Agrégation d'estimations semi-locales pour le flot optique

Denis FORTUN, Charles KERVRANN, Patrick BOUTHEMY

Inria, Centre Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France
Denis.Fortun@inria.fr, Patrick.Bouthemy@inria.fr, Charles.Kervrann@inria.fr

Résumé – La plupart des méthodes d'estimation du flot optique s'appuient sur un schéma variationnel global souffrant de plusieurs limitations dont la perte des grands déplacement de petites structures ou un certain lissage des discontinuités de mouvement. Nous présentons une nouvelle méthode d'estimation du flot optique basée sur un schéma d'agrégation à deux étapes, et conçue pour dépasser ces limitations. Dans un premier temps, des candidats semi-locaux sont estimés avec une combinaison de correspondances de patchs et d'estimations de mouvement affine. Dans un second temps, les candidats sont combinés dans une étape d'agrégation permettant d'élaborer un champ de mouvement dense global. Nous proposons deux méthodes d'agrégation distinctes : la première considère un schéma discret et sélectionne un candidat en chaque pixel avec une optimisation de type "graph-cut"; la seconde combine optimisation continue parcimonieuse et mesures de confiance, et tolère des déviations par rapport aux candidats. Les deux approches n'utilisent pas de schéma multi-résolution. Les améliorations obtenues pour les configurations complexes de mouvement mentionnées ci-dessus sont illustrées sur plusieurs exemples.

Abstract – Most existing methods for optical flow estimation rely on a global variational framework suffering from several limitations like loss of small structures with large displacements or over-smoothing of motion discontinuities. We present a novel optical flow estimation method based on a two-stage aggregation framework and designed to handle these issues. First, semi-local candidates are estimated with a combination of patch correspondences and affine motion estimations. Then, the candidates are combined in an aggregation stage creating a dense global flow field. We propose two distinct aggregation methods : the first one operates in a discrete framework and selects one candidate at each pixel with a graph-cut based optimization; the second one combines sparse continuous optimization and confidence measure and tolerates deviations from the candidates. Both approaches are free of coarse-to-fine schemes. Improvements on the aforementioned complex motion configurations are demonstrated on challenging examples.

1 Introduction

Le flot optique est une information essentielle pour un grand nombre de problèmes en vision par ordinateur comme la segmentation d'objets en mouvement, le suivi d'objets, la détection d'obstacles ou la reconnaissance d'actions. L'estimation du flot optique s'appuie sur une hypothèse de conservation d'une mesure (typiquement la conservation de l'intensité de l'image) qui ne permet pas à elle seule d'accéder aux deux composantes des vecteurs de vitesse en chaque point (*problème de l'ouverture*). Ce problème est généralement surmonté par l'introduction d'une contrainte de cohérence spatiale du champ des vitesses. On peut distinguer les méthodes *locales* et *globales* selon leur façon d'imposer cette cohérence.

L'approche *locale* initiée par [1] impose un modèle paramétrique de mouvement dans un voisinage de chaque pixel et résoud le système d'équations fournies par l'hypothèse de conservation. L'étape cruciale est alors le choix d'un voisinage approprié assurant la validité de l'approximation paramétrique. Le choix usuel d'une fenêtre de taille fixe centrée en chaque pixel [1] entraîne des erreurs dans les zones faiblement texturées et au niveau des discontinuités de mouvement. L'approche *globale* impose la cohérence spatiale au travers d'un terme additionnel de régularisation favorisant un flot lisse par morceaux [2]. L'optimisation variationnelle associée à cette approche mène aux meilleurs résultats de l'état de l'art [3], mais implique certaines limitations liées à l'utilisation de schémas multi-résolution à l'origine de la perte des grands déplacement de petites structures, et à un certain lissage des discontinuités.

Plusieurs variations de ces approches de base ont tenté d'en dépasser les limitations. Dans le cadre local, des méthodes d'adaptation de la taille du voisinage ou de sa position [4] permettent d'améliorer les résultats, mais restent loin des performances des méthodes de l'état de l'art. Dans le cadre global, le problème de la préservation des discontinuités [5] a été principalement abordé par l'élaboration de termes de régularisation sophistiqués robustes [6], la complémentarité au terme de données [7], des relations de voisinage non locales [8]. Pour s'affranchir des problèmes liés à l'utilisation de schéma multi-résolution, des stratégies ont été proposées pour intégrer la mise en correspondance de descripteurs dans des schémas combinant optimisation discrète et variationnelle [9].

Nous proposons une méthode reposant sur deux étapes successives : estimation semi-locale de vecteurs de mouvement candidats en chaque point, et agrégation de ces candidats pour créer un flot optique global. Les estimations semi-locales articulent correspondances de blocs et estimations paramétriques pour correctement appréhender les grands déplacements et atteindre une précision sous-pixellique. La diversité de tailles et de positions des fenêtres de calcul permet d'éviter le problème du choix *a priori* du voisinage pour l'estimation locale du flot

optique. Deux stratégies d'agrégation sont ensuite introduites, l'une similaire à la fusion discrète de candidats décrite dans [10], l'autre proposant une approche continue et parcimonieuse au problème de l'agrégation. Les deux options ne nécessitent pas de schéma multi-résolution.

2 Estimation du flot optique en deux étapes

Dans cette section, nous détaillons les deux étapes successives de notre méthode: estimation semi-locale de candidats et agrégation pour créer un champ de mouvement global. On considère deux images successives $I_1, I_2 : \Omega \to \mathbb{R}$ d'une séquence temporelle, et Ω désigne le domaine de l'image.

2.1 Estimation semi-locale de candidats

Ensemble de patchs semi-locaux de I_1 L'estimation des candidats repose sur la décomposition de I_1 en patchs décrite dans [11]. On note $\mathcal{P}_{s,\alpha}$ l'ensemble de patchs carrés de I_1 de taille s couvrant toute l'image avec un taux de recouvrement $\alpha \in [0, 1]$ (deux patchs voisins partagent une proportion α de leur aire). On étend cette notation à un ensemble de tailles $S = \{s_1, \ldots, s_n\}$ en définissant $\mathcal{P}_{S,\alpha} = \bigcup_{s\in S} \mathcal{P}_{s,\alpha}$. Le flot optique est estimé indépendamment dans chaque patch en deux sous-étapes décrites ci-dessous: correspondances de patchs et estimation affine.

Correspondances de patchs A la différence de [11], on n'applique pas la même décomposition en patchs à I_2, et on ne suit pas une approche variationelle. Pour chaque patch $P_1 \in \mathcal{P}_{S,\alpha}$, on détermine dans un premier temps l'ensemble $\mathcal{M}_N(P_1)$ des N patchs de I_2 les plus similaires à P_1 au sens du coefficient de corrélation normalisé (NCC). Pour chaque paire de patchs mis en correspondance $P_{1,2} = (P_1, P_2)$, avec $P_2 \in \mathcal{M}_N(P_1)$, on obtient un vecteur de translation $w_{P_{1,2}} \in \mathbb{Z}^2$ joignant les centres de P_1 et P_2.

Estimations affines Les correspondances estimées par corrélation sont des estimations pixelliques et translationnelles entre P_1 et P_2. Pour atteindre une précision sous-pixellique et autoriser des déformations plus complexes, on raffine cette première estimation grossière par l'estimation d'un champ local de mouvement affine $\delta w_{P_{1,2}} : \Omega_{P_1} \to \mathbb{R}^2$ pour chaque paire $P_{1,2}$, avec Ω_{P_1} le domaine de P_1.

Le champ $\delta w_{P_{1,2}}$ est paramétrisé par le vecteur $\theta_{P_{1,2}} = (a_1, a_2, a_3, a_4, a_5, a_6)^T$ sous la forme $\delta w_{P_{1,2}}(x) = (a_1 + a_2x_1 + a_3x_2, a_4 + a_5x_1 + a_6x_2)^T$ en chaque pixel $x = (x_1, x_2)^T$. Les paramètres $\theta_{P_{1,2}}$ sont estimés sur la base de l'hypothèse de conservation de l'intensité traduite sous la forme du potentiel

$$\rho_{data}(x, w, I_1, I_2) = \psi(I_2(x + w(x)) - I_1(x)) \qquad (1)$$

où ψ désigne la fonction robuste de Tukey. Le paramètre estimé $\hat{\theta}_{P_{1,2}}$ minimise l'énergie $E_{aff} = \int_{\Omega_{P_1}} \rho_{data}(x, w_{P_{1,2}} + \delta w_{P_{1,2}}(x), P_1, P_2)dx$. L'estimation est menée grâce au logiciel Motion2D, disponible publiquement [1] qui adopte un schéma multi-résolution et utilise la méthode IRLS (Iterative Reweighted Least Squares) pour minimiser les linéarisations successives de E_{aff} [12].

Ensemble final de candidats L'estimation semi-locale décrite ci-dessus est répétée pour chaque paire de patchs. Le recouvrement et la diversité des tailles de patch impliquent qu'un pixel $x \in \Omega$ appartient à plusieurs patchs. Un ensemble de candidats de mouvement $\mathbf{W_c}(x)$ est donc généré en chaque pixel $x \in \Omega$ et est défini par (on note $\mathcal{P}_{\mathcal{S},\alpha}(x) = \{P \in \mathcal{P}_{\mathcal{S},\alpha} : x \in P\}$):

$$\mathbf{W_c}(x) = \left\{ w_{P_{1,2}} + \delta w_{P_{1,2}}(x) : P_1 \in \mathcal{P}_{\mathcal{S},\alpha}(x), P_2 \in \mathcal{M}_N(P_1) \right\} \qquad (2)$$

L'intérêt de cet ensemble de candidats est double. D'une part, les correspondances NCC permettent de détecter les grands déplacements sans schéma multirésolution, et donc sans perdre les structures de petite taille. D'autre part, en considérant une grande variété de patchs, on évite le choix prédéfini du voisinage pour l'estimation paramétrique. La sélection du patch approprié à chaque pixel reviendra en fait à celle du vecteur candidat correspondant lors de l'étape d'agrégation.

2.2 Agrégation globale

Nous présentons deux stratégies d'agrégation visant à construire un champ de mouvement global $w_\Omega : \Omega \to \mathbb{R}^2$ à partir des candidats $\mathbf{W_c}(x)$ indépendamment estimés lors de la première étape.

Agrégation discrète L'agrégation discrète consiste à considérer l'ensemble $\mathbf{W_c}(x)$ de candidats comme une discrétisation de l'espace des vecteurs de mouvement en chaque pixel $x \in \Omega$. L'agrégation devient alors un problème d'optimisation discrète dans cet espace, défini par $w_\Omega = \arg\min_w E^d_\Omega(w, I)$ s.t. $\{w(x) \in \mathbf{W_c}(x), x \in \Omega\}$. On définit E^d_Ω par la combinaison classique du terme de données (1) modélisant l'hypothèse de conservation de l'intensité et d'un terme de régularisation modélisant la dépendance spatiale entre vecteurs de mouvement voisins et favorisant un flot lisse par morceaux. On obtient

$$\begin{aligned} E^d_\Omega(w, I) &= \sum_{x\in\Omega} \rho_{data}(x, w, I_1, I_2) \\ &+ \lambda \sum_{<x,y>} \psi(\|w(x) - w(y)\|) \end{aligned} \qquad (3)$$

où ψ désigne la fonction robuste de Tukey, $\rho_{data}(\cdot)$ est défini dans (1), $< x, y >$ est un système de cliques d'ordre 2, et λ est un coefficient de régularisation permettant d'équilibrer les deux termes.

Contrairement aux approches variationnelles [13, 14] le schéma d'optimisation discrète permet d'éviter la linéarisation

1. http://www.irisa.fr/vista/Motion2D/

du terme de données et donc l'utilisation de schémas multi-résolutions. La minimisation est réalisée avec l'algorithme fusion-move introduit dans [10], qui décompose le problème posé en une succession de problèmes d'optimisation binaire facilement résolus par des méthodes standard de "graph cut".

Agrégation continue La limitation potentielle de l'agrégation discrète est la contrainte pour les vecteurs du champ de mouvement global de prendre leurs valeurs parmi les candidats pré-définis, ce qui impose de garantir une bonne précision des candidats. Cette garantie ne peut être réalisée qu'au prix d'un temps de calcul important. Nous proposons une nouvelle approche d'agrégation conçue pour tolérer une certaine imprécision des candidats et limiter ainsi le coût de leur estimation. Cette approche est basée sur l'optimisation globale continue de l'énergie

$$\begin{aligned} E^c_\Omega(w,\boldsymbol{\alpha}) &= \int_{x\in\Omega} \left\| w(x) - \boldsymbol{\alpha}(x)^\top \mathbf{W_c}(x) \right\|_1 \\ &+ \lambda_2 \left\|\boldsymbol{\alpha}(x)\right\|_{1,\boldsymbol{\beta}(x)} + \lambda_1 \left\|\nabla w\right\|_1 dx \end{aligned} \quad (4)$$

où $\boldsymbol{\alpha}(x) = (\alpha_1(x), .., \alpha_{n(x)}(x))^\top$ est un vecteur de coefficients parcimonieux associé aux $n(x)$ candidats $\mathbf{W_c}(x) = (w_1(x), .., w_{n(x)}(x))^\top$, $\boldsymbol{\beta}(x) = \{\beta_1(x), .., \beta_{n(x)}(x)\}$ est un vecteur associant une mesure de confiance à chaque candidat, et la norme L_1 pondérée est définie par $\|\boldsymbol{\alpha}(x)\|_{1,\boldsymbol{\beta}(x)} = \sum_{i=1}^{n(x)} \beta_i(x) \left|\alpha_i(x)\right|$. Le premier terme impose une reconstruction par modèle linéaire en considérant les candidats comme un dictionnaire de mouvement. Le second terme représente la contrainte de parcimonie sur les coefficients $\boldsymbol{\alpha}(x)$. Le vecteur $\mathbf{W_c}(x)$ de candidats pouvant contenir un nombre éventuellement important de valeurs aberrantes, cet terme est important pour ne conserver l'influence que de quelques candidats pertinents. Une contrainte imposant uniquement la parcimonie des coefficients $\boldsymbol{\alpha}(x)$ serait cependant insuffisante. En effet l'optimisation de (4) par rapport à $\boldsymbol{\alpha}(x)$ viserait alors simplement à reconstruire w, et l'optimisation par rapport à w ne pourrait ensuite pas s'écarter assez de sa valeur initiale. On introduit donc une contrainte supplémentaire sur $\boldsymbol{\alpha}(x)$, liée aux données des images, en pondérant la contrainte de parcimonie par des mesures de confiance définies par :

$$\beta_i(x) = \exp\left\{ -\frac{\sum_{y\in\Omega_d} g(x,y,I_1)\rho_{data}(y, w_i(x), I_1, I_2)}{\sigma} \right\} \quad (5)$$

où Ω_d désigne le domaine discret de l'image, ρ_{data} est défini par (1) et $g(x,y,I_1) = e^{-\left(\frac{\|x-y\|_2^2}{\sigma_s} + \frac{|I_1(x)-I_1(y)|}{\sigma_g}\right)}$ est un poids similaire à un filtre bilatéral pénalisant la distance spatiale et l'écart d'intensité entre les pixels x et y (les paramètres sont fixés à $\sigma = 0.1$, $\sigma_s = 5$ et $\sigma_c = 20$). Le terme $\sum_{y\in\Omega_d} g(x,y,I_t)\rho_{data}(y, w_i(x), I_1, I_2)$ évalue la validité locale du vecteur candidat $w_i(x)$ par un filtrage bilatéral de la contrainte de conservation associée $\rho_{data}(y, w_i(x), I_1, I_2)$.

L'énergie (5) est minimisée alternativement par rapport à w et $\boldsymbol{\alpha}$. La minimisation par rapport à w avec $\boldsymbol{\alpha}$ fixe, est réalisée en résolvant les équations d'Euler-Lagrange par des itérations à point fixe [6]. Pour minimiser (5) par rapport à $\boldsymbol{\alpha}$, avec w fixe, nous avons eu recours à un algorithme glouton défini comme suit: à partir d'une configuration initiale de $\boldsymbol{\alpha}$, on suit une stratégie d'exploration des configurations possibles de $\boldsymbol{\alpha}$, et une configuration est conservée si elle entraîne une diminution de l'énergie. La stratégie d'exploration des configurations se fait en rajoutant itérativement une composante non nulle dans l'ordre décroissant des mesures de confiance.

TAB. 1: Erreurs angulaires moyennes obtenues avec SL-D, SL-C, [14] and [13] sur les séquences de la base Middlebury.

	Grove2	Grove3	Hydrangea	Urban2	Urban3
SL-D	2.19	5.43	2.47	2.47	3.42
SL-C	2.43	5.92	2.29	2.53	4.12
[14]	2.38	5.97	2.10	2.50	3.91
[13]	2.92	6.72	2.29	2.63	6.10

3 Résultats

La méthode est évaluée en termes d'erreur globale sur la base de séquences Middlebury [3]. Les améliorations locales de préservation des discontinuités et la détection des grands déplacements sont illustrées visuellement. Les paramètres sont fixés à $\mathcal{S} = \{15, 45, 115\}$, $\alpha = 0.8$ et $N = 2$.

Résultats quantitatifs globaux Le tableau 1 contient les erreurs angulaires moyennes (AAE) obtenues avec notre méthode dans les cas d'agrégation discrète (SL-D) et continue (SL-C), et avec les méthodes disponibles publiquement décrites dans [13] et [14]. Les deux versions SL-C et SL-D fournissent des erreurs plus faibles que celles de [13] pour presque toutes les séquences. Pour les séquences contenant de fortes discontinuités de mouvement ou de petits détails comme *grove2*, *grove3*, *urban2*, *urban3*, SL-C offre de meilleurs résultats que l'approche variationnelle [14].

Préservation des discontinuités La figure 1 illustre le potentiel de notre méthode pour préserver les discontinuités et les détails de mouvement. Ceci est rendu possible par la diversité des tailles de patchs introduites, détectant des régions de mouvement cohérent à plusieurs échelles. Les champs de mouvement estimés par [13] et [14] ont tendance à lisser les discontinuités et ignorer les détails du flot. En revanche, SL-D et SL-C produisent des discontinuités plus nettes et préservent des zones de détails.

Grands déplacements Dans la séquence *Backyard* (Fig. 2), le grand déplacement de la balle est perdu par le schéma multirésolution de l'approche variationnelle [13]. Avec notre approche, SL-D ou SL-C, le mouvement de la balle est recouvré de façon satisfaisante. Par rapport au résultat de la méthode décrite dans [14], conçue pour gérer les grands déplacements par correspondances de descripteurs dans un schéma variationnel, la forme de la balle est mieux préservée

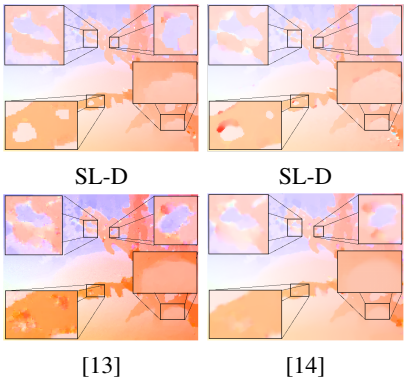

FIG. 1: Illustration visuelle sur la séquence *Grove3* des différences d'estimation au niveau de discontinuités et de détails du flot (zones agrandies dans l'image) entre SL-D, SL-C, [13] et [14].

et l'estimation est moins influencée par la région occultée.

Comparaison des agrégations discrète et continue Au niveau des discontinuités de mouvement, SL-D présente en général des transitions plus marquées que SL-C, mais génère un léger effet de blocs dû à la minimisation par graph cut (Fig. 3-(a)). Au niveau des variations lisses du flot, la restriction des vecteurs de mouvement à l'ensemble des candidats pré-estimés imposée par SL-D est insuffisante pour retrouver les déformations les plus complexes, alors que SL-C s'écarte des candidats pour assurer un champ de vitesses lisse (Fig. 3-(c)-(d)).

4 Conclusion

Cet article a présenté une nouvelle approche pour l'estimation du flot optique procédant en deux étapes : 1) estimations de vecteurs de mouvement candidats en chaque pixel à partir d'une collection de patchs; 2) agrégation des vecteurs candidats pour obtenir le champ de vitesse global. La combinaison de correspondances de patchs et d'estimation paramétrique nous permet de gérer les grands déplacements sans schéma multi-résolution. Les vecteurs de mouvement sont estimés sur le voisinage le plus approprié sans segmentation explicite du mouvement mais via l'étape d'agrégation sélectionnant les meilleurs candidats. Une méthode d'agrégation alternative à celle de [11], moins dépendante de la qualité des candidats, est proposée dans un cadre continu et parcimonieux.

Références

[1] B.D. Lucas et T. Kanade. An iterative image registration technique with an application to stereo vision. *Proc. 7th Int. J. Conf. Art. Intel.*, 1981.

[2] B.K.P. Horn et B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[3] S. Baker, D. Scharstein, JP Lewis, S. Roth, M.J. Black, et R. Szeliski. A database et evaluation methodology for optical flow. *Int. J. Comp. Vis. (IJCV)*, 92(1):1–31, 2011.

[4] P.M. Jodoin et M. Mignotte. Optical-flow based on an edge-avoidance procedure. *J. Comp. Vis. Im. Und. (CVIU)*, 113(4):511–531, 2009.

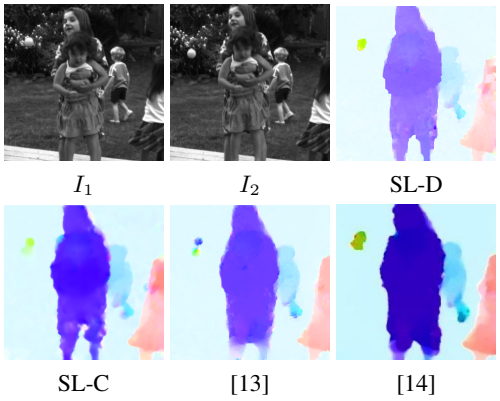

FIG. 2: Résultats obtenus sur la séquence *Backyard* illustrant le comportement de SL-D, SL-C, [13] et [14] dans le cas de grands déplacement de petits objets.

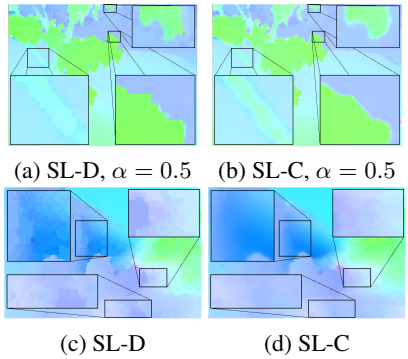

FIG. 3: Comparaison des agrégations discrète (SL-C) et continue (SL-D). (a)(b) Estimation avec un faible taux de recouvrement $\alpha = 0.5$ (séquence *Grove2*). (c)(d) Estimation d'un champ de mouvement à variation lisse (séquence *Dimetrodon*).

[5] F. Heitz et P. Bouthemy Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Trans. Patt. Anal. Mach. Intel. (PAMI)*, 15(12):1217–1232, 1993.

[6] T. Brox, A. Bruhn, N. Papenberg et J. Weickert High accuracy optical flow estimation based on a theory for warping. *Proc. Eur. Conf. Comp. Vis (ECCV)*, 25–36, 2004.

[7] H. Zimmer, A. Bruhn, J. Weickert. Optic flow in harmony. *Int. J. Comp. Vis. (IJCV)*, 93(3):368–388, 2011.

[8] P. Krähenbühl, V. Koltun Efficient Nonlocal Regularization for Optical Flow. *Proc. Eur. Conf. Comp. Vis (ECCV)*, 2464–2471, 2012.

[9] L. Xu, J. Jia et Y. Matsushita. Motion detail preserving optical flow estimation. *IEEE Trans. Patt. Anal. Mach. Intel. (PAMI)*, 34(9):1744–1757, 2012.

[10] V. Lempitsky, S. Roth, et C. Rother. Fusionflow: Discrete-continuous optimization for optical flow estimation. In *IEEE Proc. Conf. Vis. Patt. Rec. (CVPR)*, 1–8, Anchorage, 2008.

[11] D. Fortun et C. Kervrann. Semi-local variational optical flow. *Proc. Int. Conf. Im. Proc. (ICIP)*, 77–80, 2012.

[12] J. M. Odobez et P. Bouthemy Robust multiresolution estimation of parametric motion models. . *J. Vis. Comm. Im. Repr.*, 6(4):348–365, 1995.

[13] A. Chambolle et T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imag. et Vis. (JMIV)*, 40(1):120–45, 2011.

[14] T. Brox et J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Trans. Patt. Anal. Mach. Intel. (PAMI)*, 33(3):500–513, 2011.