Constrained Energy Minimization for Matching-Based Image Recognition

Tobias Gass, Phillippe Dreuw and Hermann Ney
RWTH Aachen University
Human Language Technology and Pattern Recognition
<lastname>@cs.rwth-aachen.de

Abstract

We propose to use energy minimization in MRFs for matching-based image recognition tasks. To this end, the Tree-Reweighted Message Passing algorithm is modified by geometric constraints and efficiently used by exploiting the guaranteed monotonicity of the lower bound within a nearest-neighbor based classification framework. The constraints allow for a speedup linear to the dimensionality of the reference image, and the lower bound allows to optimally prune the nearest-neighbor search without loosing accuracy, effectively allowing to increase the number of optimization iterations without an effect on runtime. We evaluate our approach on well-known OCR and face recognition tasks and on the latter outperform current state-of-the-art.

1 Introduction

For many image recognition tasks, modeling intra-class variability is a big obstacle to achieving good recognition accuracy. This especially holds in face recognition, where varying facial expressions greatly change the appearance of the image. Two approaches exist to cope with this kind of variability: First, features which are invariant to a number of local deformations such as rotation or scale can be extracted, e.g. SIFT [**?**]. As often not only local but also global variability can be found in images, e.g. due to registration errors, feature-based recognition becomes hard. A second approach to allow for strong intra-class variability is to directly incorporate domain knowledge in a distance function, which can be used in nearest-neighbor (NN) based recognition setups.

In this work, we follow the second approach by modelling a distance function which is designed to cope with arbitrary non-linear deformations: Given a query image and a reference image, we search for a deformation of the reference image so that the deformed reference becomes as similar as possible to the query.

In general, finding the optimal deformations for a two-dimensional warping (2DW) problem is NP-complete [5] due to the two-dimensional first-order dependencies. Thus, lots of effort has been put into finding feasible solutions. One technique is to *relax* the first order dependencies which results in pseudo-2D [7] or zero-order models [4]. These relaxations work very well on problems with comparably small variability, e.g. optical character recognition (OCR). Another approach is to approximate the first-order problem for example by simulated annealing or beam search. Since finding the optimal deformation can be formulated as a Markov-Random-Field (MRF) energy minimization problem, we propose to use and extend one of the more recent approaches, specifically Sequential Tree-Reweighted Message Passing (TRW-S) [6, 10] which has been shown to work very well on tasks such as stereo vision and face recognition[1] and has some desirable properties.

The remainder of this paper is structured as follows. We first give a short introduction on the 2DW and TRW-S, and propose two extensions of TRW-S greatly reducing computational costs. In the last two sections, we experimentally evaluate our algorithm, compare it to the state-of-the-art, and give an analysis and conclusions.

2 2D-Warping by Energy Minimization

In order do obtain maximum flexibility in the warping, we define the 2D warping problem as a pixel labeling problem where a complete labelling $\{w_{ij}\}$ assigns a position label w_{ij} to each pixel position ij of a query image Q. A labelling then defines an energy $E(Q, R, \{w_{ij}\})$ of Q to a reference image R and we are interested in finding the optimal labeling which gives the lowest energy $\hat{E}(Q, R) = \min_{\{w_{ij}\}} E(Q, R, \{w_{ij}\})$ and E is defined as follows.

$$E(Q, R, \{w_{ij}\}) = \sum_{ij} [\theta_{ij}(w_{ij}) + \sum_{n \in \mathcal{N}(ij)} \theta_{n,ij}(w_n, w_{ij})] \quad (1)$$

The unary data term $\theta_{ij}(w_{ij})$ is a local distance $||Q_{ij} - R_{w_{ij}}||_p$, $\mathcal{N}(ij)$ is the local neighborhood of ij and $\theta_{n,ij}$ is a pairwise interaction potential. With this formulation, finding the optimal warping corresponds to performing MAP inference on Eq. (1). Similar to [10, 1], we use TRW-S in order to find an approxima-

tion of $\hat{E}$, but instead of modelling discrete displacements we propose to model the full warping by directly using positions in the reference image as labels. Therefore, we do not separate horizontal and vertical displacements, which leads to a tighter bound.

2.1 TRW-S

Recently, TRW-S [6] has been proposed as general MRF energy minimization technique which is applicable to large problems w.r.t. nodes and label. It is an iterative algorithm which approximates the lower bound B of $\hat{E}$, which is a dual of the LP relaxation of Eq. (1). The method extends regular tree-reweighted message passing by using sequential updates which guarantee a monotonic increase of B and works on subproblems which form monotonic chains. TRW computes min-marginals $\Phi_{ij}(w_{ij})$, which are forced to be equal among subproblems, and performs re-parameterization by passing messages between neighboring nodes. Exploiting the structure of the subproblems, these computations can be efficiently combined in TRW-S.

2.2 Constrained TRW-S

In each iteration of TRW-S, a forward and backward pass over all nodes w.r.t. a total order is performed, updating forward respective backward messages between neighboring pixels. Marginals are computed along the way, but we will not go into details here because we do not change this procedure. Instead, we exemplary consider the horizontal forward message $M^{\text{fw}}_{(ij),(i+1,j)}$, which passes L values from ij to $i+1,j$, where L is the number of labels and therefore in our case the dimensionality of the reference image. Eq. (2) shows the update of the forward message from ij to $i+1,j$ w.r.t. the label $w_{i+1,j}$ and consists of minimizing over all labels w_{ij} of the sum of the corresponding min-marginal, the local interaction potential, and the respective backward message. Since this has to be computed for all labels $w_{i+1,j}$, each message update has a complexity of L^2.

$$\hat{M}^{\text{fw}}_{(ij),(i+1,j)}(w_{i+1,j}) = \min_{w_{ij}} \big\{\theta_{(ij),(i+1,j)}(w_{ij}, w_{i+1,j}) - M^{\text{bw}}_{(ij),(i+1,j)}(w_{ij}) + \Phi_{ij}(w_{ij})\big\} \quad (2)$$

It has been shown in [11] that specific constraints are necessary in order to retain the image structure in the global labeling. Namely, monotonicity constraints inhibit backward steps and thus mirroring parts of the image, and continuity constraints ensure that no large gaps appear in the labeling. These constraints can be modelled using the interaction potentials as follows. Since position constraints differ w.r.t. the relative position of neighboring pixels, we use separate interaction terms for horizontal and vertical neighborhood relations. We give the definition of the horizontal interaction term θ_h w.r.t. the constrains proposed in [11], the vertical term can be derived analogously. Assume the labels are in vectorial form with vertical and horizontal components and $\Delta_w = w_{ij} - w_{i+1,j}$ is the difference vector.

$$\theta_h(w_{ij}, w_{i+1,j}) = \begin{cases} ||\Delta_w + \binom{1}{0}|| & \text{if } \binom{0}{-1} \leq \Delta_w \leq \binom{2}{+1}, \\ \infty & \text{else.} \end{cases} \quad (3)$$

This results in at most 9 allowed combinations of w_{ij} and $w_{i+1,j}$ and it can be seen that terms with $\theta = \infty$ do not influence the minimization in Eq. (2) . Therefore, we propose to pre-calculate sets S_h, S_v for each label, where $S(w_{ij}) = \{w'_{ij} | \theta(w_{ij}, w'_{ij}) < \infty\}$. Then, minimization in Eq. (2) can be performed on this pre-calculated sets which have at most 9 elements, which reduces the complexity of each message update to $9 \cdot L$. Since message updates are the main bottleneck of TRW-S, this reduces the complexity of the algorithm by almost L, which easily gets large with growing dimensionality of the reference image as depicted in Fig. 1. Note that the calculation of sets S_h, S_v can be done during the initialisation of CTRW-S and therefore has no impact on minimization runtime despite few additional table lookups.

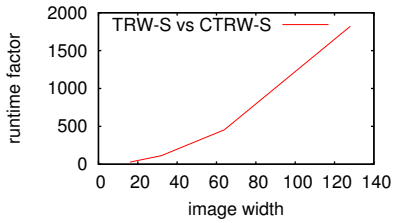

Figure 1: Relative runtime of TRW-S compared to CTRW-S with increasing width of (square) images.

In [10, 1], the authors also propose to use TRW-S for non-rigid deformations of images, but with a few key differences. Most importantly, they model relative restricted displacements, while we propose to allow full image-to-image warping. They decompose the full optimisation problem into interdependent problems of horizontal and vertical displacements, which leads to a less tight lower bound. Furthermore, they deform the images only block-wise and use different constraints on relative displacements.

2.3 Lower Bound Pruning for NN-search

As the lower bound (LB) of TRW-S is guaranteed to increase in each iteration [6], it is possible to exploit the lower bound for efficient NN-search. Consider the decision rule of the TRW-S based NN-search: $\hat{r}(Q) : Q \rightarrow C(\arg\min_R \hat{E}(Q, R))$, where $C(R)$ returns the class of R. The minimization is performed on all reference images sequentially and the class of the most

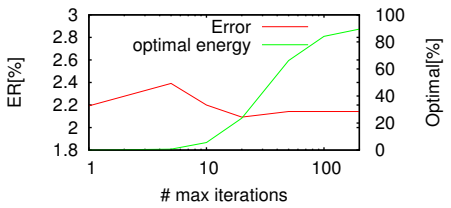

Figure 2: Error rate and percent of ϵ-optimal energy values found on the USPS database using CTRW-S with increasing number of iterations

similar reference image is returned. Using the information of the lowest distance found so far and the special properties of the lower bound given by TRW-S, it is possible to stop the energy optimization as soon as the current lower bound becomes larger than the lowest energy found so far. During NN-search for a query image Q, let σ denote the energy to the current nearest neighbor $\hat{R}$. Let $B_i(Q, R')$ denote the lower bound at iteration i of the calculation of $\hat{E}(Q, R')$ for all following reference images R'. Then, without loss of generality if $B_i(Q, R') > \sigma$, then $\hat{E}(Q, R') > \sigma$ and we can stop iterating.

Note that this LB pruning does not change the decision, but instead allows to optimize $\hat{E}$ to nearly arbitrary precision without significantly affecting overall runtime. LB pruning has the largest impact if a low energy can be found early during the NN search. This can be promoted by pre-ordering the reference observations w.r.t. a simple distance measure, e.g. Euclidean distance. The approach can easily be extended to k-NN search by keeping track of the k best distances so far.

3 Experimental Evaluation

Here, we shortly introduce the used databases and give experimental verification and analysis of the proposed approaches.

USPS. The USPS database of handwritten digits consists of 7291 images used for training and 2007 images used for evaluation. All images contain 16x16 pixels with normalized gray values. The database is known to be a hard task because of high variability of the test set, where good error rates range from 2% to 3%.

AR-Face. The AR-Face database [8] is widely used for experimental evaluation of face recognition algorithms. We follow the approach of [2] and use 7 training images and 7 test images for each of the 110 individuals in the database. Faces are detected using the Viola-Jones face detector from OpenCV, and automatically cropped and downscaled to 64x64 normalized gray values. Due to the face detection, large registration errors have to be compensated.

Labeled Faces in the Wild (LFW, cropped version) Images of different persons are crawled from the web and then paired two allow for a face-verification task of the LFW database [3]. Here, we use a recently published modified version of the LFW-database, namely the LFW-cropped database [9] where faces are cropped at the same position for all images and scaled to 64x64 pixels, discarding irrelevant image content which may accidentally influence recognition performance. The database is divided in a development set, consisting of 2200 training and 1000 test image pairs, and an evaluation set which consists of 10 folds with 5400 training and 600 test image pairs respectively. We follow the *image-restricted* approach in our evaluation, which does not take advantage of the information that the set of individuals is disjunct in development and evaluation.

3.1 Recognition Setup

For the experiments performed on the USPS and AR-Face database, we use a simple k-NN decision rule (with k=3 for USPS and k=1 for AR-Face). For the experiments on the LFW-cropped database, we calculate distances between all pairs of images and then determine a threshold on the training data, which is then used to classify pairs of images as same or not same. Feature-wise, we start our investigation of warping algorithms using simple Sobel gradients, which have been proven to work very well for image warping in OCR. For the face recognition tasks, we extract an upright SIFT (U-SIFT) descriptor [2] which is reduced by a PCA matrix and then normalized to unit length. Furthermore, we evaluate the use of different local context sizes[4]. In all experiments using message passing, we stop optimizing E if it becomes ϵ-optimal w.r.t. to the lower bound. Formally, we stop iterating when $|E(Q, R) - B_i(Q, R)| \leq \epsilon \cdot B_i(Q, R)$ with $\epsilon = 1e - 5$ or if a maximum number of iterations is exceeded.

3.2 Results

We start the experimental evaluation by showing how the maximum number of iterations allowed affects the performance both w.r.t. classification error and w.r.t. the percentage of ϵ-optimal optimizations. Fig. 2 shows both values for maximum number of iterations between 1 and 200. It can be seen that the percentage of ϵ-optimal optimizations approaches 100%, while the error rate is only slightly affected with a minimum at 20 iteration, which we will use for all following experiments. Note that for k-NN experiments, runtime is not affected by the increase in iterations.

An overview of the results and a comparison to competing methods is shown in Table 1. For all tasks, we give results for the Image Distortion Model (IDM) and the Pseudo-2D-HMM (P2DHMM) which are well-

Table 1: Results on the three databases using different deformation approaches and state-of-the-art methods.

Model	ER[%] on test data				
	USPS		AR-Face		LFW-crop.
	1x1	3x3	1x1	3x3	3x3
IDM	3.1	2.5	4.5	**3.7**	31
P2DHMM	2.5	2.7	4.2	4.5	31
TRW-S	3.8	4.3	-	-	-
CTRW-S	**2.1**	2.3	**4.0**	**3.7**	**28.1**
P2DHMD[4]		**2.1**			
Matching[2]			4.1		
MRH [9]					29.2

evaluated matching algorithms with relaxed dependencies [4]. On the USPS task, CTRW-S outperforms all competing methods using no additional local context and performs on-par with the best single method so far, the P2DHMD which is a P2DHMM with additional zero order distortions in the columns. Note that except TRW-S, all methods are far more specific in the imposed constraints and thus stronger fit to the task. Special note should also be taken of the results for regular TRW-S, which is not only slower but also too unrestricted in the possible deformations, leading to a much worse recognition performance regardless of the used context. On the AR-Face database, CTRW-S performs on par with the best other approaches. Interestingly, even the zero-order models like the IDM and the SIFT-Matching in [2] are able to achieve very good accuracy, which may be explained by the comparably small variance in the task where close matches between descriptors can be found. This changes on the LFW-cropped database, where image variability is very high and thus retaining the geometric structure of the images helps achieving good accuracy. Here, C-TRWS outperforms the current state-of-the-art on this task[1]. In general, it can be concluded that for OCR, CTRW-S is a out-of-the-box approach which achieves state-of-the-art performance with very little parameter or feature tuning. For face recognition, it can be seen that on simpler tasks, using a complex descriptor already leads to very good results while in a less controlled setting, the full geometric dependencies implemented in CTRW-S outperform all other approaches.

4 Conclusions

In this work, we proposed a general approach to image matching using MAP-inference on MRFs. To this end, we directly implemented deformation constraints into TRW-S, greatly reducing computational complexity. Additionally, we proposed to exploit the properties of the lower bound given by TRW-S in order to optimally prune NN-searches. The experimental evaluation showed very good performance on both OCR and face recognition tasks, emphasizing the generalization capabilities of our approach.

Acknowledgements. The authors thank Thomas Deselaers for for his helpful suggestions and discussions. This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

[1]Note that [9] also report a slightly higher accuracy, which has been obtained by applying the *unrestricted* setting.

References

[1] S. Arashloo and J. Kittler. Hierarchical image matching for pose-invariant face recognition. In *BMVC*, 2009.

[2] P. Dreuw, P. Steingrube, H. Hanselmann, and H. Ney. Surf-face: Face recognition under viewpoint consistency constraints. In *BMVC*, Sept. 2009.

[3] G. B. Huang, M. Mattar, T. Berg, and E. Learned Miller. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille France, 2008.

[4] D. Keysers, T. Deselaers, C. Gollan, and H. Ney. Deformation models for image recognition. *IEEE TPAMI*, pages 1422–1435, 2007.

[5] D. Keysers and W. Unger. Elastic image matching is NP-complete. *Pattern Recognition Letters*, 24(1-3):445–453, 2003.

[6] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE TPAMI*, 28:1568–1583, 2006.

[7] S. S. Kuo and O. E. Agazzi. Keyword spotting in poorly printed documents using pseudo 2-d hidden markov models. *IEEE TPAMI*, 16(8):842–848, 1994.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Feb. 2004.

[9] A. Martinez and R. Benavente. The AR face database. Technical report, CVC Technical report, 1998.

[10] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *ICB*, pages 199–208. Springer-Verlag, 2009.

[11] A. Shekhovtsov, I. Kovtun, and V. Hlaváč. Efficient mrf deformation model for non-rigid image matching. *Comput. Vis. Image Underst.*, 112(1):91–99, 2008.

[12] S. Uchida and H. Sakoe. A monotonic and continuous two-dimensional warping based on dynamic programming. In *ICPR*, volume 1, pages 521–524, Aug. 1998.