THE ROMANIAN ACADEMY

SIMION STOILOW INSTITUTE OF MATHEMATICS

Robust Estimation of Visual Correspondence Fields and Dynamic Scenes Structure

Author,
Andrei ZANFIR

Supervisor,
C.S. I Dr. Cristian SMINCHIȘESCU

PH. D. THESIS SUMMARY

Bucharest
2018

Chapter 1

Introduction

Establishing correspondences between similar structures (e.g. points) in different images is a fundamental computer problem. It exposes the need for non-local image reasoning, leveraging the underlying 3d geometry of the scene, and modeling the range of deformations that many objects – including deformable and articulated structures like people or animals – are subject to. The applications are diverse, as current emerging technologies rely on variants of image-to-image matching like optical flow, scene flow, sparse matching, semantic matching, etc. Self-driving cars need to understand visual motion, drones have to track objects and virtual reality devices need to build 3d geometrical models of the environment. Hence they require solutions to fundamental, yet challenging questions like: When do two visual primitives are the instances of the same moving 3d structures? How can two structures from the same semantic category, but with different appearance, can be matched?

In this thesis we do not only propose models and solutions for image correspondence, but we also offer insights and long-term directions for future research. Our models are designed to address the current challenges in image matching including large-displacements and fast-motions, repetitive structures, non-trivial lighting conditions, cluttered backgrounds, intra-class variability, or large occlusion areas.

For computing optical flow, we propose a sparse-to-coarse method, that combines sparse matching with a novel affine model and a geometric interpolation method, with competitive results. We show that pursuing alternative models to the classical variational approach is

beneficial and has practical relevance.

We address the problem of scene flow, the 3d counter-part of optical flow, by additionally exploiting the depth information collected from a depth sensor. By incorporating a complete energy model that balances sparse matching, a 3d rigid model hypothesis, photometric and depth consistency terms, we obtain state-of-the-art results, that can capture fast motions and sharp boundaries.

We further propose a deep-learning formulation of matching, where we design and train a model to minimize a graph-matching objective, that combines unary and pairwise node neighborhood relations. This novel approach poses considerable mathematical challenges, as the derivatives need to be passed across complete structured layers (like optimization modules with solutions based on e.g. relaxations) accurately, following matrix back-propagation rules. We construct a complex, fully trainable framework, and show improved results for both geometric and semantic matching datasets. This latter work received the best paper award honorable mention at IEEE CVPR 2018.

Chapter 2

Locally Affine Sparse-to-Dense Matching for Motion and Occlusion Estimation

This chapter is based on "Locally Affine Sparse-to-Dense Matching for Motion and Occlusion Estimation" Marius Leordeanu, Andrei Zanfir and Cristian Sminchișescu, in the IEEE International Conference on Computer Vision, Sydney, Australia, December 2013.

2.1 Introduction

Estimating a dense correspondence field between subsequent video frames is important to many visual learning and recognition tasks. Here we propose a novel sparse-to-dense matching method for motion field estimation with occlusion detection. As an alternative to the current coarse-to-fine approaches from the optical flow literature, we start from the higher level of sparse matching with rich appearance and geometric constraints, using a novel, occlusion aware, locally affine model. Then, we move towards the simpler, but denser classic flow field model, with a novel sparse-to-dense matching interpolation procedure that offers a natural transition between the sparse and dense correspondence fields. We demonstrate experimentally that our appearance features and complex geometric constraints permit the correct motion estimation even in difficult cases of large displacements

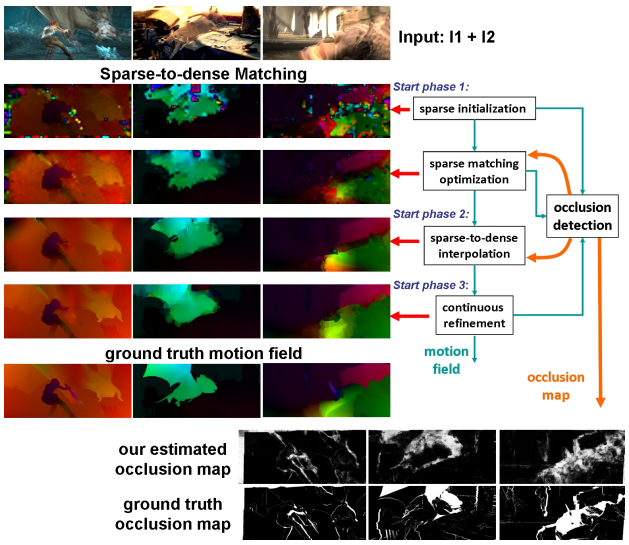

Figure 2-1: Left: motion fields estimated at different stages of our method. Right: flowchart of our approach. Sparse matching results are resized to original image size for display purposes.

and significant appearance changes. We also propose a classification method for occlusion detection that works in conjunction with sparse-to-dense matching. We validate our approach on the newly released Sintel dataset, on which we obtain state-of-the-art results.

Overview of our approach: We propose a hierarchical sparse-to-dense matching method that integrates and generalizes ideas from both the traditional coarse-to-fine variational approach and more recent methods using sparse feature matching [1, 7, 4, 3, 14, 8]. Our method consists of three main phases (Figure 2-1):

1. **Sparse Matching:** Initialize a discrete set of candidate matches for each point on a sparse grid (in the first image) using dense kNN matching (in the second image) with local feature descriptors. Use the output of a boundary detector and cues from the initial sparse matching to infer a first occlusion probability map. Then discretely optimize a sparse matching cost function using both unary data terms (from local features) and geometric relationships based on a locally affine model with occlusion constraints. Use the improved matches to refine the occlusion map.

2. **Sparse-to-Dense Interpolation:** Fix the sparse matches from the previous step.

Then apply the same occlusion sensitive geometric model to obtain an accurate sparse-to-dense matching interpolation.

3. **Dense matching refinement:** use a TV model with continuous flow optimization to obtain the final dense correspondence field. Use matching cues computed from all stages to obtain the final occlusion map.

2.1.1 Locally Affine Spatial Model

The key contribution of our paper is the locally affine spatial prior that we propose. The intuition is that points that are likely to be on the same object surface and are close to each other are also likely to have a similar displacement. The motions of such points within a certain neighborhood are expected to closely follow an affine transformation. To model this idea, we first define a geometric neighborhood system over the grid locations (see Figure 2-2), and link neighboring points (i, j) using an edge strength function e_{ij} - meant to measure the probability that two points lie on the same object surface and obey a similar affine transformation from I_1 to I_2.

We use the locations of i's neighbors, $\mathbf{p}_{N_i}^{(1)}$ in I_1, their current estimated destinations in I_2, $\mathbf{p}_{N_i}^{(2)}$, and their membership weights, to estimate a motion model that *predicts* the destination $\mathbf{p}_i^{(2)}$ of the current point i in I_2: $\tilde{\mathbf{p}}_i^{(2)} = T_{\mathcal{N}_i}(\mathbf{p}_i^{(1)})$. Then, the prediction error between the current $\mathbf{p}_i^{(2)}(\mathbf{w}_i) = \mathbf{p}_i^{(1)} + \mathbf{w}_i$ and the estimated $\tilde{\mathbf{p}}_i^{(2)}$ forms the basis of our spatial term:

$$E_S(\mathbf{w}) = \sum_i \|\mathbf{p}_i^{(2)}(\mathbf{w}) - T_{\mathcal{N}_i}(\mathbf{p}_i^{(1)})\|^2. \tag{2.1}$$

In our particular case, when $T_{\mathcal{N}_i}$ is a local affine transformation, Eq. 2.1 reduces to a second order energy that can be efficiently optimized.

For each point i we estimate its affine transformation $T_{\mathcal{N}_i} = (\mathbf{A_i}, \mathbf{t_i})$ from its neighbors' motions by weighted least squares, with weights $e_{ij},\ \forall j \in \mathcal{N}_i$. Let $2N_i \times 1$ $\mathbf{p}_{N_i} = \mathbf{p}_{N_i}^{(2)}$ be the estimated positions of its neighbors in I_2 and $\mathbf{M}$ the Moore-Penrose pseudo-inverse of the least squares problem, which depends only on the locations of the neighbors $\mathbf{p}_{N_i}^{(1)}$, and their weights.

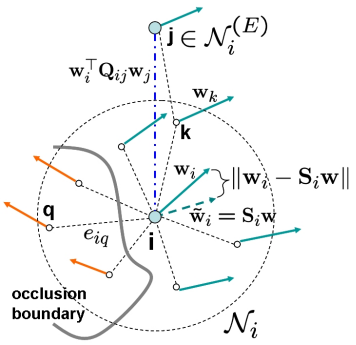

Figure 2-2: A feature i is connected to its neighbor q in $\mathcal{N}_i$ with strength e_{iq} - a function of distance, and intervening boundary and occlusion information. The current motions of neighbors in $\mathcal{N}_i$ are used to predict the motion $\tilde{\mathbf{w}}_j$ at point i, through affine mapping $\mathbf{S}_i$. Our novel quadratic affine error model effectively produces an extended neighborhood system $\mathcal{N}_i^{(E)}$ in which pairs of *connected* points (i, j) contribute to the total error with quantity $\mathbf{w}_i^\top \mathbf{Q}_{ij} \mathbf{w}_j$.

The least squares solution $\mathbf{M}\mathbf{p}_{N_i}$ gives the estimated $(\mathbf{A_i}, \mathbf{t_i})$:

$$\mathbf{A}_i = \begin{bmatrix} \mathbf{m}_1^\top \mathbf{p}_{N_i} & \mathbf{m}_2^\top \mathbf{p}_{N_i} \\ \mathbf{m}_3^\top \mathbf{p}_{N_i} & \mathbf{m}_4^\top \mathbf{p}_{N_i} \end{bmatrix} \qquad \mathbf{t}_i = \begin{bmatrix} \mathbf{m}_5^\top \mathbf{p}_{N_i} \\ \mathbf{m}_6^\top \mathbf{p}_{N_i} \end{bmatrix}. \tag{2.2}$$

Here $\mathbf{m}_k^\top$ denotes the k-th row of $\mathbf{M}$. Using $\mathbf{p}_i^{(1)} = [x_i, y_i]$, the predicted end position $\tilde{\mathbf{p}}_i^{(2)}$ of i in image I_2 is:

$$\begin{aligned} \tilde{\mathbf{p}}_i^{(2)} &= T_{\mathcal{N}_i}(\mathbf{p}_i^{(1)}) = \mathbf{A}_i \mathbf{p}_i^{(1)} + \mathbf{t}_i \\ &= \overbrace{\left[\begin{array}{c|c|c} 0 & \begin{array}{c} x_i\mathbf{m}_1^\top + y_i\mathbf{m}_2^\top + \mathbf{m}_5^\top \\ x_i\mathbf{m}_3^\top + y_i\mathbf{m}_4^\top + \mathbf{m}_6^\top \end{array} & 0 \end{array}\right]}^{\mathbf{S}_i} \mathbf{p}^{(2)}. \end{aligned} \tag{2.3}$$

Note that $\mathbf{S}_i$ does not depend on the unknown displacements (or final positions $\mathbf{p}^{(2)}$) which we want to solve for. Equation 2.1 can now be written as:

$$\begin{aligned} E_S(\mathbf{w}) &= \sum_i \|\mathbf{p}_i^{(2)} - \mathbf{S}_i\mathbf{p}^{(2)}\|^2 = \sum_i \|\mathbf{w}_i - \mathbf{S}_i\mathbf{w}\|^2 \\ &= \mathbf{w}^\top(\mathbf{I} - 2\mathbf{S}_{1\ldots n} + \sum_i \mathbf{S}_i^\top \mathbf{S}_i)\mathbf{w} \\ &= \mathbf{w}^\top \mathbf{S}\mathbf{w}, \end{aligned} \tag{2.4}$$

Table 2.1: Final results on Sintel optical flow benchmark; epe stands for end point errors, epe_m are errors for visible points, and epe_o are errors for occluded points. The last three columns show average errors for different motion magnitudes (in pixels).

Alg.	epe	epe_m	epe_o	0-10	10-40	40+
S2D	**7.88**	**3.91**	**40.16**	1.17	4.66	**48.90**
[14]	8.45	4.15	43.43	1.42	5.45	50.51
[3]	9.12	5.04	42.34	1.49	4.84	57.30
[12]-full	9.16	4.81	44.51	1.11	**4.50**	60.29
[5]	9.61	5.42	43.73	1.88	5.34	58.27
[12]++	9.96	5.41	47.00	1.40	5.10	64.14
[12]-fast	10.09	5.66	46.15	**1.09**	4.67	67.80
[13]	11.93	7.32	49.37	1.16	7.97	74.80

This energy term $E_S(\mathbf{w})$ will be used by the sparse matching and the sparse-to-dense interpolation stages to compute or update the solution $\mathbf{w}^*$.

2.2 Experimental results

We show qualitative and quantitative results in fig. 2-3 and in table 2.1.

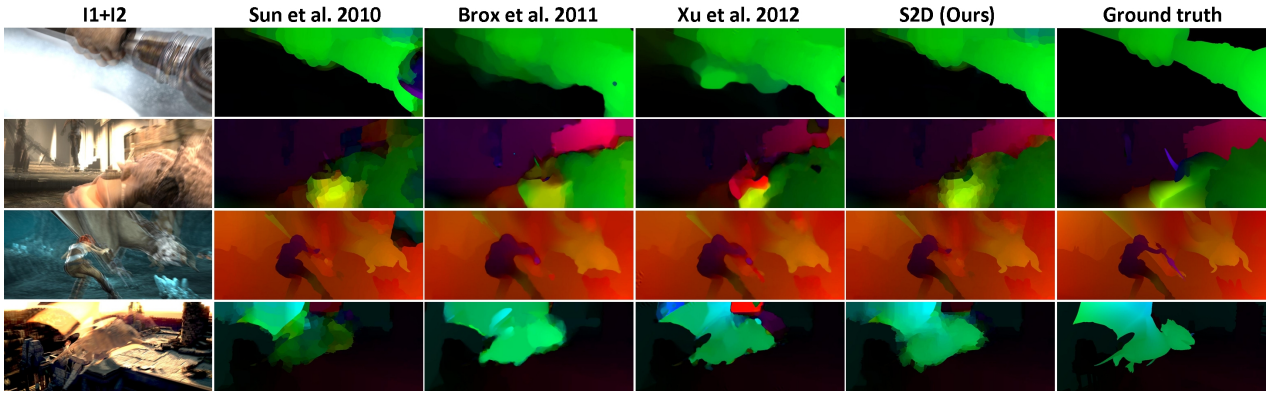

Figure 2-3: Motion field estimation results. Our algorithm can handle large displacements (in the sparse matching phase) while preserving motion details by TV continuous refinement at the last stage.

Chapter 3

Large Displacement 3D Scene Flow with Occlusion Reasoning

This chapter is based on the paper "Large Displacement 3D Scene Flow with Occlusion Reasoning" Andrei Zanfir and Cristian Sminchișescu, published in the IEEE International Conference on Computer Vision, Santiago de Chile, Chile, December 2015

3.1 Introduction

The emergence of modern, affordable and accurate RGB-D sensors increases the need for single view approaches to estimate 3-dimensional motion, also known as *scene flow*. In this paper we propose a coarse-to-fine, dense, correspondence-based *scene flow* formulation that relies on explicit geometric reasoning to account for the effects of large displacements and to model occlusion. Our methodology enforces local motion rigidity at the level of the 3d point cloud without explicitly smoothing the parameters of adjacent neighborhoods. By integrating all geometric and photometric components in a single, consistent, occlusion-aware energy model, defined over overlapping, image-adaptive neighborhoods, our method can process fast motions and large occlusions areas, as present in challenging datasets like the MPI Sintel Flow Dataset, recently augmented with depth information. By explicitly modeling large displacements and occlusion, we can handle difficult sequences which cannot be currently processed by state of the art scene flow methods. We also show

that by integrating depth information into the model, we can obtain correspondence fields with improved spatial support and sharper boundaries compared to the state of the art, large-displacement optical flow methods.

3.2 Model Formulation

We design a dense energy model that can handle large displacement and occlusion. The model is expressed in terms of several energy sub-components. It relies on geometric large-displacement correspondence anchors, local rigidity assumptions defined over large neighborhoods constructed using the RGB-D point cloud, on appearance consistency, and depth constancy terms. Geometric occlusion estimates are used in order to control the strength of the neighborhood connections and automatically mask regions for which correspondences cannot be established.

Rigid Parameter Fields. Our geometric model links the dense correspondences $\mathbf{Y}$ to the initial points $\mathbf{X}$ by assuming an underlying field of rigid parameters $\mathbf{\Theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, ..., \boldsymbol{\theta}_N^\top]^\top$, where $\boldsymbol{\theta}_i = [\mathbf{v}_i^\top, \mathbf{t}_i^\top]^\top$ are rigid parameters that constrain each neighborhood $\mathcal{N}(i)$ of a point $\mathbf{x}_i \in \mathbf{X}$. A neighborhood represents the closest K points to $\mathbf{x}_i$ in 3d, where K is chosen appropriately, according to the pyramid level.

The full energy model to optimize can be written as:

$$E(\mathbf{Y}, \mathbf{\Theta}) = E_A + \alpha E_Z + \beta E_M + \lambda E_G \tag{3.1}$$

where α, β and λ are weights estimated by validation.

Minimization. We solve (3.1) over the parameter fields $\mathbf{Y}, \mathbf{\Theta}$ by alternating variable optimization:

$$\mathbf{\Theta}^{k+1} \leftarrow \underset{\mathbf{\Theta}}{\operatorname{argmin}}\, E(\mathbf{Y}^k, \mathbf{\Theta}^k) \tag{3.2}$$

$$\mathbf{Y}^{k+1} \leftarrow \underset{\mathbf{Y}}{\operatorname{argmin}}\, E(\mathbf{Y}^k, \mathbf{\Theta}^{k+1}) \tag{3.3}$$

where k is the iteration index. We run the optimization until convergence.

3.3 Experimental results

We show quantitative and qualitative results on state-of-the-art scene flow benchmarks in table 3.1, in fig. 3-1 and in fig. 3-2.

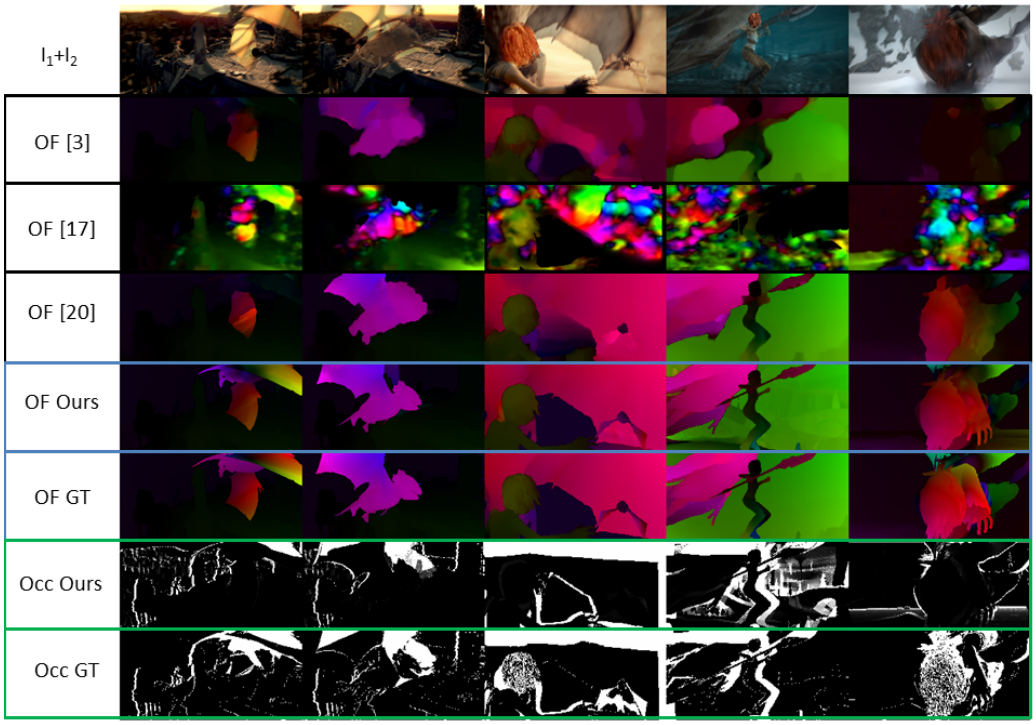

Figure 3-1: Five pairs of images and four methods illustrated on the Sintel dataset: **1st row:** input images overlaid, **2nd row:** [2], **3rd row:** [9] (N.B. this state of the art 3d scene flow method is not designed to handle large displacements, but produces excellent results for small to moderate ones), see also fig. 3-2, **4th row:** EpicFlow[10], **5th row:** our proposed approach shows sharper boundaries and improved spatial support estimates, **6th row:** ground truth optical flow. On the last two rows we show the ground truth occlusion states and our soft estimates.

Algorithms	LDOF[2]	EpicFlow[10]	Ours
Error(in pixels)	7.44	4.82	4.6

Table 3.1: Quantitative evaluation of state of the art 2d and 3d methods on the challenging Sintel dataset (*Sintel-Test92*). Our method uses 3d information and offers improved accuracy as well as accurate estimates of occlusion.

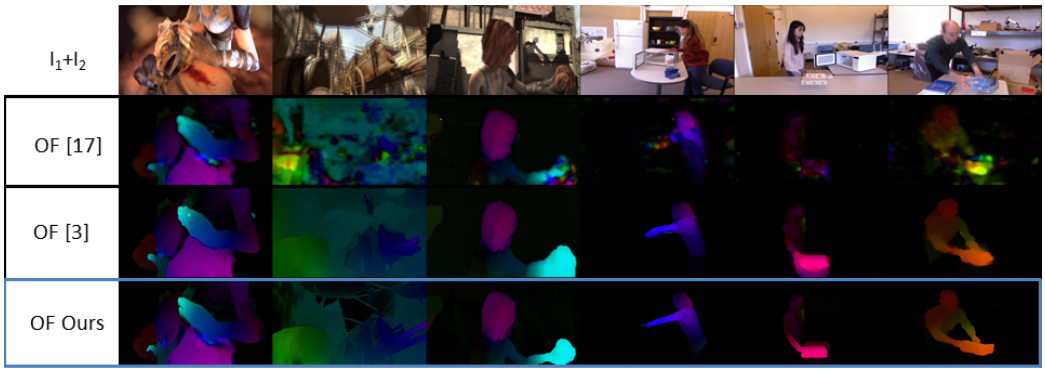

Figure 3-2: Sample scene flow estimation with results in the x-y plane for image pairs in the Sintel (first three columns) and CAD120 (last three columns). Displacements are *moderate*. We show results from 3 different methods: **1st row:** 3d scene flow[9]; **2nd row:** large-displacement 2d optical flow[2]; **3rd row:** our proposed 3d scene flow method without relying on large-displacement anchors. Notice the higher level of detail captured by our model.

Chapter 4

Deep Learning of Graph Matching

This chapter is based on the paper "Deep Learning of Graph Matching" Andrei Zanfir and Cristian Sminchișescu, published in the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake city, U.S.A, June 2018

4.1 Introduction

The problem of graph matching under node and pair-wise constraints is fundamental in areas as diverse as combinatorial optimization, machine learning or computer vision, where representing both the relations between nodes and their neighborhood structure is essential. We present an end-to-end model that makes it possible to learn all parameters of the graph matching process, including the unary and pairwise node neighborhoods, represented as deep feature extraction hierarchies. The challenge is in the formulation of the different matrix computation layers of the model in a way that enables the consistent, efficient propagation of gradients in the complete pipeline from the loss function, through the combinatorial optimization layer solving the matching problem, and the feature extraction hierarchy. Our computer vision experiments and ablation studies on challenging datasets like PASCAL VOC keypoints, Sintel and CUB show that matching models refined end-to-end are superior to counterparts based on feature hierarchies trained for other problems.

Methodologically, our contributions are associated to the construction of the different matrix layers of the computation graph, obtaining analytic derivatives all the way from the

loss function down to the feature layers in the framework of matrix back-propagation, the emphasis on computational efficiency for backward passes, as well as a voting based loss function. The proposed model applies generally, not just for matching different images of a category, taken in different scenes (its primary design), but also to different images of the same scene, or from a video.

4.2 Problem Formulation

Input. We are given two input graphs $\mathcal{G}_1 = (V_1, E_1)$ and $\mathcal{G}_2 = (V_2, E_2)$, with $|V_1| = n$ and $|V_2| = m$. Our goal is to establish an assignment between the nodes of the two graphs, so that a criterion over the corresponding nodes and edges is optimized (see below).

Graph Matching. Let $\mathbf{v} \in \{0,1\}^{nm\times 1}$ be an indicator vector such that $\mathbf{v}_{ia} = 1$ if $i \in V_1$ is matched to $a \in V_2$ and 0 otherwise, while respecting one-to-one mapping constraints. We build a square symmetric positive matrix $\mathbf{M} \in \mathbb{R}^{nm\times nm}$ such that $\mathbf{M}_{ia;jb}$ measures how well every pair $(i,j) \in E_1$ matches with $(a,b) \in E_2$. For pairs that do not form edges, their corresponding entries in the matrix are set to 0. The diagonal entries contain node-to-node scores, whereas the off-diagonal entries contain edge-to-edge scores. The optimal assignment $\mathbf{v}^*$ can be formulated as

$$\mathbf{v}^* = \underset{\mathbf{v}}{\operatorname{argmax}}\, \mathbf{v}^\top \mathbf{M}\mathbf{v}, \text{ s.t. } \mathbf{C}\mathbf{v} = \mathbf{1}, \mathbf{v} \in \{0,1\}^{nm\times 1} \tag{4.1}$$

The binary matrix $\mathbf{C} \in \mathbb{R}^{nm\times nm}$ encodes one-to-one mapping constraints: $\forall a \sum_i \mathbf{v}_{ia} = 1$ and $\forall i \sum_a \mathbf{v}_{ia} = 1$. This is known to be NP-hard, so we relax the problem by dropping both the binary and the mapping constraints, and solve

$$\mathbf{v}^* = \underset{\mathbf{v}}{\operatorname{argmax}}\, \mathbf{v}^\top \mathbf{M}\mathbf{v}, \text{ s.t. } \|\mathbf{v}\|_2 = 1 \tag{4.2}$$

The optimal $\mathbf{v}^*$ is then given by the leading eigenvector of the matrix $\mathbf{M}$. Since $\mathbf{M}$ has non-negative elements, by using Perron-Frobenius arguments, the elements of $\mathbf{v}^*$ are in the interval $[0,1]$, and we interpret $\mathbf{v}^*_{ia}$ as the confidence that i matches a.

Learning. We estimate the matrix $\mathbf{M}$ parameterized in terms of unary and pair-wise point features computed over input images and represented as deep feature hierarchies. We learn the feature hierarchies end-to-end in a loss function that also integrates the matching layer. Specifically, given a training set of correspondences between pairs of images, we adapt the parameters so that the matching minimizes the error, measured as a sum of distances between predicted and ground truth correspondences.

4.3 Approach

The work [15] introduced a factorization of the matrix $\mathbf{M}$ that explicitly exposes the graph structure of the set of points and the unary and pair-wise scores between nodes and edges, respectively,

$$\mathbf{M} = [\text{vec}(\mathbf{M}_p)] + (\mathbf{G}_2 \otimes \mathbf{G}_1)[\text{vec}(\mathbf{M}_e)](\mathbf{H}_2 \otimes \mathbf{H}_1)^\top \tag{4.3}$$

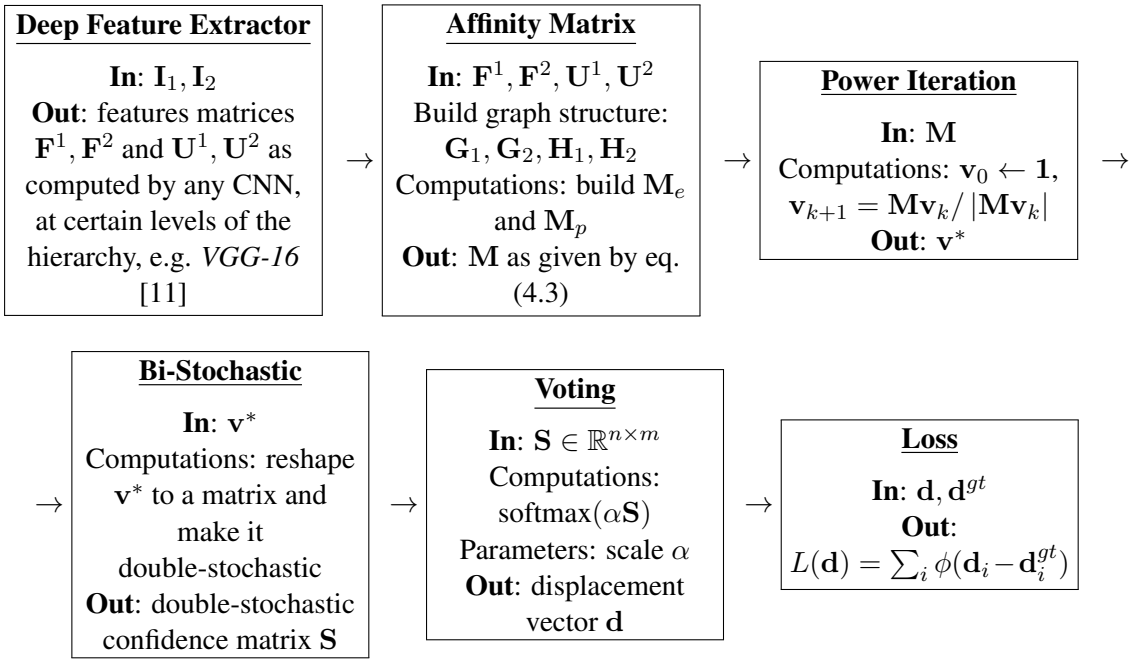

Figure 4-1: Computational pipeline of our fully trainable graph matching model. In training, gradients w.r.t. the loss function are passed through a deep feature extraction hierarchy, the factorization of the resulting affinity matrix, the eigen-decomposition solution of the matching problem, and the voting-based assignment layer.

Computing the leading eigenvector $\mathbf{v}^*$ of the affinity matrix $\mathbf{M}$ can be done using power iterations

$$\mathbf{v}_{k+1} = \frac{\mathbf{M}\mathbf{v}_k}{\|\mathbf{M}\mathbf{v}_k\|} \tag{4.4}$$

Backward pass. To compute gradients, we express the variation of the loss and identify the required partial derivatives. By employing techniques of **matrix back-propagation** [6], we can exploit the special factorization (4.3) of matrix $\mathbf{M}$, to make operations both memory and time efficient.

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{M}_e} = &\sum_k \mathbf{G}_1^\top \left(\frac{(\mathbf{I} - \mathbf{v}_{k+1}\mathbf{v}_{k+1}^\top)}{\|\mathbf{M}\mathbf{v}_k\|} \frac{\partial L}{\partial \mathbf{v}_{k+1}} \right)_{n\times m} \mathbf{G}_2 \odot \\ &\odot \mathbf{H}_1^\top \big(\mathbf{v}_k\big)_{n\times m} \mathbf{H}_2 \end{aligned} \tag{4.5}$$

$$\frac{\partial L}{\partial \mathbf{M}_p} = \sum_k \frac{(\mathbf{I} - \mathbf{v}_{k+1}\mathbf{v}_{k+1}^\top)}{\|\mathbf{M}\mathbf{v}_k\|} \frac{\partial L}{\partial \mathbf{v}_{k+1}} \odot \mathbf{v}_k \tag{4.6}$$

With careful ordering of operations, the complexities for the backward pass are now $O(\max(m^2 q, n^2 p))$ and $\Theta(pq)$.

4.4 Experimental results

We present some qualitative matching examples for 2 state-of-the-art matching datasets in fig. 4-2 and in fig. 4-3.

Figure 4-2: Four qualitative examples of our best performing network *GMNwVGG-T*, on the CUB-200-2011 test-set. Images with a black contour represent the source, whereas images with a red contour represent targets. Color-coded correspondences are found by our method. The green framed images show ground-truth correspondences. The colors of the drawn circular markers uniquely identify 15 semantic keypoints.

Figure 4-3: Twelve qualitative examples of our best performing network *GMNwVGG-T* on the PASCAL VOC test-set. For every pair of examples, the left shows the source image and the right the target. Colors identify the computed assignments between points. The method finds matches even under extreme appearance and pose changes.

Bibliography

[1] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proceedings of Computer Vision and Pattern Recognition*, 2005.

[2] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *Proceedings of Computer Vision and Pattern Recognition*, 2009.

[3] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, 2011.

[4] O. Duchenne, F. Bach, I. Kweon, and J. Ponce. A tensor-based algorithm for high-order graph matching. In *Proceedings of Computer Vision and Pattern Recognition*, 2009.

[5] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3), 1981.

[6] Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Training deep networks with structured layers by matrix backpropagation. *arXiv preprint arXiv:1509.07838*, 2015.

[7] M. Leordeanu, A. Zanfir, and C. Sminchisescu. Semi-supervised learning and optimization for hypergraph matching. In *ICCV*, 2011.

[8] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011.

[9] Julian Quiroga, Thomas Brox, Frédéric Devernay, and James Crowley. Dense semi-rigid scene flow estimation from rgbd images. In *Computer Vision–ECCV 2014*, pages 567–582. Springer, 2014.

[10] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] D. Sun, S. Roth, and M.J. Black. Secrets of optical flow estimation and their principles. In *Proceedings of Computer Vision and Pattern Recognition*, 2010.

[13] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic huber-l1 optical flow. In *British Machine Vision Conference*, 2009.

[14] L. Xu, Jiaya Jia, and Yasuyuki Matsushita. Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(9), 2012.

[15] Feng Zhou and Fernando De la Torre. Factorized graph matching. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 127–134. IEEE, 2012.