IMAGE TRANSPORT REGRESSION USING MIXTURE OF EXPERTS AND DISCRETE MARKOV RANDOM FIELDS

Fabrice Michel[1,2] *and Nikos Paragios*[1,2]

(1) Laboratoire de Mathématiques Appliquées aux Systèmes, Ecole Centrale de Paris, France
(2) Equipe GALEN, INRIA Saclay - Île-de-France, Orsay, France

ABSTRACT

The registration of multi-modal images is the process of finding a transformation which maps one image to the other according to a given similarity metric. In this paper, we introduce a novel approach for metric learning, aiming to address highly non functional correspondences through the integration of statistical regression and multi-label classification. We developed a position-invariant method that models the variations of intensities through the use of linear combinations of kernels that are able to handle intensity shifts. Such transport functions are considered as the singleton potentials of a Markov Random Field (MRF) where pair-wise connections encode smoothness as well as prior knowledge through a local neighborhood system. We use recent advances in the field of discrete optimization towards recovering the lowest potential of the designed cost function. Promising results on real data demonstrate the potentials of our approach.

Index Terms— Image registration, Machine Learning, Kernel Methods, Mixture of Experts, Discrete MRF

1. INTRODUCTION

Registration [1] is one of the most critical problems in medical imaging. It involves three components, a deformation model, a similarity metric and an optimization procedure. The task of comparing two images might seem trivial when referring to the same modality (Sum of Absolute Differences -SAD-, Sum of Square Differences -SSD- [2]), but it becomes quite challenging when seeking comparisons between different modalities. In order to address that challenge, complex statistical measurements can be considered like mutual information [3], KL-divergence [4] - with quite limited applicability range - or simulation techniques. In the latter case, using the underlying physical characteristics of the acquisition process, one aims to transport one modality to another.

The above mentioned methods are based on a single pixel joint probability model, thus a random permutation of the pixels in one of the images yields the same similarity [5]. As a result, the applicability of these methods is quite limited when considering the spectrum of multi-modal signals, like for example computed tomography to ultrasound. In order to cope with such a challenging context, machine learning techniques were recently proposed [6] towards learning a similarity measure. The central idea is to use a training set with spatial correspondences between the two modalities and learn a similarity measure through a modified support vector machine. This approach involves a tedious parameter setting and lacks tolerance to fluctuations between the prior and the images that are compared. Ideally the regression process should be modular, invariant to the position of the subject, the acquisition conditions and the nature of images to be registered; that is the objective of the paper.

Given a training set of registered examples, first we learn individual intensity-driven regression models between the population of the source and the target modality. Such regression models involve patch-based support and subspace projections using different metrics. Towards improving the performance of the individual regression model, we introduce the idea of dependencies between regression models through a mixture model [7]. Once regression models have been learned, towards improving individual prediction results, we use their output as image terms in a Markov Random Field (MRF) formulation [8], where the unknowns are the regression models per pixel. The data support comes from the conditional density of a regression model given the observation. It is combined with a smoothness term applied to the regressed imaged and a prior constraint that models spatial co-dependencies between regression models at different scales. Efficient linear programming [9] is used to determine the lowest potential of the cost function. We present the registration results obtained with our metric on a challenging data-set composed of various types of MRI images from 9 different patients. We compare our results with results we obtained by using mutual information (MI) as a metric.

2. TRANSPORT REGRESSION

Let us consider without loss of generality that a set of registered image pairs is available. The images come from two different modalities A and B. Let I_A and I_B denote the image produced by both modalities respectively. The aim of transport regression is to determine an operator f (linear or non-linear), that when applied to the source modality, will produce a new image that is comparable with the target modality.

$$f(I_A) = \widehat{I_B} \approx I_B$$

We can then consider comparing images $\widehat{I_B}$ and I_B using common uni-modality techniques. Our problem is to infer the similarity of an image acquired with modality A as if it had been issued by modality B.

We aim at finding correspondences between intensities in the source and the target image. Let Ω be the spatial domain of all images. In the remainder of this paper, we let $S(x)$ (for source) and $T(x)$ (for target) denote the input and output of the mixture, with x in Ω. The aim of transport regression is to define an operator f such that:

$$\forall x \in \Omega, \quad f\left(S\left(x\right)\right) = T(x) \quad \begin{cases} S(x) \in I_A \\ T(x) \in I_B \end{cases} \tag{1}$$

a model that does not take into account the spatial position of the observation. Such a model is compact but also ill-posed. The same origin intensity (different spatial positions) could be mapped to numerous different intensities in the target space, or

$$S(x_0) = S(x_1) \text{ and } T(x_0) \neq T(x_1) \tag{2}$$

978-1-4244-4126-6/10/$25.00 ©2010 IEEE

These situations cannot be modelled by a unique function of the input space, we refer to these situations as non-functionality. To cope with such non-functionalities we adopt a two-component approach. First we augment the information space on which the transport function is defined using a patch around the position x in the input of the function f. This augmentation of the information on each pixel drastically reduces the occurrences of the situation earlier described. Here a trade-off has to be found: the bigger the patch, the less ambiguity we will encounter, but in the same time the less general the learned function will be. When ambiguity is still found, we are going to use the neighborhood information to drive our choice.

3. MIXTURE OF EXPERTS

The mixture of experts [7] is particularly suited for problems with non-functionality. The basic idea is to infer a conditional distribution using a mixture of regression models that are local conditional probabilities. The overall conditional probability is obtained by smoothly mixing local conditional distributions.

We model the conditional probability of a target intensity t given a source patch s : $p(t|s)$. To model this probability we first build a training data-set $\mathcal{D}$ composed of all the source patches extracted in the source images and all the corresponding target intensities, or

$$\mathcal{D} = \{(s = \mathcal{P}(S(x)), t = T(x))\,|x \in \Omega\} \tag{3}$$

Where $\mathcal{P}(\cdot)$ is the patch extraction operator. The estimation of the density $p(t|s)$ is done in two complementary steps. The first step initializes the algorithm with a partitioning of $\mathcal{D}$. We are later going to train each expert of the mixture on one partition. The second performs an iterative process that updates the parameters of the conditional probability.

The partitioning is carried out through the clustering of $\mathcal{D}$. $\mathcal{D}$ is the product space of a patch space and an intensity space. A naive clustering of $\mathcal{D}$ would be biased by the patch space due to its high dimensionality. We have therefore considered to map the patch space of $\mathcal{D}$ to a single intensity space by extracting the central value of each patch. The initial partitioning is hence performed on a joint intensity space $\mathcal{D}'$ isomorphic to $\mathcal{D}$. The clustering of $\mathcal{D}'$ is carried out using a common Gaussian mixture model inferred by the Expectation-Maximization algorithm (EM), and results in K clusters. The partitioning obtained on $\mathcal{D}'$ is then applied on $\mathcal{D}$.

As a second step, each expert is trained on a partition of $\mathcal{D}$, there are as many experts as partitions. The experts are regression models. The mixture of experts consists in a sum of regression models weighted by *gating functions*. The gating functions $G_k(s)$ assess the locality on the source patch space, while the locality on the target intensity space is governed by a parameter on the regression model. These terms are essential since we use the output of all the experts to obtain the probability, the localities are used to propagate the confidence we have in each of the experts.

Here we assume that each model is a linear regression, and thus the k^{th} model tries to map the input s to the output t up to the precision of a Gaussian noise ϵ_k of zero mean and σ_k variance. The parameters β_k and σ_k will be inferred.

$$\forall k \in K, \quad s = \sum_i \beta_k^i \left[\phi(s)\right]^i + \epsilon_k \tag{4}$$

where $\left[\phi(s)\right]^i$ is the i^th coordinate of the patch $\phi(s)$. We will discuss the importance of ϕ later. Equation (4) in terms of conditional probability, translates to:

$$\forall k \in K, \quad \mathcal{N}(\sum_i \beta_k^i \left[\phi(s)\right]^i, \sigma_k) \tag{5}$$

where $\mathcal{N}$ denotes a Gaussian distribution. The parameter σ_k is the locality term for the regression model, indeed it governs the bandwidth over which the model acts. Finally the conditional probability writes:

$$p(t|s) = \sum_{k=1}^{K} G_k(s) \mathcal{N}(\sum_i \beta_k^i \left[\phi(s)\right]^i, \sigma_k) \tag{6}$$

The inference on the parameters of $G_k(s)$ (not developed here) and of the regression model, is done with EM, good approximations to the update formulas were introduced in [10]. Note that the learning of the β_k^i is done on the complete data-set and not on the cluster directly. Instead the EM procedure gives a weighting parameter to filter out the samples that do not comply within the regression model. The computation of this weight could be considered as a refinement of the initial clustering.

The function ϕ in the regression model can be used to render the regression robust to intensity shifts and noise, it is the core of the flexibility in this algorithm. Even though the results presented are made with $\phi(x) = x$, a very wide variety of functions could be used. Designing a function ϕ robust to intensity shifts and noise yields:

$$|\phi(x + \textit{shift} + \eta) - \phi(x)| \leq \epsilon(\eta)$$

where *shift* represents a constant shift in intensity and η a random noise. ϵ is a sufficiently small value driven by the intensity of η.

The transport of the image is created by sequentially selecting a patch s in the source image using the profile $p(t|s)$ that is a function of a 1D-variable with s fixed. Taking for any s the argument of the maximum $t^\star$ of $p(t|s)$ provides a good approximation of the image.

4. MRF-BASED IMAGE REGRESSION WITH MIXTURES OF EXPERTS

Yet in some cases, ambiguities arise, where the the maximum of the density fails to map to the expected density. Since the conditional distribution is made from a mixture, it has several modes which sometimes show maxima with values close to one another. Then taking the argument of the maximum of the distribution might not be optimal, and we have to define which local maximum to choose in those situations. One can overcome this limitation through the use of local constraints with respect to the reconstructed images. We consider a constraint based on the natural assumption of smoothness in the transported image.

Let us consider a set of local maxima positions that can be applied to an observation s, or a set of labels $\mathcal{L}_s = \{l_1^s, \cdots, l_{K_s}\}^s$ corresponding to the maximal number of local maxima on conditional probability learned in the previous section towards defining the transport function. The problem of generating the target image from the source can be viewed as a labelling problem. Let $l(\mathbf{x})$ be the label being associated to a local maximum at pixel $\mathbf{x}$ and $t(l(\mathbf{x}))$ be the value of t when $p(t|s)$ reaches a local maximum labelled $l(\mathbf{x})$. Then, to allow for choosing the best maximum for each pixel in the image, we want to minimize a cost directly linked to the conditional distribution:

$$V(l(\mathbf{x})) = -\log p(t(l(x))|s) \tag{7}$$

To take into account the regularity of the target image we design a cost :

$$V(l(\mathbf{x}), l(\mathbf{y})) = |t(l(\mathbf{x})) - t(l(\mathbf{y}))| \tag{8}$$

Together these costs allow us to place the problem into the framework of Markov Random Field for discrete domains, by minimizing an objective function E :

$$E(l) = \sum_{\mathbf{x} \in \Omega} V(l(\mathbf{x})) + \gamma \sum_{x,y \in \Omega} V(l(\mathbf{x}), l(\mathbf{y})) \tag{9}$$

It has to be noted that the pairwise potential features only indirectly the labels. Graph cuts [11] and message passing [12] are the most

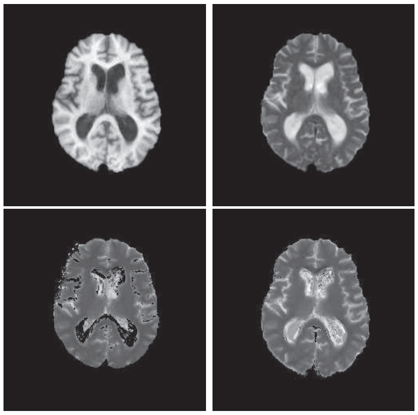

Fig. 1. Impact of the MRF on image smoothing and density maxima selection, here in the case of T1-MRI to T2-MRI regression.Top: Input image and expected result (unknown to the algorithm). Bottom: the output image before MRF regularization and after regularization

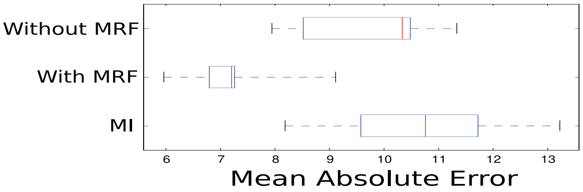

Fig. 2. Boxplot of the mean absolute difference of images following the protocol given in section 5. This registration study has been carried out with the regression from T1-MRI to T2-MRI on 7 patients, each of the lines in the diagram sums up 70 experiments. The MRF smoothing and maximum selection greatly improves the results one would obtain without its adjunction. Yet, the results without the MRF smoothing perform slightly better than the mutual information.

common methods to solve such a problem. Graph cuts are computationally efficient but can only deal with a constrained set of MRFs while the Backward-Forward Products (BFP), despite certain guarantees on the optimality of the obtained solution, are too slow. In the context of our approach, it is important to make a reasonable compromise between computational complexity and optimality of the obtained solution. The use of relaxation techniques [9], linear programming and duality seems to be a very a prominent direction. In Fig.1 we show the impact of the MRF smoothing in the cases of a regression from a T1-MRI to a T2-MRI, and a regression from a PD-MRI to a T1-MRI. It can be seen that the non-regularized version fails at finding the right intensities in the ventricles regions of the brain, the regularized version is much closer to the expected image. In Fig.2 we show a study of registration performances, on 7 patients, regression from T1-MRI to T2-MRI performances with the help of the MRF smoothing or not, have been compared. It is clear that the registration results are greatly improved by this formulation of the problem.

5. RESULTS

In order to evaluate the performance of the method, we have considered perfectly co-registered triplets of MRI images of the brain (T1, T2 and Proton Density -PD-) of 9 patients for a total of 27 volumes. All patients volumes have first been rigidly aligned to the first patient image in order to focus on deformable registration. The first two patients volumes have been used for the training data-set in 3 combinations: T1-MRI and T2-MRI, PD and T1-MRI, T1-MRI and PD-MRI. Using the training images, the transport regression model was determined. The rest of the images have been used for testing the deformable registration. For each combination (e.g. T1-MRI and T2-MRI) the same procedure has been carried out:

- Find the transformation that aligns one patient's image under one modality with the other patient's image taken under some other modality (e.g. patient 3 under T1-MRI with patient 4 under T2-MRI).
- Apply the transformation to the same patient's image, using this time the same modality as the target patient (e.g. apply transformation to patient 3 under T2-MRI).
- Compute the mean absolute error between the deformed image and the target, both under the same modality.

We have compared the performance of a simple SSD metric between the regressed image and the target, with the ones of all known statistical metrics considered in [1] [1]. This method uses discrete optimization towards the deformation of a regular grid of control points over the image. The incrementally finer resolution of the grid allows the algorithm to avoid local minima of the metric in a multi-scale approach.

For the sake of simplicity, we only show results from the conventional metrics for the Mutual Information (MI), since this method out-performs the others in this case. In order to evaluate quantitatively the performance, we show a box-plot featuring the results on the 3 chosen combinations: T1-MRI to T2-MRI (Fig.3), PD-MRI to T1-MRI registrations (Fig.4) and T1-MRI to PD-MRI (Fig.5), for 2 similarity measures: MI and our regression similarity measure (Reg. Sim.). To assess the effectiveness of our approach, we compare it to the performance in the uni-modal case (SSD). Since the regularization coefficient of the deformation is usually difficult to set, we used a range of 10 values for this coefficient. With the 7 test patients, this implies that each line of the box-plot stands for 70 experiments. These results show the clear superiority of our measure over MI based techniques and shows near uni-modal performances in some cases. The results shown are very encouraging and state the superiority of a metric that embeds a prior knowledge on image neighborhoods.

6. DISCUSSION

In this paper, we have proposed a novel approach to context-driven metric definition for registration, using machine learning techniques and MRFs. Our method is based on the concept of mixture of experts for transport functions, and the use of efficient linear programming towards recovering globally optimal solutions that encode both optimal regression and local smoothness.

The use of more complex regression models that are naturally invariant to geometric and photometric transformations is a natural extension of our method. To this end, methods that do perform ranking towards invariant patch representations will be considered. Furthermore, the integration of more complex priors learned from the data with respect to the co-dependencies between regression models could greatly improve the results. Such an approach will result on a more appropriate definition of the pairwise terms of the MRFs. Last,

[1] We have used the authors implementation from: http://www.mrf-registration.net

but not least we will focus on the application of this method in a real-time setting, like the case of virtual bronchoscope where ones seeks correspondences between pre-operative CT annotated data and live video coming from the bronchoscope.

7. ACKNOWLEDGEMENTS

Experimental images in this work were kindly provided by Prof *Christos Davatzikos*. The efficient MRF solver used for image smoothing can be found on Dr. *Nikos Komodakis* web-page[2]. The work on mixture of experts for regression has been inspired by numerous talks with Dr. *Martin de la Gorce*.

8. REFERENCES

[1] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios, "Dense image registration through mrfs and efficient linear programming," *MedIA*, vol. 12, no. 6, pp. 731–741, 2008.

[2] B. Zitová and J. Flusser, "Image registration methods: a survey," *ImaVis*, vol. 21, pp. 977–1000, 2003.

[3] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE-TMI*, vol. 16, no. 2, pp. 187–198, 1997.

[4] A. C. S. Chung, W. M. Wells III, A. Norbash, and W.E.L. Grimson, "Multi-modal image registration by minimizing kullback-leibler distance," in *MICCAI*. 2002, pp. 525–532, Springer.

[5] D. Rueckert, MJ Clarkson, DLG Hill, and DJ Hawkes, "Non-rigid registration using higher-order mutual information," in *Proceedings of SPIE*, 2000, vol. 3979, p. 438.

[6] D. Lee, M. Hofmann, F. Steinke, Y. Altun, N.D. Cahill, and B. Schlkopf, "Learning the similarity measure for multi-modal 3d image registration," in *CVPR*, 2009.

[7] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neur. Comp.*, vol. 3, no. 1, pp. 79–87, 1991.

[8] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE T-PAMI*, vol. 6, pp. 721–741, 1984.

[9] N. Komodakis, G. Tziritas, and N. Paragios, "Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies," *CVIU*, vol. 112, no. 1, pp. 14–29, 2008.

[10] Lei Xu, M. I. Jordan, and G. E. Hinton, "An alternative model for mixtures of experts," in *NIPS*, 1995, vol. 7, pp. 633–640.

[11] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE T-PAMI*, vol. 11, pp. 1222–1239, 2001.

[12] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE T-PAMI*, vol. 28, no. 10, pp. 1568, 2006.

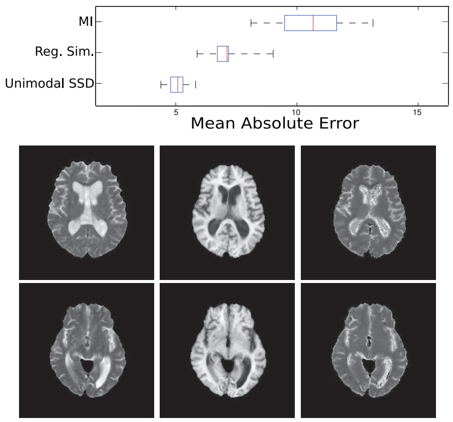

Fig. 3. Top row: boxplot of the Mean absolute differences for different metrics. Middle row: From left to right: target T2-MRI and source T1-MRI images for the registration and regressed image. Bottom row: same as Middle row with a different slice

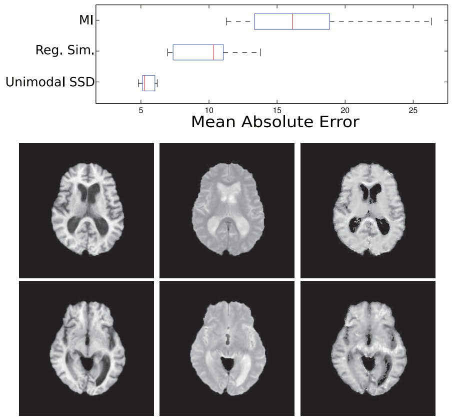

Fig. 4. same as in Fig.3 with the PD-MRI to T1-MRI registration

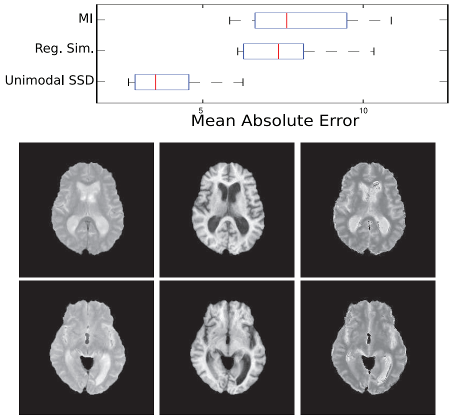

Fig. 5. same as in Fig.3 with the T1-MRI to PD-MRI registration

[2] http://www.csd.uoc.gr/ komod/FastPD/index.html