TRANSDUCTIVE ATTRIBUTES FOR SHIP CATEGORY RECOGNITION

Quentin Oliveau, Hichem Sahbi

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
{quentin.oliveau, hichem.sahbi}@telecom-paristech.fr

ABSTRACT

Ship category recognition is one of the remote sensing applications that requires designing accurate image representation and classification models. Training these models is usually a data hungry process, that requires a lot of labeled data which are usually scarce and expensive. As unlabeled data are more abundant and relatively cheaper, transductive methods exploiting these data are highly preferred.
We introduce in this paper a novel representation learning approach based on transductive attributes. The strength of our method resides in its ability to upgrade labeled training sets using attributes shared across categories, and also its ability to further push the benefit of attributes by taking into account not only the labeled data but also the abundant unlabeled ones. When plugging these learned attribute representations into transductive classifiers, we obtain a substantial gain of accuracy compared to several baselines as well as the related work, on the challenging task of ship category recognition, especially when labeled training data are very scarce.

Index Terms— image classification, representation design, maritime surveillance, ships category recognition

1. INTRODUCTION

With the spread of satellite sensors (Pléiade, GeoEye, etc.) and Unmanned Aerial Vehicles (UAV), large amount of optical remote sensing data are currently available. These large data collections require new algorithmic solutions able to automatically analyze and assign visual contents to well defined object categories (a.k.a classes). In the particular context of maritime environments, traffic monitoring and fishery control require detecting and classifying ships from tiles or continuous flows of images. Whereas many algorithms dedicated to ship detection on synthetic aperture radar and optical images have been developed (see for instance [1, 2, 3, 4, 5]), ship recognition on optical images still remains understudied and known to be challenging, especially when training data are scarce and when the underlying categories are highly variable.

Classification tasks on remote sensing data are usually divided into two major families: semantic segmentation and object classification. Semantic segmentation (including land-cover mapping) consists in scoring local image primitives – such as pixels or grid cells – using probabilistic models [6, 7], while object classification is usually achieved by first extracting and pooling image features (either handcrafted or not) [8, 9] and then assigning them to categories using variety of machine learning and classification techniques such as SVMs and deep networks [4]. In particular, classification methods relying on deep features [10], have shown very competitive results, nevertheless their success highly depends on the abundance of labeled training data, especially for fine-grained categories exhibiting large intra-class variability and high inter-class resemblance. Alternative deep feature learning solutions overcome the scarcity of labeled data using deep networks pretrained on other tasks (with larger training sets) and fine-tune them on new tasks. However, the generalization power of the resulting networks is not always sufficient especially when the new classification tasks have very few labeled training data.

In this paper, we introduce a novel fine-grained ship category recognition method based on transductive attribute learning. Attributes are defined as mid-level characteristics, with the interesting property of i) being shared across different categories and ii) learned on larger (shared) training sets (see for instance [11, 12]). Coding-based techniques, highly similar to attributes, have also been investigated in the literature for object category recognition [13]; in particular methods in [14, 15, 16], closely related to attribute learning, enforce the learned representations to be discriminant. In spite of being discriminant (while being less data hungry[1]), these representation learning techniques could further be improved if one exploits the unlabeled data in training, especially when the size of the labeled sets reaches the interesting limit of one.
In the context of remote sensing ship category recognition, we propose to learn a new discriminant attribute representations while being tranductive. Transductive learning (see for instance [17, 18, 19] and its related computer vision and remote sensing applications [20, 21, 22, 23]) basically relies on the smoothness assumption, which states that neighboring data in the original feature space should produce smooth representations as well as similar classifier outputs (and vice-versa).

[1]Less data hungry means they require less labeled data for training in contrast to other methods including deep learning.

The contribution and the advantage of the proposed method is twofold: on the one hand, it fully exploits joint attributes shared among different categories in order to assemble larger labeled (training) sets and hence learn better classifiers. On the other hand, and in order to push the benefit of attributes further, the proposed method overcomes the scarcity of the labeled data (even when used with joint attributes) by including the unlabeled data as a part of the training process. As will be shown through experiments in remote sensing ship category recognition, our proposed method learns highly effective classifiers on top of the learned representations.

2. OUR FRAMEWORK

Considering $\mathcal{I} = \{I_i\}_{i=1}^{\ell+u}$ as the union of ℓ labeled training images (belonging to C different categories) and u unlabeled test images[2]. Each image I_i is described with a feature vector $\mathbf{x}_i \in \mathbb{R}^M$ and assigned to a category label $\mathbf{y}_i \in \mathcal{C} = \{1, \ldots, C\}$ according to a well defined ground-truth. Our goal is to learn a K-dimensional representation $\alpha \in \mathbb{R}^{K \times (\ell+u)}$ (on top of the image features $\{\mathbf{x}_i\}_i$) together with classifiers $\mathbf{W} \in \mathbb{R}^{C \times K}$ that predict the categories of unlabeled images in $\mathcal{I}$. The design principle of $\mathbf{W}$ and α is discriminative and regularized, i.e., it produces different representations and classifier scores for images belonging to different categories while smoothing representations and scores for images belonging to the same classes. We implement this property using the following criteria:

Fidelity: let $\mathbf{X} \in \mathbb{R}^{M \times (\ell+u)}$ be a matrix whose i-th column corresponds to the feature vector $\mathbf{x}_i$. Our fidelity criterion aims to minimize a reconstruction error between the original image features $\{\mathbf{x}_i\}_i$ and the underlying learned representations α as

$$\min_{\alpha, \mathbf{D}} \frac{1}{2} \left\| \mathbf{X} - \mathbf{D}\alpha \right\|_F^2, \tag{1}$$

here $\mathbf{D} \in \mathbb{R}^{M \times K}$ is a dictionary and F denotes the matrix Frobenius norm.

Discrimination power: we enforce the learned representations to be discriminant by constraining classifier scores of the labeled data to be consistent with their ground truth. Defining $\mathbf{Q}$ as a $C \times \ell$ matrix whose entry $\mathbf{Q}_{ij}$ set to 1 iff the j-th example belong to the i-th class, -1 otherwise, we consider the following criterion in order to learn both the representation α and the classifier $\mathbf{W}$

$$\min_{\alpha, \mathbf{W}} \frac{1}{2} \left\| \mathbf{Q} - \mathbf{W}\alpha_\ell \right\|_F^2, \tag{2}$$

here α_ℓ denotes the representation matrix restricted on the labeled data.

Regularization: in what follows, we assume that the learned representations are distributed in a manifold (denoted $\mathcal{M}$) that encloses both the labeled and the unlabeled data. In order to capture the topology of $\mathcal{M}$, we consider a regularization criterion that produces smooth representations and similar classifier scores for neighboring data. We define a similarity matrix $\mathbf{A}$ – of size $(\ell+u) \times (\ell+u)$ – whose entry $\mathbf{A}_{ij}$ set to $\exp(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 / \sigma^2)$ iff the feature vector $\mathbf{x}_j$ belongs to the $\mathcal{N}$-neighbors of $\mathbf{x}_i$ and 0 otherwise; here σ is the parameter that controls the scaling of the weights in $\mathbf{A}$[3]. In practice, we enforce $\mathbf{A}$ to be symmetric by constraining that the memberships to the $\mathcal{N}$-neighbors are reciprocal for all the labeled and the unlabeled data. Following this setting, our regularization criterion is defined as $\frac{1}{4} \sum_{i=i}^{\ell+u} \mathbf{A}_{ij} \left(\mathbf{W}\alpha_i - \mathbf{W}\alpha_j\right)^2$ and rewritten as

$$\min_{\alpha, \mathbf{W}} \frac{1}{2} tr(\mathbf{W}\alpha\mathbf{L}\alpha'\mathbf{W}), \tag{3}$$

here $tr()$, $'$ denote the matrix trace and transpose operators respectively and $\mathbf{L} = \mathbf{B} - \mathbf{A}$ with $\mathbf{B}$ being a diagonal matrix whose ith diagonal entry set to the sum of elements of the ith row of $\mathbf{A}$.

Global model: by combining equations 1, 2 and 3, the global model that balances fidelity, discrimination power and regularization is written as

$$\begin{aligned} \min_{\alpha, \mathbf{D}, \mathbf{W}} \frac{1}{2} \left\| \mathbf{X} - \mathbf{D}\alpha \right\|_F^2 + \frac{\lambda_1}{2} \left\| \mathbf{Q} - \mathbf{W}\alpha_\ell \right\|_F^2 \\ + \frac{\lambda_2}{2} tr(\mathbf{W}\alpha\mathbf{L}\alpha'\mathbf{W}), \end{aligned} \tag{4}$$

here $\lambda_1,\ \lambda_2 \geq 0$ control the impact of discrimination power and regularization respectively, and α_ℓ, again, stands for the representation associated to the labeled data.

3. OPTIMIZATION

It is clear that the global optimization problem in (4) is not convex jointly w.r.t., $\mathbf{D}$, $\mathbf{W}$ and α. We propose to solve it iteratively following an EM-like alternate optimization. Indeed, at each iteration, we fix two of the three variables $\mathbf{D}$, $\mathbf{W}$, α and we solve the problem (4) w.r.t. the other. We repeat this process till all the variables remain unchanged from one iteration to another. In what follows, we use the superscript (t) in order to show the evolution of these variables through different iterations. Note that only the initial setting of α (i.e., $\alpha^{(0)}$) needs to be known; in practice, values of $\alpha^{(0)}$ are randomly (and uniformly) sampled from $[0, 1]$.

Dictionary and classifier update: considering fixed $\alpha^{(t)}$, we update the dictionary as $\mathbf{D}^{(t+1)} \leftarrow \arg\min_{\mathbf{D}} \frac{1}{2} \left\| \mathbf{X} - \mathbf{D}\alpha^{(t)} \right\|_F^2$. This quadratic problem has a closed-form solution given by

$$\mathbf{D}^{(t+1)} = \mathbf{X}\alpha^{(t)'} \left(\alpha^{(t)}\alpha^{(t)'}\right)^{-1}. \tag{5}$$

which results by enforcing the gradient of equation 4 to vanish w.r.t. $\mathbf{D}$. Similarly, we obtain $\mathbf{W}$ as

[2] As the purpose of this work is ship category recognition, we assume that ship snapshots are already detected on large remote sensing images using an object detection algorithm [4].

[3] In practice, σ is set to 10.

$$\mathbf{W}^{(t+1)} = \lambda_1 \widetilde{\mathbf{Q}} \mathbf{Z} \alpha^{(t)'} \left(\alpha^{(t)} \widetilde{\mathbf{L}} \alpha^{(t)'} \right)^{-1}, \quad (6)$$

with $\mathbf{Z}$ being a diagonal matrix – of size $(\ell+u) \times (\ell+u)$ – whose first ℓ diagonal entries are set to 1 and the others to 0, $\widetilde{\mathbf{L}} = \lambda_1 \mathbf{Z} + \lambda_2 \mathbf{L}$ and $\widetilde{\mathbf{Q}}$ is a zero padded matrix of size $C \times (\ell+u)$ whose first ℓ columns are set to $\mathbf{Q}$.

Updating attribute representations: solving equation 4 (w.r.t. α) is not straightforward as no direct closed-form solution exists. Assuming $\mathbf{W}^{(t)}$, $\mathbf{D}^{(t)}$, and also $\alpha^{(t)}$ known and fixed at iteration (t) (denoted simply as $\mathbf{W}$, $\mathbf{D}$, α), we find the subsequent $\alpha^{(t+1)}$ following the optimality condition about the gradient of equation 4 (w.r.t. α); this leads to a fixed-point solution $\alpha^{(t+1)} = \lim_{k \to \infty} \Phi^{(k)}$ with $\Phi^{(0)} = \alpha^{(t)}$ and

$$\Phi^{(k)} = \left(\mathbf{D}'\mathbf{D} + (\lambda_1 \mathbf{Z}_{ii} + \lambda_2 \mathbf{D}_{ii}) \mathbf{W}'\mathbf{W}\right)^{-1} \cdot \left[\mathbf{D}'\mathbf{X} + \lambda_1 \mathbf{W}'\widetilde{\mathbf{Q}}\mathbf{Z} + \lambda_2 \mathbf{W}'\mathbf{W}\Phi^{(k-1)}\mathbf{A}\right]_i, \quad (7)$$

with $[.]_i$ being the i-th column of the matrix in $[.]$, and $i \in \{1, \ldots, \ell+u\}$. In practice, we iterate over k till $\left\|\Phi^{(k+1)} - \Phi^{(k)}\right\|_F^2 \rightsquigarrow 0$.

4. EXPERIMENTS

We evaluate the performance of our representation learning framework on the Ship4 dataset illustrated in figure 1. The latter includes 320 snapshots of ships extracted from aerial images of various spatial resolutions, divided into 4 different classes (80 snapshots per class): *monohull sailboat*, *inflatable boat*, *yacht* and *multihull*. We use this Ship4 dataset for ship recognition; given an unlabeled (test) data, the goal is to predict its class membership based on the largest classifier score. Performance (accuracy) is measured as the fraction of correctly classified images over the total number of images in the unlabeled (test) set.

Features: we extract features from the Ship4 dataset by

Fig. 1. Examples from the Ship4 dataset; with high intra-class variability and high inter-class resemblance.

submitting images from this set to the VGG-16 deep network [24] pretrained on ImageNet. More specifically we rescale original images in Ship4 to 224×224 pixels and we feed them to that deep network, using the first fully connected layer (a.k.a *fc6*), resulting into a feature vector of M dimensions (with $M = 4096$). This layer was chosen as it provides the highest accuracy after classification with a soft margin linear SVM. In all these experiments, we achieve training and classification on top of these deep features and we show the behavior of our model on the unlabeled (test) set w.r.t. different parameter settings (λ_1, λ_2, ℓ, etc.).

Model analysis: first, we study the impact of different parameters (λ_1, λ_2 and the number of dimensions K) on ship classification performances. Note that the setting of K is mainly related to the discrimination power of the learned attribute representations and the classifiers (see again equation 2). Indeed, one may show that overestimating K to $\max(\ell, M)$ guarantees a zero empirical error on the labeled data, however as the actual (intrinsic) dimension of the learned attribute representation α is unknown, we choose K to be sufficiently large (but $K \ll \max(\ell, M)$); in practice, K is set between 15 and 60 and its value have no major impact on the classification accuracy. We also study the impact of the two most critical parameters (λ_1, λ_2 in equation 4) that respectively control discrimination power and regularization. The underlying performances, shown in figure 2, are obtained by fixing alternately one of the two parameters while varying the other. From these results, the impact of these two parameters on the performances is clearly important and intermediate values (shown in the figure 2) provides the best tradeoff between discrimination power and smoothing.

Comparison: we compare our framework against its two variants (shown subsequently) as well as other related algorithms. These two variants correspond to two settings where *joint attribute representation* learning and *transductive* setting are used exclusively; i.e., the first setting, referred to as *Our Discri*, is purely discriminant and learns joint attribute representations+classifiers without regularization (i.e., $\lambda_2 = 0$ in equation 4) while ii) the second setting, referred to as *Our OvA*, is transductive (i.e., $\lambda_2 > 0$) but attribute representations and classifiers are learned independently using the standard "one versus all" framework. In contrast to the first setting, the second one benefits from regularization (and hence exploit the topology of the unlabeled data) but does not take advantage of a larger (joint) labeled dataset when learning attributes and classifiers through different classes.

Besides these two baselines, we also compare our method to other representative works including soft-margin linear SVMs, RBF SVMs[4], and Transductive SVM [17] as well as the LC-KSVD attribute learning algorithm. Results given in table 1 show that the features extracted from the pre-trained deep network are already discriminant (when combined with the SVM settings and without attribute learning) as they allow us to correctly classify many unlabeled test data even with few labeled training data; however, *Our* transductive attribute framework outperforms linear and RBF-based SVMs even with small labeled (training) sets, as well as its two variants (*Our Discri*, *Our OvA*) and the LC-KSVD algorithm. Finally, we also observe that our framework is globally comparable to transductive SVMs; *the interesting limit in these experi-*

[4]the scale of the RBF kernel is tested in the range $\left[10^{-7}, \ldots, 10^{7}\right]$.

ments is when the number of labeled images per class is just 1 *which clearly shows that our transductive attribute based method provides the best performances among all the other approaches.* From all these results, it is clear that combining both transductive and attribute learning, enhances the results, especially when labeled data are very scarce.

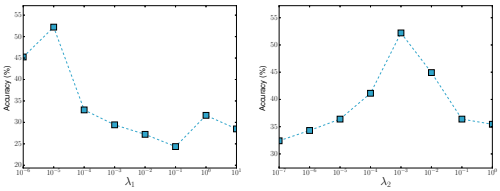

Fig. 2. Impact of λ_1 and λ_2 (in equation 4) on the performances. Again, these results are obtained by fixing alternately one of the two parameters while varying the other. These results are obtained with only one labeled sample per class and $K = 15$.

	Trans-L	J-Att-L	1	2	5	10	60
Lin. SVM	No	No	38.61	51.28	71.67	83.01	87.25
RBF SVM	No	No	29.43	43.91	74.33	82.86	91.25
TSVM[17]	Yes	No	48.41	71.15	76.33	85.00	95.00
LC-KSVD[15]	No	Yes	42.72	49.36	70.67	82.86	92.50
Our Discri	No	Yes	40.19	58.01	74.50	83.93	91.25
Our OvA	Yes	No	43.35	53.84	72.33	82.14	86.25
Our	Yes	Yes	**52.21**	58.33	75.00	84.46	92.25

Table 1. This table shows the average ship classification accuracy, w.r.t. the number of labeled images per class, for linear SVM (*Lin. SVM*), *RBF SVM* and transductive SVM (*TSVM*) as well as LC-KSVD. This table also shows the accuracy of two variants of our model (described in Section 4). The outperformance of our method is clear w.r.t. almost all these comparative techniques, especially when the number of labeled data per class is small; an interesting limit is when this number is just one. In this table "J-Att-L", "Trans-L" respectively stand for joint attribute and transductive learning.

5. CONCLUSION

We introduced in this paper a novel method for ship category recognition based on transductive attribute learning. The advantage of the proposed method is twofold: on the one hand, it fully exploits the potential of attribute learning by assembling together all the scarce labeled (training) sets across different categories, in order to learn better joint (common) attribute representations. On the other hand, the proposed method makes it possible to capture the topology of the manifold enclosing the data by considering the abundant unlabeled sets in the learning process; this results into highly discriminant attribute representations and classifiers as shown through experiments on the challenging ship category recognition task.

As a future work, we are currently investigating other assumptions about the properties of the manifold enclosing the data. For instance, one may constrain the learned attribute representations to be resilient to different geometric transformations such as rotation, in order to further improve the accuracy of our final classifiers.

6. REFERENCES

[1] Mattia Stasolla, Carlos Santamaria, Jordi J. Mallorqui, Gerard Margarit, and Nick Walker, "Automatic ship detection in sar satellite images: Performance assessment," in *IEEE Geoscience and Remote Sensing Symposium (IGARSS)*, 2015.

[2] Arnesen Tonje Nanette and Olsen Richard B, "Literature review on vessel detection," Tech. Rep., Forsvarets Forskningsinstitutt, 2004.

[3] Guang Yang, Bo Li, Shufan Ji, Feng Gao, and Qizhi Xu, "Ship detection from optical satellite images based on sea surface analysis," *IEEE GRSL*, 2014.

[4] Jiexiong Tang, Chenwei Deng, Guang-Bin Huang, and Baojun Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," *IEEE TGRS*, 2015.

[5] Feng Yang, Qizhi Xu, Feng Gao, and Lei Hu, "Ship detection from optical satellite images based on visual search mechanism," in *IEEE Geoscience and Remote Sensing Symposium (IGARSS)*, 2015.

[6] Mario Caetano, "ESA training course on land remote sensing – image classification," 2009.

[7] Thomas Martin Lillesand, Ralph W. Kiefer, and Jonathan Chipman, *Remote Sensing and Image Interpretation, 7th Edition*, John Wiley & Sons, 2015.

[8] T. Moranduzzo and F. Melgani, "A SIFT-SVM method for detecting cars in UAV images," in *IEEE Geoscience and Remote Sensing Symposium (IGARSS)*, 2012.

[9] Feng Yang, Qizhi Xu, Feng Gao, and Lei Hu, "Ship detection from optical satellite images based on visual search mechanism," in *IEEE IGARSS*, 2015.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, "Deep learning," 2016.

[11] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[12] Felix Yu, Liangliang Cao, Rogerio Feris, John Smith, and Shih-Fu Chang, "Designing category-level attributes for discriminative visual recognition," in *Proceedings of the IEEE CVPR*, Portland, OR, June 2013.

[13] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[14] Julien Mairal, Francis R. Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman, "Supervised dictionary learning," *CoRR*, vol. abs/0809.3083, 2008.

[15] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *IEEE CVPR*, 2010.

[16] Zhuolin Jiang, Zhe Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Proceedings of the 2011 IEEE Conference on CVPR*, 2011.

[17] Thorsten Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of ICPR*, 1999.

[18] Chun nam Yu, "Transductive learning of structural svms via prior knowledge constraints," in *Proceedings of the International Conference on AISTATS*, 2012.

[19] T. Joachims, "Transductive learning via spectral graph partitioning," in *Proceedings of the International Conference on Machine Learning*, 2003.

[20] Phong Vo and Hichem Sahbi, "Transductive inference and kernel design for object class segmentation," in *Proceedings of the IEEE ICIP*, September 2012.

[21] J. M. Alvarez, M. Salzmann, and N. Barnes, "Efficient transductive semantic segmentation," in *2016 IEEE WACV*, March 2016, pp. 1–9.

[22] L. Bruzzone, Mingmin Chi, and M. Marconcini, "Transductive svms for semisupervised classification of hyperspectral data," in *Proceedings of IGARSS*, 2005.

[23] Fabio N. Gttler, Dino Ienco, Pascal Poncelet, and Maguelonne Teisseire, "Combining transductive and active learning to improve object-based classification of remote sensing images," *Remote Sensing Letters*, vol. 7, no. 4, pp. 358–367, 2016.

[24] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.