Local/Global Scene Flow using Intensity and Depth Data

Julian Quiroga Frederic Devernay James Crowley

PRIMA team, **INRIA Grenoble**

julian.quiroga@inria.fr

July 8, 2013

Motivation

The **scene flow** is the 3D motion field of the scene (*Vedula* ICCV'99).

Surface Flow, Morpheo-INRIA 2011

Applications

- Action recognition
- Interaction
- 3D reconstruction
- Navigation

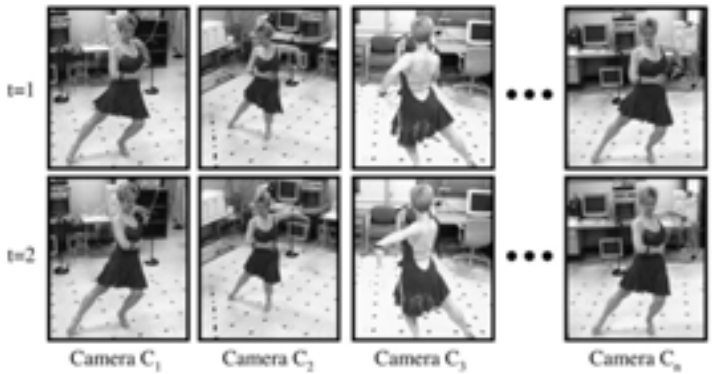
Using depth and/or color

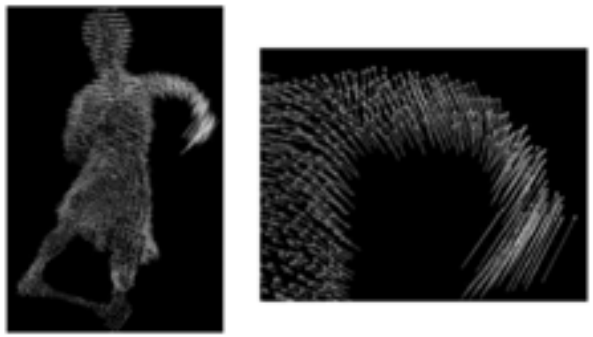
RGB-D SLAM Dataset TUM

Scene flow computation

Stereo or multiview:

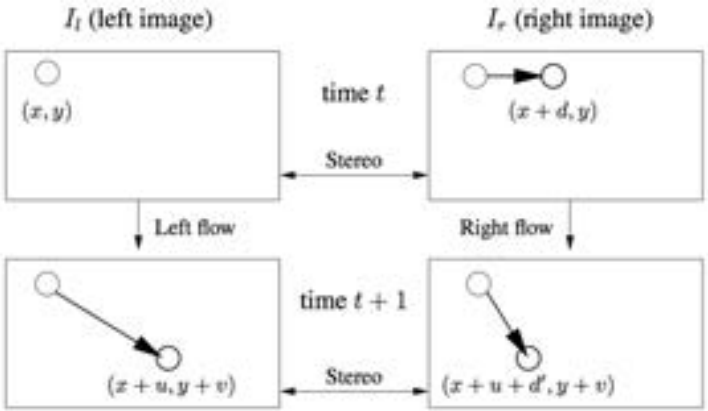
- From several optical flows (*Vedula et al.* PAMI'05)





Scene flow

- Using structure constraints (*Huguet & Devernay* ICCV'07, *Wedel et al.* ECCV'08, *Basha et al.* CVPR'10)



2 views and optical flow

Scene flow computation

Color and depth:

- Optical flow and range flow under orthography (*Spies et al.* CVIU'02, *Lukins et al.* BMVC'04)

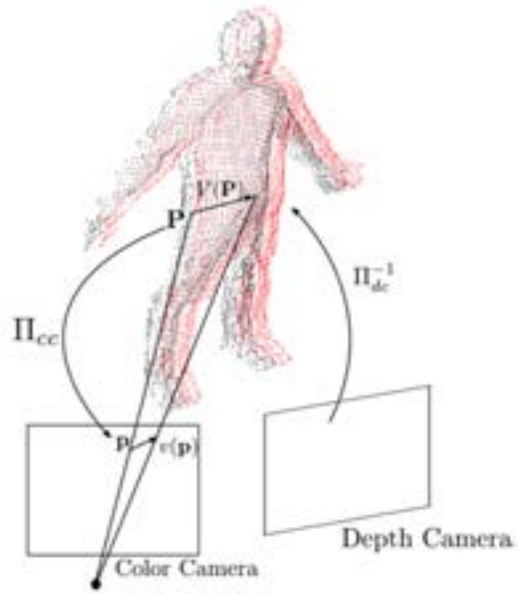
$$0 = I_X U + I_Y U + I_t$$

Optical flow equation

$$W = Z_X U + Z_Y U + Z_t$$

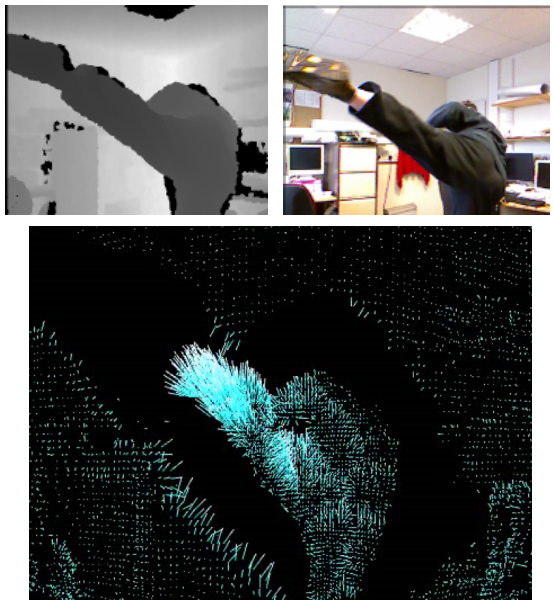
Range flow equation

- Photometric constraints (*Letouzey* BMVC'11)



Projective camera model

- Particle filtering (*Hadfield&Bowden* ICCV'11)



3D motion field

Our work

Assumptions

- Fixed camera
- Brightness and depth consistency
- Scene composed by locally-rigid moving parts

Approach

Local motion: 2D tracking of 3D surface patches in a LK framework.
Global motion: an adaptive 2D TV-regularization of the 3D motion field.
Large/small motions: multi-scale and a set of 3D correspondences.

Energy

$$E(\mathbf{v}) = E_D(\mathbf{v}) + \alpha E_M(\mathbf{v}) + \beta E_R(\mathbf{v}),$$

where $\mathbf{v} = \{v_X, v_Y, v_Z\}$.

Presentation outline

- **Motion model**
- Data term
- Regularisation term
- Sparse matching term
- Optimisation
- Experimentation
- Conclusion

Motion model

Let $\mathbf{X} = (X, Y, Z)$ be a 3D point in the camera frame. The **image flow** (u, v) induced by the 3D motion $\mathbf{v} = \{v_X, v_Y, v_Z\}$ is given by:

$$u = x' - x = \left(\frac{X + v_X}{Z + v_Z} - \frac{X}{Z}\right) = \frac{1}{Z}\left(\frac{v_X - x v_Z}{1 + v_Z/Z}\right)$$

and

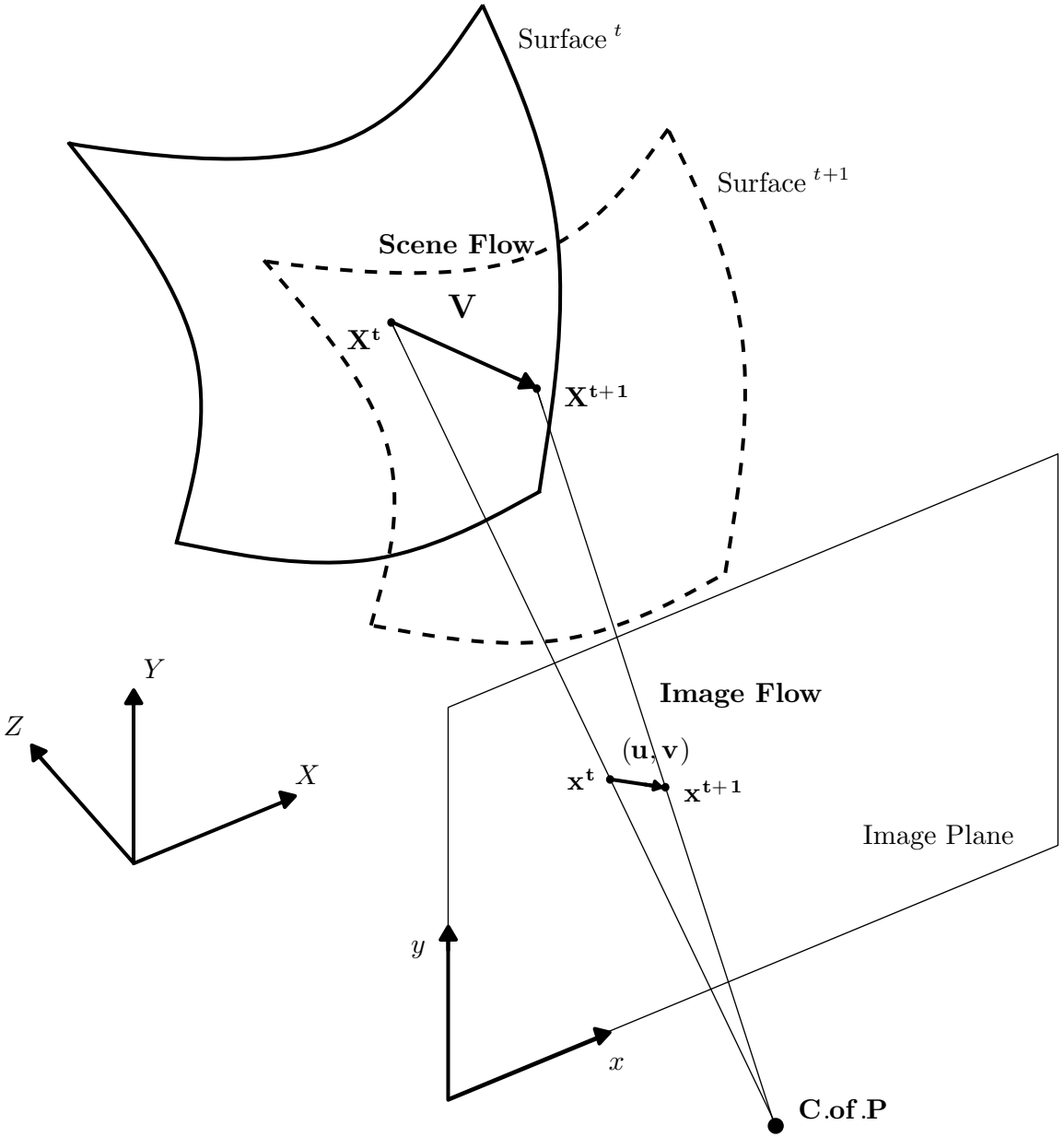
$$v = y' - y = \left(\frac{Y + v_Y}{Z + v_Z} - \frac{Y}{Z}\right) = \frac{1}{Z}\left(\frac{v_Y - y v_Z}{1 + v_Z/Z}\right).$$

where $(x, y) = \hat{\mathbf{M}}(\mathbf{X})$ and the new 3D points is $\mathbf{X}' = \mathbf{X} + \mathbf{v}$.

Using a Taylor series in the denominator term containing v_Z, we get

$$\left(\frac{1}{1 + v_Z/Z}\right) = \left(1 - \frac{v_Z}{Z} + \left(\frac{v_Z}{Z}\right)^2 - \ldots\right)$$
$$= f\left(v_Z/Z\right) \approx 1 \vee \left(1 - \frac{v_Z}{Z}\right)$$

Motion model



Surface point

$$\mathbf{X} = (X, Y, Z)^T \in \mathbf{R}^3$$

Image point

$$\mathbf{x} = (x, y)^T \in \mathbf{R}^2$$

Scene Flow

$$\mathbf{V} = (V_X, V_Y, V_Z)^T \in \mathbf{R}^3$$

$$\mathbf{V} = \mathbf{X}^{t+1} - \mathbf{X}^t$$

Warp function

$$\mathbf{x^{t+1}} = \mathbf{W}(\mathbf{x^t}; \mathbf{V})$$

$$\mathbf{W}(\mathbf{x^t}; \mathbf{V}) = \mathbf{x^t} + \begin{pmatrix} u \\ v \end{pmatrix}$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{Z^t} \begin{pmatrix} 1 & 0 & -x^t \\ 0 & 1 & -y^t \end{pmatrix} \begin{pmatrix} V_X \\ V_Y \\ V_Z \end{pmatrix}$$

Presentation outline

- Motion model
- **Data term**
- Regularisation term
- Sparse matching term
- Optimisation
- Experimentation
- Conclusion

Data term

Intensity image

Depth image

Brightness constancy assumption (**BCA**)

$$I_2(\mathbf{W}(\mathbf{x};\mathbf{v})) = I_1(\mathbf{x})$$

Depth velocity constraint (**DVC**)

$$Z_2(\mathbf{W}(\mathbf{x};\mathbf{v})) = Z_1(\mathbf{x}) + v_Z(\mathbf{x})$$

Data term

We solve for the local scene flow vector $\mathbf{v}$ that minimizes

$$\sum_{\{\mathbf{x}\}} \Psi\left(\left|\rho_I\left(\mathbf{x},\mathbf{v}\right)\right|^2\right) + \lambda\Psi\left(\left|\rho_Z\left(\mathbf{x},\mathbf{v}\right)\right|^2\right),$$

where $\Psi\left(s^2\right) = \sqrt{s^2+\varepsilon^2}$ is a differentiable approx. of the L^1 norm.

Using IRLS the scene flow increment is given by

$$\begin{aligned}\Delta\mathbf{v} = \mathbf{H}^{-1}\sum_{\{\mathbf{x}\}}\Big\{&-\Psi'\left(\rho_I^2\left(\mathbf{x},\mathbf{v}\right)\right)\left(\nabla_I\mathbf{J}\right)^T\rho_I\left(\mathbf{x}',\mathbf{v}\right)\\ &-\lambda\,\Psi'\left(\rho_Z^2\left(\mathbf{x},\mathbf{v}\right)\right)\left(\nabla_Z\mathbf{J}-(0,0,1)\right)^T\rho_Z\left(\mathbf{x}',\mathbf{v}\right)\Big\}\end{aligned}$$

where the Jacobian is defined as

$$\mathbf{J} = \frac{\partial\mathbf{W}}{\partial\mathbf{v}} = \frac{1}{Z(\mathbf{x})}\begin{pmatrix} f_x & 0 & c_x - x \\ 0 & f_y & c_y - y \end{pmatrix}.$$

Data term

The matrix H is the Gauss-Newton approximation of the Hessian

$$\mathbf{H} = \sum_{\{\mathbf{x}\}} \frac{\Psi'_{\rho_I}}{Z^2} \begin{pmatrix} I_x^2 & I_x I_y & I_x I_\Sigma \\ I_x I_y & I_y^2 & I_y I_\Sigma \\ I_x I_\Sigma & I_y I_\Sigma & I_\Sigma^2 \end{pmatrix} + \lambda \frac{\Psi'_{\rho_Z}}{Z^2} \begin{pmatrix} Z_x^2 & Z_x Z_y & Z_x (Z_\Sigma - 1) \\ Z_x Z_y & Z_y^2 & Z_y (Z_\Sigma - 1) \\ Z_x (Z_\Sigma - 1) & Z_y (Z_\Sigma - 1) & (Z_\Sigma - 1)^2 \end{pmatrix}$$

with $I_\Sigma = -\left(x I_x + y I_y\right)$ and $Z_\Sigma = -\left(x Z_x + y Z_y\right)$.

Final expression

$$E_D(\mathbf{v}) = \sum_{\mathbf{x}} \sum_{\mathbf{x}' \in N(\mathbf{x})} \Psi\left(\left|\rho_I\left(\mathbf{x}', \mathbf{v}(\mathbf{x})\right)\right|^2\right) + \lambda \Psi\left(\left|\rho_Z\left(\mathbf{x}', \mathbf{v}(\mathbf{x})\right)\right|^2\right)$$

Presentation outline

- Motion model
- Data term
- **Regularisation term**
- Sparse matching term
- Optimisation
- Experimentation
- Conclusion

Regularisation term

The regularization term is given by:

$$E_R(\mathbf{v}) = \sum_{\mathbf{x}} \omega(\mathbf{x}) \left|\nabla \mathbf{v}(\mathbf{x})\right|,$$

where we use the notation $|\nabla \mathbf{v}| := |\nabla v_X| + |\nabla v_Y| + |\nabla v_Z|$.

The decreasing positive function

$$\omega(\mathbf{x}) = \exp\left(-\alpha |\nabla Z_1(\mathbf{x})|^{\beta}\right)$$

prevent regularization of the motion field along strong depth discontinuities.

Presentation outline

- Motion model
- Data term
- Regularisation term
- **Sparse matching term**
- Optimisation
- Experimentation
- Conclusion

Matching term

Let $\left\{ \left(\mathbf{x}_1^1, \mathbf{x}_2^1 \right), ..., \left(\mathbf{x}_1^N, \mathbf{x}_2^N \right) \right\}$ be the set of correspondences, the matching term is defined as

$$E_M(\mathbf{v}) = \sum_{\mathbf{x}} p(\mathbf{x}) \Psi \left(|\delta_{3D}(\mathbf{x}, m(\mathbf{x})) - \mathbf{v}(\mathbf{x})|^2 \right)$$

with $p(\mathbf{x}) = 1$ if there is a descriptor in a region around point $\mathbf{x}$.

The matching function $m(\mathbf{x})$ gives the correspondency of each pixel $\mathbf{x}$.

The function $\delta_{3D}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{M}_{\text{cam}}^{-1} (\mathbf{x}_2 Z_2(\mathbf{x}_2) - \mathbf{x}_1 Z_1(\mathbf{x}_1))$ computes the 3D displacement for each correspondency.

Presentation outline

- Motion model
- Data term
- Regularisation term
- Sparse matching term
- **Optimization**
- Experimentation
- Conclusion

Optimization

To compute the scene flow we introduce an auxiliary flow and solve for the 3D motion field **v** that minimizes

$$E(\mathbf{v}, \mathbf{u}) = E_D(\mathbf{v}) + \alpha E_M(\mathbf{v}) + \frac{1}{2\theta} |\mathbf{v} - \mathbf{u}|^2 + \beta E_R(\mathbf{u})$$

where θ is a small constant.

1. For a fixed **v**, we solve for **u** that minimizes

$$\sum_{\mathbf{x}} \frac{1}{2\kappa} |\mathbf{u}(\mathbf{x}) - \mathbf{v}(\mathbf{x})|^2 + \omega(\mathbf{x}) |\nabla \mathbf{u}(\mathbf{x})|$$

where $\kappa = \beta\theta$. For every dimension this problem corresponds to a weighted version of the ROF model for image denoising.

Optimization

2. For a fixed $\mathbf{u}$, we solve for $\mathbf{v}$ that minimizes

$$E_D(\mathbf{v}) + \alpha E_M(\mathbf{v}) + \sum_{\mathbf{x}} \frac{1}{2\theta} |\mathbf{v}(\mathbf{x}) - \mathbf{u}(\mathbf{x})|^2$$

The scene flow increment can be computed as

$$\begin{aligned}\Delta\mathbf{v} = \mathbf{H}^{-1} \sum_{\mathbf{x}' \in N(\mathbf{x})} \Big\{ & -\Psi' \left(\rho_I^2 \left(\mathbf{x}', \mathbf{v}\right)\right) \left(\nabla_I \mathbf{J}\right)^T \rho_I \left(\mathbf{x}', \mathbf{v}\right) \\ & -\lambda \Psi' \left(\rho_Z^2 \left(\mathbf{x}', \mathbf{v}\right)\right) \left(\nabla_Z \mathbf{J} - \mathbf{D}\right)^T \rho_Z \left(\mathbf{x}', \mathbf{v}\right) \Big\} \\ & + \alpha\, p(\mathbf{x}) \Psi' \left(\rho_{3D}^2 \left(\mathbf{x}, \mathbf{v}\right)\right) \rho_{3D} \left(\mathbf{x}, \mathbf{v}\right) + \frac{1}{2\theta}(\mathbf{u} - \mathbf{v})\end{aligned}$$

where ρ_{3D} is a 3D residue defined as

$$\rho_{3D} \left(\mathbf{x}, \mathbf{v}\right) = \delta_{3D} \left(\mathbf{x}, m(\mathbf{x})\right) - \mathbf{v},$$

and $\mathbf{H}$ is the Gauss-Newton approximation of the Hessian matrix.

Optimization

The (G-N approximation) of the Hessian matrix is given by

$$\begin{aligned}\mathbf{H} = \sum_{\mathbf{x}' \in N(\mathbf{x})} & \left\{ \Psi' \left(\rho_I^2 \left(\mathbf{x}', \mathbf{v} \right) \right) \left(\nabla_I \mathbf{J} \right)^T \left(\nabla_I \mathbf{J} \right) \right. \\ & \left. + \lambda \, \Psi' \left(\rho_Z^2 \left(\mathbf{x}', \mathbf{v} \right) \right) \left(\nabla_Z \mathbf{J} - \mathbf{D} \right)^T \left(\nabla_Z \mathbf{J} - \mathbf{D} \right) \right\} \\ & + \alpha \, p(\mathbf{x}) \Psi' \left(\rho_{3D}^2 \left(\mathbf{x}, \mathbf{v} \right) \right) \mathbf{I_d} + \frac{1}{2\theta} \mathbf{I_d}\end{aligned}$$

with $\mathbf{I_d}$ the 3×3 identity matrix.

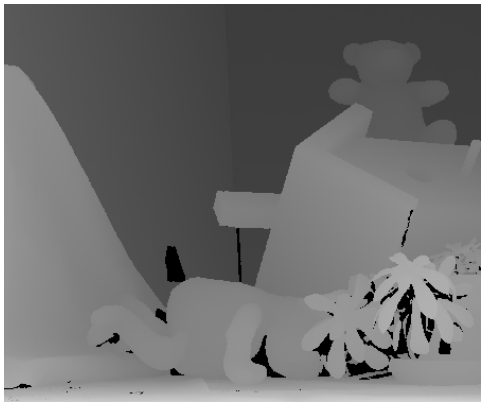
Presentation outline

- Motion model
- Data term
- Regularisation term
- Sparse matching term
- Optimization
- **Experimentation**
- Conclusion

Experimentation - Middlebury datasets

I_1

I_2



Z_1

ground truth (OF)

Details

- Images : Teddy, Cones (2 and 6)
- 5 levels of PYR decomposition
- Window size: 5×5

Error measures

- Optical flow: NRMS_{OF}, AAE_{OF}
- Scene flow: NRMS_{V}, $P10\%$

Comparisons

- LG_{SF}: proposed method
- L_{SF}: local scene flow
- TV-L^1: optical flow + depth
- ORT_{SF}: ortographic camera
- Hug_{07}: Huguet and Devernay, ICCV 2007
- Bas_{10}: Basha et al., CVPR 2010
- Had_{11}: Hadfield and Bowden, ICCV 2011

Experimentation - Middlebury datastes

	Teddy		*Cones*	
	$NRMS_{OF}$	AAE	$NRMS_{OF}$	AAE
LG_{SF}	0.0222	0.837	0.0164	0.526
TV-L^1	0.0642	1.360	0.0509	0.932
L_{SF}	0.0780	2.288	0.0577	1.991
ORT_{SF}	0.0811	0.866	0.0594	0.963
Bas_{10}	0.0285	1.010	0.0307	0.390
Hug_{07}	0.0621	0.510	0.0579	0.690
Had_{11}	0.110	5.040	0.090	5.020

Table 1 : Optical flow errors.

	Original		*Modified*	
	$NRMS_{SF}$	P10%	$NRMS_{SF}$	P10%
LG_{SF}	0.0353	97,55	0.0754	90,28
TV-L^1	0.5493	84,94	0.4662	84,85
L_{SF}	0.4415	89,07	0.3039	83,16
ORT_{SF}	0.4678	82,77	0.4999	82,34

Table 2 : Scene flow errors.

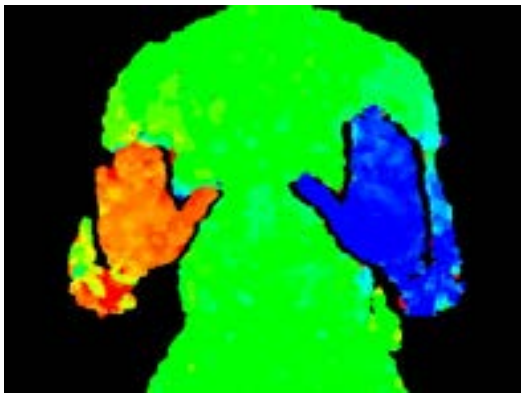
I_1

$I_2(modified)$

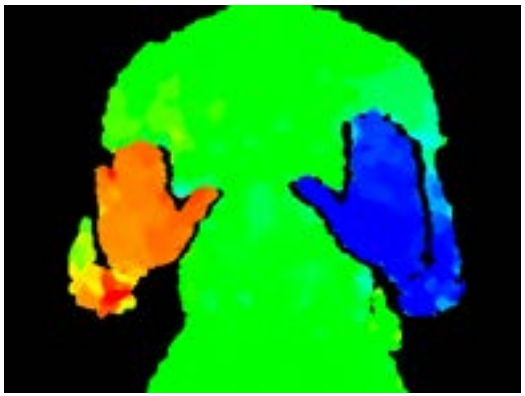
Experimentation - Kinect images

Depth velocity (V_Z)

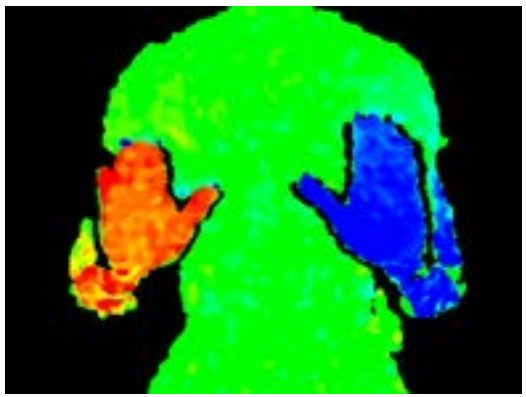
Input color frames



L_{SF}



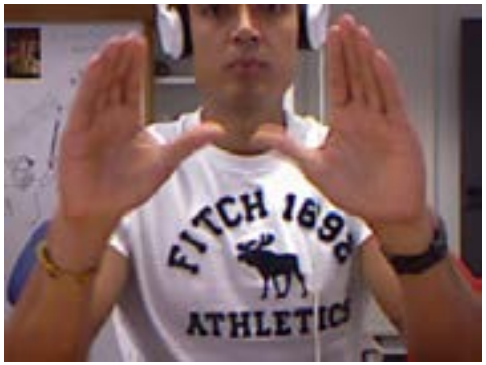
LG_{SF}



TV-L^1

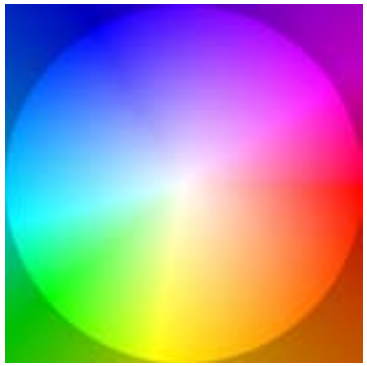
Experimentation - Kinect images

Image flow $((u, v))$



(A) Input images

(B) Input images

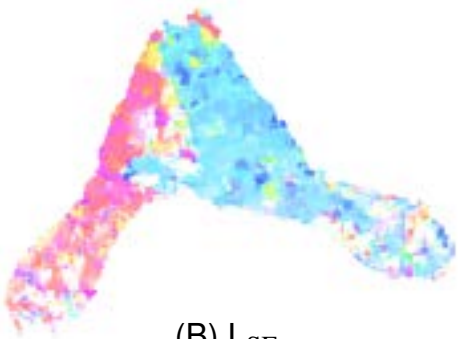


Color code

(A) L_{SF}

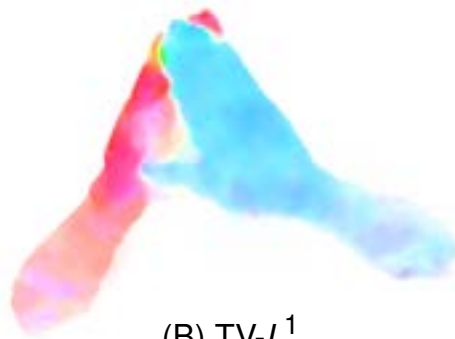
(A) LG_{SF}

(A) TV-L^1



(B) L_{SF}

(B) LG_{SF}



(B) TV-L^1

Presentation outline

- Motion model
- Data term
- Regularisation term
- Sparse matching term
- Optimization
- Experimentation
- **Conclusion**

Conclusion

- We proposed a novel approach to compute a dense scene flow using intensity and depth data.
- We combine local and global constraints to solve for the 3D motion field in a variational framework.
- Unlike previous methods, depth data is used in 3 ways: to model the motion in the image domain, to constrain the scene flow and to adapt the TV-regularization.

Current and future work

- Scene flow descriptors.
- Improvements: occlusions, large motions, noise.
- GPU implementation.
- 3D reconstruction of non-rigid objects.

The End

References

- J. Quiroga, F. Devernay, and J. Crowley, *Scene flow by tracking in intensity and depth data*, in Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 2012.
- J. Quiroga, F. Devernay, and J. Crowley, *Local scene flow by tracking in intensity and depth*, Journal of Visual Communication and Image Representation (JVCIR), April 2013.
- J. Quiroga, F. Devernay, and J. Crowley, *Local/Global scene flow*, in International Conference on Image Processing (ICIP), September 2013.