Warp that Smile on your Face: Optimal and Smooth Deformations for Face Recognition

Tobias Gass[1,3], Leonid Pishchulin[2,3], Philippe Dreuw[3] and Hermann Ney[3]

[1] Computer Vision Laboratory, ETH Zurich
gasst@vision.ee.ethz.ch

[2]Computer Vision and Multimodal Computing, MPI Informatics, Saarbruecken
leonid@mpi-inf.mpg.de

[3] Human Language Technology and Pattern Recognition Group, RWTH Aachen University
<last name>@cs.rwth-aachen.de

***Abstract*— In this work, we present novel warping algorithms for full 2D pixel-grid deformations for face recognition.**

Due to high variation in face appearance, face recognition is considered a very difficult task, especially if only a single reference image, for example a mug-shot, per face is available.

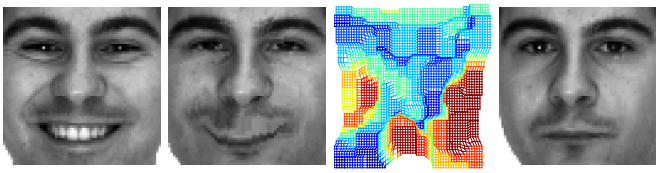
Usually model-based approaches with additional training data are used to cope with several types of variation occurring in facial imaging. Image warping contrarily yields a distance measure which is invariant with regard to several types of variation. This allows for precise recognition even using only very few reference observations. Due to the computationally complex problem of optimal 2D warping, pseudo-2D warping-based approaches in the past represented strong approximations of the original problem, and were mainly successful on data with low variability or rectified images.

We propose a novel 2D warping method which is globally optimal and makes no prior assumtions on the data variability besides two-dimensional smoothness constraints which both avoid local mirroring and gaps and significantly speed up the optimization. Furthermore, we show that occlusion handling is imperative to obtain smooth warpings in a variety of domains.

We evaluate our novel algorithm on various well known databases, such as the AR-Face and CMU-PIE database, and provide a detailed comparison to existing warping approaches. We show that by using simple relative 2D constraints, strong local features and a kernel, which is robust w.r.t. occlusions, our computationally complex approaches outperform state-of-the-art results for recognizing faces under varying expressions, occlusions and poses. Most interestingly, we achieve higher accuracy using fewer training instances per class compared to methods learning a model of the 3D shape.

I. Introduction

Recognising faces is a hard task due to the image variability both stemming from the natural variability of the face which results from temporal changes, different expressions or occlusions, and from global variations in illumination or pose (c.f. Fig. 2). Furthermore, in many domains as for example records of criminals, only a limited number of images per individual are given as references. In the most extreme case, only one frontal *mug-shot* is available, making it necessary to acquire additional external data if one aims at training a model which captures the natural variability. Since this is not always possible, we propose to use two-dimensional warping (2DW) algorithms in order to obtain a deformation-invariant dissimilarity measure which can be used for nearest-neighbour classification. In contrast to related to feature matching approaches [10, 27, 5, 35, 13, 7], our approach uses dense information and incorporates structural dependencies. This leads to smooth deformations, which can handle crucial appearance changes resulting from different poses and expressions (c.f. Fig. 1) and is also robust with regard to occlusions.



(a) test (b) (c) (d) reference

Fig. 1. Query image (a), mug-shot reference image (d), and deformed reference image (b) using our smooth global deformation algorithm. (c) shows the deformation grid, where dark blue areas correspond to small absolute deformations and dark red corresponds to strong absolute deformations.

Face recognition techniques can be divided into two groups by discriminating between methods using global or local information. Concerning local information, lots of research has been performed in finding an appropriate feature descriptor which is a priori invariant to certain transformations [20, 1, 36, 3] or can be learned from suitable training data [27, 14, 30]. On the other hand, it is possible to train global variability models using parametric shape models, such as elastic graph bunch matching [34], active shape [6] and active appearance models [8], or by imposing domain knowledge to infer 3D models from 2D images. For the latter, [37] generate virtual pose images which can be used as additional reference images and [25] learn pose variability from data and marginalise over poses.

Recently, increased research has been performed w.r.t. *image warping* methods, which stem from dynamic time warping in speech recognition. In 2D, approximations like pseudo-two-dimensional warping (P2DW) have been used for face recognition [9]. They can be calculated efficiently using decoupled hidden Markov models (HMMs), but can not find good warpings in the presence of strong non-linear variations (and can not even cope with rotations). Therefore, stronger warping methods have been proposed [2, 19, 24, 31, 11].

We will give an overview of the different approaches along with a motivation for our approach in Sec. II. Then, we recapitulate two successful alignment models which we enhance by directly implementing geometric constraints. In

Sec. IV, we directly implement geometric constraints in a tree-based warping method, leading to a novel efficient warping algorithm. Finally, we discuss a few practical aspects w.r.t. computational efficiency before evaluating our model and compare it thoroughly with other image warping methods and competing approaches.

II. Problem Statement

Aligning facial images is crucial for a large number of face recognition algorithms. Several methods for face detection and cropping [33], eye detection and subsequent rectification have been proposed. Using these methods, basic global transformations can be coped with, but it was shown that in addition strong features must be used in matching algorithms [10]. Matching algorithms try to find a matching local descriptor in the reference image for each item in a sparse or dense set of local descriptors representing the test image. In order to disallow arbitrary matchings, matches are usually searched for only in a relative local neighbourhood, which has to be defined in advance, as for example in [13]. Including relative spatial information leads to so called *warping* algorithms, where matches of local descriptors are not only found through local similarity, but also through pairwise smoothness which capture geometrically similar matchings of neighbouring pixels. Due to the loopy nature of the resulting optimisation problem, it is NP-complete to find a global optimum [16]. However, several methods exist for example using interior trust regions [24] or data driven iterations [31]. Additionally, maximum aposteriori inference (MAP) in Markov random fields (MRF) is receiving increased attention [19, 2, 11]. Here, efficient optimisation algorithms like sequential tree-reweighted message passing (TRW-S) [18] allow for finding good optima with a huge number of labels.

In order to obtain robust algorithms with suitable runtimes, approximations and heuristics are introduced by hierarchical approaches [2], splitting of horizontal and vertical displacements which leads to less tight lower bounds, or absolute restrictions of the displacement [23] which implicitly assumes prior knowledge on the variability of the data. Also, it is unclear in what magnitude using an approximation of the global optimum for a given energy function influences the classification performance. In this work, we therefore investigate in and contribute to the following topics:

- We directly compare globally optimal and approximative warping algorithms
- We present a novel warping algorithm which is globally optimal and uses structural constraints
- We evaluate several warping approaches on data with strong variations such as poses and occlusions
- We show that for warping in highly variable domains occlusions must and can be handled efficiently

We show that the direct implementation of geometry-preserving constraints in both relaxations of the dependencies, represented by tree-serial dynamic programming (TSDP) [23], and approximative alignment models, such as tree-reweighted sequential message passing (TRW-S) [18], allows to efficiently calculate alignments which lead to very good recognition results. Very similar to our approach is the work of [2], who use tree-based energy minimisation [18] for pose-invariant face recognition with local binary patterns (LBP), although using a model with a less tight lower bound, less strict constraints and no occlusion handling.

(a) test (b) reference

Fig. 2. (a) Local and global face variability caused by changes in facial expressions and partial occlusions (top row), and changes in pose (bottow row). Self-occlusions produced by e.g. closed eyes or changes in pose are common. (b) Only one reference image per person is available

III. Two-Dimensional Warping

In this section, we give a short introduction on 2DW, which can be defined as an alignment problem where each pixel ij of a test image $X \in F^{I \times J}$ is aligned to a position $w_{ij} = (u_{ij}, v_{ij})$ located in a reference image $R \in F^{U \times V}$. Here, F is an arbitrary feature descriptor of dimension D. A complete alignment $\{w_{ij}\}$ then defines a dissimilarity or energy E as follows:

$$E(X, R, \{w_{ij}\}) = \sum_{ij} \left[d(X_{ij}, R_{w_{ij}}) + \sum_{n \in \mathcal{N}(ij)} T_{n,ij}(w_n, w_{ij}) \right] . \quad (1)$$

Here, $d(X_{ij}, R_{w_{ij}})$ is the distance at feature level and T is a smoothness function in which geometrical dependencies and constraints between neighboring pixels $\mathcal{N}$ can be implemented. Given this notation, one is interested in finding the alignment which minimises E. This optimal alignment does not change when the smoothness is only computed w.r.t. horizontal and vertical predecessors, changing Eq. (1) to

$$E(X, R, \{w_{ij}\}) = \sum_{ij} \Big[d(X_{ij}, R_{w_{ij}}) + T_h(w_{i-1,j}, w_{ij}) + T_v(w_{i,j-1}, w_{ij}) \Big], \quad (2)$$

where T_h, T_v are horizontal and vertical smoothness terms. As mentioned before, directly optimizing this criterion is NP-complete and therefore suitable approximations are necessary. In the most simple approximation the smoothness function is replaced by a term $T_\Delta(ij, w_{ij})$ which penalizes the absolute deviation of ij from w_{ij} and restricts the maximum displacement in horizontal and vertical direction to Δ. We call this approach zero-order warping (ZOW), which has been introduced e.g. as Image Distortion Model in [17].

Approximative Energy Minimisation. Markov random field (MRF) inference can be used to find a minimum of Eq. (2). Here, we shortly present the tree-reweighted message passing algorithm (TRW-S) by [18], which guarantees convergence to a (local) optimum and gives a lower bound

TABLE I. Structural constraints as presented in [32].

constraints	monotonicity	continuity
horizontal	$0 \leq u_{i,j} - u_{i-1,j} \leq 2$	$\lvert v_{i,j} - v_{i-1,j} \rvert \leq 1$
vertical	$0 \leq v_{i,j} - v_{i,j-1} \leq 2$	$\lvert u_{i,j} - u_{i,j-1} \rvert \leq 1$

on the energy which can be exploited for pruning nearest-neighbor (NN) search (c.f. Sec. VI). TRW-S iteratively approximates this lower bound, which is a dual of the LP-relaxation of Eq. (2). TRW-S extends regular tree-reweighted message passing (TRW) by using sequential updates which guarantee a monotonic increase of the lower bound and work 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 neighbouring nodes. Exploiting the structure of the subproblems, these computations can be efficiently combined in TRW-S.

A. Tree-Based Optimisation

Tree-serial dynamic programming [23] relaxes 2DW by representing the two-dimensional pixel grid as a series of individual pixel neighborhood trees. Each pixel tree i^* has its own assignment stem, but shares the horizontal branches with other trees. As each tree i^* is optimized independently from the others, the optimisation problem in Eq. (2) breaks up into a series of partial tree-like ones:

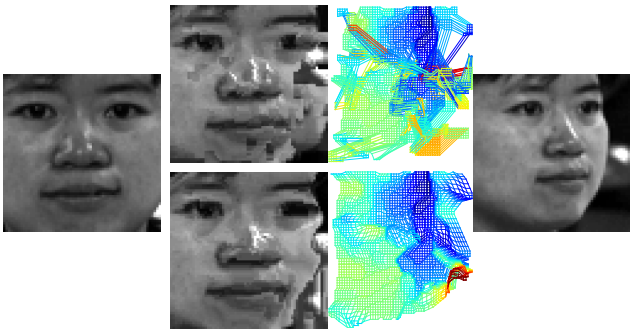
$$E_{i^*}(X, R, \{w_{ij}\}) = \sum_j \Big[\sum_i [d(X_{ij}, R_{w_{ij}}) + T_h(w_{i-1,j}, w_{ij}) + T_\Delta(ij, w_{ij})] + T_v(w_{i^*,j-1}, w_{i^*j}) \Big]. \quad (3)$$

The solution for Eq. (3) can be efficiently found by dynamic programming (DP) and the final global alignment is a composition of the alignments of the separate column stems. It should be noted that [23] impose no hard pairwise geometrical constraints in the binary smoothness function. In order to keep the complexity of the optimisation feasible, the authors penalise the absolute deviation of ij and w_{ij} by the term $T_\Delta(\cdot)$, s.t. $T_\Delta(\cdot) = 0$ iff deviation $\leq \Delta$ and $T_\Delta(\cdot) = \infty$ otherwise. Therefore, we refer to this version of TSDP as TSDP-Δ. Furthermore, simply summing up energies E_{i^*} disproportionately emphasizes the horizontal penalty terms. Hence, we use the alignments obtained by minimising Eq. (3) to compute the energy using Eq. (2).

IV. Structural Constraints

It has been shown by [32] that obeying specific hard constraints is necessary for obtaining a structure-preserving alignment. To this end, constraints ensuring the monotonicity and continuity of an alignment have been proposed which prevent large gaps and mirroring. They are conceptually similar to 0-1-2 HMMs in speech recognition, and replace the smoothness terms T_h, T_v by constrained versions T_h^c, T_v^c, which return infinity if the constraints, which are given in Tab. I, are violated.

In addition, it has been shown in [11] that using hard constraints in a warping scheme approximating the optimal solution is necessary in order to find good local optima. It is not clear, however, whether the improvements in error rates stem from the reduced search space, or whether the changed criterion itself leads to a more discriminative energy.



(a) reference (b) (c) (d) test

Fig. 3. Comparing TSDP without (top row) and with (bottom row) structural constraints on an example of the CMU-PIE poses database. Note that for poses, the alignment direction is reversed because of inaccurate cropping of the test images (c.f. Sec. VII-C). The alignment using hard structural constraints is much smoother, while the original implementation contains obvious discontinuities both in the aligned reference image and the alignment grid. The alignment has been computed using SIFT features and applied to gray images.

We propose to extend the TSDP approach by the same structural constraints. The main difference here is that TSDP computes an optimal solution for a slightly simplified criterion. This allows to better deduce the optimality of solutions derived by approximative methods. Furthermore, implementing hard geometric constraints in TSDP leads to a significant speed up and allows for much greater displacements.

We will shortly summarise the constrained TRW-S (CTRW-S) as proposed by [11], and then show how the constraints can be implemented in TSDP, leading to a more efficient formulation.

A. TRW-S with Structural Constraints

Updating messages in TRW-S involves finding a local minimum w.r.t. a pixel alignment pair (w, w'). According to [11], this minimum does not change when only pairs with $T^c_{\{h/v\}}(w, w') < \infty$ are considered, therefore these allowed pairs can be pre-computed and the minimisation can be restricted to these pairs. Since updating messages is the speed bottleneck of TRW-S with a complexity in $O((UV)^2)$, the reduction to $O(9 \cdot UV)$ due to the hard constraints given in Tab. I provides a significant speed up.

B. TSDP with Structural Constraints

In TSDP, replacing T_v, T_h with T_h^c, T_v^c allows to discard the absolute position penalty term T_Δ since the optimisation becomes much more efficient, leading to the criterion:

$$E_{i^*}(X, R, \{w_{ij}\}) = \sum_j \Big[\sum_i [d(X_{ij}, R_{w_{ij}}) + T_h^c(w_{i-1,j}, w_{ij})] + T_v^c(w_{i^*,j-1}, w_{i^*j}) \Big]. \quad (4)$$

As the optimisation breaks down into trees, DP can be used to find a global optimum. Given the hard constraints in

the smoothness terms, paths in the DP recursion containing violated constraints have an infinite cost and will be discarded at the top level. Therefore, similar to CTRW-S, we can discard any recursion containing violated constraints or conversely, just recurse through trees with allowed label combinations. DP can easily be implemented by dedicated loops, leading to a complexity in $O(IJUV)$, while the complexity of the original TSDP-Δ is in $O(IJ\Delta^4)$. The latter rapidly outgrows the former for increasing Δ, because for each alignment w_{ij} all $(2\Delta+1)^2$ possible alignments of neighboring positions have to be considered. To account for the structural constraints, we denote this version of TSDP as CTSDP. We visually compare alignments resulting from TSDP-Δ with $\Delta = 17$ and CTSDP in Fig. 3, which shows that the alignment becomes much smoother using the hard structural constraints. Here, we deform the test image (d) to best fit the reference (a) and show both the deformed test image in (b) as well as the deformed regular pixel grid in (c). It can be seen that without using structural constraints, large artifacts are visible in the deformed test image due to some very huge displacement inconsistencies which are allowed in the original TSDP formulation, although penalised. It should be noted, that in our new CTSDP warping algorithm, constraints between vertically neighbouring pixels can be violated due to the independent optimisation of the column trees. This is not very likely though, since all column trees use the same horizontal branches.

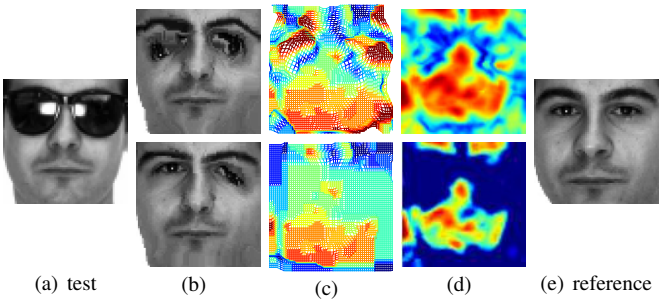
Comparing the runtime of CTSDP to CTRW-S, the optimisation of CTSDP has an equal complexity to a single forward pass of CTRW-S. Since the latter has to perform both a forward and a backward pass, and additionally multiple iterations in order to converge, the runtime of CTSDP is at least two times as fast as one single iteration of CTRW-S.

V. Occlusion Model

One big obstacle to finding a dense alignment between local features are occlusions in images. In general, occlusions can be created by something as simple as sunglasses or a scarf, but also non-visible parts of the face due to closed eyes or rotation of the head may also be present (c.f. Fig. 2). Two kinds of difficulties arise with occlusions:

1) If either the pixel at ij or w_{ij} is occluded, the local distance between this pair does not contain discriminatory information but can in fact misguide the recognition. This can be seen in Fig. 4(d), where the thresholded, blue areas are likely similar through all reference images and therefore do not contribute to the discriminative part of the distance.
2) Occluded pixels often have high distances to possible matching candidate positions and may not only enforce a bad local alignment, but also propagate it to neighboring pixels due to the structural constraints. Fig. 4(c) exemplifies this, where the warping using thresholding is much smoother in the area of the sunglasses.

It is possible to explicitly model occlusions in the warping algorithm, for instance by introducing higher-order connected graphs or by using an additional *void* or *unaligned* label. This introduces huge additional (computational) complexities and modelling difficulties. Therefore, we propose to use a local distance thresholding in order to deal with local occlusions, namely $\tilde{d}_\tau(X_{ij}, R_{w_{ij}}) = \min(\tau, d(X_{ij}, R_{w_{ij}}))$ with a threshold τ. This has several advantageous properties. On the one hand, it can be directly implemented in the local distance computation which is of minor complexity in comparison to the optimization. On the other hand, it both reduces the impact of occluded pixels on the total distance and allows the optimization algorithms to more easily align non-occluded pixels while keeping influence of the occluded pixels low. It can be seen in Fig. 4 that thresholding produces a much smoother alignment and in addition local distances are influenced much less by occluded areas, since thresholded areas should be roughly the same for all reference images. The alignment has been computed using CTRW-S, SIFT features and applied to gray images.



(a) test (b) (c) (d) (e) reference

Fig. 4. 2D warping of reference to test image without (top row) and with (bottom row) thresholding on an example from the AR-Face occlusion database. (b) is the aligned reference, (c) is the deformation grid and (d) is the local similarity map between the test and the aligned reference image. Dark blue pixels denote low similarity and red pixels denote high similarity.

VI. Practical Issues

Despite using hard-coded geometric constraints in the optimisation algorithm, the computational complexity is still high. Therefore, we shortly describe two means of accelerating the recognition.

a) Caching: In this work, we use PCA-reduced SIFT feature descriptors which will be described in more detail in Sec. VII. For speedup, we cache all pairwise distances during an initialisation phase. In addition, we extend the local distance d to include the local context of a pixel pair ij, w_{ij}. Assuming a context of 5×5, the context-size normalized local distance becomes

$$d_{5\times5}(X_{ij}, R_{w_{ij}}) = \frac{1}{25}\sum_{\Delta_x}\sum_{\Delta_y} d(X_{i+\Delta_x,j+\Delta_y}, R_{u_{ij}+\Delta_x,v_{ij}+\Delta_y}), \quad (5)$$

with Δ_x and $\Delta_y \in -2, \cdots, 2$. At image borders, the usable context and therefore the normalisation term becomes smaller. Naively replacing d with $d_{5\times5}$ in Eq. (2) leads to a huge computational overhead, since local contexts of neighboring pixels strongly overlap and local distances are computed multiple times. Therefore, we cache all local distances on the fly.

b) Pruning: As presented in [11], optimising the deformation between a query and reference image can be stopped or even completely omitted a lower bound on the energy of the current comparison surpasses the lowest energy found so far. It was shown that the lower bound of TRW-S can be exploited without loosing accuracy due to its guaranteed monotonicity. For all other optimization algorithms, a weak lower bound is the sum of the lowest local distances for all coordinates ij. This sum can be computed during the distance pre-computation and speeds up the NN search, especially if a good match can be found early.

VII. Results

In this section, we present experimental results which show that using geometric constraints alongside with the proposed occlusion model improves recognition accuracy for three major face recognition tasks. We will shortly give an overview of the used face databases and the setup used for the experiments. Then we present the results on the three different tasks, namely recognising faces from databases with mug-shot images while test images vary in occlusion, facial expression and head pose.

AR-Face. The database [21] contains frontal view face images with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). The images correspond to 126 persons: 56 women and 70 men. Each individual participated in two sessions separated by 14 days. During each session 13 pictures per person were taken under the same conditions. Similar to the work of [10], only a subset of 110 individuals for which all variability is available is used in our experiments.

CMU-PIE. The database [26] consists of over 41000 images of 68 individuals. Each person is shown under 43 different illumination conditions, 13 poses and 4 various facial expressions. We only use the subset of all persons imaged under 13 poses in neutral facial expression.

Experimental Setup. The original face images from both datasets were manually aligned by eye-center locations [12]. The images were rotated such that the eye-centre locations are in the same row and cropped to 64×64 gray valued pixels. Some sample preprocessed images from both AR-Face and CMU-PIE datasets are shown in Fig. 2 (top row) and Fig. 2 (bottom row), respectively. At each position, we extract a 128-dimensional SIFT feature descriptor [20], which is reduced to 30 dimensions using PCA as proposed by [15] which is estimated on the respective training data and subsequently normalised to unit length. Classification is performed using a NN-classifier with the deformation energy as dissimilarity measure and the L_1 norm as local feature distance. Since no separate development set is available for these databases, we estimated parameter settings for each task on the most difficult test subset. We found that for each algorithm, parameter settings generalised extremely well over all tasks, making hardly any adjustments necessary. We use our own implementations of ZOW [36], P2DW [9], TSDP [23] and CTRW-S [11] in order to have fully comparable results [1].

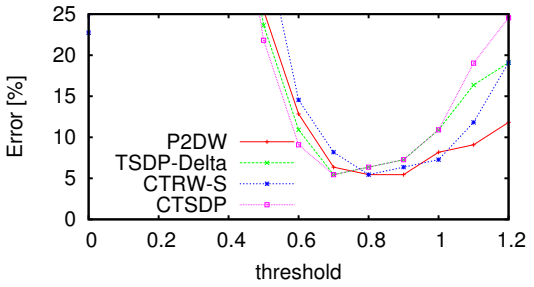


Fig. 5. Error rates on the AR-Face task, session 2, sunglasses using the original and novel formulation of TSDP and CTRW-S with different levels of distance thresholding.

A. AR-Face Occlusions

We evaluate the ability of the presented approaches to recognise partially occluded faces from the occlusion subset of the AR-Face database (c.f. Fig. 2). We use the neutral expression, non-occluded images from Session 1 as reference images and all occluded face images from Sessions 1 and 2 as test images leading to 440 test and 110 reference images.

A suitable occlusion threshold ($\tau = 0.7$) value was found on the sunglasses occlusion subset of session 2 alongside with $\Delta = 10$ for TSDP-Δ and ZOW. Results for different threshold values and algorithms are presented in Fig. 5. Interestingly, the optimal threshold value leads to a very strong decrease in error rate and is very similar for all algorithms, and therefore most probably just depends on the feature descriptor respective the distribution of the local distances. Too small thresholds lead to very high error rates since too much discriminative information is filtered out.

From Tab. II, it can be seen that warping algorithms with stricter dependencies and constraints suffer worse from occluded parts of the face. Using thresholding all algorithms despite ZOW produce nearly equally excellent recognition result with our novel warping algorithm, CTSDP, giving the best overall result. Two things should be noted:

1) CTSDP which implements structural constraints performs superior to the original version with absolute constraints despite being much more general and efficient. For TSDP-Δ, the warp range $\Delta = 10$, while CTSDP models deformations of arbitrary magnitude.
2) The surprisingly good performance of P2DW can be accounted to pre-aligned face images, in which virtually no rotations are present. Additionally, despite of the occlusions the variability is very low, allowing very good performance even for baseline warping methods.

In comparison to state-of-the-art approaches, it can be seen that our method clearly outperforms the competition. Since they are all outperformed even by the ZOW despite using zero-order-like matching algorithms themselves, it can be concluded that the SIFT descriptor in combination with occlusion modelling already provides a significant advantage.

[1] Implementations will be made available at `http://www.hltpr.rwth-aachen.de/w2d/`

TABLE II. Recognition error rates [%] on the AR-Face occlusion task using warping algorithms with and without occlusion modelling.

Model	Occlusion handling	
	no	yes
No warping	39.22	38.10
ZOW	6.79	2.46
P2DW	7.21	1.91
CTRW-S	8.27	1.69
TSDP-Δ	6.79	1.69
CTSDP	9.45	**1.48**
SURF [7]	10.54	-
DCT [10]	3.59	-
Partial Dist. [30]	*4.67	-
Stringface [4]	13.00	-
PWCM$_r$ [14]	-	16.00
LGBPHS [36]	16.00	-
SOM [29]	25	-

* used only a subset of occlusions

TABLE III. Recognition error rates [%] on AR-Face expression using the proposed warping algorithms with occlusion handling and comparison to state-of-the-art.

Model	Session 1			Session 2			Avg.
	smile	anger	scream	smile	anger	scream	
No warping	2.73	9.10	37.27	5.45	6.36	52.73	18.23
ZOW	0.00	0.00	3.64	0.91	1.82	17.27	3.94
P2DW	0.00	0.00	3.64	0.91	0.91	19.09	4.09
TSDP-Δ	0.00	0.00	3.64	0.91	1.82	17.27	3.94
CTRW-S	0.00	0.00	3.64	0.91	0.91	16.36	3.64
CTSDP	0.00	0.00	4.55	1.82	0.91	**13.64**	**3.49**
Partial Dist. [30]	0.00	3.00	7.00	12.00	14.00	37.00	12.00
Aw-SpPCA [28]	0.00	2.00	12.00	12.00	10.00	36.00	12.00
SOM [29]	0.00	2.00	12.00	12.00	10.00	36.00	12.00
SubsetModel [22]	3.00	10.00	17.00	26.00	23.00	25.00	17.00

Here, [10] uses DCT features extracted from non-overlapping blocks, LGBPHS [36] use local Gabor binary pattern histograms and [7] evaluates both SURF and SIFT features using a locally restricted matching scheme. Another two non-learning approaches, namely Stringface [4] and Partial Distance [30], employ matching procedure for recognition and hence are similar to our method: the former one is inspired by a string-based matching after representing a face as an attribute string, while the latter one uses nonmetric partial similarity measure. In opposite to our approach, the remaining two methods build a model from the training data, where SOM [29] learns a suitable self-organising map feature representation from data and PWCM$_r$ [14] learns occlusion masks in order to reconstruct invisible parts from other faces where corresponding regions are not occluded. It is worth to point out that being model-free and hence very general, our approach is able to achieve much better results, especially in comparison to the learning-based methods.

B. AR-Face Expressions

Varying expressions pose additional challenges for recognition algorithms. Not only is it necessary to correctly align the facial image, but strong non-linear deformations have to be compensated. In order to show the capabilities of the discussed warping algorithms, we use a subset of the AR-Face database containing three different expressions taken in each of the two sessions and use the neutral expression of Session 1 as reference images. A detailed quantitative analysis is presented in Tab. III. Clearly, both temporal and strong non-linear variation due to the *scream* expression poses the most difficult recognition task. All warping algorithms greatly outperform the un-aligned distance, while again the proposed extensions give small but noticeable improvements with CTSDP consistently achieving the best performance. In comparison to state-of-the-art methods the proposed extensions achieve significantly better recognition results. This is interesting since all of competitive approches except for already mentioned Partial Distance [30] method use a significantly larger amount of training data in order to learn a representative subspace via Gaussian mixtures [22] and self-organising maps [28, 30] which can not compete with the performance of the presented warping algorithms, but are probably more efficient.

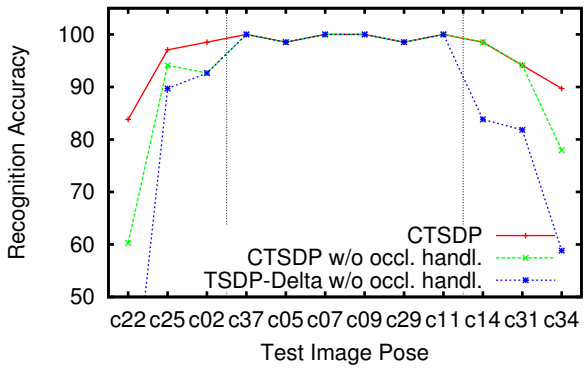


Fig. 6. Detailed plot of recognition accuracy [%] of the proposed algorithms across poses. Vertical grid lines divide between near profile (leftmost & rightmost) and near frontal (centre division) poses as used in Tab. IV. It can be seen that all algorithms perform similar on near frontal poses, while the performance an near profile poses increases with thresholding and obeyed constraints in the models.

C. Multi-Pose Recognition

One of the biggest challenges for face recognition are varying poses, which result in two-dimensional projections of three-dimensional transformations. Without considering 3D models, all variability has to be inferred in 2D space. In order to present the results of the warping algorithms, we use the CMU-PIE database where 68 subjects have been pictured from 13 viewpoints. The frontal image is used as reference image and all of the remaining 12 poses are used for testing. Two peculiarities of this database should be noted. At first, the frontal and hence reference images are much more accurately cropped, while a lot of background is visible in near profile shots. Therefore, it is best practise to reverse the alignment procedure and thus "deform" the test image to the reference image, which minimizes the impact of background pixels [2]. Second, since in profile shots only half of the face is visible, it is convenient to

TABLE IV. Average error rates [%] on CMU-PIE groups of poses of our algorithms with occlusion handling and state-of-the-art methods.

Model	near frontal	near profile	avg.
ZOW	0.49	31.61	16.05
P2DW	0.25	17.63	8.94
TSDP-Δ	0.98	25.36	13.17
CTRW-S	0.49	**6.37**	**3.43**
CTSDP	0.25	7.35	3.80
Hierarchical matching [2]	1.22	10.30	5.76
3D shape modeling [37]	0.00	**14.40	**6.55
Prob. learning [25]	* 7	* 32	19.30

Numbers with * are estimated from graphs, ** denotes missing poses and different setups

automatically generate and additionally use the left and right half of the reference face images as proposed in [2]. We use a slightly larger threshold $\tau = 1.1$ for this task because of the higher background variability. Also, Δ for ZOW and TSDP-Δ has to be increased to 17, which leads to a notable increase in complexity. For TSDP-Δ, a further increase of Δ might decrease error rate, but is infeasible because of the quadric complexity . Fig. 6 shows the results of our novel warping algorithm, CTSDP, in comparison with the original formulation and also exemplifies the impact of the occlusion handling. As can also be seen in Fig. 3, obeying strict geometric constraints leads to a large improvement in recognition accuracy shown by the superior accuracy of CTSDP in comparison to TSDP-Δ. Also, CTSDP is roughly ten times as fast as TSDP-Δ (33s vs. 350s) and also gives an 1.5 times speed increase compared to one iteration of CTRW-S (45s).

Comparing the achieved results with state-of-the-art results in Tab. IV, it is convenient to discriminate between recognition performance on the easy near-frontal poses and the much more difficult near-profile poses.

Here, it becomes clear that using geometric constraints and an occlusion model is imperative for achieving excellent recognition error rates, which is underlined by the fact that CTRW-S with occlusion handling outperforms all weaker warping algorithms as well as the hierarchical matching algorithm presented by [2], which does not use occlusion modelling. Also, they decouple horizontal and vertical displacements, which allows fast distance transforms but leads to a less tight lower bound. Zhang et al. [37] use an additional profile shot (pose 22) as reference image and generate virtual intermediate pose images using a 3D model. Since they use more training data and omit the most difficult pose in the recognition, their experiments are not entirely comparable. The method of [25] performs well despite using automatically cropped images, which is also an interesting task to be tackled with warping algorithms.

D. Discussion

The quantitative results indicate that strong neighbourhood dependencies lead to improved recognition accuracies. For an in-depth analysis, we show qualitative deformation results for faces with variability from occlusions, expressions and poses in Tab. V. For each task, we give a query image and resulting deformations to the reference image of the correct class and to a reference image of a competing class. It is clearly visible, that warping methods not using structural constraints or thresholding tend to reconstruct the query image more strictly. This obviously affects the classification performance, since it becomes harder to discriminate between the correct and competing classes. This is important, since in general the dissimilarity measure obtained by the warping algorithms is not optimised for discriminativeness, but tries to find the most similar transformation of the reference. Here it shows that a suitable geometric model as introduced by the Sakoe constraints [32] is imperative to obtain a discriminative distance measure as well as visually smooth warpings.

VIII. Conclusions

In this work, we have shown that using warping algorithms on local features provide a very general and qualitatively excellent approach for recognising faces under strong variability from just one training example. Although the computational demands are still high, the efficiency of both globally optimal as well as approximative warping methods is greatly increased by the introduction of hard geometric constraints, where our novel CTSDP warping algorithm performs on par with the CTRW-S method while being at least 1.5 times faster. In comparison with other warping approaches, it becomes clear that variability induced by projections of 3D transforms can be less reliably coped with using weaker smoothness paradigms. While results on the expressions and occlusion tasks do not vary strongly over different warping approaches, very significant differences can be observed on the CMU-PIE database. On the other hand, it became clear that, especially for the warping methods which obey the image geometry more strictly, occlusions handling is imperative and can be done efficiently using thresholding.

Acknowledgements. This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

[1] Ahonen, T., Hadid, A., and Pietikainen, M. (2004). Face recognition with local binary patterns. In *ECCV*, pages 469–481.

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

[3] Bay, H., Ess, A., Tuytelaars, T., and Gool, L. V. (2008). Surf: Speeded up robust features. *CVIU*, **110**, 346–359.

[4] Chen, W. and Gao, Y. (2010). Recognizing partially occluded faces from a single sample per class using string-based matching. In *ECCV*, volume 3, pages 496–509.

[5] Colmenarez, A. J. and Huang, T. S. (2002). Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. *IEEE TPAMI*, **24**(6), 748–763.

[6] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models—their training and application. *CVIU*, **61**(1), 38–59.

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

[8] Edwards, G. J., Cootes, T. F., and Taylor, C. J. (1998). Face recognition using active appearance models. In *ECCV*, volume 2, pages 581–595, London, UK. Springer-Verlag.

TABLE V. Qualitative Evaluation of the proposed 2D warping approaches

Task	Query	Reference	Deformed References				
			TSDP-Δ	CTSDP		CTRW-S	
			thresholding	thresholding		thresholding	
			no	no	yes	no	yes
Occlusion							
							
Pose							
							
Expression							
							

[9] Eickeler, S., Mller, S., and Rigoll, G. (1999). High performance face recognition using pseudo 2-d hidden markov models. In *ECCV*.

[10] Ekenel, H. K. and Stiefelhagen, R. (2009). Why is facial occlusion a challenging problem? In *ICB*, pages 299–308.

[11] Gass, T., Dreuw, P., and Ney, H. (2010). Constrained energy minimisation for matching-based image recognition. In *ICPR*, Istanbul, Turkey.

[12] Gross, R. (2001). `http://ralphgross.com/FaceLabels`.

[13] Hua, G. and Akbarzadeh, A. (2009). A robust elastic and partial matching metric for face recognition. In *ICCV*.

[14] Jia, H. and Martínez, A. (2008). Face recognition with occlusions in the training and testing sets. In *IEEE FG*, pages 1–6.

[15] Ke, Y. and Sukthankar, R. (2004). Pca-sift: A more distinctive representation for local image descriptors. In *IEEE CVPR*, volume 2, pages 506–513.

[16] Keysers, D. and Unger, W. (2003). Elastic image matching is np-complete. *PRL*, **24**, 445–453.

[17] Keysers, D., Deselaers, T., Gollan, C., and Ney, H. (2007). Deformation models for image recognition. *IEEE TPAMI*, **29**(8), 1422–1435.

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

[19] Liao, S. and Chung, A. C. (2010). A novel markov random field based deformable model for face recognition. In *IEEE CVPR*.

[20] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, **60**(2), 91–110.

[21] Martínez, A. and Benavente, R. (1998). The AR face database. Technical report, CVC Technical report.

[22] Martínez, A. and Zhang, Y. (2005). Subset modelling of face localization error, occlusion, and expression. In *Face Processing: Advanced Modelling and Methods*, pages 577–615.

[23] Mottl, V., Kopylov, A., Kostin, A., Yermakov, A., and Kittler, J. (2002). Elastic transformation of the image pixel grid for similarity based face identification. In *ICPR*, page 30549. IEEE Computer Society.

[24] Oxholm, G. and Nishino, K. (2010). Membrane nonrigid image registration. In *ECCV*.

[25] Sarfraz, M. S. and Hellwich, O. (2010). Probabilistic learning for fully automatic face recognition across pose. *IVC*, **28**(5), 744 – 753.

[26] Sim, T., Baker, S., and Bsat, M. (2002). The CMU pose, illumination, and expression (PIE) database. In *IEEE AFGR*.

[27] Singh, R., Vatsa, M., and Noore, A. (2009). Face recognition with disguise and single gallery images. *IVC*, **27**(3), 245–257.

[28] Tan, K. and Chen, S. (2005). Adaptively weighted sub-pattern pca for face recognition. *Neurocomputing*, **64**, 505–511.

[29] Tan, X., Chen, S., Zhou, Z., and Zhang, F. (2005). Recognizing partially occluded, expression variant faces from single training image per person with som and soft knn ensemble. In *IEEE TNN*, volume 16, pages 875–886.

[30] Tan, X., Chen, S., Zhou, Z.-H., and Liu, J. (2009). Face recognition under occlusions and variant expressions with partial similarity. *IEEE TIFS*, **4**(2), 217–230.

[31] Tian, Y. and Narasimhan, S. (2010). A globally optimal data-driven approach for image distortion estimation. In *IEEE CVPR*, pages 1277–1284.

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

[33] Viola, P. and Jones, M. (2004). Robust real-time face detection. *IJCV*, **57**(2), 137–154.

[34] Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE TPAMI*, **19**(7), 775–779.

[35] Wright, J. and Hua, G. (2009). Implicit elastic matching with random projections for pose-variant face recognition. *IEEE CVPR*, **0**, 1502–1509.

[36] Zhang, W., Shan, S., Gao, W., Chen, X., and Zhang, H. (2005). Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *ICCV*, volume 1, pages 786–791, Washington, DC, USA. IEEE Computer Society.

[37] Zhang, X., Gao, Y., and Leung, M. K. H. (2008). Recognizing rotated faces from frontal and side views: An approach toward effective use of mugshot databases. *IEEE TIFS*, **3**(4), 684–697.