Multitemporal classification without new labels: a solution with optimal transport

Devis Tuia
University of Zurich
Switzerland
devis.tuia@geo.uzh.ch

Rémi Flamary
Lagrange, CNRS, UNS, OCA
France
remi.flamary@unice.fr

Alain Rakotomamonjy
Université de Rouen
France
alain.rakoto@insa-rouen.fr

Nicolas Courty
Université de Bretagne du Sud
France
courty@univ.ubs.fr

***Abstract*—Re-using models trained on a specific image acquisition to classify landcover in another image is no easy task. Illumination effects, specific angular configurations, abrupt and simple seasonal changes make that the spectra observed, even though representing the same kind of surface, drift in a way that prevents a non-adapted model to perform well. In this paper we propose a relative normalization technique to perform domain adaptation, i.e. to make the data distribution in the images more similar before classification. We study optimal transport as a way to match the image-specific distributions and propose two regularization schemes, one supervised and one semi-supervised, to obtain more robust and semantic matchings. Code is available at http://remi.flamary.com/soft/soft-transp.html. Experiments a challenging triplet of WorldView2 images, comparing three neighborhoods of the city of Zurich at different time instants confirm the effectiveness of the proposed method, that can perform adaptation in these non-coregistered and very different urban case studies.**

I. Introduction

Providing labeled information for each remote sensing image acquired is not an option for efficient multitemporal processing. Especially now that satellites have daily revisit periods, the amount of images acquired surpasses the annotation capacity of human operators. Even though approaches to annotate less but better exist [1], the need of labeled examples in each image is a requirement that prevents remote sensing automatic classification to meet the users' expectations.

If no labels are available for the newly acquired image, one must make the best use of the labeled information that is available for other similar scenes. The idea might seem tempting, but re-using labeled information as such generally leads to catastrophic results: from one acquisition to the other, the spectra are distorted by either the acquisition parameters, the atmospheric conditions or by the differences in scale/appearance of the objects in different geographical regions [2]. The compensation for such distortions, or *shifts* is one of the lively areas of machine learning, *domain adaptation* [3], [4].

One natural way to perform adaptation is to bring the data distributions in the source and target domain closer, such that the model trained with the (possibly adapted) source samples can be used directly to predict class memberships in the adapted target domain. To increase similarity between domains, one can resort to global methods based on projections such as KPCA [5] or TCA [6] or adapting the distribution locally, for example exploiting graph matching techniques [7].

In this paper we tackle the problem of adapting remote sensing data distributions as the minimization of a *transportation* cost. Optimal transportation (OT) was first introduced in the 19th century as the way to find a minimal effort solution to the transport of masses of dirt into a series of holes. By making the parallel between masses and probability distributions, OT has recently gained interest in the machine learning community and has been used to tackle domain adaptation problems of label propagation in graphs [8] or classification of visual categories [9].

We propose to adapt distributions between couples of remote sensing images with regularized optimal transport: we apply two forms of regularizations, namely an entropy-based regularization [10] and a class-based regularization [9] to a series of classification problems involving very high resolution images acquired by the WorldView2 satellite. We study the effect of the two regularizers on the quality of the transport.

Section II introduces the concepts of optimal transport and the regularizers considered. Section III presents the data and the setup of the experiments, whose results are discussed in Section IV. Section V concludes the paper.

II. Optimal transport for domain adaptation

Consider two data sources (domains): first is a source domain, X_s, that is labeled. X_s is composed of a set of examples $\mathbf{X}_s = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{n_s}$, where $\mathbf{x}_i^s \in \mathbb{R}^d$ are the pixel values in each image band (possibly also expanded with spatial filters) and y_i is the pixel label that can take one of C discrete values, each one corresponding to a land cover class c. Second is a target domain, X_t, that is unlabeled and composed of a set of pixels $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t} \in \mathbb{R}^d$. We want to estimate class memberships in X_t by using models trained with the labeled set $\mathbf{X}_s$. We assume that the two domains are different, i.e. a model trained on X_s cannot predict well on X_t without a phase of adaptation and that this difference can be at least partially compensated by a data transformation. In this paper, we chose to explore a class of transformations that minimizes a transportation cost from one domain to another. This type of transformation is related to the notion of *optimal transport*.

978-1-4673-7119-3/15/$31.00 ©2015 IEEE

A. Discrete optimal transport

Let $\mathcal{P}(X_s)$ and $\mathcal{P}(X_t)$ be the set of probability measures over the two domains. In the discrete case, the empirical distributions of the probability measures related to $\mathbf{X}_s$ and $\mathbf{X}_t$ are

$$\mu_s = \sum_{i=1}^{n_s} w_i^s \delta_{\mathbf{x}_i^s}, \quad \mu_t = \sum_{i=1}^{n_t} w_i^t \delta_{\mathbf{x}_i^t} \tag{1}$$

where $\delta_{\mathbf{x}_i}$ is the Dirac at location $\mathbf{x}_i$. w_i^s and w_i^t are probability masses associated to the i-th sample, and belong to the probability simplex, *i.e.* $\sum_{i=1}^{n_s} w_i^s = \sum_{i=1}^{n_t} w_i^t = 1$. They are typically computed by running a density estimation, or set as uniform weights $w_i^s = 1/n_s$ and $w_i^t = 1/n_t$ for each sample in either domain.

We are looking for a transformation $\mathbf{T} : X_s \rightarrow X_t$ that matches the source and the target domains. To be efficient for domain adaptation, this transformation must preserve the label information, i.e. it must preserve the conditional distribution $P_t(y|\mathbf{x}^t) = P_s(y|\mathbf{T}(\mathbf{x}^s))$. To do so, we restrict only the transformations that match the marginal distributions defined in (1), i.e. transformations, for which the marginal distribution of the source samples becomes the marginal observed in the target domain. Such transformations can be obtained from a probabilistic coupling γ between $\mathcal{P}(X_s)$ and $\mathcal{P}(X_t)$. In the discrete case the set of probabilistic couplings between those two distributions is the transportation polytope $\mathcal{P}$ defined as

$$\mathcal{P} \quad = \quad \left\{ \boldsymbol{\gamma} \in (\mathbb{R}^+)^{n_s \times n_t} | \, \boldsymbol{\gamma} \mathbf{1}_{n_t} = \mu_s, \boldsymbol{\gamma}^T \mathbf{1}_{n_s} = \mu_t \right\} \tag{2}$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. Among all possible couplings $\gamma \in \mathcal{P}$, optimal transport chooses the one that minimizes the transportation cost from the source distribution to the target one. This cost is generally linked with a metric of the embedding space. The Kantorovitch formulation of the optimal transport [11] then reads:

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\arg\min} \, \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F \tag{3}$$

where $\mathbf{C} \geq 0$ is the cost function matrix of term $C(i,j)$ related to the energy needed to move a probability mass from $\mathbf{x}_i^s$ to $\mathbf{x}_j^t$. This cost can be, as in this study, the Euclidian distance between the two locations, i.e. $C(i,j) = ||\mathbf{x}_i^s - \mathbf{x}_j^t||^2$. Once the optimal transport matrix $\boldsymbol{\gamma}_0$ has been found, we can transform the source elements $\mathbf{X}_s$ in a target domain-dependent version $\hat{\mathbf{X}}_s$:

$$\hat{\mathbf{X}}_s = \text{diag}((\boldsymbol{\gamma}_0 \mathbf{1}_{n_t})^{-1}) \boldsymbol{\gamma}_0 \mathbf{X}_t. \tag{4}$$

In this equation, the transported source points are expressed as barycenters of samples from $\mathbf{X}_t$ with weights found in the coupling matrix $\boldsymbol{\gamma}_0$. Using the mapped labeled samples, $\hat{\mathbf{X}}_s$, a model can be trained and used to predict the class membership of samples $\mathbf{X}_t$.

B. Regularized optimal transport

The OT solution can be obtained using efficient linear programming solvers, but it is also prone to overfitting due to small sample sizes or outliers. Regularization can be used to prevent such unwanted behaviours. In the following, we test two forms of regularization:

- **OT-Sink**: an entropy-regularized transportation [10], preventing too sparse solutions by regularizing the entropy of the transportation matrix, $h(\boldsymbol{\gamma})$:

$$\boldsymbol{\gamma}_0^\lambda = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\arg\min} \, \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F - \frac{1}{\lambda} h(\boldsymbol{\gamma}), \tag{5}$$

where $h(\boldsymbol{\gamma}) = -\sum_{i,j} \gamma_{i,j} \log(\gamma_{i,j})$. Such regularization allows more connections to remain during transport, since it favours transportation matrices γ with high entropy (i.e. many non-zero elements).

- **OT-labreg**: a class-regularized transportation ł[9] that forces samples of the same label in the source domain to remain close during transportation:

$$\boldsymbol{\gamma}_0^{\lambda_c} = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\arg\min} \, \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F - \frac{1}{\lambda} h(\boldsymbol{\gamma}) + \eta \sum_j \sum_c ||\boldsymbol{\gamma}(\mathcal{I}_c, j)||_q^p, \tag{6}$$

where $\mathcal{I}_c$ contains the index of the lines of the source elements of the c-th class and $|| \cdot ||_q^p$ denotes the ℓ_q norm to the power of p, which promotes group-sparsity, i.e. that the coefficients of each column of $\boldsymbol{\gamma}$ are active for all the samples of a single class, and 0 everywhere else. This regularizer tends to associate each unlabeled target example to source examples of the same class. No constraint is applied on the samples in the target domain, as it is supposed to be unlabeled.

Note that other forms of regularization have also been proposed in recent literature: for example, authors in [12] proposed a regularization based on the graph Laplacian preserving local relationships during the transport.

III. Data and setup of experiments

A. Data

We consider domain adaptation between three very high resolution images acquired by WorldView2 in 2010 and 2011. The images show three neighborhoods of the city of Lausanne, Switzerland. Images and acquisition dates are shown in Fig. 1. Two of them (`Ma`, `Mo`) are part of a same acquisition (2010), but depict neighborhood with different spatial structures. The third (`Pr`) comes from the 2011 acquisition and has similar spatial structures as `Mo`. Therefore, we can qualitatively rank the difficulty of the domain adaptation problems involved as the couples (`Pr,Mo`) < (`Ma,Mo`) < (`Ma,Pr`). We consider the original 8 WorldView2 bands, plus a series of opening and closing morphological filters (with circular structuring elements of size $3, 5, 7$ pixels) as input space. We consider the following six classes: residential buildings, commercial building, roads, meadows, trees and shadows. The ground truth available is illustrated in the bottom row of Fig. 1 and the number of pixels per class is reported in Tab. I.

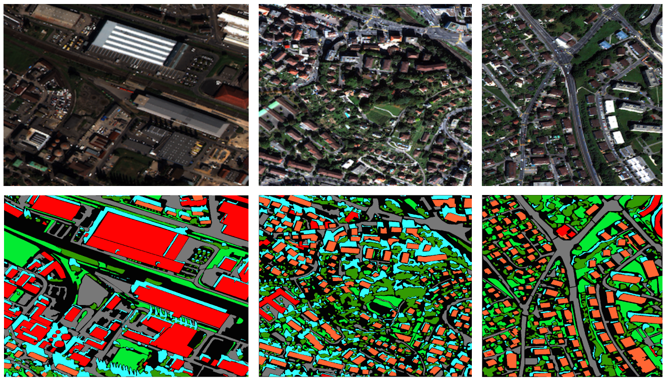
B. Experimental setup

Each image is taken in turn as the source domain and used to predict in the two others. When an image is taken as source domain, 100 pixels per class are selected for both the definition of the transport and the classification, which is obtained by

TABLE II

NUMERICAL PERFORMANCE OF THE OPTIMAL TRANSPORT METHODS, ASSESSED WITH THE AVERAGE OVERALL ACCURACY OVER FIVE REALIZATIONS (± STANDARD DEVIATION).

	Training on X^S							Training on X^T
	No adapt.	KPCA ł[5]	TCA [6]	GM [7]	**OT-Sink**	SSTCA [6]	**OT-labreg**	
# labels X^S	-	0	0	0	0	600	600	-
`Mo` → `Pr`	59.67±2.92	56.21±3.09	49.83±2.22	60.30±1.66	**66.81**±1.38	55.88±6.89	65.09±0.82	84.17±0.39
`Pr` → `Mo`	57.12±4.66	53.82±4.25	50.52±2.19	62.07±1.45	71.95±0.97	57.36±5.75	**72.37**±1.08	81.43±0.83
`Ma` → `Mo`	45.62±1.79	46.46±3.13	47.47±1.95	43.92±1.70	59.88±1.12	49.27±4.47	**70.66**±1.98	82.12±0.74
`Mo` → `Ma`	33.74±1.69	32.61±3.08	31.52±2.50	37.14±2.31	50.75±2.67	39.58±2.57	**55.41**±2.38	80.73±0.40
`Ma` → `Pr`	46.84±1.21	45.12±1.41	46.02±1.47	43.87±1.58	57.47±2.33	46.12±1.55	**66.60**±2.69	83.74±0.37
`Pr` → `Ma`	26.40±5.27	23.71±4.16	21.49±4.14	38.49±1.89	49.60±1.98	31.41±3.23	**54.90**±1.47	80.79±0.59



Malley (`Ma`)
(1124 × 1516)
29 . 09 . 2010

Montelly (`Mo`)
(1064 × 1248)
29 . 09 . 2010

Prilly (`Pr`)
(1040 × 1032)
02 . 08 . 2011

Fig. 1. The three WorldView2 images used in the experiments, along with their ground truths (class legend: commercial buildings, residential buildings, meadows, trees, roads, shadows), pixel sizes and acquisition dates.

1-NN classification. In the target domain, we extract both 600 unlabeled samples as target domain set $\mathbf{X}_t$ and a set of additional independent pixels for testing (therefore: 289792 for `Ma`, 152666 for `Pr` and 187510 for `Mo`). To compute the probability masses in the source domain, w_i^s, weighting of the examples has been adjusted to match the proportions of the classes in the target domain. Matching these proportions is of importance because of the specific form of the class-label regularizer in (6) and of the bi-stochastic nature of the transportation matrix $\boldsymbol{\gamma}$. Indeed, if the empirical prior probability of classes in the source and target domains are not equal, the regularizer will necessarily match some source and target examples of different classes (which is to be avoided). If these prior probabilities are known, ideal weights for the source examples are $w_i^s = p_t^c / n_s^c$, where p_t^c corresponds to the prior probability in the target domain of the class c to which the sample $\mathbf{x}_i$ belongs to and n_s^c is the observed number of samples of class c in the source domain. In practice, we consider an empirical estimation of p_t^c. In the following, we will also study the importance of the accuracy of such estimation.

We compare optimal transportation with several state of art methods (see Tab. II). As an upper bound on performance, we also report the results of a model trained directly in the target domain using 600 labeled pixels (100 per class).

IV. RESULTS

Table II summarizes the numerical results obtained. In all the adaptation experiments considered the proposed optimal transportation methods outperform significantly both the case without adaptation and the state of the art methods. The competing methods cannot cope with local shifts in the PDF and fail at improving classification performance in the target domain in most of the cases. The only exception seems to be graph matching (GM), that can improve the ‘no adaptation’ baseline in four out of the six cases. But even in these cases, the improvement is much smaller than in the case of optimal transport, for which we observe improvements of $7 - 23\%$ in the **OT-Sink** case and of $6 - 28\%$ in the **OT-labreg** case. Including information about the labels in the optimization of the transportation plan seems to boost the performance of the subsequent classifier, since it prevents samples from different classes to be transported into similar regions of the target space, thus easing the work of the discriminant classifier.

For all methods, we are still 10-25% less accurate than a classifier trained in the target domain directly (last column of Tab II), but still the increase with respect to the classifier without adaptation and most state of the art methods is striking. Moreover, the problems considered are particularly hard, as we are transporting classifiers between images with different acquisition geometry and conditions (like most recent domain adaptation studies), but also not coregistered and with differences in the scaling and structural properties of the objects imaged. Facing such a challenging setup, optimal transport seems to be a first step in the good direction for efficient domain adaptation under strong spectral and structural deformation between the domains.

TABLE I

NUMBER OF LABELED PIXELS AVAILABLE FOR EACH DATASET.

Class	Color in Fig. 1	Dataset		
		Prilly	Montelly	Malley
Residential	🟧	151271	179695	24898
Meadows	🟩	148604	47865	143674
Trees	🟩	116343	177203	63956
Roads	⬛	141353	104582	294687
Shadows	🟦	39404	218189	194321
Commercial	🟥	13692	22506	437633

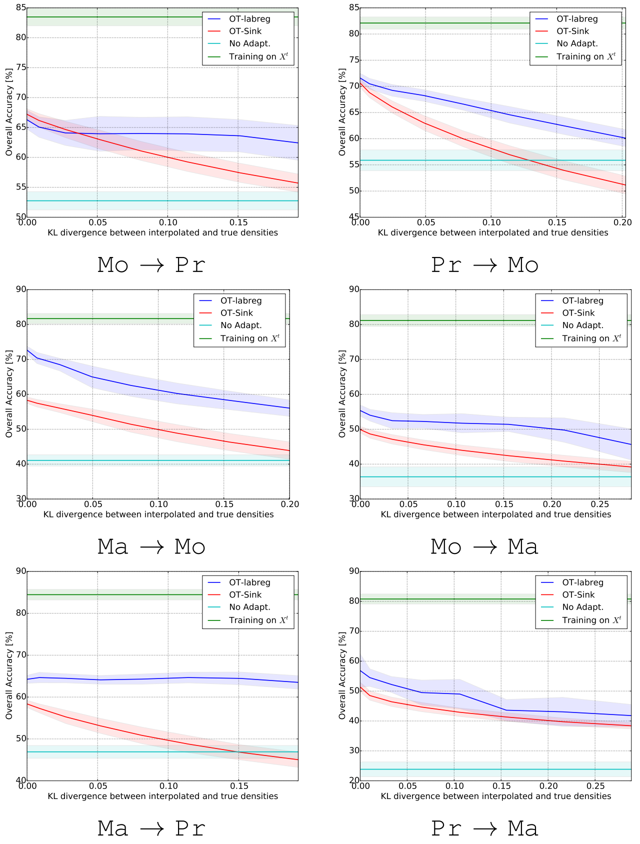


Fig. 2. Evolution of the 1-NN classification accuracy with respect to the quality of the estimation of the target class proportions. For each panel, the leftmost values correspond to the results obtained knowing the true target class proportions, while the rightmost values correspond the an uniform distribution in the target domain.

Figure 2 illustrates the influence of the choice of the weight estimation of the masses. In these graphs, we observe the evolution of overall accuracy over a path interpolated from the true target distribution (leftmost value, $w_i^s = p_t^c/n_s^c$, see Section III-B) to an uniform distribution (rightmost value, $w_i^s = 1/n_s^c$), which might seem a natural choice when no information is available beforehand. We sampled 10 paths between these two points on the classes probabilities simplex, and computed their Kullback Leibler (KL) divergence with respect to the true target distribution (x-axis in the figure, corresponding to the distance between distributions on the simplex). In the figure we report an averaged result on an interpolation along the KL path. From the results, we can observe that an estimation of the true target class distribution leads to a significant boost in the results for both regularizers. But with the exception of the Mo → Pr case, the optimal transport methods always outperform the situation with no adaptation (cyan horizontal line). **OT-labreg** also keeps a constant advantage with respect to the unsupervised **OT-Sink**: the decrease is generally similar for both methods, but **OT-labreg** shows larger variances and stronger drops in performance when approaching the uniform distribution: this is understandable, since it implies that the proportion of target classes are uniform, which is clearly not the case (see also Table I). Using a discriminative alignment as the one promoted by **OT-labreg** will force the transformation to preserve the class-specific masses observed in the source, while in reality they should be different in the target domain. Using some knowledge (at least partial) of the class distribution in the target seems to be unavoidable to get the most from the proposed class-regularized optimal transport.

V. Conclusion

We proposed to use optimal transport for the adaptation and matching of data distributions issued from multitemporal remote sensing images. Optimal transport can match distributions coming from images that are not registered and have undergone strong deformations. We also tested two types of regularization to improve the quality of the matching. The first (**OT-Sink**) allows for a transportation plan that is less abrupt (it decreases sparsity of the transport solution), while the second (**OT-labreg**) includes class-regularization and forces labeled samples in the source domain to be transported to nearby locations in the target. When applied on a challenging multidate adaptation problem, OT outperformed all state of the art methods, both in the unsupervised and semisupervised settings. We also provided a study on the impact and necessity of estimating the class proportions in the target domain, and showed that for OT to exploit all its potential, an estimate of the class proportions in the target domain is necessary. In the future we will study ways to obtain such estimation.

References

[1] M. M. Crawford, D. Tuia, and L. H. Hyang, "Active learning: Any value for classification of remotely sensed data?," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 593–608, 2013.

[2] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification," *IEEE Signal Proc. Mag.*, vol. 31, pp. 45–54, 2014.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, October 2010.

[4] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: a survey of recent advances," *IEEE Signal Proc. Mag.*, in press.

[5] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comp.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[6] S. J. Pan and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, pp. 199–210, 2011.

[7] D. Tuia, J. Muñoz-Marí, L. Gómez-Chova, and J. Malo, "Graph matching for adaptation in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 329–341, 2013.

[8] J. Solomon, R. Rustamov, G. Leonidas, and A. Butscher, "Wasserstein propagation for semi-supervised learning," in *International Conference on Machine Learning (ICML)*, 2014, pp. 306–314.

[9] N. Courty, R. Flamary, and D. Tuia, "Domain adaptation with regularized optimal transport," in *European Conference on Machine Learning (ECML)*, Nancy, France, 2014.

[10] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transportation," in *NIPS*, pp. 2292–2300. 2013.

[11] L. Kantorovich, "On the translocation of masses," *C.R. (Doklady) Acad. Sci. URSS (N.S.)*, vol. 37, pp. 199–201, 1942.

[12] S. Ferradans, N. Papadakis, J. Rabin, G. Peyré, and J.-F. Aujol, "Regularized discrete optimal transport," in *Scale Space and Variational Methods in Computer Vision, SSVM*, 2013, pp. 428–439.