Masterarbeit im Fach Informatik

Matching Algorithms for Image Recognition

Der Fakultät für
Mathematik, Informatik und Naturwissenschaften der
RHEINISCH-WESTFÄLISCHEN TECHNISCHEN HOCHSCHULE AACHEN

Lehrstuhl für Informatik 6
Prof. Dr.-Ing. H. Ney

vorgelegt von:
Leonid Pishchulin
Matrikelnummer 284102

Gutachter:
Prof. Dr.-Ing. H. Ney
Prof. Dr. B. Leibe

Betreuer:
Dipl.-Inform. Tobias Gass
Dipl.-Inform. Philippe Dreuw

Januar 2010

Erklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe. Alle Textauszüge und Grafiken, die sinngemäß oder wörtlich aus veröffentlichten Schriften entnommen wurden, sind durch Referenzen gekennzeichnet.

Aachen, im Januar 2010

Leonid Pishchulin

Abstract

We analyze the usage of matching algorithms for image recognition. We focus on the approaches which aim at finding nonlinear deformations of an entire image. Zero-Order Warping (ZOW), Pseudo 2D Hidden Markov Model (P2DHMM) and Tree-Serial Dynamic Programming (TSDP) are studied. The effects of different constraints and parameter settings are analyzed. Furthermore, a new version of the TSDP and extensions for the P2DHMM are proposed. The proposed approaches allow to compensate for large disparities and additionally intend to preserve the monotonicity and continuity of the warping.

The problem of local and global image variability occurs in many image recognition tasks and is a typical issue in the domain of face recognition. Many local deformations caused by changes in facial expression and pose make conventional distance functions fail. Additionally, face registration errors worsen the performance of most holistic methods.

The P2DHMM is limited to a column-to-column mapping and is sensitive to registration errors, while the previously known version of the TSDP restricts the absolute displacement of each pixel. We propose to extend the P2DHMM by allowing deviations from a column which preserve the first-order dependencies between the pixels. Furthermore, we propose to relax some of the constraints imposed on the warping to cope with registration errors. A new version of the TSDP algorithm is proposed which relaxes the absolute constraints and intends to preserve the monotonicity and continuity of the warping.

The proposed extensions are compared to the already known methods. Experimental results on the AR Face and the Labeled Faces in the Wild dataset show that the proposed approaches can outperform state-of-the-art methods.

Acknowledgment

I would like to express my gratitude to the people who supported me during the preparation of this thesis.

My foremost thanks go to Prof. Dr.-Ing. Hermann Ney, head of the Chair of Computer Science VI at the RWTH Aachen University for his continuous interest in this work, numerous ideas and many fruitful discussions.

I would like to thank Prof. Dr. Bastian Leibe for agreeing to take the time to evaluate this thesis as a co-supervisor and for his interest in my work. I am very grateful to him for he has opened for me an exciting and inspiring world of computer vision.

I would like to thank Tobias Gass and Philippe Dreuw for supervising this work, their continuous support, many ideas and helpful feedback. Furthermore, I am very grateful to them for proofreading this thesis and introducing to me nicer ways of writing scientific texts.

Also, I would like to thank the other members of the image group at the Chair of Computer Science VI, especially Jens Forster and Harald Hanselmann for their help, useful suggestions and interesting discussions.

My special gratitude is due to my parents and grandparents for supporting and encouraging me through all the years. Finally, I would like to thank my brother for his interest and support.

Contents

Chapter 1

Introduction

One of the goals of computer vision is to automatically interpret general digital images of arbitrary scenes. Given an image of a scene, one challenge is to determine whether or not the image contains some specific object, feature, or activity. This task can easily be solved by a human, but is a difficult problem for a machine, since it requires reasoning from various image attributes and extensive amounts of knowledge representation.

One of the challenging issues in image recognition is the modeling of the local and global image variability which is typical for different categories of images and often occurs in many image recognition tasks. For instance, the problem of high local image variability is a typical issue in the domain of face recognition. Strong changes in facial expression and pose make the images of the same person look completely dissimilar. Moreover, illumination and viewing conditions may vary, and registration errors can occur. Finally, a face can be partially occluded by some objects, such as sunglasses or a scarf, which makes the problem of face recognition difficult even for humans. Therefore, sophisticated approaches are needed which are able to cope with many local changes in image content.

In general, one of the ways to approach the problem of image recognition is to find a suitable representation of the image data. Therefore, image content can be described by different cues. Appearance-based approaches (e.g. [Roth & Winter 08]) and model-based approaches (e.g. [Jain & Zhong$^+$ 98, Rucklidge 97]) exist. For appearance-based approaches, only appearance is used which is captured by different two-dimensional views of the object of interest. Based on the applied features, these methods can be sub-divided into global and local approaches. Global, or holistic approaches try to cover the information content of the whole image. They vary from simple statistical measures, such as histograms of features [Schiele & Crowley 00], to more sophisticated dimensionality reduction techniques, i.e., subspace methods, such as principal component analysis (PCA) [Jolliffe 02], or independent component analysis (ICA) [Hyvarinen & Karhunen$^+$ 01]. The main idea behind the last group of methods is to project the original data into a subspace which represents the data optimally according to a predefined criterion. However, due to the global representation of the data, holistic approaches have major limitations, such as sensitivity to intra-class variability, partial occlusions and registration errors. In contrast to holistic approaches, local appearance-based approaches try to find local representations of an image. They search for distinctive regions which are characterized by e.g. edges or corners, and describe the regions by means of a local feature descriptor [Mikolajczyk & Schmid 05]. Among a great variety of local feature descriptors, the Scale Invariant Feature Transform (SIFT) [Lowe 04] is the most widely used one. It offers scale and rotation invariant properties and can handle significant changes in viewpoint and illumination. Due to the local representation of the image data, local approaches are more robust to intra-class variability and partial occlusions than holistic approaches.

However, although some of the global and local appearance based approaches additionally consider the local geometric information most of them are unable to take into account the whole geometric structure of the image content. In some specific tasks, such as face recognition, appearance based approaches can simply fail due to poor-contrast, adverse viewing conditions or registration errors. Contrary, in model-based recognition problems, the image content undergoes explicit geometric transformation which maps the image data onto a coordinate system, e.g. an image plane [Rucklidge 97]. In other words, if a test image has to be compared with some reference image, first the reference image is deformed through a geometric transformation and then the deformed reference image can directly be compared with the test image by e.g. the Euclidean distance. Linear and non-linear types of image deformation are known which correspond to linear and nonlinear geometric transformation, respectively. The image deformation of each particular type can be found through the minimization of a corresponding cost function. For instance, linear image deformations which correspond to affine transformations can be efficiently located through the minimization of the Hausdorff distance [Rucklidge 97]. Another example of the approach which is also able to compensate for small linear image deformations is the Tangent distance [Simard & LeCun^{+} 93]. It has been successfully used in the task of handwritten character recognition [Keysers & Dahmen^{+} 00], while the Hausdorff distance was applied in the domain of document processing [Son & Kim^{+} 08]. However, in more challenging tasks, such as face recognition, the ability to locate linear image deformations is not sufficient. Besides registration errors which can still be compensated by global affine transformations, many local changes occur, such as variations in facial expression and illumination. Therefore, distance functions based on nonlinear image deformation models are needed which are able to cope with high local image variability.

Numerous nonlinear image deformation models of different complexity have been proposed in the literature [Smith & Bourgoin^{+} 94, Kuo & Agazzi 94],
[Uchida & Sakoe 98, Keysers & Deselaers^{+} 07]. Each deformation model underlies a matching algorithm which finds the deformation of an entire image through the min-

imization of a cost function. Recognition by flexible matching of a test image to the given reference images is one of the most promising techniques to achieve a high recognition accuracy. Different matching algorithms have been successfully used in the tasks of handwritten character recognition [Ronee & Uchida$^+$ 01], [Keysers & Gollan$^+$ 04b], document processing [Kuo & Agazzi 94], or medical image analysis [Keysers & Gollan$^+$ 04a, Keysers & Deselaers$^+$ 07], but no direct performance comparison in more challenging tasks, such as face recognition, is provided. Moreover, some of the approaches have major limitations which become obvious when image resolution increases. Additionally, no performance comparison of the approaches in combination with the SIFT [Lowe 04] local feature descriptor has been reported so far.

In this work, we investigate existing matching algorithms based on nonlinear image deformation models, namely Zero-Order Warping (ZOW) [Smith & Bourgoin$^+$ 94], Pseudo 2D Hidden Markov Model (P2DHMM) [Kuo & Agazzi 94] and Tree-Serial Dynamic Programming (TSDP) [Mottl & Kopylov$^+$ 02]. We analyze limitations of these approaches and propose improved methods which are intended to overcome the shortcomings of the P2DHMM and TSDP by changing some of the constraints imposed on image deformation. We provide a thorough comparison of the matching algorithms for the task of face recognition and experimentally show the superior performance of the improved approaches over original ones. We also investigate the advantages of using sophisticated local feature descriptors, such as the SIFT descriptor, in combination with the presented approaches. Furthermore, we evaluate the matching algorithms on the AR Face [Martinez & Benavente 98] and Labeled Faces in the Wild [Huang & Mattar$^+$ 08] dataset and show that the proposed approaches are able to outperform many state-of-the-art methods.

This work is organized as follows: First, in Chapter 2, the problem of image warping is formulated and a couple of existing approaches are described. Next, in Chapter 3, problems of existing approaches are analyzed and a new version of the TSDP and improvements of the P2DHMM are proposed. In Chapter 4, an overview of our recognition system is given. Finally, in Chapter 5, thorough performance comparison of the matching algorithms on the AR Face and the LFW database is presented.

Chapter 2

Theoretical Background

Various matching algorithms based on nonlinear image deformation models have already been discussed in the literature. They are distinguishable by underlying image deformation models and employed optimization strategies. Each particular model is characterized by an assumption about dependencies between the pixel displacements during the image deformation. In the most general case, the displacement of each pixel depends on the displacements of its direct neighbors [Moore 79, Levin & Pieraccini 92], which results in the first-order two-dimensional deformation model. However, as shown in [Keysers & Unger 03], the optimization of this problem is NP-complete. In [Uchida & Sakoe 98], the authors propose to *approximate* the optimization of the first-order two-dimensional model by using beam search. In this case, the complexity is reduced, but the obtained solution is probably suboptimal. Another way to reduce the complexity is to *relax* the first-order two-dimensional model by making further assumptions about dependencies between the pixel displacements. A couple of relaxations have been proposed in the literature. In the P2DHMM [Kuo & Agazzi 94], the first-order dependencies are reduced to neighbors along one dimension (hence pseudo 2D). The P2DHMM has been successfully used in the domain of document processing [Kuo & Agazzi 94] and in the task of handwritten digit recognition [Keysers & Deselaers$^+$ 07]. In the work presented in [Mottl & Kopylov$^+$ 02], the authors propose the TSDP approach where the first-order dependencies are retained between the displacements of neighboring pixels in individual pixel neighborhood trees. The TSDP has been applied to the problem of face identification [Mottl & Kopylov$^+$ 02]. No dependencies between the pixel displacements are assumed in the ZOW [Smith & Bourgoin$^+$ 94, Uchida & Sakoe 05] which has been independently described in the literature several times due to its simplicity and efficiency. The ZOW has been successfully used for handwritten character recognition [Smith & Bourgoin$^+$ 94, Keysers & Gollan$^+$ 04b, Keysers & Deselaers$^+$ 07] and radiograph recognition [Keysers & Gollan$^+$ 04a].

In this chapter, the mentioned image deformation models, as well as matching algorithms based on these models are described in detail. First, the problem of image warping is formulated. Next, the first-order two-dimensional model is described in more detail and the advantages of its relaxations over the approximative optimiza-

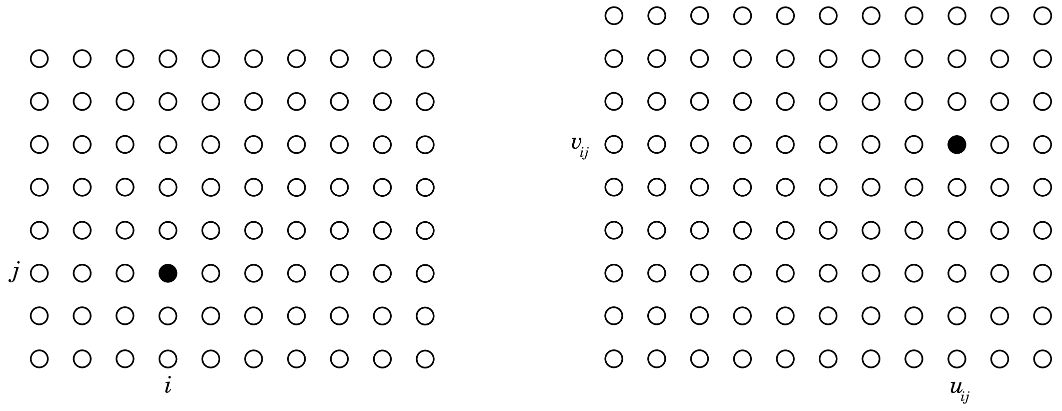


Figure 2.1: Illustration of warping function

tion are explained. Finally, relaxations of the first-order two-dimensional model are discussed.

2.1 Problem Formulation

In this section, we recapitulate the problem of image warping. We refer to the work [Uchida & Sakoe 98] for more details.

For the discussion of the matching algorithms the following notation will be used: given a test image $X = \left\{x_{ij}\right\}$ of dimension $I \times J$ and a reference image $R = \left\{r_{uv}\right\}$ of dimension $U \times V$ with pixel grids $\left\{(i,j)\right\}$ and $\left\{(u,v)\right\}$, respectively, an image warping is denoted as a mapping

$$(i,j) \rightarrow w_{i,j} = (u_{i,j}, v_{i,j}) \tag{2.1}$$

of each pixel (i,j) of the test image to some pixel (u,v) of the reference image. An example of such a mapping is shown in Figure 2.1. The result of the warping is a deformed reference image $R_{\{w_{i,j}\}} = \left\{r_{w_{ij}}\right\}$ of dimension $I \times J$ with pixel grid $\left\{w_{i,j}\right\}$. According to Equation (2.1), the test image must be mapped completely, while the reference image can have some pixels “unused” in the final result.

The warping is found through the optimization of the following combined criterion:

$$\min_{\left\{w_{i,j}\right\}} \left\{ \sum_{i,j} [d_{i,j}(w_{i,j}) + T(w_{i-1,j}, w_{i,j-1}, w_{i,j})] \right\}, \tag{2.2}$$

where the set $\left\{ w_{i,j} \right\}$ of all possible warpings can be restricted by certain constraints which depend on chosen deformation model.

The optimized criterion in Equation (2.2) consists of two parts. The first part of the criterion is an accumulated pixel distance $\sum_{i,j} d_{i,j}(w_{i,j})$ between the test image X and the deformed reference image $R_{\{w_{i,j}\}}$. As distance $d_{i,j}(w_{i,j})$ any distance function can be used, for example the Euclidean distance

$$d_{i,j}(w_{i,j}) = ||x_{i,j} - r_{w_{i,j}}|| \tag{2.3}$$

The second part of the cost function is a warping penalty term $T(\cdot)$ which is used to smooth the deformation. The penalty term can also be used to model the first-order dependencies between the pixel displacements. In the most general case, the displacement of each pixel depends on the displacements of all its direct neighbors. This kind of dependency is shown in Figure 2.2(a). However, without loss of generality, only displacements of the left $(i-1, j)$ and bottom $(i, j-1)$ neighbors of each pixel (i, j) can be taken into account (see e.g. [Ney & Dreuw$^+$ 09] for more details). In other words, the warping $(i, j) \rightarrow w_{i,j}$ depends only on the warpings $(i-1, j) \rightarrow w_{i-1,j}$ and $(i, j-1) \rightarrow w_{i,j-1}$. This assumption is demonstrated in Figure 2.2(b). Consequently, the penalty term $T(\cdot)$ can further be decomposed into corresponding vertical $T_v(\cdot)$ and horizontal $T_h(\cdot)$ components.

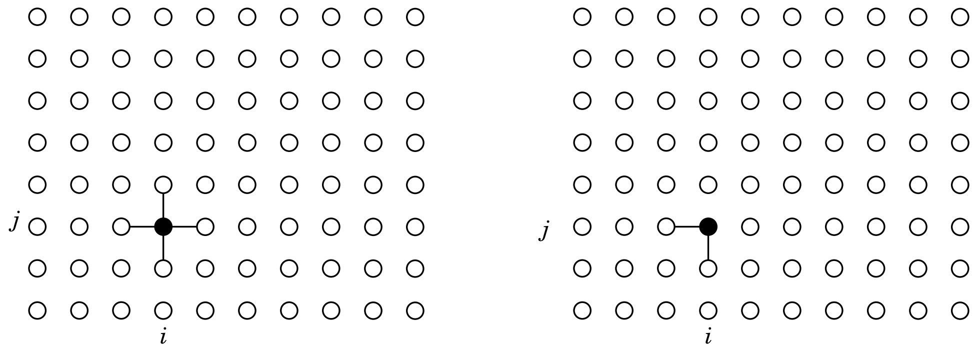
$$T(w_{i-1,j}, w_{i,j-1}, w_{i,j}) = T_h(w_{i-1,j}, w_{i,j}) + T_v(w_{i,j-1}, w_{i,j}) \tag{2.4}$$

At this point, it should be remarked that the notion of neighborhood relations between the pixels in the deformed reference image is different from those accepted for the test image. For every two pixels (i, j) and $(i-1, j)$ being row neighbors in the test image, warped pixels $w_{i,j}$ and $w_{i-1,j}$ lay not necessarily nearby in the deformed reference image. The same can be said about column neighbors (i, j) and $(i, j-1)$. Too unconstrained deformation may lead to the appearance of large gaps and discontinuities in the final warping. An appropriate definition of the penalty function $T(\cdot)$ prevents the pixels $w_{i-1,j}$, $w_{i,j-1}$ and $w_{i,j}$ from being placed far from each other which results in the smoothness of the deformation.

The most intuitive way to define a penalty term is to set it being proportional to some convex function. For example, the Euclidean distance can be used:

$$\begin{aligned} T_h(w_{i-1,j}, w_{i,j}) &= \alpha||\ w_{i,j} - w_{i-1,j} - (1,0)\ || && (2.5) \\ T_v(w_{i,j-1}, w_{i,j}) &= \alpha||\ w_{i,j} - w_{i,j-1} - (0,1)\ || && (2.6) \end{aligned}$$

According to Equations (2.5) and (2.6), no penalty cost arises, if $w_{i-1,j}$ and $w_{i,j-1}$ are left and bottom neighbors of $w_{i,j}$, respectively. If not, the penalty cost increases.



(a) Full dependency

(b) Decomposition

Figure 2.2: Illustration of neighborhood dependencies between the pixels in the test image

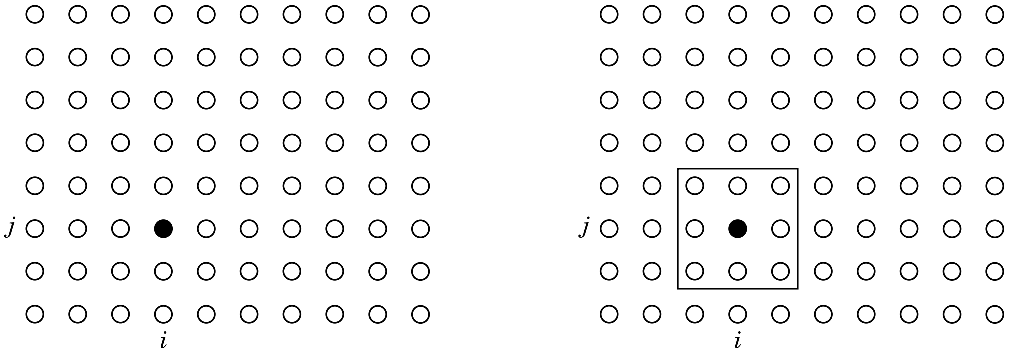
Penalty terms can be understood as soft constraints. They still allow strong deformations of the reference image although making them rather improbable. Along with the soft constraints hard constraints can be introduced. Hard constraints of two types exist: absolute position constraints and relative position constraints. The absolute position constraints prohibit large absolute displacements of pixels in the deformed reference image with respect to the test image. For each pixel (i, j) of the test image, the absolute position constraints restrict the possible locations of the pixel $w_{i,j}$ by a warping range W. Formally, it can be expressed as follows:

$$\begin{aligned} \mid u_{i,j} - i \mid & \leq W, & (2.7) \\ \mid v_{i,j} - j \mid & \leq W & (2.8) \end{aligned}$$

However, these equations are valid only in the case, when both the test and reference images have the *same* sizes, i.e. $I = U$ and $J = V$. Figure 2.3 illustrates possible warpings of pixels under the absolute position constraints when $W = 1$. As the value of W is usually small, the absolute position constraints reduce the complexity of the optimization, as we explain in Section 2.3.1.

The relative position constraints affect the ordering of pixels in the final warping with respect to each other. In general, for each two row neighbors $(i-1, j)$ and (i, j) in the test image, at most one pixel can be skipped between the pixels $w_{i-1,j}$ and $w_{i,j}$ in the deformed reference image. This is expressed through the following constraints:

$$\begin{aligned} \mid u_{i,j} - u_{i-1,j} \mid & \leq 2, & (2.9) \\ \mid v_{i,j} - v_{i-1,j} \mid & \leq 2 & (2.10) \end{aligned}$$



Test image

Reference image

Figure 2.3: Illustration of absolute position constraints ($W = 1$)

Figure 2.4 exemplifies the effect of the constraints in Equations (2.9) and (2.10) onto the neighborhood relations between the pixels. As it can be seen, the number of all possible predecessors $w_{i-1,j}$ for the pixel $w_{i,j}$ is at most 25.

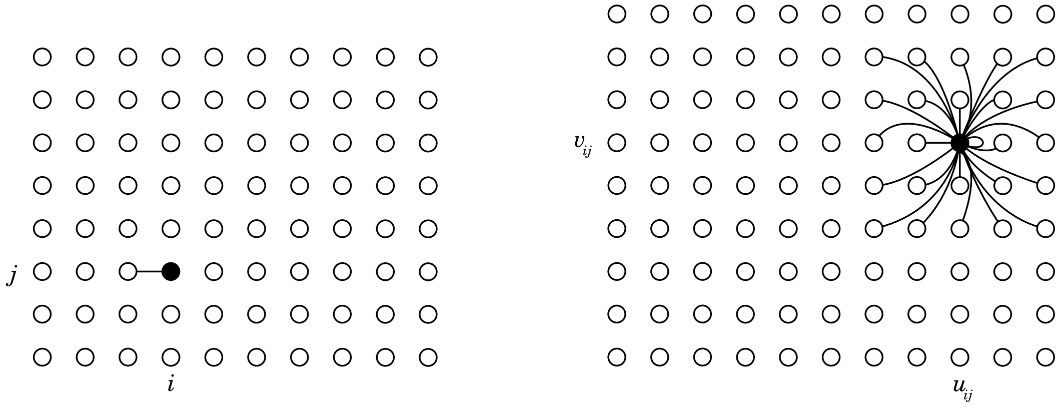
Similar constraints are imposed on the warping of the pixels $(i, j-1)$ and (i, j):

$$\begin{aligned} \mid u_{i,j} - u_{i,j-1} \mid &\leq 2, & (2.11) \\ \mid v_{i,j} - v_{i,j-1} \mid &\leq 2 & (2.12) \end{aligned}$$

Definition of the relative position constraints in Equations (2.9 - 2.12) is a general form of the monotonicity and continuity constraints. Monotonicity requirement prevents some regions of the deformed reference image from being mirrored, while continuity requirement prohibit large gaps between the pixels in the warping. The monotonicity and continuity constraints are defined as follows:

$$\begin{aligned} 0 \leq u_{i,j} - u_{i-1,j} &\leq 2, & (2.13) \\ \mid u_{i,j} - u_{i,j-1} \mid &\leq 1, & (2.14) \\ 0 \leq v_{i,j} - v_{i,j-1} &\leq 2, & (2.15) \\ \mid v_{i,j} - v_{i-1,j} \mid &\leq 1 & (2.16) \end{aligned}$$

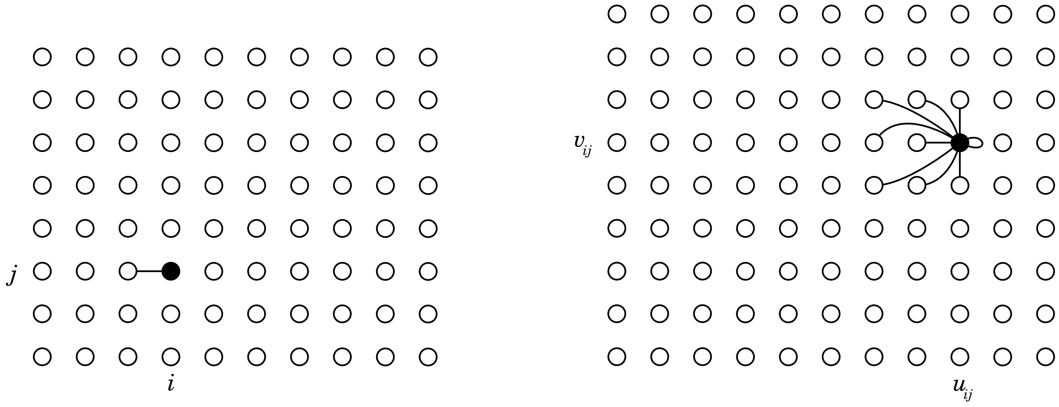
According to Equations (2.13 - 2.16), at most one pixel can be skipped in vertical and horizontal directions, and no backward steps are allowed. Figures 2.5 and 2.6 show the neighborhood relations between the pixels under the monotonicity and continuity constraints. It can be seen that for each pixel $w_{i,j}$, the number of allowed predecessors is reduced to 9. Additionally to the explained types of constraints, boundary



Test image

Reference image

Figure 2.4: Illustration of relative position constraints: row neighbors



Test image

Reference image

Figure 2.5: Illustration of monotonicity and continuity constraints: row neighbors

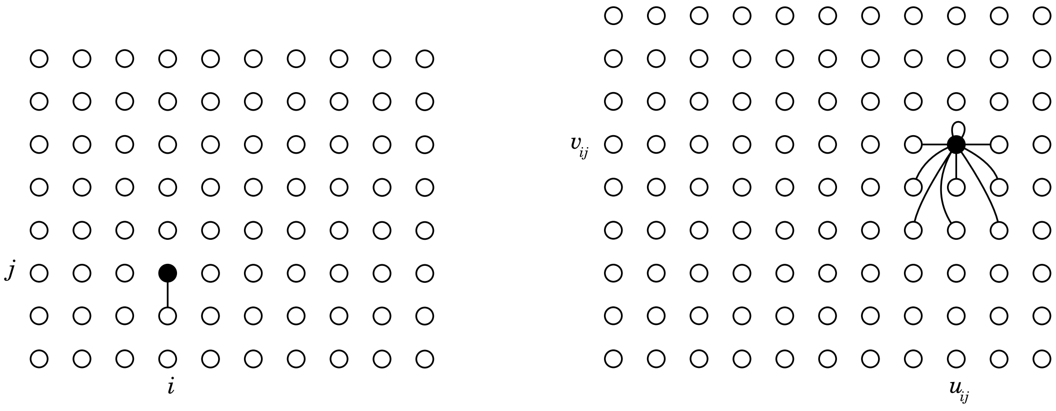


Figure 2.6: Illustration of monotonicity and continuity constraints: column neighbors

constraints can be enforced:

$$u_{1,j} = v_{i,1} = 1,\ u_{I,j} = U,\ v_{i,J} = V \tag{2.17}$$

According to Equation (2.17), boundary pixels of the test image are allowed to be mapped only to the boundary pixels of the reference image.

2.2 First-Order 2D Warping

According to the first-order two-dimensional deformation model, the displacement of each pixel depends on the displacements of its direct neighbors. Figure 2.7 shows the dependencies between the pixels in this case. First-Order 2D Warping (2DW) [Uchida & Sakoe 98] was proposed to find an exact solution of the image warping under the first-order two-dimensional deformation model. The 2DW can be understood as an extension of one-dimensional dynamic programming [Sakoe & Chiba 90] to two dimensions. The warping is found through the optimization of the criterion in Equation (2.2). The optimization is performed under the monotonicity and continuity constraints (c.f. Eq. (2.13 - 2.16)), and the boundary constraints (c.f. Eq. (2.17)).

In contrast to the one-dimensional case, which allows polynomial solutions, optimization of the criterion in Equation (2.2) is NP-complete [Keysers & Unger 03]. There are two ways to overcome this hurdle. One way is to find an *approximate* solution of the original problem. For instance, it can be done by an application of dynamic programming based algorithms accelerated by beam search [Uchida & Sakoe 98] or by

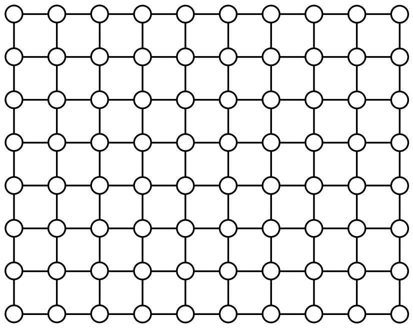


Figure 2.7: Illustration of first-order dependencies between the pixels

performing a piecewise linear two-dimensional warping [Ronee & Uchida^{+} 01]. Both approaches showed good performance in the task of optical character recognition. However, the application of them in more challenging tasks, such as face recognition, is impractical, since the complexity of both approaches increases dramatically with the increase of image resolution. Another way to reduce the complexity is to *relax* the first-order two-dimensional deformation model. In this case, an *exact* solution of the image warping under relaxed deformation model can be searched efficiently. In the following, a couple of relaxations of the original two-dimensional problem are discussed.

2.3 Relaxations

In general, a relaxation simplifies the complexity of two-dimensional warping by making assumptions about dependencies between pixel displacements. At the same time, the purpose of the relaxation is also to find an appropriate representation of the problem, such that an exact solution exists and can be efficiently searched. In the following, we explain three relaxations of the first-order two-dimensional deformation model. First, we describe the ZOW [Smith & Bourgoin^{+} 94] since it is simple and efficient. Then, the P2DHMM [Kuo & Agazzi 94] is explained, as this approach models the first-order dependencies between the pixels. Finally, we describe the TSDP [Mottl & Kopylov^{+} 02] which offers all advantages of dynamic programming on trees.

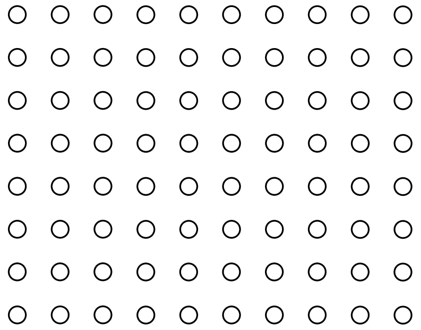


Figure 2.8: Illustration of zero-order dependencies

2.3.1 Zero-Order Warping

Zero-Order Warping (ZOW) [Smith & Bourgoin+ 94, Keysers & Gollan+ 04b], [Uchida & Sakoe 05] is an extreme relaxation of the first-order two-dimensional problem. It neglects the first-order dependencies between pixel displacements and imposes only the absolute position constraints on the warping (c.f. Eq. (2.7) and (2.8)). In this case, no neighborhood relations between the pixels are taken into account, as shown in Figure 2.8. Therefore, this is a zero-order deformation model.

The position of each pixel $w_{i,j}$ in the deformed reference image depends only on the pixel (i,j) in the test image. The deformation of the reference image can partially be smoothed by introducing absolute position penalties [Keysers & Deselaers+ 07]

$$T((i,j),w_{i,j}) \quad = \quad \alpha || \, (i,j) - w_{i,j} \, ||, \tag{2.18}$$

which penalize the absolute displacement of each pixel in the deformed image with respect to the test image. The warping is found through optimization of the following criterion:

$$\min_{\left\{w_{i,j}\right\}} \left\{ \sum_{i,j} [d_{i,j}(w_{i,j}) + T((i,j),w_{i,j})] \right\} \tag{2.19}$$

As a locally optimal decision is made independently for each pixel, the optimization criterion in Equation 2.19 can be rewritten as follows:

$$\sum_{i,j} \min_{w_{i,j}} [d_{i,j}(w_{i,j}) + T((i,j),w_{i,j})] \tag{2.20}$$

The optimal solution of the problem in Equation (2.20) is found efficiently in a straightforward way.

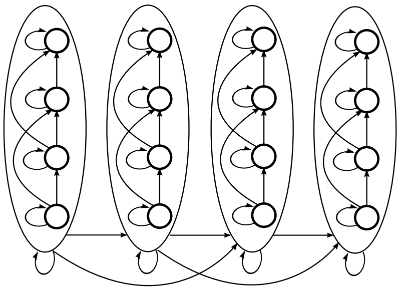


Figure 2.9: Schematic overview of the P2DHMM

As reported in [Keysers & Gollan$^+$ 04b], the performance of the ZOW is significantly improved when the local context of each pixel is included. In Chapter 4, this procedure is explained in more detail.

Complexity. The simplicity and efficiency of the ZOW make it attractive for image warping problems. Indeed, its computing time is increased by only about a factor of $(2W+1)^2$ in comparison to the Euclidean distance calculation, while the overall complexity of the algorithm is $IJ(2W+1)^2$.

2.3.2 Pseudo 2D Hidden Markov Model

The ZOW has an evident drawback: due to the lack of the first-order dependencies between pixel displacements, unrealistic deformations of the reference image can be produced. This allows for unwanted good mappings between the images from different classes, which can result in the increase of the recognition error rate. Pseudo 2D Hidden Markov Model (P2DHMM) [Kuo & Agazzi 94] is intended to overcome this shortage. The P2DHMM is an extension of the one-dimensional Hidden Markov Model (HMM), which is widely used for the alignment of one-dimensional signals in the area of speech recognition. The P2DHMM relaxes the first-order two-dimensional deformation model, but retains the first-order dependencies between the pixels in a column and across the whole columns. Figure 2.9 shows a schematic overview of the P2DHMM. As it can be seen from the picture, the positions of the pixels in a column correspond to the states, while the whole columns are represented as superstates.

The P2DHMM requires the mapping in Equation (2.1) to be rewritten as follows:

$$i \rightarrow u_i, \qquad (2.21)$$

$$(i,j) \rightarrow v_{i,j} \qquad (2.22)$$

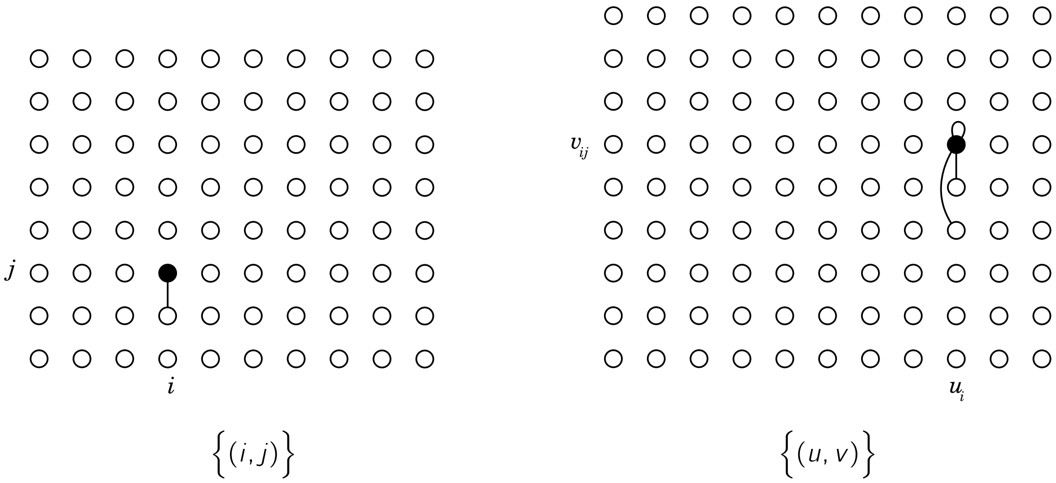


Figure 2.10: Example of permitted warpings in the P2DHMM

This means that a column in the test image is completely mapped to a singe column in the reference image. No deviations from a column are allowed. Formally, this is expressed through the constraint

$$u_{i,j} \quad = \quad u_{i,j-1} \equiv u_i \tag{2.23}$$

The first-order dependencies between the pixels in a column are expressed through the monotonicity constraint in Equation (2.15). In this case, permitted warpings of column neighbors are shown in Figure 2.10.

The first-order dependencies across the columns are expressed as follows:

$$0 \quad \leq \quad u_i - u_{i-1} \quad \leq 2 \tag{2.24}$$

In the work presented in [Keysers & Deselaers$^+$ 07], the boundary constraints are imposed on the warping.

The warping of pixels in each column is found independently from other columns. The alignment of columns is performed by means of the HMM which allows a skip of at most one pixel. An example of such alignment is shown in Figure 2.11. As it can directly be seen from the picture, each pixel (i,j) can only be mapped to those pixels (u,v), which are available according to the boundary constraints. Therefore, the boundary constraints help to reduce the complexity. At the same time, the boundary constraints are not restrictive if a disparity between the test and the reference image is small. Otherwise, the boundary constraints can be prohibitive. This problem is discussed in more detail in Chapter 3.

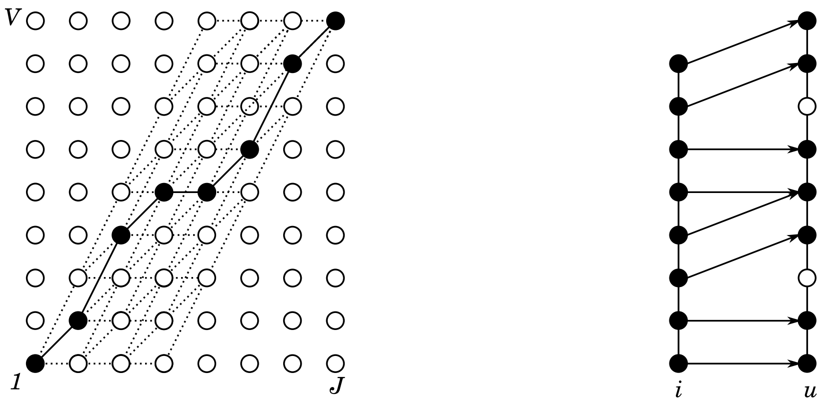


Figure 2.11: Example of a column alignment by means of the HMM ($skip = 1$)

The requirement that a column in the test image is completely mapped to a single column in the reference image has two consequences: First, only entire columns can be skipped in the warping. Second, dependencies between the vertical displacements of the pixels in neighboring columns are neglected.

Under the considerations discussed above, the optimized criterion in Equation (2.2) can be rewritten as follows [Keysers & Deselaers+ 07]:

$$\min_{\{u_i, v_{ij}\}} \left\{ \sum_{i,j} [d_{i,j}(u_i, v_{i,j}) + T(u_{i-1}, u_i, v_{i,j-1}, v_{i,j})] \right\} \tag{2.25}$$

The exact solution for such a relaxed problem is found efficiently by means of the following two-phase algorithm [Keysers & Deselaers+ 07]: in the first phase, for each pair (i, u_i) of columns, the best mapping $(i,j) \to (u_i, v_{i,j})$ is found by using dynamic programming within columns. This results in the following scores:

$$D(i, u_{i-1}, u_i) = \min_{\{v_{i,j}\}} \left\{ \sum_{j} [d_{i,j}(u_i, v_{i,j}) + T(u_{i-1}, u_i, v_{i,j-1}, v_{i,j})] \right\} \tag{2.26}$$

Then, the order of columns in the final warping is determined by an application of dynamic programming to the scores in Equation (2.26). As a result, the mapping $i \to u_i$ is found:

$$\min_{\{u_i\}} \left\{ \sum_{i} [D(i, u_{i-1}, u_i)] \right\} \tag{2.27}$$

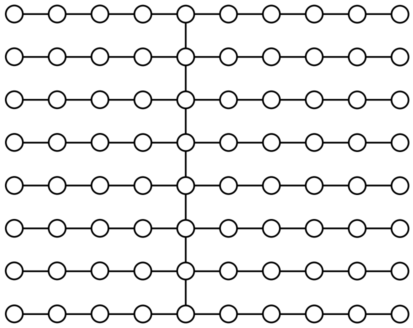


Figure 2.12: Example of an individual pixel neighborhood tree in the TSDP

Complexity. For each two columns i and u, the algorithm performs less than $3JV$ pixel comparisons in total, where 3 is the number of possible transitions between a pixel and its predecessor. As a skip of at most one column is allowed, the upper bound of the overall complexity is $IU(3+3JV)$. Due to the boundary constraints, this upper bound is never achieved. Additionally, the cost of the distance computation between each pixel (i,j) and (u,v) has to be considered. Again, due to the boundary constraints, this cost is less than $IUJV$.

2.3.3 Tree-Serial Dynamic Programming

Tree-Serial Dynamic Programming (TSDP) [Mottl & Kopylov^{+} 02] is another relaxation of the first-order two-dimensional deformation model. Instead of treating a rectangular pixel lattice as a collection of columns, as it is done in the P2DHMM, this approach represents the two-dimensional pixel grid as a series of individual pixel neighborhood trees. Each tree is obtained from the grid by removing vertical edges beyond the assignment stem which is a particular column. Although each tree has its own stem, it shares the same horizontal branches with other trees and is defined on the whole pixel set. An example of such a tree is illustrated in Figure 2.12.

Under the explained assumption, the original two-dimensional optimization problem in Equation (2.2) breaks up into a series of partial tree-like ones, each of which can be efficiently solved by dynamic programming.

The algorithm proceeds as follows: for each column i^* a tree is constructed. The warping of the pixels composing the stem of the tree is found through optimization of

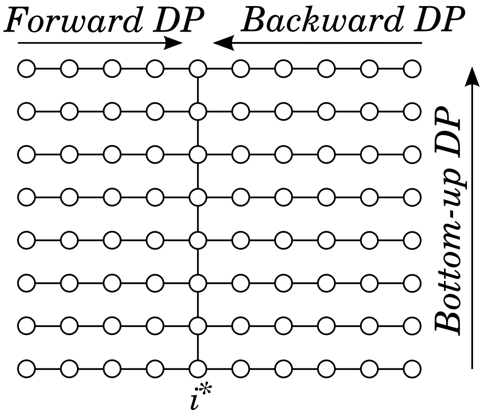


Figure 2.13: Schematic overview of the TSDP algorithm

the following criterion:

$$i^* : \min_{\{w_{ij}\}} \left\{ \sum_{i,j} d_{ij}(w_{ij}) + \sum_{j} \left[T_v(w_{i^*,j-1}, w_{i^*j}) + \sum_{i} T_h(w_{i-1,j}, w_{ij}) \right] \right\} \tag{2.28}$$

The optimization is performed in two steps. First, for each j, $j = 1..J$, the scores

$$D_{i^*,j}(w_{i^*,j}) = \min_{\{w_{ij}\}:w_{i^*,j}=w} \left\{ \sum_{i} \left[d_{ij}(w_{ij}) + T_h(w_{i-1,j}, w_{ij}) \right] \right\} \tag{2.29}$$

are computed. These scores characterize the optimal warping of the pixels in corresponding trees' branches. Then, the solution $\left\{ w_{i^*,j} \right\}$, $j = 1..J$, of the problem in Equation (2.28) is obtained by optimizing the criterion

$$\min_{\{w_{i^*,j}\}} \left\{ \sum_{j} \left[D_{i^*,j}(w_{i^*,j}) + T_v(w_{i^*,j-1}, w_{i^*,j}) \right] \right\} \tag{2.30}$$

The solution of the last problem contributes to the final warping $\left\{ w_{i,j} \right\}$.

For efficient computation of the scores in Equation (2.29), forward and backward dynamic programming procedures are applied independently to each row of the pixel grid. For that purpose, the objective function in Equation (2.29) is transformed as follows:

$$\begin{aligned} D_{i^*,j}(w_{i^*,j}) = d_{i^*,j}(w_{i^*,j}) \quad & + \min_{\{\hat{w}_{i',j'}\}:i'<i^*,j'=j} \left\{ F_{i^*-1,j}(\hat{w}_{i^*-1,j}) + T_h(\hat{w}_{i^*-1,j}, w_{i^*,j}) \right\} \\ & + \min_{\{\hat{w}_{i',j'}\}:i'>i^*,j'=j} \left\{ B_{i^*+1,j}(\hat{w}_{i^*+1,j}) + T_h(w_{i^*,j}, \hat{w}_{i^*+1,j}) \right\} \end{aligned} \tag{2.31}$$

where

$$
\begin{aligned}
F_{i,j}(w_{i,j}) &= \min_{\{\hat{w}_{i',j'}\}i'<i,j'=j} \Big\{ F_{i-1,j}(\hat{w}_{i-1,j}) + T_h(\hat{w}_{i-1,j}, w_{i,j}) \Big\}, \\
B_{i,j}(w_{i,j}) &= \min_{\{\hat{w}_{i',j'}\}i'>i,j'=j} \Big\{ B_{i+1,j}(\hat{w}_{i+1,j}) + T_h(\hat{w}_{i,j}, w_{i+1,j}) \Big\}
\end{aligned}
$$

Figure 2.13 gives a schematic representation of the algorithm.

The lack of dependencies between the trees' branches allows efficient optimization of pixel displacements. However, as each branch is optimized independently from the others, no constraints are imposed on their alignment with respect to each other. Therefore, the neighboring branches can strongly intersect or lie far from each other, which can negatively affect the deformation of the reference image. This drawback of the TSDP is discussed in a more detail in Chapter 3.

In the work presented in [Mottl & Kopylov$^+$ 02], the authors propose to allow a skip of more than one pixel at once in the warping. At the same time, in order to keep the computational complexity within reasonable limits, the absolute position constraints are imposed. The number of pixel skips is controlled by the warping range and can be $2W + 2$ pixels at most. In the following, we refer to this version of the TSDP as Tree-Serial Dynamic Programming with Many Skips (TSDP-W*), where W^* is stated for the dependency of the number of skips on the warping range. Permission of more pixel skips comes at a price for restriction of the warping by the absolute position constraints which make the performance of the TSDP-W* be strongly dependent on the warping range. This major shortcoming of the explained approach is discussed in Chapter 3.

Complexity. According to the absolute position constraints, $(2W + 1)^2$ possible displacements of each pixel $w_{i,j}$ are allowed. For each of these displacements, $(2W+1)^2$ displacements of the predecessor pixel are taken into account. Therefore, the compound complexity of the forward, backward and bottom-up run of dynamic programming is $3IJ(2W + 1)^4$. Additionally, the cost of the distance computation between each pixel (i, j) and (u, v) has to be considered, which is $IJ(2W + 1)^2$.

Chapter 3

Matching Algorithms

As we mention in Chapter 2, the P2DHMM and the TSDP-W* relaxations of the first-order two-dimensional deformation model have evident drawbacks. The major shortcomings of both approaches stem from the constraints enforced on the warping.

In case of the P2DHMM, two requirements can be unnecessarily restrictive, namely

- the boundary constraints and
- the column-to-column mapping,

while the performance of the TSDP-W* can drop due to

- the absolute position constraints and
- the lack of dependencies between the trees' branches.

In this chapter, we analyze these drawbacks and propose new approaches intended to overcome them. First, we explain the shortcomings of the P2DHMM and the TSDP-W* in more detail. Then, we propose methods which intend to overcome these drawbacks.

3.1 Drawbacks of Existing Approaches

In this section, we analyze the drawbacks of existing approaches. First, we discuss the boundary constraints and the restriction caused by the column-to-column mapping in the P2DHMM. Then, we analyze the shortcomings of the TSDP-W* which stem from the absolute position constraints and the lack of dependencies between the trees' branches.

3.1.1 P2DHMM

Boundary Constraints. According to the boundary constraints (c.f. Eq. (2.17)), the boundary pixels of the test image must be mapped to the boundary pixels of the reference image. As we explain in Chapter 2, this requirement is not restrictive, if a

Test image

Reference image

Figure 3.1: Example of a significant disparity between the test and the reference image due to a registration error

disparity between the test and the reference image is small. However, the boundary constraints can be prohibitive if e.g. registration errors occur or if the scale of the test and the reference image is different. Figure 3.1 exemplifies the situation when a registration error causes large disparity between the images. In this case, the P2DHMM would map a part of the face in the test image to the background in the reference image, which would negatively affect the deformation of the reference image.

We propose an approach which intends to overcome this drawback of the P2DHMM in two different ways. In the first version of the approach, we propose to allow a skip of more pixels and columns at once, while in the second version we propose to completely eliminate the boundary constraints. We analyze both possibilities and present the approach in Section 3.2.

Column-to-column Mapping. The next shortcoming of the P2DHMM is caused by the restriction that a column in the test image is completely mapped to a single column in the reference image (c.f. Eq. (2.23)). According to this constraint, only complete columns of the reference image can be skipped in the warping. However, as we show in Chapter 5, this requirement can be unnecessarily restrictive. Therefore, this constraint has to be relaxed by allowing additional deviations from a column.

In the work presented in [Keysers & Deselaers$^+$ 07], the authors propose to permit *zero-order* one pixel distortions from a column, which allows to obtain additional flexibility with a slight raise of the complexity. However, in this case, the decision on the horizontal deviation of each pixel is only locally optimal. As a result, the continuity constraint in Equation (2.16) can be violated.

Therefore, we propose to model the deviations from a column in accordance with the *first-order* dependencies between the pixels. Additionally, we do not restrict the deviations to be at most one pixel and allow to adjust their extent. Both ideas underlie the approach which is presented in Section 3.3.

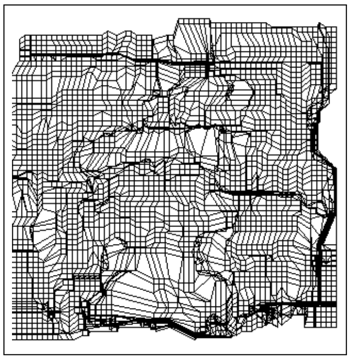


Figure 3.2: Example of deformed pixel grid containing discontinuities (TSDP-W*)

3.1.2 TSDP-W*

Absolute Position Constraints. The major drawback of the TSDP-W* stems from the absolute position constraints (c.f. Eq. (2.7) and (2.8)) since they make the performance of the approach be strongly dependent on the warping range. This limitation results in the problem of selecting an appropriate warping range. As generally the strength of a deformation is not known in advance, the optimal warping range must be chosen empirically, which leads to a large number of experiments and possible overfitting. At the same time, the complexity of the TSDP-W* grows dramatically with the increase of the warping range since in this case the number of pixel skips increases as well. Furthermore, a large number of pixel skips can result in unrealistic deformations of the reference image and consequent increase of the recognition error rate. Although the warping can be smoothed to some degree by an introduction of the relative position penalties, no hard continuity constraints are enforced. Figure 3.2 exemplifies the deformed pixel grid computed by the TSDP-W*. As it can clearly be seen in the picture, the grid contains significant discontinuities.

Taking these considerations into account, we propose to replace the absolute position constraints with the monotonicity and continuity constraints. We present our approach in Section 3.4.

Lack of dependencies between the trees' branches. The TSDP-W* finds the warping of pixels in individual pixel neighborhood trees, as we explain in Section 2.3.3. In each tree, the warping of pixels in a tree's branch is found independently from other branches. Figure 3.3 exemplifies a warping of an individual pixel neighborhood tree computed by the TSDP-W*. As it can directly be seen from the picture, the lack of dependencies between the branches results in a bad alignment of the branches with respect to each other.

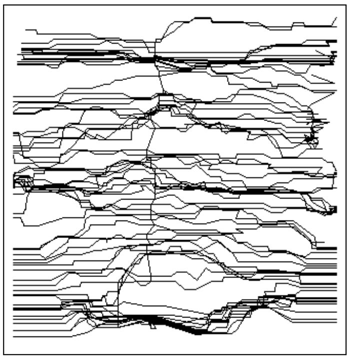


Figure 3.3: Example of a warping of an pixel neighborhood tree (TSDP-W*)

One way to approach this problem is to additionally consider the vertical dependencies between the displacements of pixels in the neighboring branches. However, in this case, the notion of trees would not exist any more. Moreover, the whole optimization problem would be similar to the 2DW, which is known to be NP-Complete (c.f. Section 2.2).

Therefore, we propose to iteratively improve the alignment of the branches by additionally considering the dependencies between the pixels in a branch and their bottom neighbors, and the positions of bottom neighbors are taken from the previous iteration. In this case, the alignment of trees' branches can be improved without raise of the complexity of a single iteration. The explained idea underlies the approach which is presented in Section 3.5.

3.2 Pseudo 2D Warping

The boundary constraints can be prohibitive, as we explain in Section 3.1. Here, we propose an approach which intends to overcome the negative effects of the boundary constraints in two different ways: by allowing a skip of more pixels and columns at once and by eliminating the boundary constraints. At the same time, our approach retains all advantages of the P2DHMM. Therefore, we call it Pseudo 2D Warping (P2DW).

First, in the P2DW, we propose to reduce the negative effect of the boundary constraints by allowing a skip of more pixels and columns at once than it is permitted in the P2DHMM. In this case, the P2DW first maps the boundary pixels of the test image to the boundary pixels of the reference image, and then is able to skip the areas of the reference image which are unrelevant for the warping.

Formally, we express the permission of a skip of more pixels at once in each column

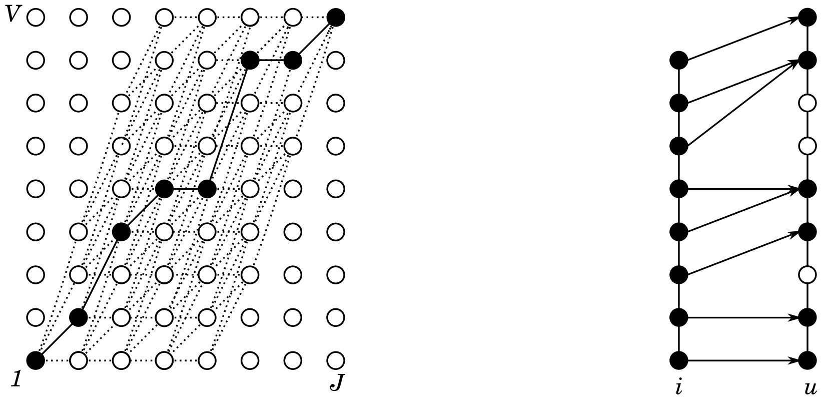


Figure 3.4: Example of a column alignment by means of the HMM ($skip = 2$)

through the following constraint:

$$0 \leq v_{i,j} - v_{i,j-1} \quad \leq \quad N + 1, \tag{3.1}$$

where N is a number of pixels to skip. This is a generalization of the constraint in Equation (2.15).

We propose to extend the HMM to allow a skip of at most N pixels at once. Figure 3.4 shows an example of a column alignment by means of the HMM which permits a skip of at most two pixels, i.e. when $N = 2$. As it can directly be seen from the picture, for each pixel, the number of possible warpings is increased.

We express the permission of a skip of more columns at once in the warping as follows:

$$0 \quad \leq \quad u_i - u_{i-1} \quad \leq N + 1, \tag{3.2}$$

where N is a number of columns to skip. This is a generalization of the constraint in Equation (2.24). Logically, with the increase of the parameter N, the complexity of the P2DW also increases.

However, skips of many pixels and columns at ones can cause discontinuities in the deformed pixel grid. On the other hand, the parameter N is not known in advance and have to be determined empirically, which can lead to overfitting. Therefore, we propose a second version of the P2DW in which the boundary constraints are completely eliminated.

Without boundary constraints, a skip of more than one pixel or column is unnecessary since large skips are only needed to omit unrelevant areas of the image right after matching the boundaries. Therefore, we propose to allow a skip of at most one

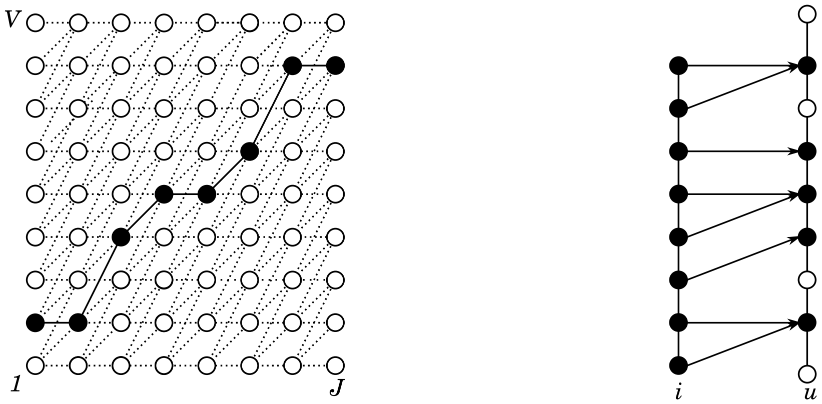


Figure 3.5: Example of a column alignment without boundary constraints

pixel or column in the warping, similarly to the P2DHMM. Figure 3.5 exemplifies an alignment of columns without boundary constraints. As it can clearly be seen from the picture, for each pixel, the number of possible warpings is maximal. Logically, the elimination of the boundary constraints increases the complexity of the P2DW.

As we explain in Section 2.3.2, the dependencies between the vertical displacements of the pixels in neighboring columns are ignored. Therefore, in the P2DW, we propose the separation of the penalty term in optimization criterion (c.f. Eq. (2.25)) in the following way:

$$T(u_{i-1}, u_i, v_{i,j-1}, v_{i,j}) \quad = \quad T(v_{i,j-1}, v_{i,j}) + J \cdot T(u_{i-1}, u_i) \tag{3.3}$$

The first component $T(v_{i,j-1}, v_{i,j})$ in Equation (3.3) describes the first-order dependencies between the displacements of the pixels in a column. The second penalty term $T(u_{i-1}, u_i)$ expresses the first-order dependencies between the displacements of entire columns. As a skip of an entire column has more effect on the warping than a skip of a pixel, we propose to weight the second penalty term with the number J of pixels in a column. The separation of the penalty term reflects the pseudo two-dimensional character of the P2DW. At the same time, this separation allows to introduce the penalties across the columns.

Taking into account the explained changes, we rewrite the optimization criterion (c.f. Eq. 2.25) in the following way:

$$\min_{\left\{u_i, v_{ij}\right\}} \Big\{ \sum_{i,j} [d_{i,j}(u_i, v_{i,j}) + T(v_{i,j-1}, v_{i,j}) + J \cdot T(u_{i-1}, u_i)] \Big\} \tag{3.4}$$

The optimization procedure is similar to the one described in Section 2.3.2). First, the scores

$$D(i, u_i) = \min_{\left\{v_{i,j}\right\}} \left\{ \sum_j [d_{i,j}(u_i, v_{i,j}) + T(v_{i,j-1}, v_{i,j})] \right\} \tag{3.5}$$

are computed. They express the best alignment of each pair of columns (i, u_i). Then, the ordering of columns in the final warping is found through optimization of the following criterion:

$$\min_{\left\{u_i\right\}} \left\{ \sum_i [D(i, u_i) + J \cdot T(u_{i-1}, u_i)] \right\} \tag{3.6}$$

Complexity. The complexity of the P2DW is increased compared to the P2DHMM. When the boundary constraints are eliminated, the complexity of the presented approach equals to $3IU(1 + JV)$, which is the upper bound for the complexity of the P2DHMM. When the boundary constraints are retained and a skip of at most N pixels and columns is allowed, the complexity of the P2DW is less than $(N + 2)IU(1 + JV)$. Additionally, the cost of distance computation has to be considered, which is exactly $IUJV$ in the first case, and less than $IUJV$ in the second case, respectively.

3.3 Strip Extension of the Pseudo 2D Warping

According to the constraint in Equation (2.23), a column in the test image is completely mapped to a single column in the reference image. As we explain in Section 3.1, this constraint can be too restrictive. Therefore, it can be relaxed by allowing additional *zero-order* one pixel deviations [Keysers & Deselaers$^+$ 07]. We propose to model the deviations in accordance with the monotonicity and continuity constraints. In the proposed approach, a column in the test image is mapped to a *strip* around a column in the reference image such that the *first-order* dependencies between the pixels in the strip are retained. Hence, we call this approach First-Order Strip Extension of the Pseudo 2D Warping (P2DW-FOSE). Additionally, we propose to allow larger deviations from a column than in the approach described in [Keysers & Deselaers$^+$ 07].

Formally, we express a deviation $u_{i,j}$ from a column u_i through the following constraint:

$$| \; u_{i,j} - u_i \; | \quad \leq \quad \frac{\Delta - 1}{2}, \tag{3.7}$$

where $\Delta = 1, 3, 5, ...$, is the width of the strip around a column u_i. In the case when $\Delta = 1$, no deviations are allowed, and the P2DW-FOSE is equal to the P2DW. The increase of the strip's width can result in the additional flexibility of the warping.

In a strip, we propose to model the first-order dependencies between the pixel displacements in accordance with the monotonicity and continuity constraints. Figure

3.6 exemplifies possible warpings of a pixel $w_{i,j}$ when $\Delta = 3$. As it can clearly be seen from the picture, only in the first case all 9 predecessors are considered, while in the second and in the third case the number of predecessors is reduced to 6 due to the constraint in Equation (3.7).

The first-order dependencies between the displacements of entire strips which correspond to the columns u_{i-1} and u_i, respectively, are expressed through the constraint in Equation (2.23). Therefore, a skip of at most one column is allowed.

Taking into account the explained changes, we propose to rewrite the optimization criterion in the following form:

$$\min_{\left\{u_i, w_{i,j}\right\}} \left\{ \sum_j [d_{i,j}(w_{i,j}) + T(u_{i-1}, u_i, w_{i,j-1}, w_{i,j})] \right\}, \tag{3.8}$$

where

$$T(u_{i-1}, u_i, w_{i,j-1}, w_{i,j}) = T(u_i, u_{i,j}) + T_v(w_{i,j-1}, w_{i,j}) + J \cdot T(u_{i-1}, u_i) \tag{3.9}$$

The first penalty term $T(u_i, u_{i,j})$ in Equation (3.9) penalizes the deviation $u_{i,j}$ of a pixel $w_{i,j}$ from the column u_i within the corresponding strip. The second penalty term $T_v(w_{i,j-1}, w_{i,j})$ models the first-order dependencies between the pixel displacements within a strip, while the last penalty term $T(u_{i-1}, u_i)$ is used to model the first-order dependencies between the displacements of the strips defined around the columns u_{i-1} and u_i, respectively. We propose to weight the last penalty term with the number J of pixels in a column.

The optimization of the criterion in Equation (3.8) is performed similarly to the algorithm described for the P2DHMM. First, the scores

$$D(i, u_i) = \min_{\left\{w_{i,j}\right\}} \left\{ \sum_j [d_{i,j}(w_{i,j}) + T(u_i, u_{i,j}) + T_v(w_{i,j-1}, w_{i,j})] \right\} \tag{3.10}$$

are computed. Then, the final solution is found through the optimization of the criterion in Equation (3.6).

In the P2DW-FOSE, we propose to overcome the negative effects of the boundary constraints through their complete elimination, as it is done in the second version of the P2DW.

Complexity. The complexity of the P2DW-FOSE is increased compared to the P2DW due to the permission of additional deviations from a column. For each pixel (i, j) in a column i in the test image, all pixels (u, v) in a strip around a column u in the reference image are considered, which is ΔJV. For each warping $w_{i,j}$ of the pixel (i, j), the warpings of at most 9 possible predecessors are taken into account (c.f. Fig. (3.6)).

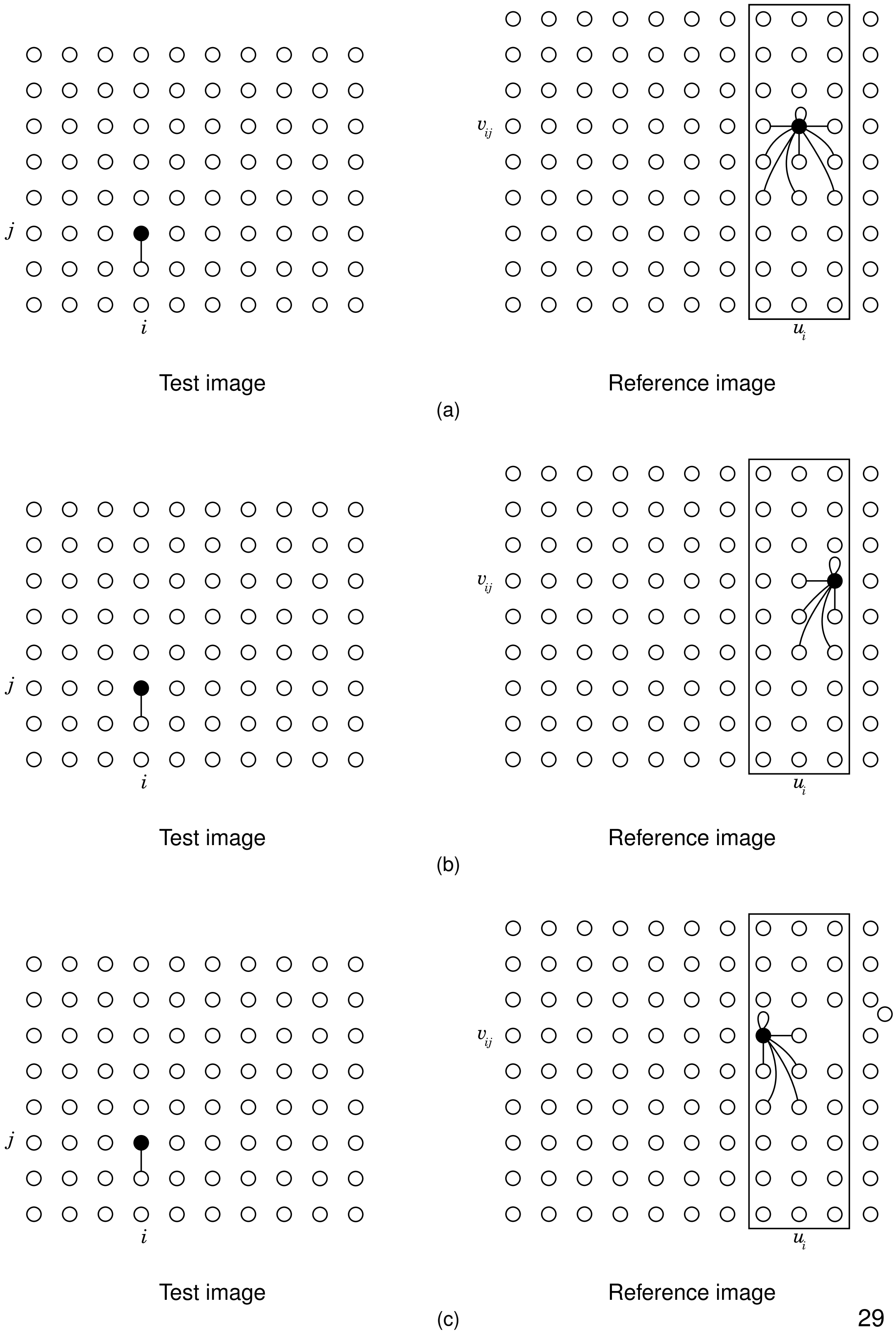


Figure 3.6: Illustration of permitted warpings in the P2DW-FOSE ($\Delta = 3$)

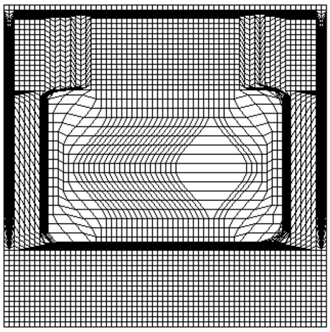


Figure 3.7: Example of violated constraints in the TSDP-1

Therefore, for each pair (i, u) of columns, less then $9\Delta JV$ comparisons are performed. Finally, as a skip of at most one column is allowed, the upper bound for the overall complexity of the P2DW-FOSE is $IU(3 + 9\Delta JV)$. It is summed up with the complexity of distance computation, which is $IUJV$.

3.4 Tree-Serial Dynamic Programming with One Skip

The main drawback of the TSDP-W* stems from the absolute position constraints, as we discuss in Section 3.1. Therefore, we propose to replace the absolute position constraints through the monotonicity and continuity constraints. In this case, a skip of at most one pixel is allowed. Therefore, we call the proposed approach Tree-Serial Dynamic Programming with One Skip (TSDP-1).

As we propose to eliminate the absolute position constraints, the performance of our approach does not depend on the warping range. Therefore, the TSDP-1 is able to compensate for deformations of various strength with less risk of overfitting.

Due to the monotonicity and continuity constraints, the TSDP-1 computes more restricted warping compared to the TSDP-W*, which makes unrealistic deformations of the reference image less probable and can result in the decrease of the recognition error rate.

However, the monotonicity and continuity constraints between the pixels in the neighboring columns can be violated, as each pixel tree is optimized independently of the others. Therefore, the deformation of the reference image may contain discontinuities. Figure 3.7 exemplifies the deformed pixel grid computed by the TSDP-1. It can clearly be seen from the picture that noticeable gaps occur. The described problem is a general drawback of the tree-serial concept. This shortcoming cannot be overcome without significant increase of the complexity. However, as we explain in Section 3.5, the amount of the violated constraints can be reduced.

Complexity. The overall complexity of the TSDP-1 sums up from the complexity of the forward, backward and bottom-up runs of dynamic programming, and the complexity of the distance computation. During the forward run of dynamic programming, for each pixel (i, j) in the test image each pixel (u, v) in the reference image is considered, which results in $IJUV$ comparisons. Additionally, for each displacement of a pixel $(u_{i,j}, v_{i,j})$ in the deformation of the reference image, the displacements of 9 possible predecessors are taken into account. Therefore, the complexity of the forward run is $9IJUV$. The complexities of the backward and bottom-up runs are similar to the complexity of the forward run. Hence, the compound complexity of the optimization is $3 \cdot 9IJUV$. Finally, the complexity of the distance computation is $IJUV$.

3.5 Iterative Tree-Serial Dynamic Programming

The lack of dependencies between the trees' branches in the TSDP can result in a bad alignment of branches with respect to each other, as we explain in Section 3.1. Therefore, we propose an approach which intends to overcome this drawback. In the proposed approach, the alignment of the trees' branches is iteratively improved by including the vertical dependencies between the pixels in a branch and their bottom neighbors during the forward-backward run, and positions of the bottom neighbors are taken from the previous iteration. We call the proposed approach Iterative Tree-Serial Dynamic Programming (ITSDP). In each iteration, we propose to compute the warping similarly to the TSDP-1 approach.

Formally, for each pixel (i,j) in a branch, we denote its warping as $w_{i,j}$ and the warping of its bottom neighbor as $\widetilde{w}_{i,j-1}$. The warping $\widetilde{w}_{i,j-1}$ is computed in the previous iteration. We propose to model the first-order dependency between the warpings $w_{i,j}$ and $\widetilde{w}_{i,j-1}$ by introducing the vertical penalty $T_v(\widetilde{w}_{i,j-1}, w_{i,j})$. As we propose to use the warping of the bottom neighbor which is computed in the previous iteration, the complexity of a single iteration does not increase. At the same time, all advantages of the tree structure are preserved. According to the explained changes, we denote a new type of a tree as an extended individual pixel neighborhood tree. An example of such a tree is shown in Figure 3.8.

In order to compute the warping in each iteration of the ITSDP, we propose to use the extended TSDP-1 which includes the penalties $T_v(\widetilde{w}_{i,j-1}, w_{i,j})$ during the forward-backward run of dynamic programming. Formally, we extend the optimization criterion (c.f. Eq. (2.28)) in the following way:

$$
\begin{aligned}
i^* : \min_{\{w_{ij}\}} \Big\{ \sum_{i,j} d_{ij}(w_{ij}) \quad + \quad & \sum_{j} \Big[T_v(w_{i^*,j-1}, w_{i^*j}) \\
& + \sum_{i} [T_h(w_{i-1,j}, w_{ij}) + T_v(\widetilde{w}_{i,j-1}, w_{i,j})] \Big] \Big\}
\end{aligned}
\tag{3.11}
$$

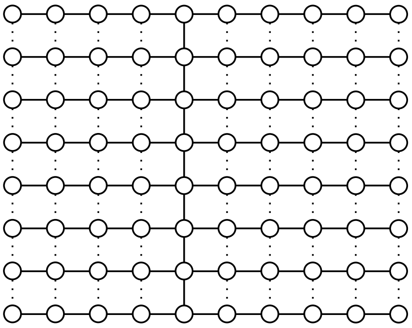


Figure 3.8: Illustration of extended pixel neighborhood tree in the ITSDP

Consequently, we extend the function in Equation (2.29) as follows:

$$D_{i^*,j}(w_{i^*,j}) = \min_{\{w_{ij}\}:w_{i^*,j}=w} \left\{ \sum_i \left[d_{ij}(w_{ij}) + T_h(w_{i-1,j}, w_{ij}) + T_v(\widetilde{w}_{i,j-1}, w_{ij}) \right] \right\} \quad (3.12)$$

Like every iterative procedure, the presented approach requires an appropriate initialization. Therefore, we propose to use the TSDP-1 for the computation of the initial warping.

The application of the proposed approach results in better alignment of branches with respect to each other, as we show in Chapter 5. In addition, the ITSDP can help to partially overcome the problem of violated constraints which is explained in Section 3.4. Experimental evaluation shows that with each new iteration of the proposed approach, the amount of violated constraints in the warping is significantly reduced.

Complexity. As we previously explained, the complexity of a single iteration of the presented approach is similar to the complexity of the TSDP-1, which is $3 \cdot 9IJUV$. Therefore, the compound complexity of the optimization procedures in the ITSDP is $M \cdot 3 \cdot 9IJUV$, where M is a number of iterations. Additionally, the complexity of the distance computation is $IJUV$.

Chapter 4

System Overview

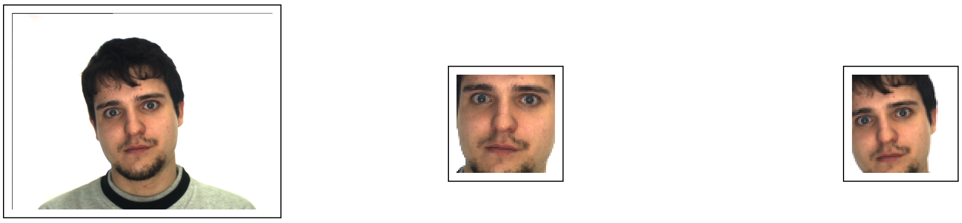
In this chapter, an overview of the recognition system is given. The system consists of a preprocessing step, feature extraction and decision rule, similar to the work [Keysers & Deselaers$^+$ 07].

The preprocessing step is performed to highlight the image content which is relevant for recognition. At the same time, the undesirable areas of an image are eliminated. For example, if the matching algorithms are used for face recognition, the only area of an image containing a face is relevant. Therefore, during the preprocessing step faces are extracted and then scaled to a needed resolution. Since preprocessing usually reduces the image data, it also helps to decrease the computing time of the recognition system.

During the feature extraction the image data is transformed into a set of feature descriptors. Interest point based feature extraction [Mikolajczyk & Schmid 05] and grid-based feature extraction [Dreuw & Steingrube$^+$ 09] are known. In the first case, the local feature descriptors are computed in a sparse way around interest points, while in the second case the local feature descriptors are extracted at all points of a regular grid, which provides a dense description of the image content. As matching algorithms proceed on the whole image pixel grid, we perform the grid-based feature extraction similar to the work [Dreuw & Steingrube$^+$ 09].

The local feature descriptors characterize a position in an image through its local neighborhood. As the local feature descriptors are distinctive and at the same time robust to changes in viewing conditions, they are especially suited for matching algorithms. Many different local descriptors have been proposed in the literature [Mikolajczyk & Schmid 05]. The Sobel features [Duda & Hart 73] has been shown to be a simple and efficient descriptor [Keysers & Gollan$^+$ 04b]. The SIFT descriptor [Lowe 04] is one of the best local feature descriptors since it is distinctive and relatively fast to compute. The SIFT has been used successfully for the tasks of face recognition [Zhang & Chen$^+$ 08, Dreuw & Steingrube$^+$ 09] and face authentication [Bicego & Lagorio$^+$ 06]. The discrete cosine transform (DCT) has been used as a feature extraction step in various studies on face recognition [Ekenel & Stiefelhagen 06, Hanselmann 09].

In the following, the recognition system is described in more detail. First, cropping



(a) Original photograph (b) Manually cropped face (c) Automatically cropped face

Figure 4.1: Example of the (a) original image, (b) a manually cropped and rectified face, (c) an automatically cropped face

and scaling procedures as parts of the preprocessing step are described. Then, the Sobel features, the SIFT and the DCT local feature descriptors are briefly explained. Finally, the decision rule is presented.

4.1 Preprocessing

Cropping. Cropping is defined as a process of removal of the outer parts of an image and extraction of some image region. Cropping helps to accentuate the image content which is relevant for recognition. As matching algorithms proceed on the whole pixel grid and deform the entire image, undesirable image areas, such as e.g. background has to be preliminarily eliminated. The decision on which region to preserve is made by a human or a detection procedure. An example for the latter case is the Viola and Jones Face Detector [Viola & Jones 04] which is used to determine the regions containing faces. However, in the case of automatic detection, registration errors can occur. Figure 4.1 shows an original photograph of a person, as well as examples of manually and automatically detected face image. As it can be seen in Figure 4.1(c), a registration error occurs, whereas manual detection is more precise (c.f. Fig. 4.1(b)).

The accuracy of cropping has direct influence on the performance of the matching algorithms discussed in this work. If both the test and the reference image are precisely extracted, the strength of the deformation is mostly determined by local changes in the image content. However, if registration errors occur, the disparity between the test and the reference image increases. Therefore, the warping requirements become more demanding and the reference image has to undergo stronger deformation.

Scaling. Cropping can noticeable reduce the size of an image. However, the resolution of an extracted region may still be high. In this case, application of the matching algorithms can be impractical due to the high computational complexity. Therefore, the size of extracted images must further be reduced. It can be done by scaling, which is the process of resizing a digital image through changing its pixel resolution. As we are only interested in reducing the image size, downsampling is performed. During this process, new pixel values are found through interpolation of old ones. Various interpolation schemes exist, ranging from simplistic procedures, such as the nearest neighbor algorithm, to sophisticated strategies, e.g. cubic spline interpolation [Moler 04].

Downsampling of the images allows to significantly reduce the computing time of the recognition system since the complexity of the discussed matching algorithms directly depends on the image resolution. However, as downsampling reduces the amount of an image data, much discriminative information is getting lost. This results in better matchings between the images from different classes and consequently leads to the increase of the recognition error. Therefore, the choice of an appropriate image scale must be viewed as a compromise between the quality of recognition and the computation complexity. As is shown in the work presented in [Hanselmann 09], 64×64 image resolution is optimal for the matching algorithms. In this case, the computing time is still feasible, while the quality of recognition is significantly improved in comparison to the results obtained on the 32×32-resoluted images.

4.2 Local Feature Descriptors

Local feature descriptors characterize a position in an image through its local neighborhood. They should be distinctive and at the same time robust to local deformations. One straightforward way to include the local image context is to use local subimages (e.g., of size 3×3 pixels) which are extracted around the position concerned. The optimal size of these patches depends on the image resolution and hence should be determined for each task individually. The inclusion of local context allows to smooth the warping and makes unrealistic deformations of the reference image less probable. As reported in the work presented in [Keysers & Gollan$^+$ 04b], the performance of the ZOW was significantly improved by the inclusion of local context. However, local image context is not distinctive by itself. Therefore, it should rather be combined with other local feature descriptors.

In the following, a couple of local feature descriptors are briefly explained. First, simple and efficient, the Sobel features [Forsyth & Ponce 02] are described. Then, the DCT-Features [Ekenel & Stiefelhagen 06] are explained which are compact and distinctive. Finally, the SIFT [Lowe 04] descriptor is briefly explained which is distinctive and relatively fast to compute.

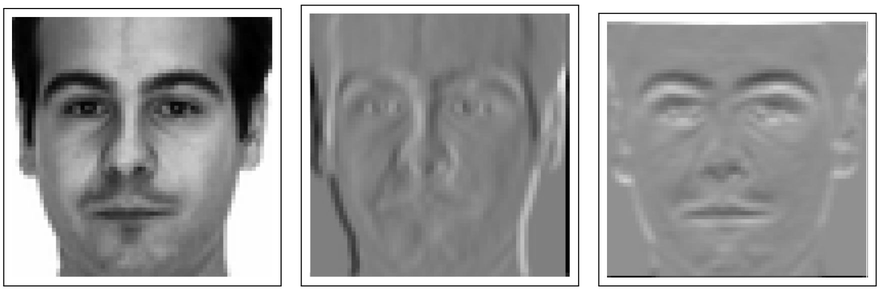
-1	0	1
-2	0	2
-1	0	1

(a)

-1	-2	-1
0	0	0
1	2	1

(b)

Figure 4.2: Illustration of horizontal (a) and vertical (b) Sobel filter



(a) (b) (c)

Figure 4.3: Example of original image (a), horizontal (b) and vertical (c) gradient image

4.2.1 Sobel Features

The Sobel feature [Duda & Hart 73] is a two-dimensional vector of derivatives which are computed by the horizontal and vertical Sobel filter from the image values. Figure 4.2 shows both filters, while the result of their application is exemplified in Figure 4.3. Due to the form of the filters, the obtained Sobel features include local context of 3×3 pixels and offer invariant properties with respect to the absolute image brightness. Moreover, the Sobel filters highlight the image structure, which partially prevents the matching algorithms from finding unrealistic image deformations. As reported in the work presented in [Keysers & Gollan$^+$ 04b], the Sobel features can significantly improve the performance of the ZOW.

4.2.2 Discrete Cosine Transform (DCT)

The DCT local feature descriptor is a vector of DCT coefficients, which are computed by the discrete cosine transform. In the local appearance based approach [Ekenel & Stiefelhagen 06], an image is divided into 8x8 pixel blocks, and the discrete cosine transform is performed on each block. The obtained DCT coefficients are ordered by a zig-zag pattern. The first coefficient represents the average intensity of the image, whereas the second and the third coefficients capture the average vertical and horizontal intensity changes, respectively. According to the feature selection strategy, K coefficients are selected resulting in K-dimensional local feature vector. As DCT descriptors should be invariant to illumination variations, three feature selection strategies are suggested: selecting the first K DCT coefficients (dct-all), removing the first DCT coefficient and selecting the first K coefficients from the remaining ones (dct-0), or removing the first three DCT coefficients and selecting the first K coefficients from the remaining ones (dct-3). As it was shown in [Ekenel & Stiefelhagen 06], dct-0 selection strategy provided the lowest recognition error rates under strong changes in illumination conditions.

Recently, in the work presented in [Hanselmann 09], the DCT local feature descriptors were used in combination with the ZOW and P2DHMM matching algorithm. It was shown that allowing an overlap of 8×8 pixel image blocks on which the DCT is performed helps to noticeable improve the performance of both approaches. It can be explained by the fact that with the increasing overlap more local information can be captured. The best results were achieved for dct-3 feature selection strategy when overlap of blocks was maximal. Therefore, in this case, the DCT local feature descriptors are invariant to illumination variations and robust to local deformations.

4.2.3 Scale Invariant Feature Transform (SIFT)

The SIFT descriptor is a 128-dimensional vector which stores the gradients of 4x4 locations around a pixel in a histogram of 8 main orientations [Lowe 04]. The gradients are aligned with respect to the main direction which makes the SIFT descriptor rotation invariant. As the vector is computed in different Gaussian scale spaces, the SIFT descriptor is also scale invariant.

However, in this work, the scale invariant properties of the SIFT descriptor are not relevant since the changes in scale are insignificant. Moreover, as is shown in [Dreuw & Steingrube^{+} 09], an upright version of the SIFT (U-SIFT) can outperform the rotation invariant descriptor in the task of face recognition. In Chapter 5, we investigate the impact of using the upright version of the SIFT descriptor on the performance of the discussed matching algorithms.

4.3 Recognition by Warping

As the main emphasis in this work lies on different matching algorithms based on various deformation models, simple decision rule for recognition is chosen. Similar to the work [Keysers & Deselaers$^+$ 07], a Nearest Neighbor (NN) scheme is used which has been shown to provide good results in various applications. Formally, given a test image X and a reference data set of images $R_{k1}, ..., R_{k,N_k}$ for classes $k = 1, ..., K$, NN decision rule performs an assignment:

$$X \rightarrow \widehat{k}(X) = arg\min_{k} \Big\{ \min_{n=1,\dots,N_k} D(X, R_{k,n}) \Big\}, \tag{4.1}$$

where $D(X, R_{k,n})$ is the cost of the warping of the reference image $R_{k,n}$ with respect to the test image X. This decision rule does not introduce new parameters and allows to directly observe the performance of the matching algorithms in terms of the recognition error rate.

In the case of the TSDP-W* and TSDP-1, a *recognition score* $D_{rec}(X, R_{k,n})$ is used instead of $D(X, R_{k,n})$. The recognition score is computed after the warping is found. Formally, it is defined as follows:

$$D_{rec}(X, R_{k,n}) = \sum_{i,j} \Big[d_{ij}(w_{ij}) + T_v(w_{i,j-1}, w_{i,j}) + T_h(w_{i-1,j}, w_{i,j}) \Big] \tag{4.2}$$

The score sums up from the accumulated pixel distances between the test image and the deformed reference image and the accumulated penalty costs.

Chapter 5

Experimental Evaluation

In this chapter, we study the performance of the different matching algorithms from two points of view. First, we qualitatively evaluate the discussed approaches on synthetic and real examples and visually compare the warpings. Then, we perform a quantitative evaluation of the presented matching algorithms on two face datasets: the AR Face [Martinez & Benavente 98] and the Labeled Faces in the Wild (LWF) [Huang & Mattar$^+$ 08] database. The AR Face dataset contains face images taken under strictly controlled conditions. In contrast, the LFW database was designed to study unconstrained face recognition problems. We provide the results of the evaluation on both datasets in terms of the recognition error rate.

5.1 Qualitative Evaluation

For the qualitative evaluation of the matching algorithms we selected two types of images: synthetic examples and face images. We use simple synthetic images to study the robustness of the presented approaches to linear image deformations caused by rotation, translation and scaling. We compare the warpings produced by the approaches. In the case of face images we analyze the ability of the matching algorithms to cope with local changes, in particular with differences in facial expression. We compare image deformations computed by the approaches and analyze the visual quality of the results (e.g. presence of discontinuities in the warping, appearance of artefacts). As we also aim at obtaining a warping which characterizes the similarity of images showing the same person and dissimilarity of the pictures in the opposite case, we additionally test the ability of the matching algorithms to compute "good" unwanted inter-class deformations, which can negatively affect the recognition results.

In the following, we evaluate the approaches on both types of images. First, we provide a comparison of the matching algorithms on simple synthetic images. Then, we present the results obtained on face images.

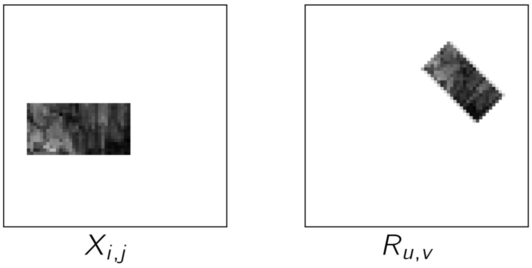
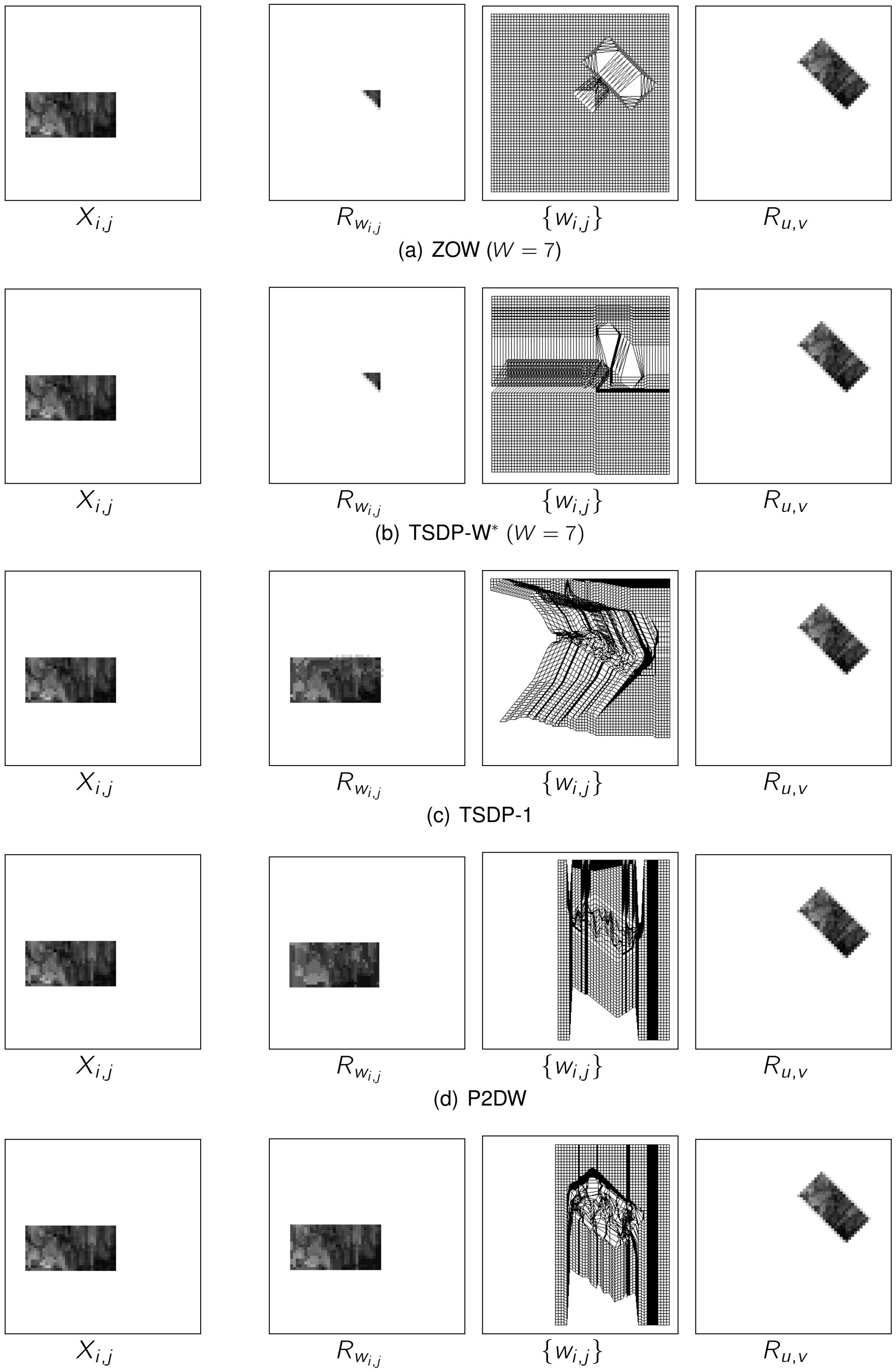


Figure 5.1: Synthetic examples

5.1.1 Synthetic Examples

Here, we qualitatively evaluate the matching algorithms discussed in this work on synthetic examples. We analyze the ability of the approaches to cope with linear image deformations, namely rotation, translation and scaling. For that purpose, we created a test and a reference synthetic image of dimension 64×64, each showing a dark textured rectangle on a white background. Figure 5.1 shows the synthetic images. The rectangle in the test image $X_{i,j}$ is 32×16 pixels large and horizontally oriented . Downscaling, shift and rotation of the rectangle in the test image results in the rectangle shown in the reference image $R_{u,v}$. As we explain in Section 4.1, difference in the background of the test and reference image can negatively affect the performance of the matching algorithms. Therefore, we leave the white background of the synthetic images unchanged.

Evaluation. For the experiments, we empirically optimize penalty weights. For the TSDP-W*, we select the maximal possible warping range such that the computing time of this approach is not significantly increased compared to the TSDP-1 and P2DW-FOSE. Informal experiments show that this is the warping range of 7 pixels. In order to make the comparison of both approaches with the absolute position constraints fair, we select the same warping range for the ZOW. In the case of the P2DW-FOSE, we empirically find that the strip width of 5 pixels is optimal. The obtained results are shown in Figure 5.2. It can directly be seen in Figures 5.2(a) and 5.2(b) that both the ZOW and TSDP-1 cannot find an appropriate deformation of the reference image since the warping range is too small, which clearly shows that the absolute position constraints are prohibitive. In contrast, the approaches whose performance does not depend on the warping range are able to compensate for a large disparity between the images. The warped pixel grid computed by the TSDP-1 (c.f. Fig. 5.2(c)) reveals the traces of the image deformation where the rectangle in the reference image is shifted down and left, rotated and upscaled at the same time. The dark artefacts correspond to the areas where several pixels in the test image



(a) ZOW ($W = 7$)

(b) TSDP-W* ($W = 7$)

(c) TSDP-1

(d) P2DW

(e) P2DW-FOSE ($\Delta = 5$)

Figure 5.2: Comparison of matching algorithms on synthetic examples

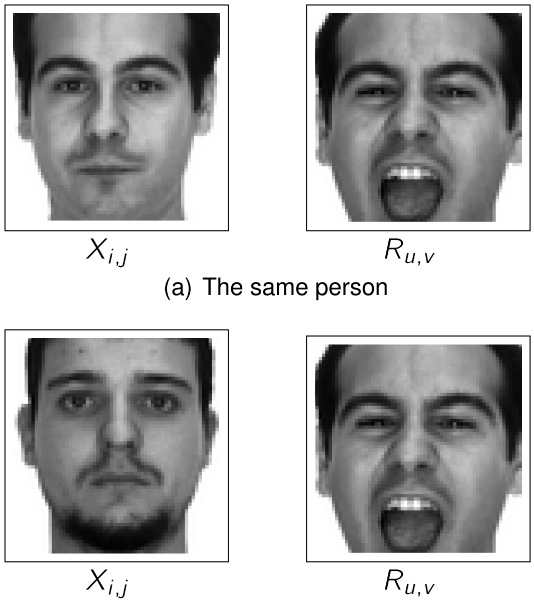
are mapped to the same pixel in the reference image. The same artefacts can be seen in the deformed pixel grids computed by the P2DW and P2DW-FOSE (c.f. Fig. 5.2(d) and 5.2(e), respectively). However, both warped grids look completely different compared to the TSDP-1. As no boundary constraints are imposed, both the P2DW and P2DW-FOSE skip unrelevant areas of the reference image and completely deform the rest. The deformation is performed by shifting the entire columns containing the parts of the rectangle in the vertical direction. It can also be seen that the shift of some columns is especially strong compared to their neighbors. This is possible since the vertical displacements between the pixels in the neighboring columns are ignored. In the case of the P2DW-FOSE, slight deviations from columns appear in the area containing the rectangle.

Although the deformations of the reference image computed by the last three approaches look very similar to the test image, slight differences in the texture can be seen. As both the TSDP-1 and P2DW-FOSE are less restricted than the P2DW, they are able to reconstruct the texture of the deformed reference image with higher accuracy. Indeed, the quality of reconstruction performed by the P2DW is visibly worse due to the limitation stemming from the column-to-column mapping.

Conclusion. According to the results obtained in these experiments, the absolute position constraints are clearly shown to be prohibitive: both the ZOW and TSDP-W* are not able to compensate for a large disparity between the images when the warping range is not large enough, while the increase of the warping range would significantly raise the computing time of the TSDP-W*, as we show in Section 5.2.5. In contrast, the TSDP-1, P2DW and P2DW-FOSE can find the deformation of the reference image which looks very similar to the test image. The limitation of the P2DW caused by the requirement of the column-to-column mapping is obvious, as this approach is not able to reconstruct the texture of the rectangle as good as the P2DW-FOSE. Finally, only the TSDP-1 deforms the entire reference image, while the P2DW and P2DW-FOSE mainly deform the area containing the rectangle.

5.1.2 Real Examples

In order to compare the qualitative performance of the discussed matching algorithms on real examples, we evaluate the approaches on face images. We analyze the ability of the approaches to cope with changes in facial expression, as this is one of the most challenging problems in face recognition [Ekenel & Stiefelhagen 09, Wolf & Hassner^{+} 08]. For that purpose, we selected two photographs of the same person with different facial expressions. The first photograph which is a test image $X_{i,j}$ represents a neutral facial expression. The second photograph being a reference image $R_{u,v}$ shows a screaming face. Figure 5.3(a) exemplifies the described images.



(a) The same person

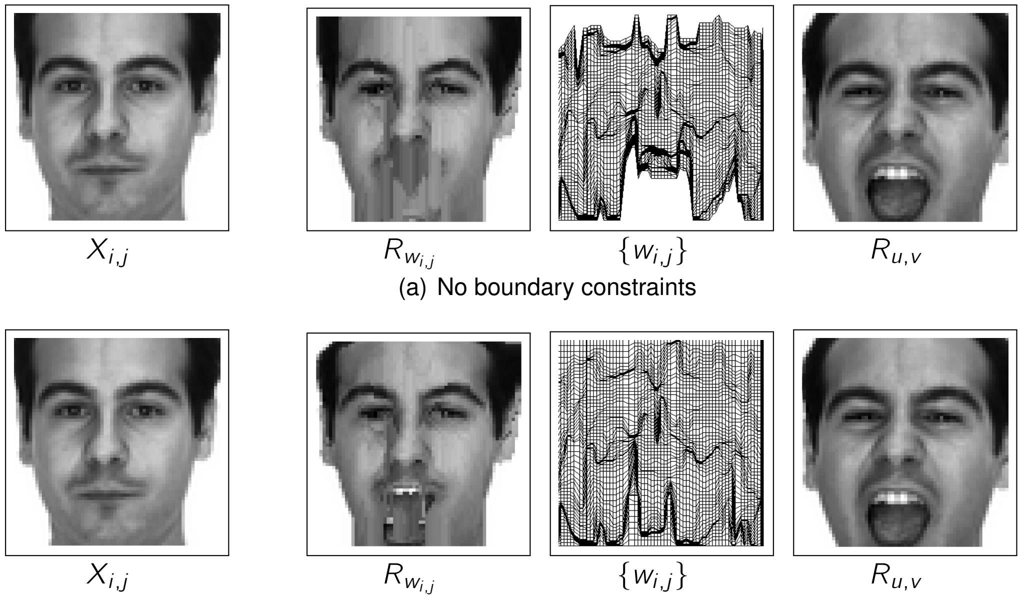
(b) Different persons

Figure 5.3: Illustration of face images used for evaluation of matching algorithms

As the boundary constraints can be too restrictive in this experiments, we additionally study the effects of the boundary constraints in the P2DW on the warping.

We also analyze the ability of matching algorithms to compute unwanted "good" inter-class deformations, which can negatively affect the recognition. Therefore, we additionally selected another test image which shows a face of a different person. Figure 5.3(b) exemplifies the photographs of different persons. For all experiments, we include a local context of 3×3 pixels to smooth the deformation of the reference image.

Boundary Constraints. As we explain in Section 3.1, the boundary constrains can be too restrictive. To verify this supposition, we study the effects of the boundary constraints on the warping. For that purpose, we apply the P2DW with and without boundary constraints to the images of the same person. The results are shown in Figure 5.4. As it can be seen in Figure 5.4(a), when no boundary constraints are enforced, the deformed pixel grid $\{w_{i,j}\}$ contains evident offsets from the image borders. The offsets are especially large in the central area, where a couple of columns are shifted up. A large difference between the vertical positions of some neighboring columns can be explained by the fact that the vertical displacements between the pixels in neighboring columns are ignored. This allows the P2DW to compensate for the



(a) No boundary constraints

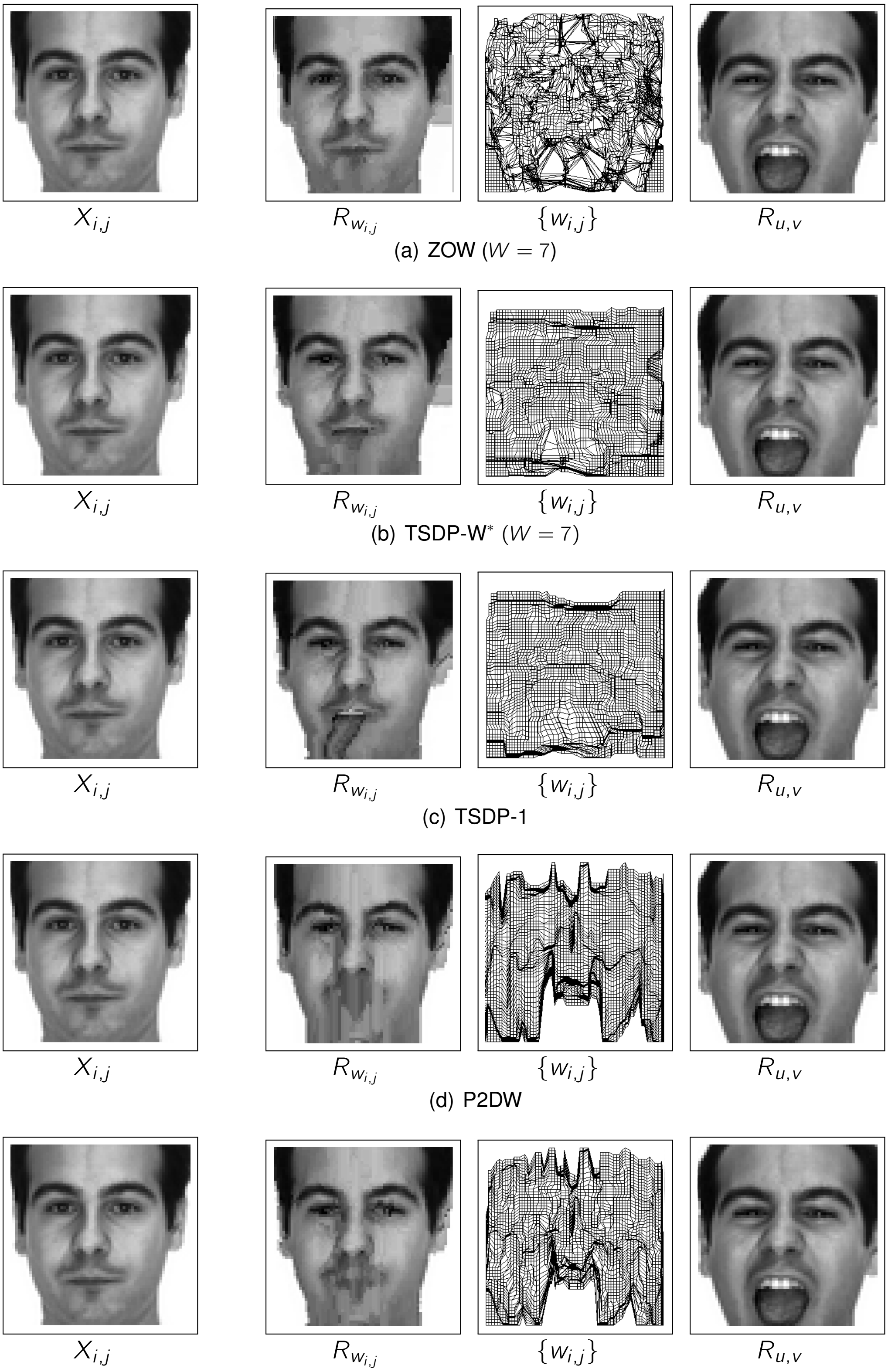
(b) Boundary constraints

Figure 5.4: Effects of boundary constraints in the P2DW on the warping (the same person)

presence of teeth by starting the warping of some columns directly above the mouth of the screaming face. If the boundary constraints are enforced during the warping, the borders of both images are matched, as shown in Figure 5.4(b). Since a skip of at most one pixel is allowed, the P2DW with the boundary constraints cannot completely omit the area containing the teeth. As a result, white artefacts are obvious in the deformation $R_{w_{i,j}}$ of the reference image $R_{u,v}$. According to these results, it can be seen that the boundary constraints can negatively affect the warping.

Intra-class Deformations. After studying the effects of the boundary constraints, we evaluate the matching algorithms on the same pair of images. Figure 5.5 shows the results obtained by the application of each algorithm. As shown in Figure 5.5(a), the deformation $R_{w_{i,j}}$ of the reference image $R_{u,v}$ produced by the ZOW looks similar to the test image. However, the deformed pixel grid contains evident discontinuities due to the unconstrained deformation of the reference image within the warping range. In particular, large gaps can be seen in the area containing the mouth of the person. The TSDP-W* also computes a deformation of the reference image which is similar to the test image (c.f. Fig. 5.5(b)), but in contrast to the ZOW, the amount of discontinuities in the deformed pixel grid is reduced. This can be explained by the fact that relative position penalties are included into the optimization procedure, which smoothes the warping. The TSDP-1 computes more restricted deformation of the reference image due to the monotonicity and continuity constraints imposed on the warping. As a result, the deformed pixel grid in Figure 5.5(c) contains very few discontinuities. The TSDP-1 is not able to completely reconstruct the area containing the mouth and a couple of obvious artefacts occur in the deformed reference image $R_{w_{i,j}}$. The first artefact is the white stripe which remains from the teeth and the second artefact is the dark area which remains from the mouth. The TSDP-1 cannot completely omit the corresponding areas of the reference image since a skip of at most one pixel is allowed during the forward-backward and the bottom-up optimization procedures. In contrast, the P2DW and the P2DW-FOSE are able to skip the whole area containing the mouth since the vertical displacements between the pixels in neighboring columns are neglected. Both the P2DW and the P2DW-FOSE produce similar results which are shown in Figures 5.5(d) and 5.5(e), respectively. The deformation of the reference image computed by the P2DW has evident vertical artefacts due to the constraint that a column in the test image is completely mapped to a single column in the reference image. The vertical artefacts are smoothed in the warping computed by the P2DW-FOSE since additional deviations from a column are allowed. According to the obtained results, all approaches are able to compensate for changes in facial expression for the selected pair of images. However, the warpings computed by the TSDP-1, P2DW and P2DW-FOSE look more restricted than ones found by the ZOW and the TSDP-W*.

Inter-class Deformations. In order to test the ability of the matching algorithms to compute unwanted good inter-class deformations, we evaluate the discussed approaches on a pair of images which contain the faces of different persons. The results are presented in Figure 5.6. As it can be seen from the pictures, both the ZOW and TSDP-W* compute a deformation of the reference image which is very similar to the test image (c.f. Fig. 5.6(a) and 5.6(b), respectively). Due to the lack of the continuity constraints, both approaches produce unrestricted warping within the warping range, which results in visually good inter-class deformations. In contrast, the TSDP-1, P2DW and P2DW-FOSE produce a deformation of the reference image which is unsimilar to the test image. This is shown in Figures 5.6(c), 5.6(d) and 5.6(e). According to the results presented in Figure 5.6, the approaches with the monotonicity and continuity constraints are unable to find good inter-class image deformations. In contrast, the approaches with the absolute position constraints compute deformations of the reference image which look very similar to the test image although both the test and the reference image show different persons.



(a) ZOW ($W = 7$)

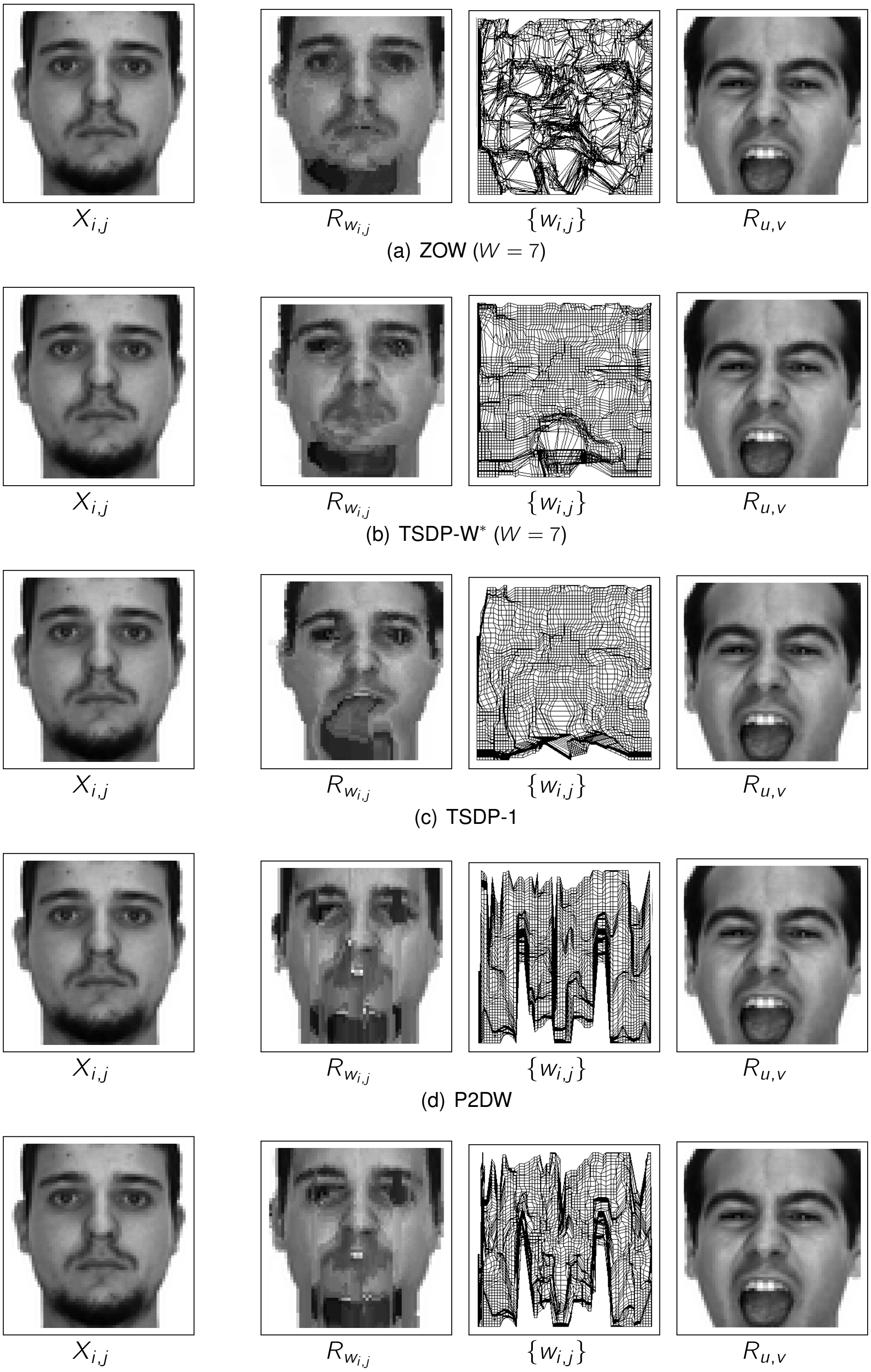
(b) TSDP-W* ($W = 7$)

(c) TSDP-1

(d) P2DW

(e) P2DW-FOSE ($\Delta = 3$)

Figure 5.5: Comparison of matching algorithms (the same person)



(a) ZOW ($W = 7$)

(b) TSDP-W* ($W = 7$)

(c) TSDP-1

(d) P2DW

(e) P2DW-FOSE ($\Delta = 3$)

Figure 5.6: Comparison of matching algorithms (different persons)

Conclusion. According to the results of the qualitative evaluation of the discussed matching algorithms on face images, all approaches are able to cope with the changes in facial expression. However, the warpings computed by the TSDP-1, P2DW and P2DW-FOSE look more restricted than ones produced by the ZOW and TSDP-W*. Moreover, in the case when the test and the reference image show different persons, both the ZOW and the TSDP-W* compute the deformation of the reference image which looks very similar to the test image. This can negatively affect the recognition. At the same time, the approaches which impose the monotonicity and continuity constraints on the warping cannot find unwanted good inter-class deformations.

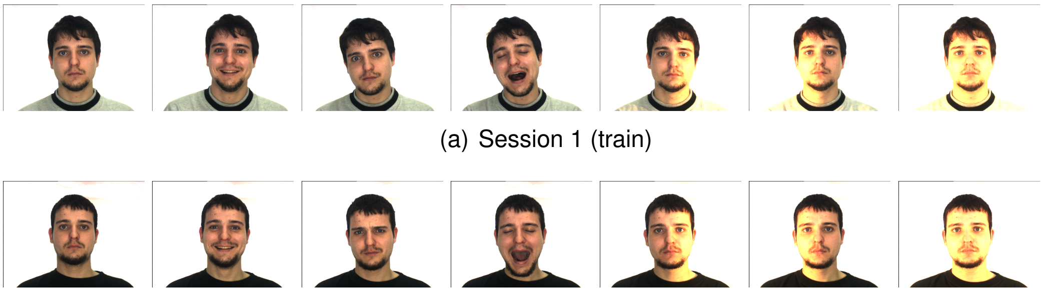
5.2 Quantitative Evaluation

In this section, we present the results of the quantitative evaluation of the presented matching algorithms. As one of the objectives of this work is to compare the performance of the matching algorithms in the task of face recognition, we selected two face datasets: AR Face [Martinez & Benavente 98] and Labeled Faces in the Wild (LWF) [Huang & Mattar$^+$ 08] database. The AR Face dataset consists of the frontal view face images taken under strictly controlled conditions and is a standard benchmark for face recognition approaches. We selected the AR Face database for the thorough evaluation of the matching algorithms because of many local deformations appearing in the images due to changes in facial expression. As we also aim at comparing the performance of the matching algorithms in the task of unconstrained face recognition, we selected the LWF dataset where the number of face variations is uncontrolled.

In the following, we evaluate the matching algorithms on both datasets. First, we use the AR Face database to study the effects of different parameter settings on the recognition error rate. Than, we report the performance of the matching algorithms on the LFW dataset.

5.2.1 AR Face Database

The AR Face Database [Martinez & Benavente 98] contains frontal view face images with different facial expressions, illumination conditions, and occlusions. 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 [Ekenel & Stiefelhagen 09], only a subset of 110 individuals is used in experiments. Moreover, we discarded the images with partially occluded faces since we do not aim at studying the ability of the matching algorithms to reconstruct missing details. Only images of faces showing smile, anger, scream and neutral expression, as well as those with illumination from



(a) Session 1 (train)

(b) Session 2 (test)

Figure 5.7: Example subset of face images of the same person (AR Face database)

left, right and both sides simultaneously, are chosen. Figure 5.7 represents an example subset of face images of one person. In total, 770 images from the first session are used for training, and the same number of pictures taken in the second session is selected for testing.

For our experiments, we use the AR Face database in two conditions: with automatically and manually cropped faces.

5.2.2 Automatically Cropped Faces

Face registration errors can have a large impact on the overall recognition quality [Rentzeperis & Stergiou$^+$ 06]. The aim of this section is to show that the matching algorithms are not only able to cope with variability in facial expression and illumination, but are also robust to registration errors. For that purpose, the faces are detected by means of the OpenCV implementation of the Viola and Jones face detector [Viola & Jones 04], cropped and scaled to 64×64 pixel resolution, similar to the work [Dreuw & Steingrube$^+$ 09]. Examples of automatically cropped face images are represented in Figure 5.8. As we first aim at studying the performance of different warping approaches which is not extremely facilitated by a sophisticated local feature descriptor, we select simple Sobel features. We extract the vertical and horizontal Sobel features from each image, as we describe in Chapter 4.

For the evaluation of the matching algorithms, we use the L_1-Norm as a distance function since it was shown in informal experiments to lead to better recognition results compared to the Euclidean distance.

In the following, we study the impact of different parameter settings on the recognition error rate in order to obtain the best possible performance of the matching algorithms. First, we analyze the effects of penalty and local context. Then, we evaluate the warping range and study the impact of the boundary constraints. Additionally, we show the effects of iterations. Finally, we summarize the best recognition results.

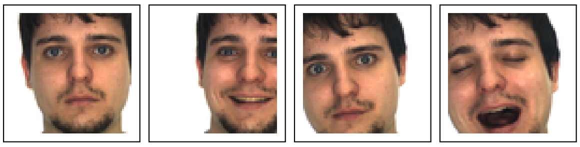


Figure 5.8: Examples of automatically cropped face images (AR Face database)

Table 5.1: Effects of penalties

Matching algorithm	**ER [%]**			
	no pen.	pen. $\sim d_{abs}$	pen. $\sim d_{euc}$	pen. $\sim d_{euc^2}$
ZOW ($W = 7$)	29.0	20.3	**19.9**	24.5
TSDP-W* ($W = 7$)	29.0	5.2	**4.9**	9.2
TSDP-1	11.4	6.4	**6.2**	9.1
P2DW	9.6	9.0	**9.0**	9.0

5.2.3 Penalty

We start the evaluation of the matching algorithms with studying the effects of penalties on the recognition error rate. First, we do not include penalty in the optimization criteria. Recognition results for selected algorithms are shown in the second column of Table 5.1. The best result is achieved by the P2DW. In comparison to the other approaches, the P2DW produces the most restricted warping. Therefore, unrealistic deformations of the reference image are rather improbable, which positively affects the recognition error rate. The next fact which is worth attention is that both the ZOW and TSDP-W* perform similarly. Although the optimization strategies are different, both approaches impose absolute position constraints on the warping. Within the warping range, the decision on the warping of each pixel is affected by a pixel distance and a warping penalty. However, if the penalty is not included in the optimization procedure, the warping of each pixel is found through the minimization of the pixel distance. Hence, both algorithms find the same deformation of the reference image.

The inclusion of the warping penalty in the optimization criteria helps to improve the results. As it can be seen from Table 5.1, the most significant reduction of the recognition error rate is achieved for the TSDP-W*. The penalty function based on the Euclidean distance allows to reduce the error rate from 29% down to 4.9%. Experiments show that using this kind of the penalty function leads to the best results for all approaches. The lowest error rates are emphasized in bold in Table 5.1. The quadratic Euclidean distance and the L_1-Norm, if used as penalty functions, also lead to significant decrease of the error rate. However, in the case of the L_1-Norm, the results are insignificantly worse compared to the Euclidean distance, while the use of

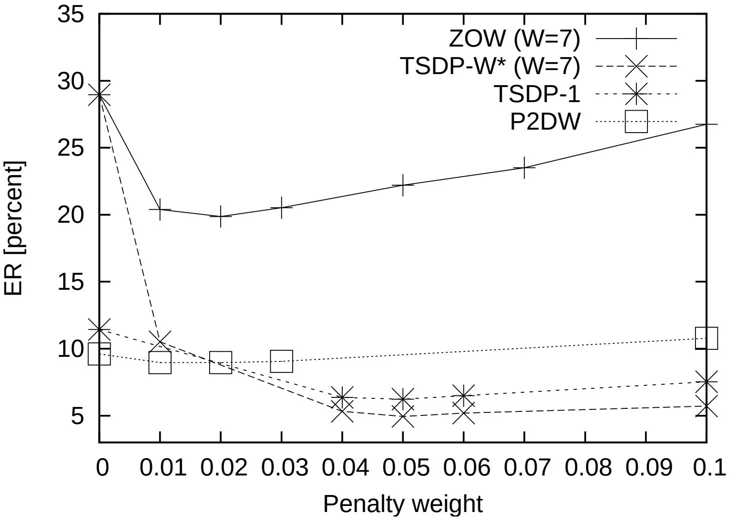


Figure 5.9: Effects of penalty weights, pen. $\sim d_{euc}$

the quadratic Euclidean distance leads to the noticeable increase of the recognition error rate. This can be explained by the fact that the quadratic Euclidean distance provides significantly larger penalty cost for the distortions of each pixel than the other two functions. Therefore, the deformation of the reference image is too rigid. The choice of the penalty function does not affect the performance of the P2DW, if a skip of at most one pixel and column is allowed. In this case, the arguments of both penalty functions, within and across the columns, can only take values of 0 and 1. According to the obtained results, we select the Euclidean distance as a penalty function for all further evaluations of the matching algorithms.

In order to find the optimal penalty weight, we study the effects of the penalty weights on the recognition error rate. In the extreme case, if the cost of each pixel displacement is infinitely large, no deformation of the reference image is produced. On the other hand, the decrease of the penalty weights reduce the influence of penalties on the warping. This dependency is shown in Figure 5.9 for the case when the Euclidean distance is used as a penalty function. As it can be seen from the picture, we find the optimal weight by choosing a good heuristic value and additionally trying previous and next values on the axis with a step of 0.01. It can also be seen in Figure 5.9 that for both tree-serial approaches the optimal penalty weights are similar. Finally, the choice of the penalty weight on this interval strongly affects the performance of the ZOW, but causes insignificant changes of the error rate in the case of the P2DW.

Table 5.2: Effects of local context

Matching algorithm	Local context	ER [%]
ZOW ($W = 10$)	1x1	18.9
	3x3	11.2
	5x5	**7.5**
TSDP-W* ($W = 7$)	1x1	**4.9**
	3x3	5.8
	5x5	6.5
TSDP-1	1x1	**6.2**
	3x3	6.5
	5x5	7.4
P2DW	1x1	9.0
	3x3	8.2
	5x5	**7.5**
P2DW-FOSE ($\Delta = 3$)	1x1	7.8
	3x3	**6.2**
	5x5	6.4

5.2.4 Local Context

The inclusion of local context allows to smooth the warping and makes unrealistic deformations of the reference image less probable. As we aim at finding the optimal local context for each approach, we study the effects of local context on the recognition error rate. For that purpose, we vary the size of local context and report the performance of the matching algorithms in each particular case. We consider local contexts of 3×3 and 5×5 pixels. Additionally, we provide the performance of the matching algorithms in the case when no local context (1×1) is taken into account. For each local context and each approach, we empirically find the optimal penalty weight. The results are listed in Table 5.2. As it can be seen from the table, the inclusion of local context leads to an improvement of the recognition results for all approaches except the TSDP-W* and the TSDP-1 for which the results get worse. It can be explained by the specifics of the optimization strategy in both methods. Due to the forward-backward and the bottom-up runs of dynamic programming in the TSDP-W* and TSDP-1, more context information is implicitly considered in comparison to the other approaches. Therefore, for both approaches, the explicit inclusion of local context can make the deformation of the reference image more rigid, which negatively affects the recognition error rate.

The inclusion of local context helps to significantly improve the performance of the ZOW, as many unrealistic deformations of the reference image can be avoided. For both the P2DW-FOSE and P2DW, the inclusion of local context is also advantageous.

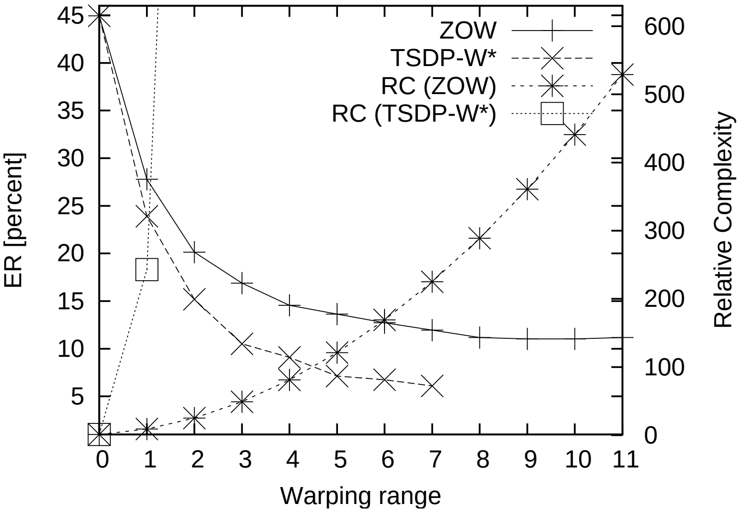


Figure 5.10: Effects of warping range on recognition error rate and relative complexity of both approaches (relative to L_1-Norm computation)

However, in comparison to the ZOW, the relative improvement of the performance of both approaches is less significant. This can be explained by the fact that the warping computed by the P2DW-FOSE and P2DW is more constrained than one found by the ZOW. Therefore, for both the P2DW-FOSE and P2DW, the computation of an unrealistic image deformation is rather improbable, even without inclusion of the local context.

5.2.5 Warping Range

The performance of the ZOW and TSDP-W* strongly depends on the warping range. At the same time, the complexity of the ZOW grows quadratically with the increase of the warping range, while the complexity of the TSDP-W* raises to the power of four. Therefore, we study effects of the warping range on the recognition error rate and additionally provide the complexity of both approaches for each warping range relatively to the complexity of the L_1-Norm computation. We gradually increase the warping range and optimize the penalty weights in each particular case. Local context of 3×3 pixels is included for each approach since it helps to significantly improve the recognition results obtained by the ZOW. The performance of both approaches is represented in Figure 5.10. As it can be seen from the picture, with the increase of the warping range, the recognition error decreases for both the ZOW and TSDP-W*. It can be explained by the fact that both approaches are able to compensate

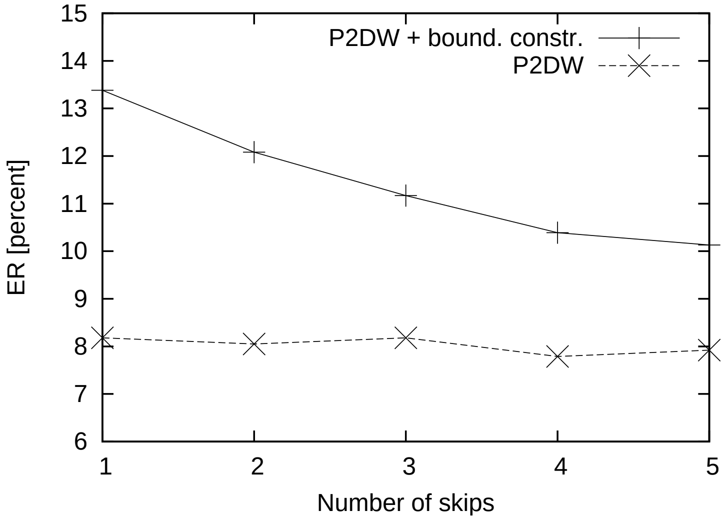


Figure 5.11: Effects of boundary constraints

for stronger image deformations. At the same time, the graphic shows that when the warping range is large enough, no further increase of it helps to noticeably reduce the error. It can also be seen in Figure 5.10 that the complexity of the TSDP-W* grows dramatically with the increase of the warping range since the number of pixel skips also increases. Therefore, we assume the warping range of 7 pixels to be optimal for the TSDP-W* although a further increase of it can insignificantly decrease the error rate. For the ZOW, the optimal warping range is 10 pixels. It should be emphasized here that the warping range determined for each approach is only optimal on the AR Face database with automatically cropped faces since this parameter is extremely task dependent. At the same time, optimizing the warping range on the test data in order to obtain low recognition error rate may possibly lead to overfitting.

5.2.6 Boundary Constraints

The results of the qualitative evaluation of the matching algorithms in Section 5.1.2 show that the boundary constraints can negatively affect the warping. Here, we study effects of the boundary constraints on the recognition error rate. For that purpose, we evaluate the P2DW with and without boundary constraints. In both cases, we gradually increase the number of pixel and column skips. Additionally, we include a local context of 3×3 pixels. Figure 5.11 shows the obtained results. As it can clearly be seen in the picture, if boundary constraints are imposed on the warping, the increase of the number of pixel and column skips helps to reduce the recognition error rate.

Table 5.3: Effect of monotonicity requirement

Matching algorithm	**ER [%]**	
	general rel. constr.	mon. & con. const.
TSDP-1	6.5	**6.2**

This observation justifies our consideration that a large number of pixel and column skips at once is needed to compensate for registration errors. In this case, the P2DW maps the boundary pixels in the test image to the boundary pixels in the reference image and then starts the warping of ”right” pixels by skipping unrelevant areas of the reference image. However, it can directly be seen from Figure 5.11 that the complete elimination of the boundary constraints helps to significantly reduce the recognition error, even if a skip of at most one pixel and column is allowed. Further increase of the number of pixel and column skips does not result in noticeable decrease of the error rate. This can be explained by the fact that in most cases, a skip of at most one pixel and column is enough to compensate for small local deformations.

5.2.7 Monotonicity Requirement

As we explain in Section 3.4, the TSDP-1 imposes the monotonicity and continuity constraints (c.f. Eq. 2.13 - 2.16) on the warping. However, another way to restrict the deformation is to enforce the general relative position constraints (c.f. Eq. 2.9 - 2.12) which require continuity, but no monotonicity of the warping. Therefore, we study the performance of the TSDP-1 in both cases. Additionally, we optimize the penalty weights for each type of constraints. The results are shown in Table 5.3. It can directly be seen from the table that the monotonicity and continuity constraints help to reduce the recognition error. In order to understand the reasons, we analyze images which are incorrectly recognized by the TSDP-1 with the general relative position constraints. Figure 5.12(a) shows example gradient images, while the warped grid computed by the TSDP-1 with the general relative position constraints is shown in Figure 5.12(b) on the left side. As it can clearly be seen in the picture, the general relative position constraints are violated in the left and right bottom corners of the deformed pixel grid due to independent optimization of pixel trees. We analyze one of the pixel trees which causes violations of the constraints. The tree is shown in Figure 5.12(c) on the left side. The tree’s stem is additionally emphasized in bold. As it can be seen in the picture, the ordering of pixels in the tree’s stem is not monotone. The TSDP-1 with the general relative position constraints starts the tree in the middle of the reference image and then finds the warping of each next pixel which is located below the warping of the current pixel. In other words, the algorithm performs steps back instead of moving forward. According to the general relative position constraints,

Table 5.4: Effect of iterations

Matching algorithm	ER [%]
TSDP-1	**6.2**
ITSDP ($N = 5$)	7.3

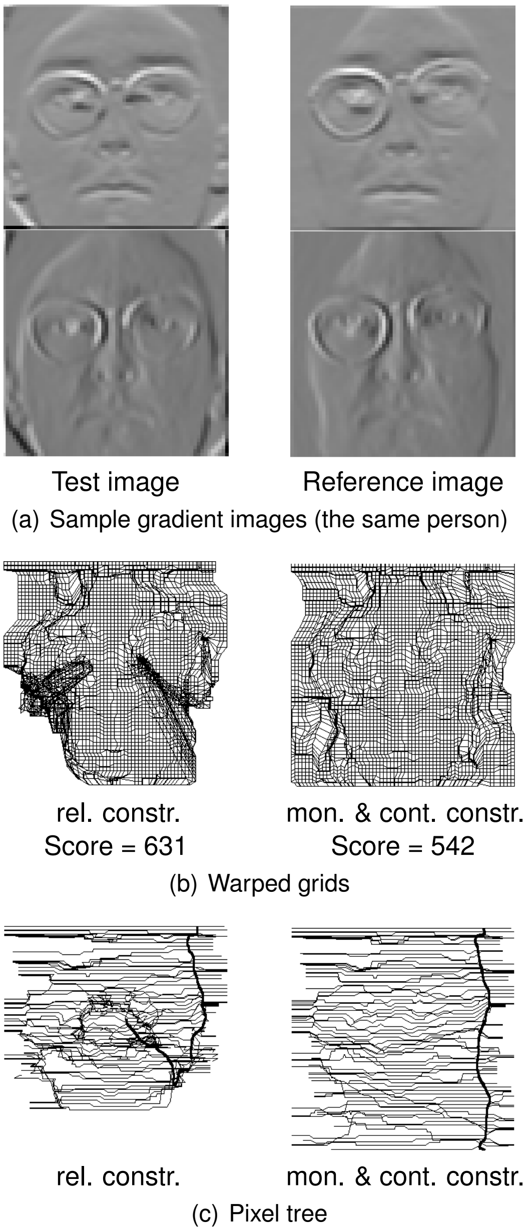
such transitions between a pixel and its successor are permitted. Figure 5.12(b) demonstrates that due to the independent optimization of each pixel tree and non-fulfillment of the monotonicity requirement the pixels in neighboring columns can be located far from each other. This leads to an increase of local distances between such pixels, and consequently, the recognition score grows (c.f. Eq. 4.2). Therefore, even for two images from the same class, the recognition score can be larger than for the images which belong to different classes.

However, if the monotonicity requirement is imposed on the warping, no backward transitions are allowed. This is clearly seen on the right side of Figures 5.12(b) and 5.12(c) which represent the warped grid and the corresponding pixel neighborhood tree, respectively, computed by the TSDP-1 with the monotonicity and continuity constraints. As a result of the monotonicity requirement, the quality of warping is improved. This leads to a decrease of the recognition score. At the same time, the monotonicity requirement further constraints the warping, which makes unrealistic deformations of the reference image rather improbable.

5.2.8 Iterative Reestimation

In order to verify the supposition that iterative reestimation by the ITSDP helps to improve the recognition results, we study effects of iterations on the recognition error rate. According to informal experiments, in most cases the warping does not significantly change after 5 iterations. Hence, we select this number of iterations for recognition. The obtained result is compared to the performance of the TSDP-1 in Table 5.4. As it can be seen from the table, iterations do not help to reduce the recognition error rate. Contrary, the recognition error rate increases. To understand the reasons, we analyze some images which are incorrectly recognized by the ITSDP after 5 iterations. A pair of sample images showing the same person is represented in Figure 5.13(a).

First, we study the visual effects of iterations on the warping. Figure 5.13(b) shows the initial warped grid and one of the pixel trees computed by the TSDP-1 for the same pair of images. As it can clearly be seen from the picture, some of the tree's branches strongly intersect due to the lack of vertical dependencies between pixels. This results in a bad alignment of rows in the deformed grid. Additionally, a violation of the constraints is apparent in the upper part of the grid. We analyze the structure



Test image Reference image

(a) Sample gradient images (the same person)

rel. constr.
Score = 631

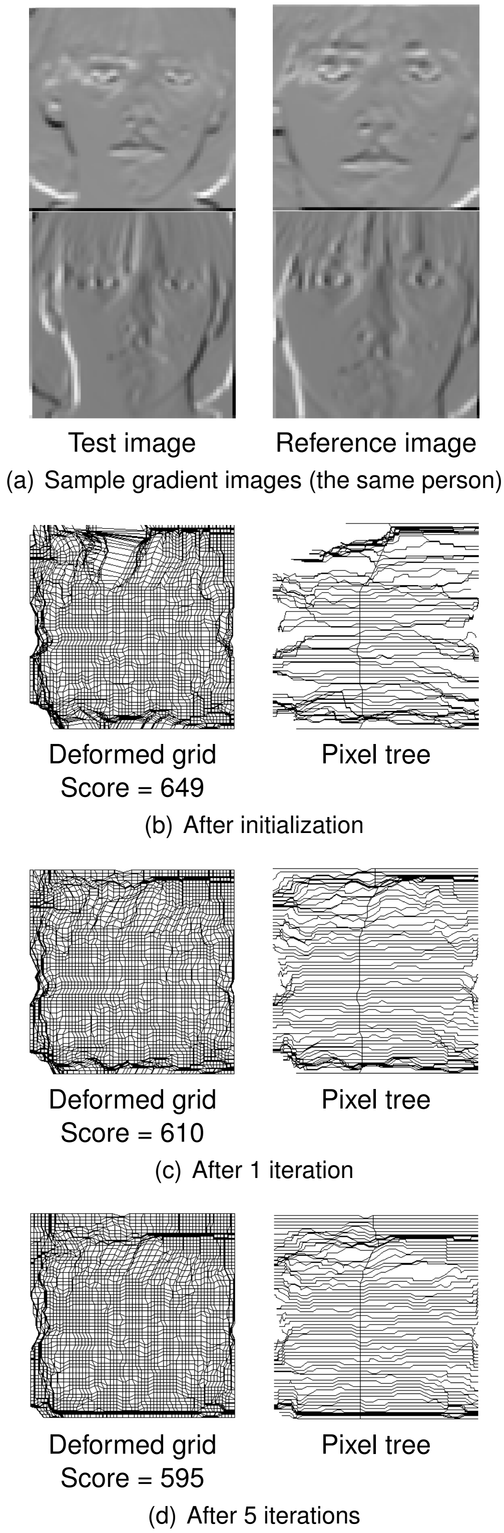
mon. & cont. constr.
Score = 542

(b) Warped grids

rel. constr. mon. & cont. constr.

(c) Pixel tree

Figure 5.12: Examples of warpings computed by the TSDP-1



(a) Sample gradient images (the same person)

(b) After initialization

(c) After 1 iteration

(d) After 5 iterations

Figure 5.13: Effects of iterations on warping (the same person)

of the tree's branches after one and five iterations. It can directly be seen in Figure 5.13(c) that the alignment of the tree's branches with respect to each other is noticeably improved already after the first iteration. This results in an improvement of the warped grid's structure. Moreover, it can be clearly seen that the number of violated constraints is significantly reduced, which leads to a decrease of the recognition score shown under the corresponding pixel grid. The increase of the number of iterations results in further improvements in the alignment of branches. Figure 5.13(d) shows the individual pixel neighborhood tree and the deformed pixel grid after five iterations. However, the relative improvement in the last case is less noticeable in comparison to the first iteration. According to the results obtained for this pair of images, iterations help not only to improve the alignment of the trees' branches, but also to reduce the amount of violated constraints in the warping.

During the iterations, the warping is significantly improved not only for images from the same class, but also for those which belong to different classes. This is shown in Figure 5.14. It can directly be seen that an improvement of the warping results in lower recognition score. It can happen that the recognition score which adequately characterizes the similarity between images before the iterations are performed (compare the scores in Figure 5.14(b)) does not reflect the class membership already after the first iteration (compare the scores in Figure 5.14(c)). This leads to an increase of the recognition error rate.

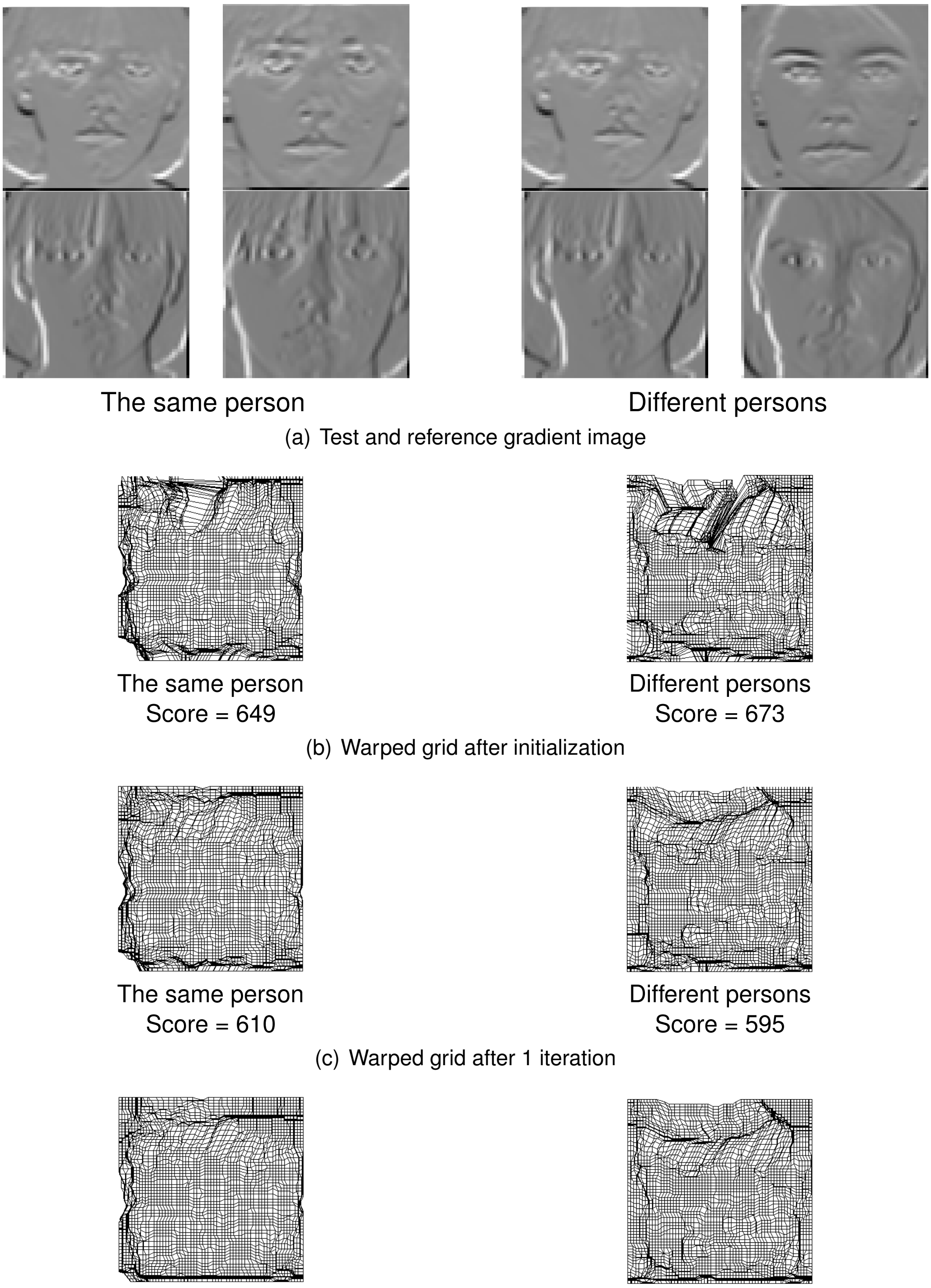
Finally, we analyze the warping cost of each particular pixel tree. According to the optimized criterion in Equation 3.12, this cost must decrease or at least remain the same after each new iteration. However, the experiments on a pair of sample images show that although the cost of warping of each particular tree decreases after the first iteration, it grows during next iterations. Hence, there is no guarantee of the convergence of the iterative procedure.

According to the obtained results, an improvement of the warping does not necessarily lead to a decrease of the recognition error rate. Therefore, the iterative procedure needs further analysis.

5.2.9 Summary of the Best Results

Here, we present the summary of the best results produced by each approach. We empirically optimize the size of the warping range for the ZOW and TSDP-W*. For the P2DW-FOSE, we empirically find the optimal strip width.

Additionally, the results are compared to a version of Sequential Tree-Reweighted Message Passing (TRW-S) [Kolmogorov 06] which we implemented by ourself. The TRW-S is a well-known iterative technique for discrete energy minimization. In contrast to the discussed approaches, the TRW-S is an *approximation* of the two-dimen-



(a) Test and reference gradient image

(b) Warped grid after initialization

(c) Warped grid after 1 iteration

(d) Warped grid after 5 iterations

Figure 5.14: Effects of iterations on warping (different persons)

Table 5.5: Comparison of complexities and computing times, 3×3 local context (AR Face automatically cropped)

Matching algorithm	ER[%]	Complexity			CTF
		distance computation	column level	global level/ forward-backward	
ZOW $(\Delta = 21)$	11.2	-	-	$IJ\Delta^2$	1
TSDP-W* $(\Delta = 15)$	**5.8**	$IJ\Delta^2$	$\Delta^4 IJ$	$2\Delta^4 IJ$	79
TSDP-1	6.5	$IJUV$	$9IJUV$	$18IJUV$	63
P2DW	8.2	$IJUV$	$3IJUV$	$3IU$	13
P2DW-FOSE $(\Delta = 5)$	6.0	$IJUV$	$9\Delta IJUV$	$3IU$	69
CTRW-S	8.4	$IJUV$	-	$(N)18IJUV$	56

sional image warping. In the implementation of the TRW-S which we use in our experiments pruning is performed which uses the ability of the approach to compute a lower bound for the warping cost. The warping between a pair of images is not computed if the lower bound of the warping cost is higher compared to the already found cost of deformation for another pair of images. Informal experiments showed that the average number of iterations (averaged over all image comparisons) is one. Additionally, in the implementation of the TRW-S, the monotonicity and continuity constraints are imposed on the warping. In the following, we refer to this version of the TRW-S as Constrained Sequential Tree-Reweighted Message Passing (CTRW-S).

First, we provide the comparison of the complexities and computing times of the approaches. Then, we report the best possible performance. Finally, we discuss some issues concerning the obtained results.

Complexity. Here, we compare the complexities and computing times of the approaches. For fairer comparison, we set the size of local context to be the same for all matching algorithms. As we show in Section 5.2.4, the inclusion of local context of 3×3 pixels is advantageous for 3 approaches out of 5. Therefore, we select the 3×3 local context. In order to directly compare the warping range and the strip's width, we uniformly express the warping range through Δ, where $\Delta = 2W + 1$. The results are listed in Table 5.5. Columns 3 to 5 of the table show the overall complexity of the approaches. The complexity sums up from the cost of the distance computation, the complexity of the optimization within the columns and the complexity of the optimization across the columns. The last summand is replaced by the cost of the forward-backward procedure in the case of the TSDP-1 and TSDP-W*. It can clearly be seen from the 3rd column of Table 5.5 that the cost of the distance computation is the same in the TSDP-1, P2DW and P2DW-FOSE, as these approaches impose neither boundary, nor absolute position constraints on the warping. At the same time, the cost of the distance computation in the TSDP-W* depends of the warping range. The column-level complexity of the P2DW is the lowest one in comparison to the P2DW-FOSE and TSDP-1, as is shown in the 4th column, due to the restriction of the column-to-column mapping. It can directly be seen from the 5th column of Table 5.5 that the complexities of the global optimization in the P2DW and P2DW-FOSE are similar, while the complexity of the forward-backward run in the TSDP-1 exceeds the global-level complexities of both approaches. This can be explained by the fact that during the forward-backward run in the tree-serial approaches the optimization is performed on the pixel level, while the global optimization in the pseudo-2D methods is done across the entire columns. The complexity of each single iteration of the CTRW-S is similar to the complexity of the forward-backward run in the TSDP-1, as the CTRW-S also imposes the monotonicity and continuity constraints on the warping and performs two passes through the entire image.

The last column in Table 5.5 shows the computing time factor (CTF) with respect to the ZOW. It can be observed that the TSDP-W* is the slowest approach. Although the warping range of 7 pixels intends to reduce the complexity of the TSDP-W*, the number of allowed pixel skips in this case is 16, which significantly increases the computing time. In contrast, both the TSDP-1 and P2DW-FOSE allow a skip of at most one pixel, but do not impose the absolute position constraints. The computing times of both approaches do not differ much compared to the TSDP-W*. As expected, the P2DW is the second fastest approach after the ZOW due to the low complexity on the column level. As the average number of iterations in the CTRW-S is one due to the lower bound pruning, the computing time of this approach is insignificantly lower compared to the TSDP-1.

In order to give a clue about the absolute computing times of the approaches, we provide the results for the P2DW-FOSE. In this case, the average absolute computing time is 20 seconds per image comparison (measured on Intel Core 2 2400 GHz). Taking into account that 770×770 warpings have to be computed, we face with over 3300 hours of computing time for the whole database.

Best Performance. As we show in Section 5.2.4, the optimal size of the local context depends on the selected approach. Hence, we provide the recognition error rates with respect to the optimal local context for each matching algorithm. The results are listed in Table 5.6. Surprisingly, the TSDP-W* with the warping range of 7 pixels performs best, while the TSDP-1 provides significantly higher recognition error rate. This can be caused by either permission of larger pixel skips in the TSDP-W*, or by the restriction stemming from the absolute position constraints. We analyze both possible reasons below. The second best approach is the P2DW-FOSE. Expectedly, its performance in much better compared to the P2DW. The warping computed by the

Table 5.6: Summary of the best results (AR Face automatically cropped)

Matching algorithm	Best local context	ER [%]
ZOW ($W = 10$)	5x5	7.5
TSDP-W* ($W = 7$)	1x1	**4.9**
TSDP-1	1x1	6.2
P2DW	5x5	7.5
P2DW-FOSE ($\Delta = 5$)	3x3	6.0
CTRW-S	3x3	8.4

P2DW-FOSE is more flexible since additional deviations from a column are allowed, which positively affects the recognition error rate. At the same time, the restriction of the column-to-column mapping worsens the performance of the P2DW leading to similar recognition results compared to much more simple ZOW. The explanation of the unexpectedly good performance obtained by the ZOW is in the large portion of the local context. The inclusion of local context significantly improves the recognition results provided by the ZOW. Unexpectedly, the performance of the CTRW-S is below the level of the recognition quality achieved by the other approaches. The CTRW-S computes the most restricted warping among all approaches presented in Table 5.6. This approach imposes the monotonicity and continuity constraints which in contrast to the TSDP-1, are never violated. At the same time, the warping produced by the P2DW-FOSE is less restricted since the vertical dependencies between the pixels in neighboring strips are neglected. Therefore, the explanation of the worst performance of the CTRW-S can be in too constrained warping. This problem can possibly be overcome by allowing more pixel skips.

Why is the TSDP-W* better than the TSDP-1? According to the results presented in Table 5.6, the TSDP-W* performs surprisingly better than the TSDP-1. We analyze the reasons of this phenomenon. As we explain in Chapter 3, there are two principal points which distinguish the TSDP-W* from the TSDP-1: restriction by the absolute position constraints and permission of many pixel skips in the warping. We study, which of these points makes the difference in the recognition error rates.

First, we analyze whether the **absolute position constraints** facilitate better performance of the TSDP-W*. For that purpose, we reduce the number of pixel skips in the TSDP-W* by enforcing the monotonicity and continuity constraints on the warping and leave the size of the warping range unchanged. The obtained results are presented in Table 5.7. It can directly be seen from the table that when the same number of pixel skips is allowed in both approaches, introduction of the warping range leads to an increase of the recognition error rate. Hence, better performance of the TSDP-

Table 5.7: Effect of the monotonicity and continuity constraints in the TSDP-W*

Matching algorithms	**ER [%]**
TSDP-1	**6.2**
TSDP-W* ($W = 7$ + cont.& mon. constr)	6.9

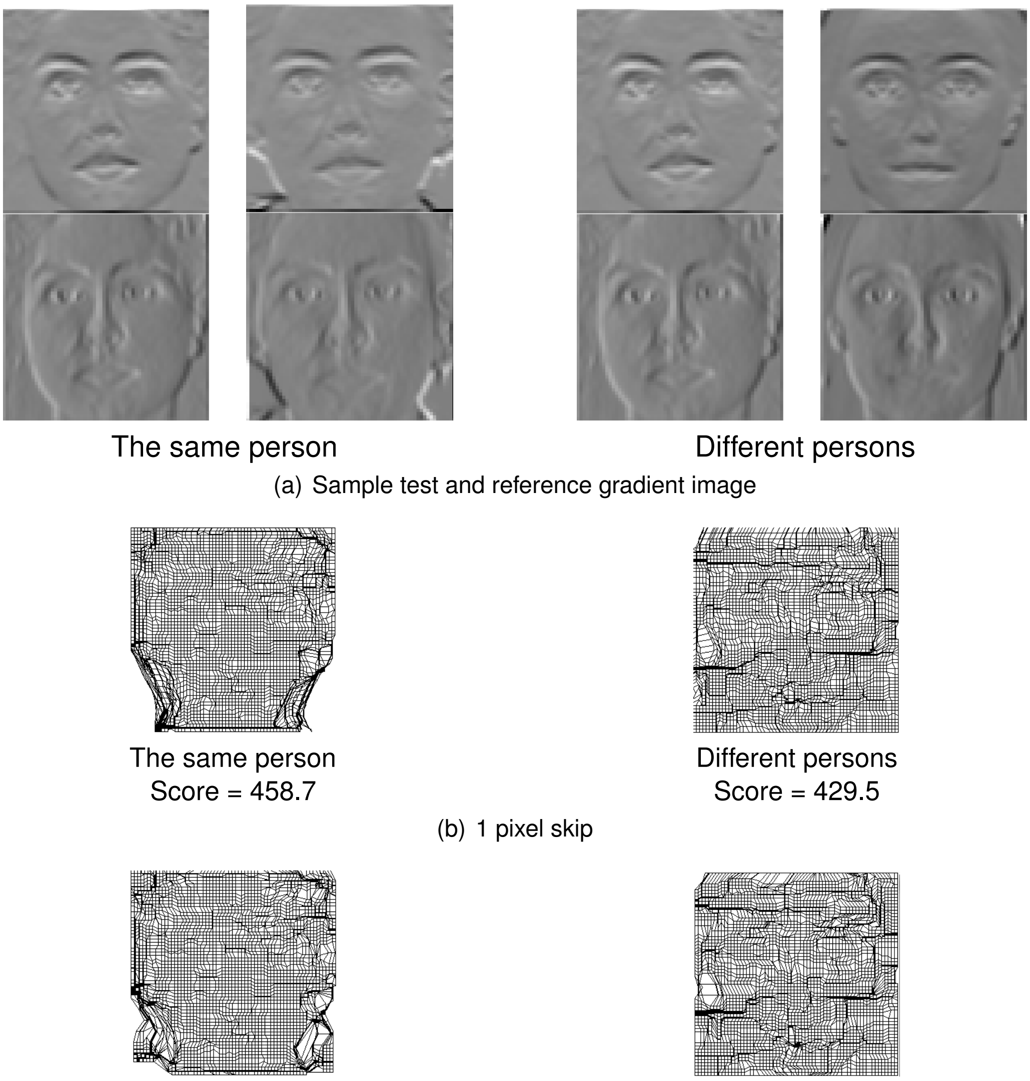
W* can not be explained by the absolute position constraints which are shown to be too restrictive.

Next, we study effects of **many pixel skips** on the warping. We select sample gradient images, which are incorrectly recognized by the TSDP-1, but accurately classified in the case of the TSDP-W*. The gradient images are shown in Figure 5.15(a). First, we look at the results, if a skip of at most one pixel is allowed. The deformed pixel grids are shown in Figure 5.15(b). As it can clearly be seen in the picture, in the case of the same person, the deformed pixel grid contains areas of noticeable violations of constraints (bottom left an right corners) where the parts of some border columns are located closer to the image center. As a result of the violated constraints, the recognition score increases. According to the recognition scores shown below the corresponding warpings, the images in Figure 5.15(b) are recognized incorrectly. We increase the number of allowed pixel skips. Figure 5.15(c) demonstrates the results obtained for the same pairs of images when a skip of at most three pixels is allowed. As it can directly be seen in the picture, for the images showing the same person, the amount of violated constrains is reduced. As a result, the recognition score is also decreased. The recognition score for the second pair of images does not significantly change. Therefore, for these sample images, permission of many pixel skips helps to reduce the number of violated constraints. This can lead to a decrease of the recognition error rate. However, if larger pixel skips are allowed, the complexity of the algorithm significantly increases and the computing time becomes unfeasible. Hence, we do not study the effects of larger pixel skips in the TSDP-1 on the recognition error rate.

According the these considerations, better performance of the TSDP-W* is not due to the absolute position constraints which are shown to be prohibitive, but is probably enhanced by the permission of many pixel skips. At the same time, the performance of the TSDP-1 is worsened by the problem of violated constraints which often occur in the warping computed by this approach.

5.2.10 Comparison of Features

Here, we study how the choice of a local feature descriptor affects the performance of the matching algorithms. For that purpose, we select three different descriptors, namely the Sobel features, the DCT features and the SIFT. The Sobel features are



The same person

Different persons

(a) Sample test and reference gradient image

The same person
Score = 458.7

Different persons
Score = 429.5

(b) 1 pixel skip

The same person
Score = 424.4

Different persons
Score = 430.2

(c) 3 pixel skips

Figure 5.15: Effects of many pixel skips in the TSDP-1 on the warping

Table 5.8: Comparison of features (AR-FACE automatically cropped)

Matching algorithm	Best local context	ER [%]		
		Sobel	DCT	PCA U-SIFT
ZOW ($W = 10$)	5x5	7.5	4.7	3.1
P2DW-FOSE ($\Delta = 5$)	3x3	6.0	3.1	**2.5**
L_1-Norm	-	44.3	30.3	22.2

shown in the previous sections to be simple and efficient. Compact and distinctive, the DCT local feature descriptor has been used in various studies on face recognition, while the SIFT seems to be the most widely used descriptor nowadays. However, no direct comparison of these descriptors in combination with the matching algorithms has been performed so far.

For our experiments, we extract DCT and SIFT local feature descriptors from the automatically cropped images. For the DCT descriptor, we allow a block overlap of 7, employ the dct-3 feature selection strategy and retain 5 coefficients, as presented in [Hanselmann 09]. We compute the upright version of the SIFT (U-SIFT) similar to the work [Dreuw & Steingrube$^+$ 09], as it was shown to outperform the rotation invariant descriptor in the task of face recognition. Additionally, we carry out a PCA-reduction of the U-SIFT to the 30 components with the largest Eigenvalues.

For performance comparison we select the ZOW since it is fast and efficient and the P2DW-FOSE, as this approach shows the best recognition results on the Sobel features among the matching algorithms with the monotonicity and continuity constraints. In order to show the absolute improvement of recognition results obtained by the matching algorithms, we provide the recognition error rates for the case when solely the L_1-Norm is used. We optimize the size of local context and penalty coefficients for each approach. Additionally, the optimal warping range for the ZOW and the strip width for the P2DW-FOSE are found. The obtained results are listed in Table 5.8. It can directly be seen from the table that the choice of the DCT features helps to significantly decrease the recognition error for both approaches compared to the Sobel features. It can be explained by the fact that the DCT features are more distinctive and robust to changes in illumination. The U-SIFT descriptor allows to further improve the recognition results. The overall performance improvement obtained by the U-SIFT is more than a factor of two for both approaches in comparison to the Sobel features. At the same time, the ranking of matching algorithms remains the same, i.e. the P2DW-FOSE outperforms the ZOW whatever local feature descriptor is selected.

Comparing the performance of the matching algorithms to the L_1-Norm, we observe a significant decrease of the recognition error rate in each case. Conventional

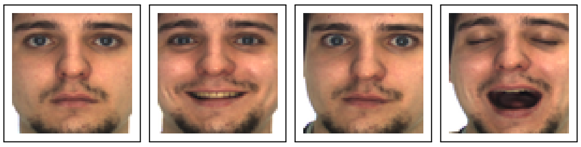


Figure 5.16: Examples of manually cropped face images (AR Face database)

distance function fails due to the occurrence of many local changes in face images, while the matching algorithms are able to cope with local deformations of the image content.

According to the obtained results, the combination of the matching algorithms with the U-SIFT descriptor provides the most significant reduction of the recognition error rate, while the choice of the DCT features leads to worse results compared to the U-SIFT.

5.2.11 Manually Cropped Faces

As we explain in Section 4.1, the accuracy of cropping has a direct impact on the performance of matching algorithms. Here, we study the best possible performance of the discussed approaches on the AR Face Database with optimal alignments of images. For that purpose, the original face images have been manually aligned by the eye-center locations [Gross]. The images were rotated such that the eye center locations are in the same row. Then, faces were manually cropped and scaled to a 64×64 resolution. Examples of manually cropped face images are represented in Figure 5.16.

Performance Reporting. For our experiments, we select the U-SIFT local feature descriptor since it is shown to provide lower recognition error rates in comparison to the Sobel and DCT features. We extract the U-SIFT descriptor, as we explain in Section 5.2.10. We optimize the penalty weights and the size of local context for each approach. The best warping ranges for the ZOW and the TSDP-W* are selected and the optimal strip width for the P2DW-FOSE is chosen. The obtained results are listed in Table 5.9. It can directly be seen from the table that all matching algorithms are able to achieve very low recognition error rates and the number of incorrectly classified images is at most 6 out of 770. The P2DW-FOSE approach shows the best performance with 0.52% of the recognition error. However, the difference between all matching algorithms is insignificant.

Outstanding performance demonstrated by all matching algorithms is mostly due to two reasons: almost perfect manual alignment of the images and the choice of an efficient local feature descriptor. The U-SIFT descriptor already helps to achieve a

Table 5.9: Summary of the best results (AR Face manually cropped)

Matching algorithm	Best local context	ER [%]	N incorr. recog.
ZOW ($W = 4$)	3x3	0.78	6
TSDP-W* ($W = 4$)	3x3	0.78	6
TSDP-1	5x5	0.65	5
P2DW	3x3	0.65	5
P2DW-FOSE ($\Delta = 5$)	3x3	**0.52**	**4**
L_1-Norm	-	3.38	26
Aw-SpPCA [Tan & Chen 05]		6.43	49
DCT [Ekenel & Stiefelhagen 06]		4.70	36
SURF-Face [Dreuw & Steingrube$^+$ 09]		0.25	2

low recognition error rate when used solely in the combination with the L_1-Norm (c.f. Table 5.9), while its usage together with the matching algorithms allows to significantly improve the recognition quality.

We compare the obtained results with the current state of the art on the AR Face Database. We select three approaches which were shown to provide low recognition rates, namely the local appearance based approach [Ekenel & Stiefelhagen 06], Adaptively weighted Sub-Pattern PCA (Aw-SpPCA) [Tan & Chen 05] and the recent SURF-Face [Dreuw & Steingrube$^+$ 09] method. The first approach employs the DCT features and is explained in more detail in Section 4.2. In the Ac-SpPCA, an image is divided into sub-images and the PCA is performed on each sub-image. Then the contributions of each part are adaptively computed and used for classification. The SURF-Face approach [Dreuw & Steingrube$^+$ 09] uses the SURF and the SIFT local feature descriptors in combination with a nearest neighbor matching under viewpoint consistency constraints.

The performance of the state of the art approaches is shown Table 5.9. It can clearly be seen that all matching algorithms outperform holistic Aw-SpPCA and DCT approaches. Although both methods employ local features, conventional distance function is used for the global comparison of the images. Therefore, the performance of both Aw-SpPCA and DCT should rather be compared to the result obtained by the combination of L_1-Norm with the U-SIFT descriptor. In contrast, the recently proposed SURF-Face approach finds local matchings between the images, which makes this method more robust to local deformations compared to the Aw-SpPCA and DCT. As a result, the SURF-Face approach obtains the performance on par with the matching algorithms, which shows that the methods proposed in this work are competitive with the current best performance on the AR Face database.

5.2.12 Conclusion

According to the results obtained on the AR Face Database, various parameter settings such as e.g. penalty, size of local context and warping range, as well as different types of constraints imposed on the warping have a large impact on the performance of the matching algorithms. The approaches are shown to be able to cope with changes in facial expression and illumination. Moreover, the results obtained on the automatically cropped images suggest that the matching algorithms are also robust to registration errors. Combination of the approaches with more sophisticated U-SIFT local feature descriptor helps to significantly improve the performance of the matching algorithms compared to the Sobel features. The results on the manually cropped images show that the approaches proposed in this work outperform existing methods and are able to achieve extremely low recognition error rates. However, as all the parameters are chosen to optimize the performance of the matching algorithms on the test data, risk of overfitting is high. Therefore, the discussed matching algorithms have to be evaluated on another database, where the test data is used only for performance reporting.

5.3 Labeled Faces in the Wild

Labeled Faces in the Wild (LFW) [Huang & Mattar$^+$ 08] is a database of face photographs designed to study the problem of face recognition in an unconstrained environment. The LFW dataset contains more than 13000 face images which have several challenging problems, such as variable facial expressions, changes in pose and scale, in-plane rotations, non-frontal illumination, as well as registration errors. The face images were collected from the web. Each face was detected by the Viola and Jones Face detector [Viola & Jones 04]. The detected region was automatically expanded to capture the entire head, then cropped and rescaled to 250x250 pixel resolution. Each face was manually labeled with the name of the person pictured.

As we explain in Section 4.1, presence of a background in the images can negatively affect the performance of the matching algorithms. Therefore, in our experiments, we used the cropped version of the LFW (LFWcropped) similar to the work presented in [Sanderson & Lovell 09]. In the LFWcropped, closely cropped faces were extracted using a fixed bounding box placed in the same location in each LFW image. The extracted faces were downscaled to a 64×64 pixel resolution. Figure 5.17 exemplifies the original and closely cropped images. In comparison to the AR Face database with automatically cropped faces, the number and severity of face variations is the LFWcropped dataset is uncontrolled, which significantly complicates the recognition.

We perform the experiments on the LFWcropped dataset in accordance with a

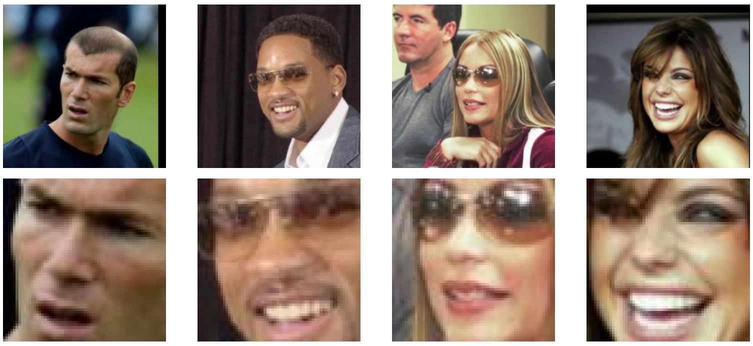


Figure 5.17: Examples of original and closely cropped image (LFW dataset)

prescribed protocol [Huang & Mattar$^+$ 08] where the task is to determine for each pair of previously unseen faces whether the images show the same person (matched pair) or two different persons (mismatched pair). The data is organized into two views: *view 1* which is used for model selection and algorithm development, and *view 2*, aimed at the final performance reporting. The goal of this separation is to use the final test set as seldom as possible before reporting to minimize overfitting. View 1 consists of two subsets of the database: one for training, containing 2200 pairs of face images, and one for testing, containing 1000 pairs. We use the training subset to determine the optimal penalty weights and the size of local context for each approach, as well as to choose the decision threshold which suggests whether the images show the same person. The threshold is optimized to obtain the highest accuracy which is averaged over the classification accuracies for pairs of images. View 2 consists of 6000 pairs divided into 10 subsets. The performance is reported using the estimated mean accuracy and the standard error of the mean from 10 folds of the view 2. According to the protocol, the performance must be computed using a leave-one-out 10-fold cross validation scheme, i.e. each fold in turn is used as test data, while the other 9 folds are used as training data. We assume that the decision threshold selected on the train data of view 1 is also optimal for view 2. Hence, we only use each of the 10 subsets for testing.

As we show in Section 5.2.10, the performance of matching algorithms is significantly improved in the case of the PCA-reduced U-SIFT local feature descriptor compared to the Sobel and DCT features. Therefore, we choose the PCA-reduced U-SIFT descriptor for the LFW experiments.

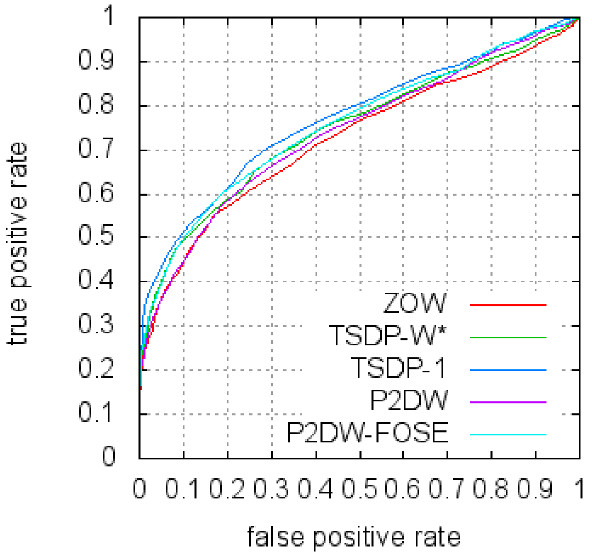


Figure 5.18: ROC curves for pair matching (LFWcropped dataset)

5.3.1 Model Selection

We empirically select the optimal penalty weight and size of local context for each matching algorithm using the training data of view 1. Additionally, we determine the optimal decision threshold which leads to the highest average accuracy. For each value of the threshold, the accuracy is characterized by a point on a ROC curve for pair matching. Therefore, we choose the threshold which corresponds to the point providing the highest average accuracy. The ROC curves for all approaches are shown in Figure 5.18. For instance, the optimal threshold value for the TSDP-1 corresponds to the point when the fractions of true and false positives are 0.67 and 0.24, respectively.

Having determined the optimal parameters on the training data, we report the performance of the approaches on the test data of view 1. The obtained results are listed in Table 5.10. The optimal size of local context is 5×5 pixels for each approach. Although each value of the U-SIFT already contains much local information, additional inclusion of local context is advantageous in the presence of extremely high local variability of the image content which is characteristic for the images in the LFWcropped dataset. The best performance on the test data of view 1 is provided by the TSDP-1. The accuracy of the TSDP-1 is significantly higher, compared to the other approaches, which can probably be explained by the ability of the TSDP-1 to compensate for stronger in-plane face rotations. This consideration is additionally

Table 5.10: Summary of the best results on the test data of view 1 (LFWcropped)

Matching algorithm	**Best local context**	**Accuracy [%]**
ZOW ($W = 8$)	5x5	69.6
TSDP-W* ($W = 8$)	5x5	70.3
TSDP-1	5x5	**72.0**
P2DW	5x5	68.0
P2DW-FOSE ($\Delta = 7$)	5x5	70.2

Table 5.11: Summary of the best results on view 2 (LFWcropped)

Matching algorithm	**Best local context**	**Mean accuracy [%]**	**Standard error [%]**
ZOW ($W = 8$)	5x5	69.05	0.46
TSDP-W* ($W = 8$)	5x5	69.37	0.44
TSDP-1	5x5	**71.95**	0.46
P2DW	5x5	68.72	0.43
P2DW-FOSE ($\Delta = 7$)	5x5	70.12	0.45
L_1-Norm	1x1	62.50	0.43
CTRW-S	3x3	71.77	0.30
MRH [Sanderson & Lovell 09]	-	70.38	0.48

confirmed by the fact that in the case of the P2DW-FOSE, the strip width is relatively large, which allows to cope with rotation to some extent. However, even the width of 7 pixels seems to be not enough, while its further increase would inevitably raise the complexity with probably insignificant improvement of the performance. Having the least ability among all approaches to compensate for face rotations, the P2DW performs worst. Both TSDP-W* and ZOW achieve better recognition results, compared to the P2DW, but cannot provide the accuracy on par with the TSDP-1 due to the absolute position constraints imposed on the warping.

5.3.2 Performance on the Evaluation Set

Here, we provide the performance of the matching algorithms on view 2 of the dataset. For each approach, we select the penalty weight and the size of local context, as well as the decision threshold which lead to the best results on the training data of view 1. The performance is reported using the mean and standard error of the average accuracies from 10 folds of the dataset. The obtained results are listed in Table 5.11. It can be observed that the TSDP-1 provide the highest mean accuracy among the approaches discussed in this work. This is expectedly since the TSDP-1 is able to compensate for stronger deformations caused by in-plane face rotations and changes in pose in comparison to the other approaches. At the same time, this result is consistent with the performance of the TSDP-1 obtained on the test data of view 1. As the performance of the TSDP-1 depends on the minimal set of parameters, the risk of overfitting low. In contrast, the performance of the ZOW and especially TSDP-W* significantly drops compared to the results in Table 5.10. As the warping range for both approaches is chosen using the training data of view 1, it must not be necessarily optimal on the test data of view 2. Furthermore, the difference between both approaches is not statistically significant since the standard errors of the mean overlap. This clearly shows that too restrictive absolute position constraints make very simple ZOW and much more sophisticated TSDP-W* perform similarly. The recognition result obtained by the P2DW-FOSE is consistently lower than the performance of the TSDP-1 due to the reasons explained previously. Expectedly, the P2DW performs worst due to the limitation stemming from the column-to-column mapping.

We provide the recognition results obtained by the L_1-Norm in combination with the U-SIFT descriptor. As it can be seen from Table 5.11, the relative decrease of the recognition error rate in the case of all matching algorithms is less significant than it is on the AR Face database. It can be explained by considerably greater complexity of the LFWcropped compared to the AR Face dataset.

In order to compare the recognition results obtained by the matching algorithms with other approaches, we provide the performance of CTRW-S briefly explained in Section 5.2.9 and Multi-Region Histograms (MRH) [Sanderson & Lovell 09]. The MRH is the state-of-the-art approach on the LFWcropped dataset. It is a face matching algorithm which describes each face in terms of multi-region probabilistic histograms of visual words, followed by a distance computation between the histograms. The authors report the performance of the MRH for two cases: when a raw distance is used and when the distance between the histograms is additionally normalized by comparing each face from a pair in view 2 with some faces in view 1 and using the fact that both views are disjunct. Although the normalization of the distance helps to improve the performance of the MRH, the assumption that the train and the test sets are disjunct must not necessarily hold in the general case. According to this consideration, the performance of the MRH with the normalized distance rather corresponds to the image-unrestricted training paradigm [Huang & Mattar$^+$ 08], while the performance of the matching algorithms is reported for the image-restricted case. Therefore, we include the results obtained by the MRH with raw distance.

The performance of both the CTRW-S and the MRH is shown in Table 5.11. It can be seen that the TSDP-1 outperforms both approaches. However, the difference between the CTRW-S and the TSDP-1 is not statistically significant, as the standard errors of the mean overlap. Similarly to the TSDP-1, the CTRW-S does not impose absolute position constraints on the warping and is not restricted by the strip's width as the P2DW-FOSE, which allows to compensate for stronger rotations and changes in pose. At the same time, the holistic MRH approach cannot achieve the accuracy of both best performed the TSDP-1 and CTRW-S. The explanation can be in the lower robustness of the MRH to small local deformations since this approach deals with relatively large image parts. However, the performance of the MRH is equivalent to the result obtained by the P2DW-FOSE. Additionally, the MRH approach is much more scalable compared to the matching algorithms due to efficient fast histogram approximation, easy parallelization and use of a simple distance function.

5.3.3 Conclusion

Experiments on the recent and difficult LFWcropped dataset show that the matching algorithms are able to cope with several concurrent and uncontrolled factors, such as variations in facial expression, illumination, pose and scale, as well as in-plane rotations and registration errors. The obtained results suggest that all matching algorithms perform well providing about 70% accuracy, while the proposed TSDP-1 approach can outperform the other methods achieving 71.95% accuracy. These results show that more general the TSDP-1 approach has significant advantage over the other more restricted matching algorithms when the recognition problem becomes more difficult and unconstrained. Moreover, the TSDP-1 obtains the performance on par with the version of well-known TRW-S approach and is able to outperform the recently proposed MRH method.

Chapter 6

Conclusions

In this work, we studied the problem of two-dimensional image warping and the application of known and improved matching algorithms on the task of face recognition. In particular, we analyzed the Zero-Order Warping (ZOW), Pseudo 2D Hidden Markov Model (P2DHMM) and Tree-Serial Dynamic Programming with Many Skips (TSDP-W*) approaches which are based on the relaxations of the original problem. Limitations of the P2DHMM and TSDP-W* were analyzed and ways to overcome shortcomings of both approaches were proposed. We showed that the proposed approaches can outperform many current methods.

The main contributions of this work are as follows:

- The matching algorithms based on the relaxations of the original problem of two-dimensional image warping were compared.
- We showed that in some cases, the constraints imposed on the warping were either too hard, or too soft.
- We presented extensions for existing approaches intended to improve the performance of the matching algorithms for face recognition.
  - We proposed the Pseudo 2D Warping (P2DW) approach which intends to overcome the limitation of the P2DHMM caused by the boundary constraints
  - The First-Order Strip Extension of the Pseudo 2D Warping (P2DW-FOSE) extension of the P2DHMM was presented which relaxes the restriction of column-to-column mapping by modelling additional deviations from a column in accordance with the monotonicity and continuity requirement
  - The Tree-Serial Dynamic Programming with One Skip (TSDP-1) approach was proposed which intends to overcome the shortcoming of the TSDP-W* stemming from the absolute position constraints. The proposed approach relaxes the absolute position constraints and requires the monotonicity and continuity of the warping at the same time.
  - We showed that the warping computed by the TSDP-1 could further be qualitatively improved by an iterative procedure.

- We qualitatively evaluated the matching algorithms on synthetic examples and face images. We showed the superior performance of the improved approaches over the original ones
- The quantitative evaluation of the matching algorithms on the AR Face and Labeled Face in the Wild (LFW) dataset was performed
  - The effects of different constraints and parameter settings were shown.
  - The choice of features was investigated. We showed that the combination of the matching algorithms with the upright SIFT (U-SIFT) local feature descriptor can significantly improve the performance compared to the Sobel and DCT features.
  - Experiments on the AR Face database showed that the proposed approaches are robust and could outperform many of the generic methods.
  - The evaluation on the LFW dataset (unconstrained recognition problem) showed that the proposed approaches could outperform the state-of-the-art method.

Summarizing, we showed that the matching algorithms based on nonlinear image deformation models could successfully be used for face recognition. We demonstrated that the proposed approaches were able to cope with strong changes in facial expression, illumination and pose. Additionally, the presented approaches were shown to be robust to registration errors. Furthermore, the proposed approaches could outperform many generic approaches. The lowest recognition error rate of 0.52% on the AR Face database was obtained by the P2DW-FOSE (state-of-the-art SURF-Face [Dreuw & Steingrube$^+$ 09] provided 0.25%), while the TSDP-1 approach performed best on the LFWcropped dataset (unconstrained environment) having achieved 71.95% accuracy (state-of-the-art MRH [Sanderson & Lovell 09] obtained 70.38%).

Experiments on the AR Face database showed that the choice of features and accuracy of cropping can significantly facilitate the performance of the matching algorithms. Moreover, the better a local feature descriptor, the smaller the difference in the recognition results provided by the different warping approaches. The boundary constraints were shown to be prohibitive, especially if registration errors occurred. The proposed P2DW approach, intended to overcome the negative effects of the boundary constraints, reduced the recognition error rate on the AR Face dataset from 13.9% down to 8.2% (automatically cropped images, Sobel features). The recognition error rate was further reduced to 6.0% by the proposed P2DW-FOSE approach which relaxes the constraint of column-to-column mapping. The inclusion of large local context for the ZOW, as well as the choice of an appropriate warping range for the ZOW and TSDP-W* was shown to be extremely important for good performance of these approaches. We demonstrated on the synthetic examples that both the ZOW and TSDP-W* fail when the warping range is not large enough. Furthermore, the results obtained on the LFWcropped database suggested that the restriction by the absolute position constraints could make the sophisticated TSDP-W* approach perform similar to the very simple ZOW. The presented TSDP-1 approach which relaxes the absolute position constraints was able to compensate for stronger image deformations. At the same time, the proposed approach was unable to find unwanted good inter-class deformations due to the requirement of the monotonicity and continuity of the warping. Moreover, the TSDP-1 increased the accuracy on the LFWcropped dataset from 69.37% obtained by the TSDP-W* up to 71.95% which is the absolute best performance among the matching algorithms. Finally, experiments on the AR Face and the LFWcropped dataset showed that the more difficult and unconstrained the recognition problem becomes, the greater the advantage of the more general TSDP-1 over the other more limited matching algorithms.

However, experiments on both datasets also disclosed the shortcomings of the proposed approaches. The P2DW-FOSE obtained moderate performance on the LFWcropped database due to the restriction by the strip's size which did not allow the approach to compensate for strong in-plane face rotations. The performance of the TSDP-1 on the AR Face database (automatically cropped images, Sobel features) was worsened by many violated constraints between the pixels occurred due to independent optimization of the trees. This problem could possibly be overcome to some extend by allowing more pixel skips in the TSDP-1, which has to be evaluated. However, the computational complexity of the approach would inevitably raise in this case. Relatively high computational complexity which grows dramatically with the increase of the image resolution is the general problem of the proposed approaches. This makes the application of the approaches impractical to images which resolution is higher than 64×64 pixels, which encourages further research on the ways to reduce the computing time of the proposed methods. For instance, it can be done by determining a lower bound of a deformation cost and stopping the warping procedure if the lower bound exceeds some predefined threshold similar to the CTRW-S. Another way to reduce the complexity is to divide the images into small non-overlapping regions and to find the warping within the regions following by optimization across the entire regions by means of the ZOW. We feel that using such method might also lead to an improvement of the recognition results if the warping within the regions is performed in accordance with the first-order two-dimensional image deformation model.

Appendix A

Appendix

A.1 Software

Throughout this work, the W2D-Software originated from the work [Gollan 03] was continuously extended to new matching algorithms. A few example invocations are shown below.

- P2DW-FOSE experiment on the AR Face database with manually cropped faces (PCA-reduced U-SIFT, d_{euc}, 3×3 local context, pen. $\sim d_{euc}$, weight $= 0.03$, $\Delta = 2 \cdot 2 + 1$) (c.f. Table 5.9)

  ```
  ./w2d
  usift-64x64-0.5-1-train-PCA30-patchNorm.ff
  usift-64x64-0.5-1-test-PCA30-patchNorm.ff
  -MP2DW_FOSE -Dfeature:absrec3x3:1,penalty:pen4:0.03 -B2 -Q2
  ```

- TSDP-1 experiment on the LFWcropped dataset, view 2, fold 1 (PCA-reduced U-SIFT, d_{euc}, 5×5 local context, pen. $\sim d_{euc}$, weight $= 0.1$) (c.f. Table 5.11)

  ```
  ./w2d
  usift-skip1-fold01-test-left-PCA30-patchnorm.ff
  usift-skip1-fold01-test-right-PCA30-patchnorm.ff
  -MTSDP_1 -Dfeature:absrec5x5:1,penalty:pen4:0.1 -Q2
  ```

List of Figures

List of Tables

Glossary

2DW	First-Order Two-Dimensional Warping
CTRW-S	Constrained Sequential Tree-Reweighted Message Passing
HMM	Hidden Markov Model
ITSDP	Iterative Tree-Serial Dynamic Programming
P2DHMM	Pseudo 2D Hidden Markov Model
P2DW	Pseudo 2D Warping
P2DW-FOSE	First-Order Strip Extension of the Pseudo 2D Warping
TRW-S	Sequential Tree-Reweighted Message Passing
TSDP	Tree-Serial Dynamic Programming
TSDP-1	Tree-Serial Dynamic Programming with One Skip
TSDP-W*	Tree-Serial Dynamic Programming with Many Skips
ZOW	Zero-Order Warping

Bibliography

[Bicego & Lagorio^{+} 06] M. Bicego, A. Lagorio, E. Grosso, M. Tistarelli: On the Use of SIFT Features for Face Authentication. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, 35, Washington, DC, USA, 2006. IEEE Computer Society.

[Dreuw & Steingrube^{+} 09] P. Dreuw, P. Steingrube, H. Hanselmann, H. Ney: SURF-Face: Face Recognition Under Viewpoint Consistency Constraints. In *British Machine Vision Conference*, London, UK, Sept. 2009.

[Duda & Hart 73] R.O. Duda, P.E. Hart: *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, 1973.

[Ekenel & Stiefelhagen 06] H.K. Ekenel, R. Stiefelhagen: Analysis of Local Appearance-Based Face Recognition: Effects of Feature Selection and Feature Normalization. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, 34, Washington, DC, USA, 2006. IEEE Computer Society.

[Ekenel & Stiefelhagen 09] H.K. Ekenel, R. Stiefelhagen: Why Is Facial Occlusion a Challenging Problem? In *ICB '09: Proceedings of the Third International Conference on Advances in Biometrics*, pp. 299–308, Berlin, Heidelberg, 2009. Springer-Verlag.

[Forsyth & Ponce 02] D.A. Forsyth, J. Ponce: *Computer Vision: A Modern Approach*. Prentice Hall, us ed edition, August 2002.

[Gollan 03] C. Gollan: Nichtlineare Verformungsmodelle für die Bilderkennung. Diploma Thesis. September 2003.

[Gross] R. Gross: http://ralphgross.com/FaceLabels.

[Hanselmann 09] H. Hanselmann: Face Recognition using Distortion Models. Bachelor Thesis. July 2009.

[Huang & Mattar^{+} 08] G.B. Huang, M. Mattar, T. Berg, E. Learned Miller: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained

Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille France, 2008. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie.

[Hyvarinen & Karhunen$^+$ 01] A. Hyvarinen, J. Karhunen, E. Oja: *Independent Component Analysis*. Wiley-Interscience, May 2001.

[Jain & Zhong$^+$ 98] A.K. Jain, Y. Zhong, M.P. Dubuisson-Jolly: Deformable template models: a review. *Signal Process.*, Vol. 71, No. 2, pp. 109–129, 1998.

[Jolliffe 02] I.T. Jolliffe: *Principal Component Analysis*. Springer, second edition, October 2002.

[Keysers & Dahmen$^+$ 00] D. Keysers, J. Dahmen, T. Theiner, H. Ney, L.F. Informatik: Experiments with an Extended Tangent Distance. In *In Proceedings 15th International Conference on Pattern Recognition*, pp. 38–42, 2000.

[Keysers & Deselaers$^+$ 07] D. Keysers, T. Deselaers, C. Gollan, H. Ney: Deformation Models for Image Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 29, No. 8, pp. 1422–1435, 2007.

[Keysers & Gollan$^+$ 04a] D. Keysers, C. Gollan, H. Ney: Classification of Medical Images Using Non-Linear Distortion Models. In *In Boildverarbeitung fr̈ die Medizin*, pp. 366–370. Springer-Verlag, 2004.

[Keysers & Gollan$^+$ 04b] D. Keysers, C. Gollan, H. Ney: Local Context in Non-Linear Deformation Models for Handwritten Character Recognition. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4*, pp. 511–514, Washington, DC, USA, 2004. IEE Computer Society.

[Keysers & Unger 03] D. Keysers, W. Unger: Elastic image matching is NP-complete. *Pattern Recognition Letters*, Vol. 24, No. 1-3, pp. 445–453, 2003.

[Kolmogorov 06] V. Kolmogorov: Convergent Tree-Reweighted Message Passing for Energy Minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, pp. 1568–1583, 2006.

[Kuo & Agazzi 94] S.S. Kuo, O.E. Agazzi: Keyword Spotting in Poorly Printed Documents using Pseudo 2-D Hidden Markov Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 16, No. 8, pp. 842–848, 1994.

[Levin & Pieraccini 92] E. Levin, R. Pieraccini: Dynamic Planar Warping for Optical Character Recognition. In *Proceedings of the International Conference on Accoustics, Speech and Signal Processing*, Vol. 3, pp. 149–152, 1992.

[Lowe 04] D.G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, Vol. 60, No. 2, pp. 91–110, 2004.

[Martinez & Benavente 98] A. Martinez, R. Benavente: The AR face database. Technical report, CVC Technical report, 1998.

[Mikolajczyk & Schmid 05] K. Mikolajczyk, C. Schmid: A Performance Evaluation of Local Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 27, No. 10, pp. 1615–1630, 2005.

[Moler 04] C.B. Moler: *Numerical Computing with Matlab*. Society for Industrial Mathematics, January 2004.

[Moore 79] R. Moore: A Dynamic Programming Algorithm for the Distance between Two Finite Areas. Vol. 1, No. 1, pp. 86–88, 1979.

[Mottl & Kopylov$^+$ 02] V. Mottl, A. Kopylov, A. Kostin, A. Yermakov, J. Kittler: Elastic Transformation of the Image Pixel Grid for Similarity Based Face Identification. In *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 3*, 30549, Washington, DC, USA, 2002. IEEE Computer Society.

[Ney & Dreuw$^+$ 09] H. Ney, P. Dreuw, T. Gass, L. Pishchulin: Image Recognition and 2D Warping. Lecture Notes. 2009.

[Rentzeperis & Stergiou$^+$ 06] E. Rentzeperis, A. Stergiou, A. Pnevmatikakis, L. Polymenakos: Impact of Face Registration Errors on Recognition. In *AIAI*, pp. 187–194, 2006.

[Ronee & Uchida$^+$ 01] M.A. Ronee, S. Uchida, H. Sakoe: Handwritten Character Recognition Using Piecewise Linear Two-Dimensional Warping. In *Proceedings of the 6th International Conference on Document Analysis and Recognition*, pp. 39–43, 2001.

[Roth & Winter 08] P.M. Roth, M. Winter: Survey of Appearance-Based Methods for Object Recognition. Technical report, Graz University of Technology, Austria, 2008.

[Rucklidge 97] W.J. Rucklidge: Efficiently Locating Objects Using the Hausdorff Distance. *Int. J. Comput. Vision*, Vol. 24, No. 3, pp. 251–270, 1997.

[Sakoe & Chiba 90] H. Sakoe, S. Chiba: Dynamic programming algorithm optimization for spoken word recognition. Vol., pp. 159–165, 1990.

[Sanderson & Lovell 09] C. Sanderson, B.C. Lovell: Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. In *ICB '09: Proceedings of the Third International Conference on Advances in Biometrics*, pp. 199–208, Berlin, Heidelberg, 2009. Springer-Verlag.

[Schiele & Crowley 00] B. Schiele, J.L. Crowley: Recognition without Correspondence using MultidimensionalReceptive Field Histograms. *Int. J. Comput. Vision*, Vol. 36, No. 1, pp. 31–50, 2000.

[Simard & LeCun$^+$ 93] P. Simard, Y. LeCun, J.S. Denker: Efficient Pattern Recognition Using a New Transformation Distance. In *Advances in Neural Information Processing Systems 5, [NIPS Conference]*, pp. 50–58, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[Smith & Bourgoin$^+$ 94] S.J. Smith, M.O. Bourgoin, K. Sims, H.L. Voorhees: Handwritten Character Classification Using Nearest Neighbor in Large Databases. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 16, No. 9, pp. 915–919, 1994.

[Son & Kim$^+$ 08] H.J. Son, S.H. Kim, J.S. Kim: Text image matching without language model using a Hausdorff distance. *Inf. Process. Manage.*, Vol. 44, No. 3, pp. 1189–1200, 2008.

[Tan & Chen 05] K. Tan, S. Chen: Adaptively weighted sub-pattern PCA for face recognition. *Neurocomputing*, Vol. 64, pp. 505–511, 2005.

[Uchida & Sakoe 98] S. Uchida, H. Sakoe: A Monotonic and Continuous Two-Dimensional Warping Based on Dynamic Programming. In *Proc. 14th ICPR*, pp. 521–524, 1998.

[Uchida & Sakoe 05] S. Uchida, H. Sakoe: A Survey of Elastic Matching Techniques for Handwritten Character Recognition. *IEICE - Trans. Inf. Syst.*, Vol. E88-D, No. 8, pp. 1781–1790, 2005.

[Viola & Jones 04] P. Viola, M. Jones: Robust real-time face detection. *International Journal of Computer Vision*, Vol. 57, No. 2, pp. 137–154, 2004.

[Wolf & Hassner$^+$ 08] L. Wolf, T. Hassner, Y. Taigman: Descriptor Based Methods in the Wild. In *Real-Life Images workshop at the European Conference on Computer Vision (ECCV)*, October 2008.

[Zhang & Chen$^+$ 08] L. Zhang, J. Chen, Y. Lu, P. Wang: Face Recognition Using Scale Invariant Feature Transform and Support Vector Machine. In *ICYCS '08: Proceedings of the 2008 The 9th International Conference for Young Computer Scientists*, pp. 1766–1770, Washington, DC, USA, 2008. IEEE Computer Society.