The Fast and the Flexible: Extended Pseudo Two-Dimensional Warping for Face Recognition

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

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

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

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

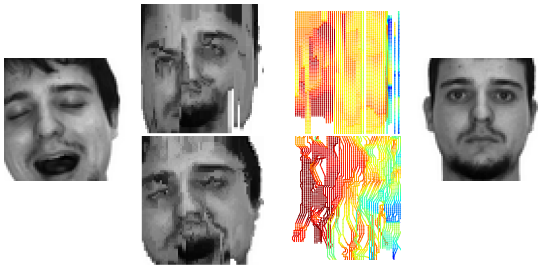
Abstract. In this work, we propose a novel extension of pseudo 2D image warping (P2DW) which allows for joint alignment and recognition of non-rectified face images. P2DW allows for optimal displacement inference in a simplified setting, but cannot cope with stronger deformations since it is restricted to column-to-column mapping. We propose to implement additional flexibility in P2DW by allowing deviations from column centers while preserving vertical structural dependencies between neighboring pixel coordinates. In order to speed up the recognition we employ hard spacial constraints on candidate alignment positions. Experiments on two well-known face datasets show that our algorithm significantly improves the recognition quality under difficult variability such as 3D rotation (poses), expressions and illuminations, and can reliably classify even automatically detected faces. We also show an improvement over state-of-the-art results while keeping computational complexity low.

1 Introduction

Fully automatic reasoning about similarity of facial images is a hard task in computer vision. Strong changes in expression and pose, as well as affine transformations stemming from automatic face detection all contribute to rich intra-class variability which is difficult to tell apart from inter-class dissimilarity.

Many methods approach the problem of intra-class variability by extracting local features from interest points or regular grids and matching them between images. The similarity is then based on the quality or the number of found matches [2, 3, 18, 22]. No geometrical dependencies between matches are considered, which makes these methods fast. However, descriptors must be chosen or trained to carry as much discriminatory information as possible which makes these methods prone to overfitting on a certain task. Even more task-specific are methods like Elastic Bunch Graph Matching [21], where faces are represented as labelled graphs, and the approach of [23] who obtain pose projections by creating 3D head models from two training images per class.

Recently, increased research focus has been put on finding geometrically smooth, dense correspondences between images, which is alleviated by the availability of relatively fast, approximative energy minimization techniques for (loopy)



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

Fig. 1: The reference image (d) is aligned to the query image (a) using P2DW (top row) and the proposed P2DW-FOSE approach (bottom row). The aligned reference image (b) shows vertical artifacts for P2DW while the proposed approach allows for much better alignment due to flexible warping; (c) shows respective warping grids

graphs [1, 5, 9]. The complexity of these approaches is high, and the impact of the approximative optimization on the classification performance remains unclear. Contrarily, relaxing the *first-order* dependencies between neighbouring pixels leads to optimally solvable problems. [14] developed a pseudo-2D hidden Markov model (P2DHMM), where column-to-column mappings are optimised independently, leading to two separate 1D alignment problems. This idea has been extended to trees [13], allowing for greater flexibility compared to P2DHMMs at the cost of great computational complexity.

In this work, we present a novel algorithm for finding dense correspondences between images. Our approach is based on the ideas of pseudo-2D warping (P2DW) motivated by [4, 10, 14]. We show that the restriction to column-to-column mapping is insufficient for recent face recognition problems and extend the formulation to allow strip-like deviations from a central column while obeying first-order smoothness constraints between vertically neighbouring pixels (c.f. Fig. 1). This leads to an efficient formulation which is experimentally shown to work very well in practise.

We will first introduce a general formulation of two-dimensional warping (2DW) before discussing P2DW and introducing our novel algorithm. Then, we will present an experimental evaluation and finally provide concluding remarks.

2 Image Warping

In this section, we briefly recapitulate the two-dimensional image warping (2DW) as described in [19]. In 2DW, an alignment of a reference image $R \in F^{U \times V}$ to a test image $X \in F^{I \times J}$ is searched so that the aligned or *warped* image $R' \in F^{I \times J}$ becomes as similar as possible to X. F is an arbitrary feature descriptor. An alignment is a pixel-to-pixel mapping $\{w_{ij}\} = \{(u_{ij}, v_{ij})\}$ for each position $(i, j) \in I \times J$ to a position $(u, v) \in U \times V$. This alignment defines a dissimilarity E as follows:

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

where $d(X_{ij}, R_{w_{ij}})$ is a distance between corresponding pixel descriptors and $T_h(\cdot)$, $T_v(\cdot)$ are horizontal and vertical smoothness functions implementing first-order dependencies between neighboring pixels. An alignment is obtained through minimization of the energy function $E(X, R, \{w_{ij}\})$. Unfortunately, finding a global minimum for such energy functions was shown to be NP-complete [7] due to cycles in the underlying graphical model representing the image lattice.

2.1 Pseudo Two-dimensional Warping (P2DW)

In order to overcome the NP-completeness of the problem, P2DW [4, 10, 14] decouples horizontal and vertical displacements of the pixels. This decoupling leads to separate one-dimensional optimization problems which can be solved efficiently and optimally. In this case, the energy function (1) is transformed as follows:

$$\begin{aligned} E(X, R, \{w_{ij}\}) &= \sum_{ij} \Big[d(X_{ij}, R_{w_{ij}}) + T_v(v_{ij}, v_{i,j-1}) + T_h(u_i, u_{i-1}) \Big] \\ &= \sum_{i} J \cdot T_h(u_i, u_{i-1}) + \sum_{ij} \Big[d(X_{ij}, R_{w_{ij}}) + T_v(v_{ij}, v_{i,j-1}) \Big] \ , \quad (2) \end{aligned}$$

where the horizontal smoothness is only preserved between entire columns by the slightly changed term T_h. Dynamic programming (DP) techniques have been used to separately find optimal alignments between column matching candidates, and then perform an additional DP optimization in order to find the globally optimal column-to-column mapping [4].

3 Extended Pseudo-2D Warping

The simplification of horizontal dependencies not only reduces complexity of P2DW, but also decreases the flexibility of the approach since all pixels in a column are forced to have the same horizontal displacement. An example of such an alignment is demonstrated in Fig. 1(b) (top row) revealing the inability of P2DW to cope with rotation. Furthermore, scan-line artifacts are clearly visible. Column-to-column mapping degrades discriminative qualities of P2DW, which can lead to an overall decrease of recognition performance. In the following we present a flexible extension of P2DW which intends to overcome the explained shortcomings with a reasonable raise of complexity.

Strip extension. In order to overcome the limitations of the column-to-column mapping in P2DW, we propose to permit horizontal deviations from the column centers. This allows for more flexible alignments of local features within a *strip* of neighbouring columns rather than within a single column. The degree of flexibility is controlled through parameter Δ restricting the maximal horizontal deviation. This parameter is task-dependent and can be adjusted in each particular case. Setting Δ to 0 results in the original P2DW, while large values of Δ allow to compensate for noticeable image misalignments.

Especially in the last case it is important to enforce structure-preserving constraints within a strip, since otherwise one facilitates matching of similar but non-corresponding local features, which degrades the discriminative power. Therefore, we propose to model horizontal deviations from column centers while retaining the first-order dependencies between alignments in a strip, which results in a **f**irst-**o**rder **s**trip **e**xtension of P2DW (P2DW-FOSE). The first-order dependencies are modeled by hard structure-preserving constraints enforcing monotonicity and continuity of the alignment. This type of constraints was introduced in [19] in order to prevent mirroring and large gaps between aligned neighbouring pixels. Formally these constraints are expressed as follows:

$$0 \leq v_{i,j} - v_{i,j-1} \leq 2, \quad \mid u_{i,j} - u_{i,j-1} \mid \leq 1 \ . \tag{3}$$

The constraints (3) can easily be implemented in the smoothness penalty function T_v by setting the penalty to infinity if the constraints are violated. In order to decrease the complexity, we hardcode the constraints in the optimization procedure, which prevents the computation of all alignments by considering only those permitted by the constraints. This helps to greatly reduce the number of possible alignments of a coordinate given the alignments of its neighbours.

Energy function. According to the explained changes, we rewrite Eq. (2) as

$$\begin{aligned} E(X, R, \{w_{ij}\}) = &\sum_{i} J \cdot T_h(u_i, u_{i-1}) \\ &+ \sum_{ij} \Big[d(X_{ij}, R_{w_{ij}}) + T_{cv}(w_{ij}, w_{i,j-1}) + T_{\Delta}(u_i, u_{i,j}) \Big] \ . \end{aligned} \tag{4}$$

Here, T_{Δ} penalizes the deviations from the central column u_i of a strip, and $T_{\Delta} = \infty$ if $|u_i - u_{i,j}| > \Delta$; T_{cv} is the smoothness term with continuity and monotonicity constraints. In comparison to P2DW, minimization of (4) is of slightly increased complexity which is linearly dependent on the choice of parameter Δ.

Absolute displacement constraints. In order to reduce the overall complexity of the proposed approach, we restrict the absolute displacement between ij and its matching candidate w_{ij} [16]. Formally these constraints are expressed as

$$0 \leq \mid i - u_{i,j} \mid \leq W, \quad \mid j - v_{i,j} \mid \leq W \ . \tag{5}$$

The warp-range parameter W can be adjusted for each task. It can be relatively small assuming pre-aligned faces, while more challenging conditions of misaligned faces require sufficiently large W. Absolute displacement constraints help to reduce the complexity from $O(IJUV\Delta)$ to $O(IJW^2\Delta)$ providing a significant speed-up even for a large W which is viewed as an accuracy/complexity trade-off.

Fig. 1(b) (bottom row) exemplifies the advantages of the proposed approach over the original P2DW. It can clearly be seen that the deviations from columns allow to compensate for local and global misalignments, while the implemented monotonicity and continuity constraints preserve the geometrical structure of the facial image. Both improvements lead to a visibly better quality of alignment.

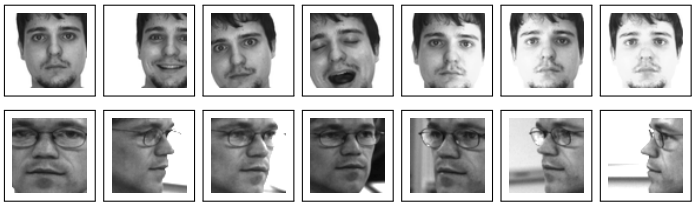


Fig. 2: Sample images from AR Face (top row) and CMU-PIE (bottom row) datasets. Faces in the top row were detected by VJ, faces in the bottom were manually aligned.

We accentuate that preserving structural constraints within a strip does not guarantee global smoothness, since strips are optimised independently. The latter can lead to intersecting paths in neighbouring columns, especially for large Δ.

4 Results

We evaluate the proposed algorithm on two challenging databases with varying expressions, illuminations, poses and strong misalignments.

AR Face. Following [3], we use a subset of 110 individuals of the AR Face [12]. We use four different expressions and three illuminations, all fully taken in two sessions two weeks apart. The first session is for training, the second for testing. Simulating a real world environment we detect and crop the faces automatically to 64x64 pixels using the Viola&Jones (VJ) detector [20]. See Fig. 2 for samples.

CMU-PIE. The CMU-PIE [17] database consists of over 41000 images of 68 individuals. Each person is imaged under 43 different illumination conditions, 13 poses and 4 various facial expressions. In order to evaluate our algorithm on 3D transformations, we use a subset of all individuals in 13 poses with neutral facial expression. The original face images were manually aligned by eye-centre locations [6] and cropped to 64×64 resolution. Fig. 2 shows sample images.

Experimental Setup. We extract an 128-dimensional SIFT [11] descriptor at each position of the regular pixel grid. As proposed by [8], we reduce the descriptor to 30 dimensions by means of PCA estimated on the respective training data and subsequently normalize each descriptor to unit length. We use a NN classifier for recognition directly employing the obtained energy as dissimilarity measure and the L_1 norm as local feature distance. Similar to [5], we include a context of 5×5 neighboring pixels in the distance, which is also thresholded with an empirically estimated threshold value of $\tau = 1$. This makes our approach robust to unalignable pixels. Additionally, we speed up the computation of the alignments using local distance caching, and track the smallest energy obtained to stop if it is surpassed by a rough lower bound on the current energy [5]. For comparison, we use our own re-implementation of P2DW [4].

Evaluation on the AR Face database. First, we show the effects of strip width on the recognition error. Fig. 3 shows the error rate for increasing Δ where the biggest improvement is seen at $\Delta = 1$. Although the error decreases further

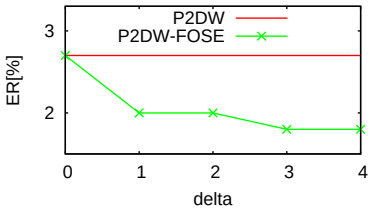


Fig. 3: Error rate on automatically detected faces for different strip widths Δ, where $\Delta = 0$ is equivalent to the P2DW.

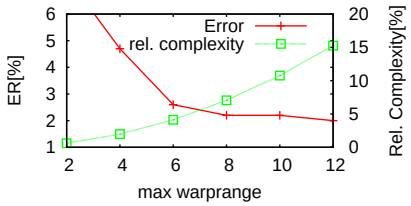


Fig. 4: Error rate on VJ detected faces and relative complexity compared to P2DW-FOSE with different warping ranges.

afterwards, the return is diminishing quickly. This gives rise to two interpretations: on the one hand, it seems most important to allow (even slight) horizontal movements of individual pixels. On the other hand, big strip widths increase the chance of intersecting column paths, making the deformation less smooth.

In order to study means of speeding up the recognition, we fix $\Delta = 3$ (c.f. Fig. 3) and vary the warp-range parameter W restricting the maximum absolute displacement. Fig. 4 shows the influence of W on both recognition accuracy and computational complexity. As the total number of possible alignments grows quadratically with increasing W, the recognition error decreases until the accuracy of the unconstrained version is reached (c.f. Fig. 3). For $W = 8$, the relative complexity is 7.1%, corresponding to a speed up by a factor of 15 (in comparison to $W = \infty$) while leading to only a slight increase of the error.

In Tab. 1, we summarise our findings and compare relative run-times and performance of the proposed approach with basic methods and results from the literature. The last column shows a computing-time factor (CTF) relative to P2DW, which therefore has a CTF of 1 (26 s per image). It can be seen that increasing the flexibility of P2DW by means of the proposed strip extension greatly improves the accuracy. The proposed speedup allows us to use 64x64 pixels resolution, while the energy minimization technique presented in [5] operates on 32x32 pixels due to much higher complexity. Our method also greatly outperforms state-of-the-art feature matching approaches [2, 3, 18] which are though more efficient. Moreover, [3, 18] used manually pre-registered faces.

Evaluation on CMU-PIE database. To demonstrate the robustness of our approach w.r.t. to pose deformation, we evaluate our algorithm on the pose subset of the CMU-PIE database, using the frontal image as reference and the remaining 12 poses as testing images. As the reference is much more accurately cropped compared to the testing images (see Fig. 2 (bottom row)), we reverse the alignment procedure and align the test image to the reference one. This helps to minimize the impact of background pixels in the test images. We also follow [1] and additionally use left and right half crops of the reference image. We choose $\Delta = 3$ and set no absolute constraints for P2DW-FOSE, as this setup was shown to lead to the best performance on the AR Face database. Recognition results on the CMU-PIE database are listed in Tab. 2. In order to highlight the specific difficulties of the task, we divide the test data in near frontal and near profile

Table 1: Results for VJ-detected faces and comparison of run-times.

Model	ER [%]	CTF
No warping	22.3	-
P2DW	2.7	1
P2DW-FOSE	**1.8**	2.3
$+ W = 8$	2.0	**0.2**
CTRW-S [5]	3.7	0.4
SURF-Face [2]	4.15	-
DCT [3]	4.70*	-
Av-SpPCA [18]	6.43*	-

* with manually aligned faces

Table 2: Average error rates [%] on CMU-PIE groups of poses by our algorithms.

Model	near frontal	near profile	avg.
No warping	40.69	86.27	63.48
P2DW	0.25	17.63	8.94
P2DW-FOSE	0.25	10.39	**5.32**
Hierarch. match. [1]	1.22	10.39	5.76
3D shape mod. [23]	0.00	**14.40	**6.55
Prob. learning [15]	* 7	* 32	19.30

* estimated from graphs, ** missing poses

Table 3: Qualitative evaluation of the proposed approach.

Query	Ref	Deformed Ref.		Query	Ref	Deformed Query	
		P2DW	P2DW-FOSE			P2DW	P2DW-FOSE
							

poses. For the former, most approaches are able to achieve error rates near to 0%, while the latter is very difficult. A clear improvement is achieved compared to P2DW, and we also obtain the best result compared to the literature, where [1] uses a much more complex warping algorithm and [23] even use an additional profile shot as training data in order to generate a 3D head model. [15] uses automatically cropped images, which make the task even harder.

Tab. 3 shows qualitative results on an expression and pose image: in both cases the alignment by our method is much smoother compared to P2DW.

5 Conclusion

In this work, we have shown that a flexible extension of pseudo-2D warping helps to significantly improve recognition results on highly misaligned faces with different facial expressions, illuminations and strong changes in pose. Interestingly, even small deviations from the strict column-to-column mapping allow for much smoother alignments, which in turn provides more accurate recognitions. One interesting result from our evaluation is that it pays of to sacrifice a little of the global smoothness for tractable run-time on higher-resolution images. Also, we show that our globally optimal solution to a simplified problem outperforms an hierarchical approximation of the original problem, which might suffer from local minima. We believe this is an important road to explore, since quite often problems in computer vision are made tractable by introducing heuristics such as hierarchies without clearly investigating the impact of the hidden assumptions.

References

[1]Arashloo, S., Kittler, J.: Hierarchical image matching for pose-invariant face recognition. In: BMVC. (2009)
[2]Dreuw, P., Steingrube, P., Hanselmann, H., Ney, H.: Surf-face: Face recognition under viewpoint consistency constraints. In: BMVC. (2009)
[3]Ekenel, H.K., Stiefelhagen, R.: Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization. In: CVPRW, Washington, DC, USA (2006) 34
[4]Eickeler, S., Mller, S., Rigoll, G.: High performance face recognition using pseudo 2-d hidden markov models. In: ECCV. (1999)
[5]Gass, T., Dreuw, P., Ney, H.: Constrained energy minimisation for matching-based image recognition. In: ICPR, Istanbul, Turkey (2010) in press
[6]Gross, R.: `http://ralphgross.com/FaceLabels`
[7]Keysers, D., Unger, W.: Elastic image matching is np-complete. Pattern Recognition Letters **24** (2003) 445–453
[8]Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: CVPR (2). (2004) 506–513
[9]Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE TPAMI **28** (2006) 1568–1583
[10]Kuo, S.S., Agazzi, O.E.: Keyword spotting in poorly printed documents using pseudo 2-d hidden markov models. IEEE TPAMI **16**(8) (1994) 842–848
[11]Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004) 91–110
[12]Martinez, A., Benavente, R.: The AR face database. Technical report, CVC Technical report (1998)
[13]Mottl, V., Kopylov, A., Kostin, A., Yermakov, A., Kittler, J.: Elastic transformation of the image pixel grid for similarity based face identification. In: ICPR. (2002)
[14]Samaria, F.: Face Recognition Using Hidden Markov Models. PhD thesis, Cambridge University (1994)
[15]Sarfraz, M.S., Hellwich, O.: Probabilistic learning for fully automatic face recognition across pose. Image and Vision Computing **28**(5) (2010) 744 – 753
[16]Smith, S.J., Bourgoin, M.O., Sims, K., Voorhees, H.L.: Handwritten character classification using nearest neighbor in large databases. IEEE TPAMI **16**(9) (1994) 915–919
[17]Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: AFGR. (2002)
[18]Tan, K., Chen, S.: Adaptively weighted sub-pattern pca for face recognition. Neurocomputing **64** (2005) 505–511
[19]Uchida, S., Sakoe, H.: A monotonic and continuous two-dimensional warping based on dynamic programming. In: ICPR. (1998) 521–524
[20]Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision **57**(2) (2004) 137–154
[21]Wiskott, L., Fellous, J.M., Krger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE TPAMI **19** (1997) 775–779
[22]Wright, J., Hua, G.: Implicit elastic matching with random projections for pose-variant face recognition. CVPR **0** (2009) 1502–1509
[23]Zhang, X., Gao, Y., Leung, M.K.H.: Recognizing rotated faces from frontal and side views: An approach toward effective use of mugshot databases. IEEE TIFS **3**(4) (2008) 684–697