ist, die von einer Trajektorien des Labels  $i$  eingenommen werden, und  $\Omega \subset \mathbb{R}^2$  den Bildbereich beschreibt. Anschaulich gesprochen würde  $\tilde{u}$  eine Menge von Binärbildern beschreiben, bei dem jedes  $\tilde{u}_i$  für ein Label steht. Der Einfachheit halber wird sich im Folgenden auf die Arbeit mit Einzelbildern beschränkt. Jedoch versichern die Autoren, dass die Methoden auch ohne weiteres auf die Berechnung der ganzen Videosequenz erweitert werden kann.

Gesucht ist nun eine Funktion  $u := (u_1, \dots, u_n) : \Omega \rightarrow \{0, 1\}^n$ , die nahe an den Labels bleibt, die bereits in den Punkten in  $L := \bigcup_{i=1}^n L_i$  verfügbar sind. Dies wird erreicht, indem man versucht die Energie

$$E_{\text{data}}(u) := \frac{1}{2} \int_{\Omega} c \sum_{i=1}^n (u_i - \tilde{u}_i)^2 dx \quad (10)$$

zu minimieren. Dabei ist  $c : \Omega \rightarrow \{0, 1\}$  eine Indikatorfunktion, oder auch charakteristische Funktion mit den Werten 1 bei allen Punkten in  $L$  und 0 an den übrigen. Dadurch werden die durch den Datenterm der Energiefunktion festgelegten Bedingungen auf die Punkte beschränkt, die auch tatsächlich zu einer Trajektorie gehören. Alle anderen Punkte können jeden beliebigen Wert annehmen.

Um die übrigen Punkte dazu zu bringen, nur spezielle Werte anzunehmen, wird eine Regularisierungsfunktion

$$E_{\text{reg}}(u) := \int_{\Omega} g \psi \left( \sum_{i=1}^n |\nabla u_i|^2 \right) dx \quad (11)$$

eingeführt. Sie sorgt dafür, dass die Regionen einerseits kompakt und mit minimaler Durchmesser bleiben, aber auch, dass bevorzugt in Richtungen mit homogenen Bereichen propagiert wird. Erstes wird durch die regularisierte Norm der totalen Variation (TV Norm)  $\psi(s^2) := \sqrt{s^2 + \varepsilon^2}$  mit  $\varepsilon := 0.001$  erreicht. Anschaulich betrachtet, wird für ein minimales Auftreten von Kanten  $|\nabla u_i|$  in jedem der binären Labelbilder  $u_i$  gesorgt und gleichzeitig eine Lösung ohne Kanten verhindert. Zweiters wird durch die Diffusionsfunktion  $g : \omega \rightarrow \mathbb{R}^+$

$$g(|\nabla I(x)|^2) := \frac{1}{\sqrt{|\nabla I(x)|^2 + \varepsilon^2}} \quad (12)$$

erreicht, die dafür sorgt, dass die Labelkanten vorzugsweise dort liegen, wo auch ein großer Farbgradient  $|\nabla I(x)|$  liegt.

Die konvexe Kombination der Energien aus (10) und (11) ergibt

$$\begin{aligned} E(u) := & \frac{\alpha}{2} \int_{\Omega} c \sum_{i=1}^n (u_i - \tilde{u}_i)^2 dx \\ & + (1 - \alpha) \int_{\Omega} g \psi \left( \sum_{i=1}^n |\nabla u_i|^2 \right) dx \end{aligned} \quad (13)$$

mit  $\sum_i u_i(x) = 1$ ,  $\forall x$ , mit dem Steuerungsparameter  $\alpha \in [0, 1]$ , der in Abhängigkeit zur Glaubwürdigkeit der von den Trajektorien stammenden Labels gewählt werden kann. Für  $\alpha \rightarrow 1$  wird die Minimierungsfunktion zur Interpolation, anderenfalls liegt eine Approximation vor, die es erlaubt, fehlerhafte Labels zu korrigieren.

#### 4.1 Minimierung

Um eine Variationsrechnung mit den Binärfunktionen  $u_i$  durchzuführen zu können, muss zunächst das Problem relaxiert betrachtet werden. Dazu wird vereinbart, dass  $u_i$  jeden Wert im Intervall  $[0, 1]$  annehmen kann. Diese Art von Relaxation wurde bereits in zahlreichen ähnlichen Problemen vorgeschlagen [7, 17, 13]. Die Euler-Lagrange Gleichungen, für die relaxierte Energiefunktion, sieht dann wie folgt aus:

$$\begin{aligned} 0 = & \alpha c(u_i - \tilde{u}_i) \\ & - (1 - \alpha) \operatorname{div} \left( g \psi' \left( \sum_{i=1}^n |\nabla u_i|^2 \right) \nabla u_i \right) \quad \forall i \end{aligned} \quad (14)$$

Dieses nichtlineare System wird mittels Fixpunkt Iterationsschema gelöst, wobei der nichtlineare Faktor  $\psi'(s^2) = (s^2 + \varepsilon^2)^{-\frac{1}{2}}$  in jeder Iteration konstant gehalten wird. Das daraus resultierende lineare System wird mit Hilfe des Überrelaxationsverfahrens (SOR) gelöst. Die Bedingung  $\sum_i u_i(x) = 1$ ,  $\forall x$  wird in jedem Fixpunktiterationsschritt sicher gestellt, indem eine Normalisierung gemäß [8] durchgeführt wird. Das relaxierte Ergebnis wird schließlich nach  $\{0, 1\}^n$  projiziert, wobei

$$u_i := \begin{cases} 1, & \text{if } i = \operatorname{argmax}_i \{u_i\} = 1, \dots, n \\ 0, & \text{else} \end{cases} \quad (15)$$

gilt.

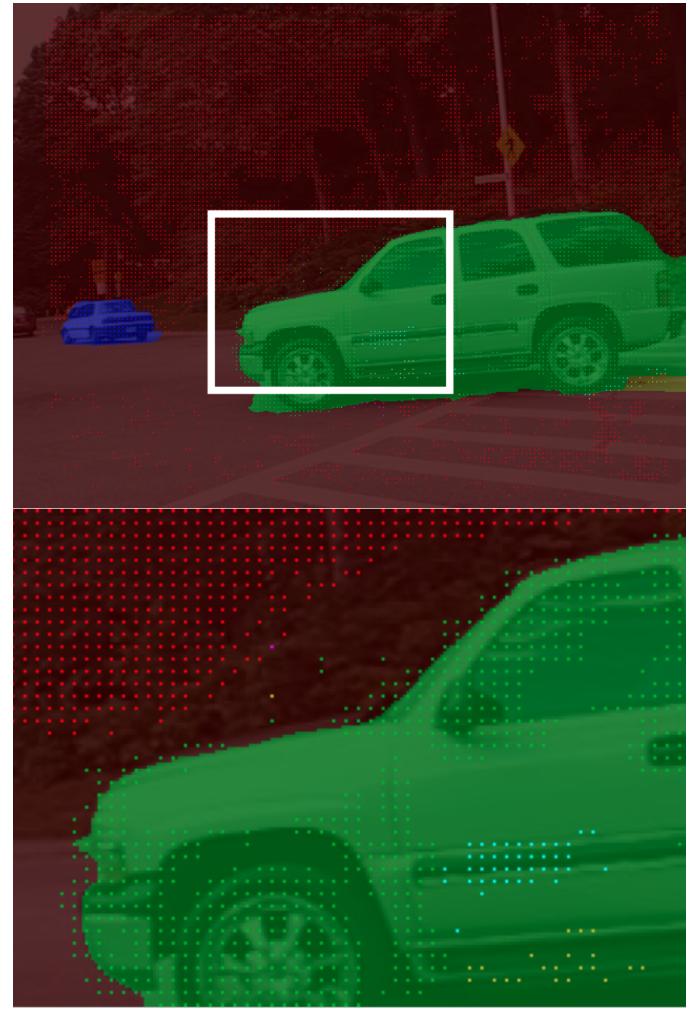


Abb. 4. Die hellen, einzelnen Punkte sind die Quellpunkte, die durch die Trajektorien erhalten wurden. Die Hintergrundfarbe stellt das endgültige Label dar. Oben: Gesamtbild. Unten: Bildausschnitt, zeigt wie selbst die ursprünglich fehlerhaften Quellpunkte korrigiert werden.

#### 5 MEHRSTUFIGES, VARIATIONALES MODEL

Die Euler-Lagrange Gleichung aus (14) kann man auch als nichtlinearen Diffusionsprozess interpretieren. Dabei dienen die Punkte, die zu einer Trajektorie gehören, als Quellpunkte, um ihre entsprechende Labelinformation zu verbreiten. Punkte in unmittelbarer Nachbarschaft wirken wie Gegenpole: Je nachdem wie groß die Labelmasse der Punkte in der Nachbarschaft ist, sowie welcher Wert für  $\alpha$  gewählt wurde, kann ein Quellpunkt  $x \in L$  sein Label in der finalen Lösung  $u$  auch ändern, obwohl er immer noch ein Quellpunkt für die dortige, ursprüngliche Labelmasse ist (Abbildung 4).

Besonders in homogenen Bereichen ist die Dichte dieser Quellpunkte gering. Dies bedeutet, dass die Information über sehr viel größere Distanzen propagiert werden muss, meistens behindert durch Rauschen und unwichtige Strukturen. Um dieses Problem zu lösen, soll nun ein mehrstufiges Modell zum Einsatz kommen. Die feinste Ebene dieser Hierarchie entspricht dem zuvor besprochenen, einstufigen Modell. Weitere Ebenen nutzen die Information von Superpixelen, die durch die Anwendung des Verfahrens aus [1] berechnet werden. Abbildung 5a zeigt eine kontinuierliche Hierarchie, während Abbildung 5b sie als entsprechende diskrete Graphstruktur zeigt.

Jedes Level  $k$ ,  $k = 0, \dots, K$  repräsentiert eine kontinuierliche Funktion, die in  $M^k$  Superpixel  $\Omega_m^k$ ,  $m = 1, \dots, M^k$  unterteilt sind. Für  $k = 0$  gibt es die Funktionen  $u^0$  und  $I^0$  wie sie schon im einstufigen Modell definiert sind. Für  $k > 0$  gibt es die entsprechenden stückweisen

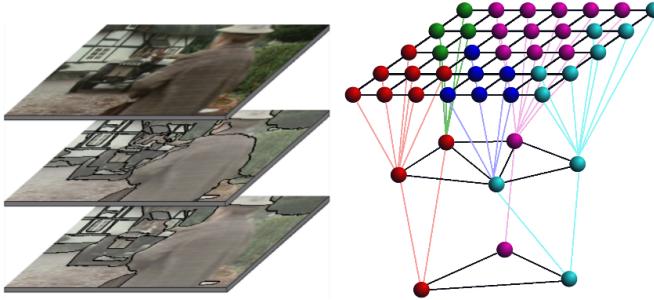


Abb. 5. Stellt die das Schema des mehrstufigen Models da. Links (a): Kontinuierliches Model. Jede Ebene ist eine kontinuierliche Funktion. Gröbere Ebenen sind stückweise konstant entsprechend ihrer Superpixel-Unterteilung. Rechts (b): Illustrierung als diskreter Graph

konstanten Funktionen  $u^k$  und  $I^k$ , wobei  $I^k(x) = \frac{1}{\Omega_m^k} \int_{\Omega_m^k} I^0(x') dx'$  den mittleren Farbwert aller Punkte des zugehörigen Superpixels  $\Omega_m^k$  annimmt. Die Idee hinter diesen Hilfsfunktionen auf größeren Ebenen ist es, einen Diffusionsprozess zu definieren, der sich besser der Bildstruktur auf unterschiedlichen Maßstäben anpasst.

Die Energiefunktion aus (13) wird nun wie folgt erweitert:

$$\begin{aligned} E(u) := & \frac{\alpha}{2} \int_{\Omega} \rho c \sum_{i=1}^n (u_i^0 - \tilde{u}_i)^2 dx \\ & + (1-\alpha) \sum_{k=0}^K \int_{\Omega} g^k \psi \left( \sum_{i=1}^n |\nabla u_i^k|^2 \right) dx \\ & + (1-\alpha) \sum_{k=1}^K \int_{\Omega} g_l^k \psi \left( \sum_{i=1}^n |u_i^k - u_i^{k-1}|^2 \right) dx, \end{aligned} \quad (16)$$

wobei  $u := (u_1^0, \dots, u_n^0, u_1^1, \dots, u_n^1, \dots, u_1^K, \dots, u_n^K)$  nun die Labelfunktion für die gesamte Hierarchie bezeichnet. Der erste Teil des Terms ist identisch mit dem des einstufigen Models, mit Außnahme der Gewichtungsfunktion  $\rho : \Omega \rightarrow \mathbb{R}$ , auf die später noch eingegangen wird. Labelquellen existieren nur auf der feinsten Ebene. Der dritte Term existiert, um eine Verbindung zwischen den Ebenen herzustellen und genau diese Information von der feinsten Ebene auf die übrigen zu propagieren. Die Ebenendiffusivitäten  $g_l^k$  haben die selbe Bedeutung wie die räumliche Diffusivität  $g^k$ , nur mit dem Unterschied, dass sie die Farbwertdistanz zwischen den Ebenen verwenden

$$g_l^k(x) := \frac{1}{\sqrt{|I^k(x) - I^{k-1}(x)|^2 + \varepsilon^2}}, \quad \varepsilon = 0.001 \quad (17)$$

im Gegensatz zum Bildgradient  $|\nabla I|$ . Der zweite Term in (8) ist die einfache Erweiterung des entsprechenden Terms aus dem einstufigen Model.

Was bringen nun die zusätzlichen Ebenen? Die Superpixel auf den größeren Ebenen führen zu Bereichen mit konstanten Werten für  $I^k$ . Folglich gilt  $\nabla I^k = 0$  innerhalb eines jeden Superpixels, was wiederum zu einer unendlichen Diffusivität  $g^k$  führt. Mit anderen Worten: Innerhalb eines Superpixels wird die Labelinformation über das gesamte Superpixel mit unendlicher Geschwindigkeit propagiert. Durch die Verbindung zwischen den Ebenen hat dieser Effekt auch auf die Punkte der nächst feineren Ebene Auswirkungen. Anstatt den gesamten räumlichen Weg über die feine Ebene, erschwert durch verrauschte Pixel und schwache Strukturen, zurück legen zu müssen, kann die Information nun eine Abkürzung über die größeren Ebenen nehmen, wo dieses Rauschen bereits entfernt wurde.

Diese Hierarchie kommt mit dem großen Vorteil einher, dass es nicht notwendig ist, ein richtiges Rauschlevel festzulegen, das unter Umständen gar nicht global existiert. Stattdessen werden mehrere

Ebenen dieser Superpixelhierarchie verwendet und in das Modell integriert. Theoretisch wäre es vorteilhaft, so viele Ebenen wie möglich zu verwenden. Im praktischen Einsatz jedoch empfiehlt es sich, aus Gründen des Berechnungsaufwandes, diese Anzahl möglichst klein zu halten. Die Autoren empfehlen, dass bereits drei Ebenen ausreichen, um bessere Ergebnisse zu erzielen.

Der letzte noch zur Erklärung ausstehende Teil von (16) ist die Gewichtungsfunktion  $\rho$ , die es erlaubt ein höheres Gewicht an spezielle Trajektorien zu verteilen. Der Anreiz dafür liegt darin, dass der verwendete Ansatz von [5] dazu tendiert, falsche Labels an Objektgrenzen auf Grund von Ungenauigkeiten des optischen Flusses zu liefern. Daher macht es Sinn, den Einfluss von Trajektorien zu erhöhen, die weiter weg von Objektgrenzen liegen, wohingegen die Labels nahe an den Rändern weniger Einfluss erhalten sollten. Da aber zu diesem Zeitpunkt die Objektgrenzen noch nicht bekannt sind, werden die Distanzwerte durch Approximation der euklidischen Distanz zu den Superpixelgrenzen  $\partial\Omega_m$  auf der grössten Ebene angenommen, welche wiederum sehr effizient berechenbar sind [9]. Basierend auf diesen Distanzen wird

$$\rho(x) := \frac{dist(x, \partial\Omega_m)}{\frac{1}{|\Omega_m|} \sum_{x \in \Omega_m} dist(x, \partial\Omega_m)} \quad (18)$$

definiert, mit  $x \in \Omega_m$ . Dies schließt auch eine Normalisierung der Distanz über die Größe und Form des Superpixels ein. In großen, homogenen Bereichen, gerade wo der optische Fluss die meisten Probleme hat, erhöht sich  $\rho$  langsam mit zunehmender Distanz. In kleinen Superpixeln, die auf texturstarke Regionen hindeuten, sind selbst Punkte nah an den Rändern stark gewichtet.

## 5.1 Minimierung

Die Minimierung des mehrstufigen Models ist der des einstufigen Models sehr ähnlich. Die Euler-Lagrange Gleichungen von (16) für die Ebenen  $k > 0$  sind

$$\begin{aligned} \mathcal{D}_i^k := & -\text{div} \left( g^k \psi' \left( \sum_{i=1}^n |\nabla u_i^k|^2 \right) \nabla u_i^k \right) \\ & + \left( g_l^k \psi' \left( \sum_{i=1}^n |u_i^k - u_i^{k-1}|^2 \right) |u_i^k - u_i^{k-1}|^2 \right) \\ & - \left( g_l^{k+1} \psi' \left( \sum_{i=1}^n |u_i^{k+1} - u_i^k|^2 \right) |u_i^{k+1} - u_i^k|^2 \right) = 0 \end{aligned} \quad (19)$$

für alle  $i = 1, \dots, n$ , die ein nichtlineares System mit Variablen auf mehreren Ebenen bilden. Unter Benutzung dieses Terms und den Neumann-Randbedingungen  $u_i^{-1} = u_i^0$  und  $u_i^{K+1} = u_i^K$  für alle  $i$ , ergibt sich daraus die Euler-Lagrange Gleichung für  $k = 0$

$$0 = \alpha \rho c (u_i^0 - \tilde{u}_i) + (1-\alpha) \mathcal{D}_i^0. \quad (20)$$

Auch hier lassen sich Fixpunktiterationen zusammen mit SOR anwenden. Abbildung 6 zeigt Zwischenschritte dieser iterativen Methode.

## 6 ZUSAMMENFASSUNG

Die Arbeit von P. Ochs und T. Brox stellt ein variationales, hierarchisches Model vor, um aus dünnen, unvollständigen Sets von Labels dichte Segmente zu generieren. Der variationale Ansatz optimiert dabei gleichzeitig kontinuierliche Funktionen auf mehreren Superpixel-Ebenen, und ist, nach Aussagen der Autoren, der erste variationale Ansatz, der auf mehreren Ebenen arbeitet. Bei verschiedenen Experimenten stellen die Autoren fest, dass ihre Segmentierung sogar noch die Genauigkeit der unvollständigen Eingabesets verbessert. Ziel der Arbeit war es, die Forschung in Richtung der unbeaufsichtigten Objektsegmentierung und somit letztendlich auch dem vollständig unbefeuerten Lernens voranzubringen.

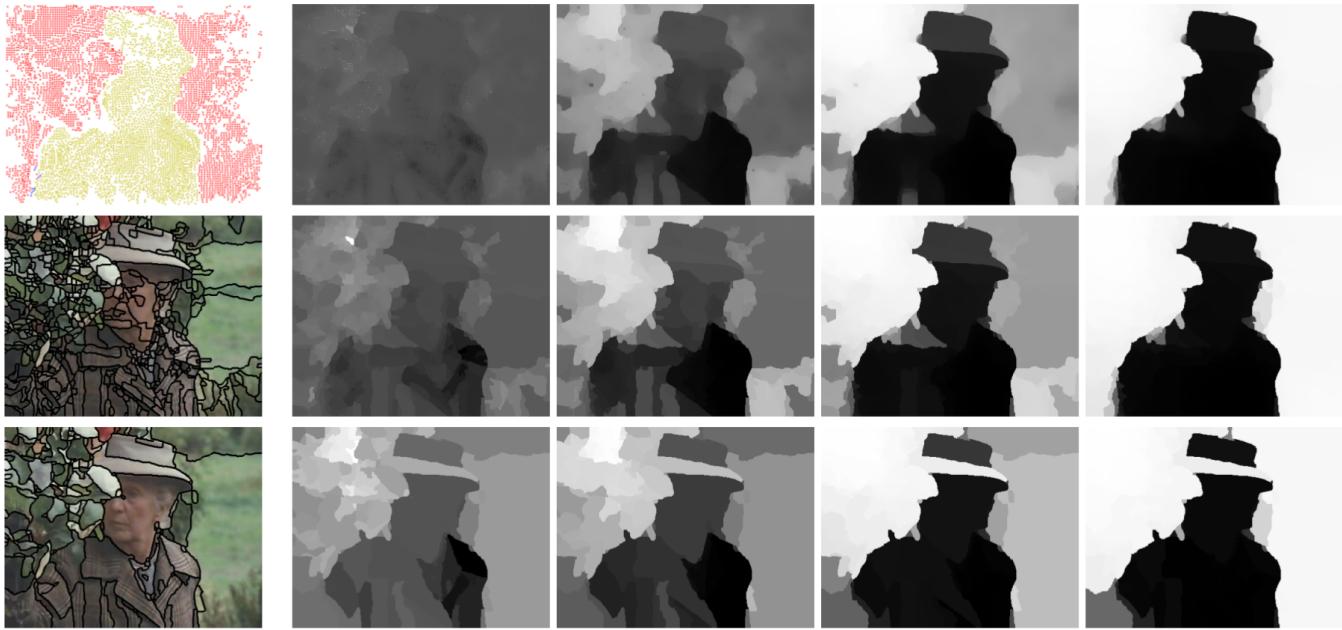


Abb. 6. Links, von oben nach unten: Quellpunkte der Trajektoriencluster, Superpixelbereiche in zwei Stufen. Rechts: Die Entwicklung der Labelfunktion  $u_1^k$  auf allen drei Ebenen gleichzeitig. Zwischenstufen nach 30, 300, 3000 und 30000 Iterationen.

## LITERATUR

- [1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898–916, 2011.
- [2] X. Bai and G. Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.
- [3] Y. Boykov and G. Funka-Lea. Graph cuts and efficient nd image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006.
- [4] W. Brendel and S. Todorovic. Video object segmentation by tracking regions. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 833–840. IEEE, 2009.
- [5] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *Computer Vision–ECCV 2010*, pages 282–295. Springer, 2010.
- [6] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(3):500–513, 2011.
- [7] T. F. Chan, S. Esedoglu, and M. Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal on Applied Mathematics*, 66(5):1632–1648, 2006.
- [8] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the  $l_1$ -ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.
- [9] P. Felzenszwalb and D. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell University, 2004.
- [10] I. Galic, J. Weickert, M. Welk, A. Bruhn, A. Belyaev, and H.-P. Seidel. Image compression with anisotropic diffusion. *Journal of Mathematical Imaging and Vision*, 31(2-3):255–269, 2008.
- [11] L. Grady. Random walks for image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1768–1783, 2006.
- [12] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2141–2148. IEEE, 2010.
- [13] J. Lellmann, F. Becker, and C. Schnorr. Convex optimization for multi-class image labeling with a novel family of total variation based regularizers. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 646–653. IEEE, 2009.
- [14] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3369–3376. IEEE, 2011.
- [15] C. Nieuwenhuis, B. Berkels, M. Rumpf, and D. Cremers. Interactive motion segmentation. In *Pattern Recognition*, pages 483–492. Springer, 2010.
- [16] Y. Ostrovsky, E. Meyers, S. Ganesh, U. Mathur, and P. Sinha. Visual parsing after recovery from blindness. *Psychological Science*, 20(12):1484–1491, 2009.
- [17] T. Pock, A. Chambolle, D. Cremers, and H. Bischof. A convex relaxation approach for computing minimal partitions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 810–817. IEEE, 2009.
- [18] B. L. Price, B. S. Morse, and S. Cohen. Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 779–786. IEEE, 2009.
- [19] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1219–1225. IEEE, 2009.
- [20] E. S. Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.
- [21] M. Unger, T. Pock, W. Trobin, D. Cremers, and H. Bischof. Tvsseg–interactive total variation based image segmentation. In *BMVC*, pages 1–10. Citeseer, 2008.
- [22] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Multiple hypothesis video segmentation from superpixel flows. In *Computer Vision–ECCV 2010*, pages 268–281. Springer, 2010.