

Performance of Image Registration and its Extensions for Interpolation of Facial Motion

Stella Grasshof, Jörn Ostermann

Institut für Informationsverarbeitung (TNT)
<http://www.tnt.uni-hannover.de>

Abstract. We compare the performance of an intensity based nonparametric image registration algorithm and extensions applied to frame interpolation of mouth images. The mouth exhibits large deformations due to different shapes, additionally some facial features occlude others, e.g. the lips cover the teeth. The closures and disclosures represent a challenging problem, which cannot be solved by the traditional image registration algorithms.

The tested extensions include local regularizer weight adaptation, incorporation of landmarks, self-occlusion handling and penalization of folds, which have all been examined with different weight parameters.

Since the performance of these algorithms and extensions turns out to be superior in case of mouth closures, we provide an algorithm for the automatic selection of deformable template and static reference image for the registration procedure. Subjective tests show that the quality of results for interpolation of mouth images is enhanced by this proposal.

Keywords: frame interpolation, facial motion, nonparametric image registration, variational, optical flow, numeric optimization

1 Introduction and Motivation

The mouth is the most deformable part of the human face, which requires a very flexible setting to rebuild realistic animations of human faces. Additionally humans are very sensitive to unnatural facial movements especially at the mouth part. This implies that the standard error measures and training sets of the optical flow community is not well suited, as they tolerate local deviations of the warp. These pixelwise errors might already lead to very disturbing results if observed by a human which makes subjective evaluation inevitable in the field of facial animation. The challenge in facial motion is to cover large local deformations as well as incorporating occlusions and disclosures which occur due to mouth closing and opening in a sequence.

Several parametric models have been introduced to model the facial movements and rebuild them. One well known model with good performance is the AAM (Active Appearance Models), in [1] the authors take advantage of the asymmetry of image registration and apply an AAM for face tracking, successfully. However one drawback of the AAM model is that it needs manual work to

set some features points in a training set. Additionally it has been shown that it is not scale invariant.

In [2] the authors introduce an image based facial animation system which creates an animation by concatenation of given image sequences. At transitions there might occur large visual differences, which are overcome by the use of morphing, which requires feature points for each mouth image.

Our goal is to improve facial motion interpolation to obtain intermediate frames of a sequence with no feature points and by applying as few manual tasks as possible, but still allow for large local deformations. We focus on nonparametric image registration with an intensity based variational framework for the estimation of deformations which will transform one image completely towards a second one. No parametrization of our deformation function will be assumed, to allow for a high degree of flexibility. We then estimate one two-dimensional vector for each pixel position.

We explore the standard nonparametric intensity based image registration techniques for our application and then introduce extensions to fit the specific task, which then are compared to the baseline methods. We then introduce an automatic choice of deformable and static image for the image registration process which takes advantage of the apparent asymmetry included in the algorithm and thus improves the results.

The paper is structured as follows: In section 2 we introduce the basics of nonparametric intensity based image registration. In section 3 we then describe how it can be used to estimate intermediate frames. We present some extensions in section 4 followed by their experimental evaluation in section 6, which are then discussed in 7.

2 Image Registration

In this section we will give a very brief introduction to the general image registration framework, which we later refer to for our applied adaptations. We used the approaches described in [3, 4] and refer to this books for more details.

Given two gray-scale images, the traditional image registration task is to find a deformation φ such that one image becomes as similar as possible to the other. This problem statement incorporates one image to be a static *reference* image R , while the other will be a deformable *template* T to match the reference. For our application this means that we want to transform one frame T towards the next known frame R . This is a very imprecise formulation which will be refined in the following to introduce our mathematical framework.

Generally the *images* are modeled as functions, which assign one scalar, positive gray-value to each location in the twodimensional space: $R, T : \mathbb{R}^2 \rightarrow \mathbb{R}$. Additionally we define the images to be greater than zero on a predefined image domain $\Omega \subset \mathbb{R}^2$ and zero elsewhere. To obtain these functions from the provided pixel images, we used a cubic BSpline interpolation, such that the images can be evaluated at any position rather than only their pixel positions itself.

To quantify the *similarity* of two images the sum of squared pixelwise differences (sum of squared differences) is used.

The *deformation* φ is defined as a vectorfield, mapping one location $x = (x_1, x_2)$ in the deformed template image T_φ to a position $\varphi(x)$ in T

$$\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}^2, \quad \varphi(x_1, x_2) = (\varphi_1(x_1, x_2), \varphi_2(x_1, x_2)) \quad (1)$$

The grayvalue of the deformed template image T_φ at the position x is determined by evaluation of the template image T at the position $\varphi(x)$:

$$T_\varphi(x_1, x_2) = T(\varphi(x_1, x_2)) \quad (2)$$

This approach is known as backwards transformation with *Euler* coordinates, while the opposite direction is known as *Lagrange* (see [3] for details). Now that all parts of the introduced problem have been defined, it can be described as the following minimization task:

$$\min_{\varphi} \frac{1}{2} \int_{\Omega} (T(\varphi(x)) - R(x))^2 dx \quad (3)$$

Due to the fact that the deformation φ is not restricted, it could simply be any function, which makes (3) an ill-posed problem which is not solvable uniquely. Therefore at least one penalty term is included which restricts the deformation φ to have certain properties, e.g. penalizing large values in the derivatives of the deformation, to ensure the estimated function φ will be smooth. These *smoothers* are added with an additional positive weighting parameter to the distance measure:

$$\min_{\varphi} \frac{1}{2} \int_{\Omega} (T(\varphi(x)) - R(x))^2 dx + \alpha S(u), \quad \alpha \in \mathbb{R}^+, \quad u(x) = \varphi(x) - x \quad (4)$$

For our tests we choose the well known, simple *diffusive* regularizer, which is defined as

$$S(u) := \frac{1}{2} \sum_{k,l=1}^2 \int_{\Omega} (\partial_{x_k} u_l(x))^2 dx = \frac{1}{2} \sum_{k=1}^2 \int_{\Omega} \|\nabla u_k(x)\|_2^2 dx \quad (5)$$

For our application we also tested the *curvature* regularizer (penalizing large values of the second derivatives) [3, 4], but found the differences to be negligible in our experiments.

2.1 Numeric Optimization Framework

Since there is no analytical solution to (4), we apply the well-known first-discretize-then-optimize framework [3]. To discretize the functions involved in (4) we use a cell-centered grid x_c , with grid points centered at the pixel positions, on the image domain Ω . We use finite differences to estimate the involved derivatives.

The discretized optimization function (4) is formulated as a least squares problem, which favors the use of the Gauss-Newton method to iteratively determine an estimate for the unknown deformation φ [5]. This also holds for the extensions which we introduce later.

In each optimization step the new estimate of the deformation field is calculated by a linesearch procedure, which makes use of the preceding estimate and adds a new direction, weighted by a specific stepsize. While the Gauss-Newton Algorithm gives the direction, we apply the Armijo Algorithm to determine the stepsize.

The optimization procedure delivers the transformation φ evaluated at the initial grid points x_c , which were chosen to be the pixel positions. The advantage of this procedure is that we receive an intensity value for each pixel position of the deformed image T_φ , such that no further interpolation tasks have to be performed on T_φ to obtain one grayvalue for each pixel position.

3 Using Image Registration for Frame Interpolation

Image registration has been introduced as finding a one-way deformation, which fully deforms one image to match another one. On the other hand frame interpolation tasks need intermediate images and therefore an intermediate deformation estimate. In a frame interpolation setting the assumption that one pixel performs a linear movement between two frames is common and reasonable, which means the positional change from x to $\varphi(x)$ is expected to be linear. We therefore introduce the *displacement* vector u corresponding to a given transformation φ as in (4)¹

$$u_\varphi(x) = \varphi(x) - x \quad (6)$$

Assuming the transformation φ to deform T to R completely, the corresponding displacement u_φ describes the full motion. If we desire to compute half of the motion from T to R we get

$$u_{\varphi_{\text{half}}}(x) = \frac{u_\varphi(x)}{2} = \frac{\varphi(x) - x}{2} \quad (7)$$

The corresponding deformation is determined by

$$\varphi_{\text{half}}(x) = u_{\varphi_{\text{half}}}(x) + x = \frac{1}{2}(\varphi(x) + x) \quad (8)$$

We are now able to define intermediate deformations by adapting the denominator in (8) to the specific task.

¹ Please note that the definition of these terms differs in the literature.

4 Extensions for Frame Interpolation for Facial Motion

In the application of facial motion interpolation, especially in the case of the mouth, we face problems like large local deformations as well as occlusions and disclosures which lead to discontinuities. In the traditional image registration task the problem of disappearing and reappearing image parts is not modeled. We therefore need to adapt the classic optimization function to better fit our needs. In this section we describe some extensions for the basic image registration framework. Their performance will be discussed in section 6.

Local regularizer weight adaptation We expect the deformations in the mouth part to be larger than in the surrounding facial parts (e.g. chin, nose, cheeks), so we propose to incorporate this knowledge by demanding lower smoothness and setting lower regularizer weights in our region of interest to enable greater movements, but still restrict the surrounding parts. Applied to the image registration problem in (4) we change the constant registration weight α to a position dependent map $\alpha(x)$, which is then included in our smoother function (5)

$$S(u, \alpha) = \frac{1}{2} \sum_{k,l=1}^2 \int_{\Omega} \alpha(x) (\partial_{x_k} u_l(x))^2 dx \quad (9)$$

In (4) the term $\alpha \cdot S(u)$ is replaced by $S(u, \alpha)$ of (9).

Landmarks Though the use of landmarks has been proven to improve image registration results [6, 7], the issue remains that they must be provided with high accuracy for each image and often need to be set manually. We additionally face the problem that we choose to not imply any parametrical representation of our transformation φ , which means that we only know the values of our transformation at the grid point positions. Inspired by [7] where the authors incorporate an individual landmark inaccuracy for each landmark pair, we assume that the landmarks are located at pixel positions (grid points), which avoids the interpolation of the deformation φ for subpixel locations. Landmarks allow us to influence the direction of the deformation estimation process. Defining r_k and t_k , $k = 1, \dots, n$ as the corresponding twodimensional landmarks, i.e. specific grid points, for reference and template image, we then introduce the landmark penalty for any deformation φ as

$$\mathcal{L}(\varphi) = \frac{1}{2} \sum_{k=1}^n (\varphi(r_k) - t_k)^2 \quad (10)$$

For the experiments, this term is then added to (4) with an additional positive weighting parameter.

Self-Occlusion Handling As a matter of fact transforming an open mouth with teeth to a closed mouth without teeth implies occlusion of the teeth. Therefore

a paper for Gay-Bellile [8] caught our attention as it deals with self-occlusion estimation in a nonrigid image registration setting. The authors propose a binary self-occlusion map, by defining a pixel as occluded if the minimum of the squared directional derivative of the deformation φ is below a certain threshold, i.e. if two gridpoints are mapped onto one another. Each occluded pixel will be ignored in the distance measure. Additionally a folding penalty is introduced with the goal to favor shrinkage at the self-occlusion boundaries. The corresponding term is based on the observation that the sign of left and right derivative of the transformation φ differs at these places.

The performance of the introduced algorithm is demonstrated on image sequences of deformed surfaces, where the total transformations are of a more global character than in our problem setting. In [9] the authors point out the limits of their work are reached by deformations which violate the smoothness constraint. Occlusions of teeth in a sequence of opening and closing mouth movements will lead to nonsmooth deformations.

Penalizing Folds We observed that at places where folds are

$$\varphi_1(x_1^i, x_2^j) > \varphi_1(x_1^l, x_2^j), \quad x_1^i < x_1^l, \quad \varphi_2(x_1^i, x_2^j) > \varphi_2(x_1^i, x_2^l), \quad x_2^j < x_2^l$$

The condition for no folds is defined as

$$\varphi_1(x_1^i, x_2^j) < \varphi_1(x_1^l, x_2^j), \quad \forall x_1^i < x_1^l, \quad \varphi_2(x_1^i, x_2^j) < \varphi_2(x_1^i, x_2^l), \quad \forall x_2^j < x_2^l \quad (11)$$

This might include the equal sign for the case of shrinkage. It basically means that the deformation φ must preserve the relative neighborhood of the original grid to avoid folds. Assuming that no global deformations (e.g. rotations) occur, we therefore claim φ_k to increase monotonously in its k th component, to not contain folds. We therefore agree with [8] here that the deformation is not injective at places where folds occur. But instead of only focusing on boundaries of these areas, we propose to penalize the entire area of the fold by its current extent. We therefore define our folding penalty as

$$\mathcal{F}(\varphi) = \frac{1}{2} \sum_{k=1}^2 \int_{\Omega} g(x) \cdot (\partial_k \varphi_k(x))^2 dx, \quad g(x) = \begin{cases} 1 & , \text{ if } \partial_k \varphi_k(x) < 0 \\ 0 & , \text{ else} \end{cases} \quad (12)$$

In contrast to (5), we do not require mixed derivatives. The incorporated threshold inside of the function g of (12) might be altered.

We have proposed adaptations for the traditional image registration optimization function itself. We now propose to incorporate former knowledge of the two involved images by automatically choosing template T and reference image R to benefit from the asymmetry of image registration.

5 The Automatic Choice of Deformable Template and Static Reference Image

Within a frame interpolation task the intuitive approach to obtain an estimate for frame i is to deform frame $i - 1$ towards $i + 1$ and then apply the intermediate transformation φ_{half} of (8) to frame $i - 1$, which is a highly asymmetric procedure. We argue that the choice of deformable and static image should not be arbitrary. On the other hand a symmetric treatment of the images would not take advantage of the present asymmetry of known mouth movements. Comparing two arbitrary images of a speaking mouth sequence, one is likely to behold more information than the other, in terms of visibility of teeth and tongue.

We therefore aim to find a qualitative vote (weight) for the amount of inner mouth pixels for each frame, to automatically choose deformable and static image. The pixelwise, automatic segmentation of the lips is known to be very complicated and error-prone, due to low color difference between lips, skin and tongue, and the presence of facial hair.

We experimented with thresholds in the colorspace RGB, YUV, HSV, CieLab and combinations. In [10] the hue channel of the HSV colorspace is used to roughly determine the lip region. The authors point out that gaps and noise is visible, which we can confirm. We in contrast observed the a-channel of CieLab to be more reliable and less noisy. To decide whether a pixel belongs to the inner part of the mouth (if any), we use the following algorithm:

1. Choose bounding box of mouth area for one image of the sequence²
2. Determine CieLab and YUV colorspace for each frame
3. Define pixel as lip pixel if a-channel (CieLab) is bigger than a threshold λ_a ³
4. Eliminate Outliers (only if the bounding box was too large or inaccurate)
5. Define convex hull of lip-pixels as mouth area
6. Define each pixel inside the mouth area as a inner mouth pixel if it is no lip pixel (a-channel is below λ_a) and:
 - L-channel of CieLab is below a threshold λ_L (allows detection of dark areas) OR
 - V-channel of YUV is below a threshold λ_v (Find low red values to obtain teeth area, which is more reliable than looking for large values in the L-channel)⁴
7. Define number of inner mouth part pixels as weight of this specific mouth image

The drawback of this method is of course that the three thresholds λ_a , λ_L , λ_v have to be chosen manually. The advantage of this algorithm is that it must

² Our sequences have the mouth at the same positions for all frames already. But for sequences where this is not the case this task could be easily replaced by an automatic algorithm, e.g. OpenCV.

³ It should be preferred to underestimate the lips region rather than to overestimate it.

⁴ Surprisingly we found teeth could not be detected well by selecting pixels over a certain threshold for the L-channel.

only be done once for each face model. Additionally we do not require the results to be highly accurate because we do not use it for a segmentation tasks, which would require higher precision, so this approach tolerates errors.

Now that a weight for each image is available, we choose the image of the two as deformable template T , which has a higher weight (amount of inner mouth pixels) compared to the second image.

6 Experimental Results

Our experiments focus on the evaluation of the introduced methods for large local deformations of the mouth. Therefore we choose several sequences of two speakers with a frontal viewpoint and low illumination changes. This results in image sequences with low variation in mouth position, which enables to focus on local deformations rather than global ones, which would be inevitable if head movements were present in the sequence. We filmed the complete head including the upper parts of the shoulders, though we will focus on the mouth parts in the following. For testing a total of seven sequences from two individuals were applied, which originally consist of 56 to 212 frames, where the mean frame number is 116.5 frames.

Measuring Quality of Results How is an image registration result judged to have a good quality if there is no ground truth? Of course the deformed template image should match the reference image, because that was what the algorithm was supposed to do, but additionally we desire the transformation to be as smooth as possible. Searching for the smoothest transformations in the results is not reasonable as it would naturally lead to the unity transform. On the other hand focusing on the transformations with the lowest resulting distance measure, we observed that noisy, erroneous grayvalues occur in the middle of the mouth part. This happens due to local minima of the optimization function for small weights α which lead to nonsmooth estimations results for the deformation φ .

Pixelwise errors will disturb the human observer even more than bigger, yet more homogeneous deviations which affect the total mouth shape instead of only isolated pixels. Unfortunately, these pixelwise errors are exactly what we obtain if we focus on the minimum available optimization function value. With this conclusion we realize that there is no objective measure available yet to value the realism of a transformation applied to a mouth image. We therefore focus on visual evaluation, which is a sufficient tool for now, as the presented results have easily visible limitations. We tested different settings of weights for all extensions introduced in section 4 and present the best results. The algorithms are applied on grayscale images, but to make the errors apparent, we will show our results of estimated deformations applied to color images.

Preprocessing It has been shown that a reasonable preprocessing of the input images may influence the quality of the transformation result immensely. To inspect the effect of the preprocessing method on our estimations, we tested none,

smoothing and bilateral filtering. We additionally checked the performance of using the smoothed input images first, and then use this result as an initialization for the original images registration process, which was meant to be an alternative to the multiresolution framework, which we tested additionally. We observed that the latter serves the best results and leads to similar outputs as the multiresolution framework.

Multiresolution A common and well known practice to estimate large deformations is to compute a transformation on lower resolutions, then upscale it and use it as a starting value for the next higher resolution step. This is especially effective for large deformations, for which the estimation procedure might be trapped in local minima of the optimization function otherwise. We tested different resolutions and steps sizes, but found that the results are very similar and that big deformations are still not recovered better than estimations which have been done on the images with original size.

Local Regularizer Weight Adaptation As expected, we observe that weakening the local weight adaptation leads to higher deformations in the desired areas, which on the other hand leads to very unsmooth transformations with large local deviations, causing noisy errors in the deformed images.

Landmarks Including the landmark penalty term (10), we expected the landmarks to influence the direction estimation in our numeric optimization framework positively. We observe that with increasing weighting parameter, the landmarks will more probably overlap, but this on the other hand will harm the smoothness of the total deformation. Vice versa a low landmark weighting parameter might lead to smoother deformations, but the overlap of the provided landmarks will be weak, such that the use of the landmark penalty in the current setting does not lead to a convincing improvement.

Self-Occlusion Handling We discovered that this framework does not estimate the occluded areas of our deformed grids correctly. The shrinkage term has been tested separately and additionally to the self-occlusion estimation, but both approaches do not improve our results compared to the simple image registration setting. We assume this to be the effect of apparent discontinuities in the estimated deformation φ .

Folding Penalty Our folding penalty (12) leads to smoother deformations with increasing weight parameter. The drawback is that it thus weakens the total amount of deformation, meaning that the estimated transformation φ becomes similar to the identity transform with increasing weight parameter. This will diminish areas of shrinkage which might have been desired. We state that the penalty punishes folds without explicitly punishing or favoring shrinkage, though the latter might be a desirable feature.

We tested our folding penalty on deformations φ which induce folds in the deformed grid and observed that it can detect folds successfully and robust

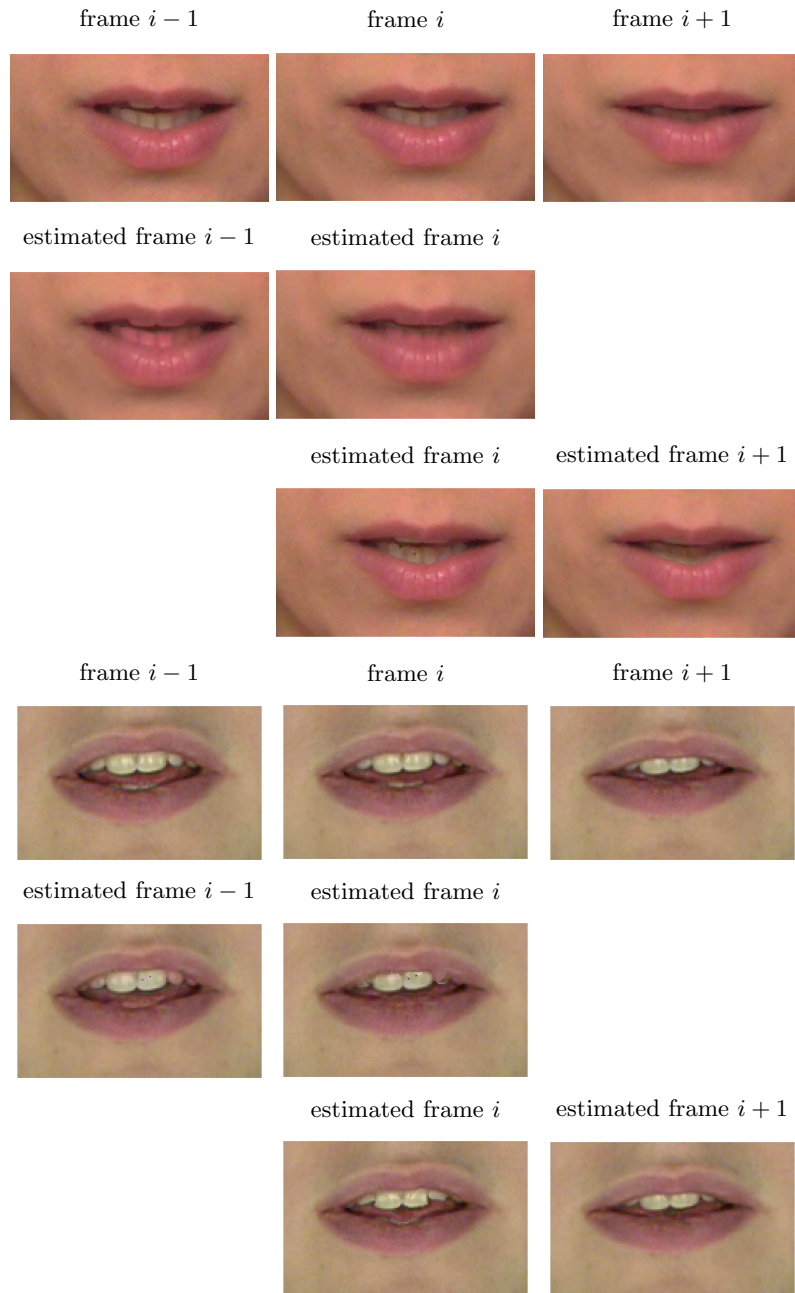


Fig. 1: Illustration of direction dependency for deformations, showing that the wrong direction choice leads to pink teeth for two examples taken from a total of seven different sequences. Top row shows three consecutive original frames. The second row shows the result if the deformation φ is estimated by using $i + 1$ as template and $i - 1$ as reference image. For the bottom row the choice was the other way around, i.e. from open mouth $i - 1$ to closed mouth $i + 1$.

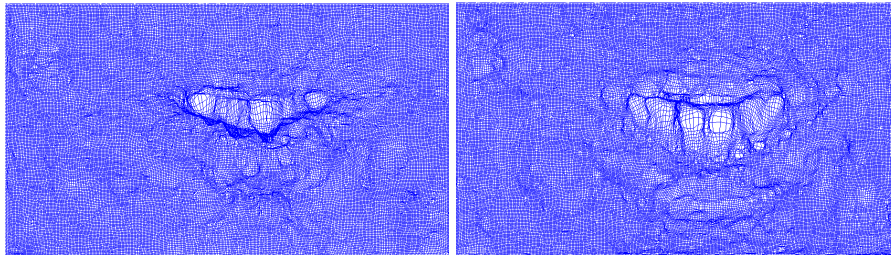


Fig. 2: Deformed grids corresponding to the results, displayed in the upper part of figure 1. The left image shows the deformation resulting from estimating the transformation from a closed to an open mouth, while the right image displays the result of applying the other direction.

without inclusion of areas with shrinkage. This might serve as an advantage for other applications.

Automatic Choice of Deformation Direction Our results verify the expectation formulated in section 5 that the result is improved by the choice of a reasonable transformation direction. Figure 1 shows the original consecutive frames at the top row. To enable the calculation of the intermediate image and images in between, we performed the image registration task from frame $i - 1$ to frame $i + 1$ (second row) and vice versa (third row). The resulting fully transformed template images are displayed below the corresponding original reference image. The third row contains the results obtained by automatically choosing template T and reference R and matches the corresponding original frames at the top better than the second row for both sequences. The second row contains the results of image registration from a closed to an open mouth, which then results in pink teeth rather than white. This effect is hardly noticeable in grayscale images. It occurs because the color information which is needed for the teeth is not available in the chosen template image for 1, as it only holds colors for skin and lips. In bottom example of figure 1 the difference between the original frames $i - 1$ and $i + 1$ is less than in the upper example, but the reasonable registration direction leads still to a better result. Figure 2 shows the corresponding deformed grids for the upper case of image 1. It can be seen that the grid for the worse estimate on the left contains folds, which shows that the estimated deformation has not the smooth character which we desire.

7 Discussion

We applied nonparametric image registration techniques to frame interpolation tasks of mouth images and tested different methods on seven sequences of two individuals. Given two images, known methods did not lead to satisfying results, however we showed that the right choice of deformable template image T and static reference image R is able to increase the quality of the results by

taking advantage of known asymmetry of image contents during the process of opening and closing of mouths. Our introduced simple method for automatic image selection is able to decrease the necessary manual work considerably, as only three color thresholds have to be selected for each individual rather than several landmarks for at least 20 mouth images, which was the case before. Our proposed automatic direction selection was also applied to interpolate a 25Hz mouth sequence to 50Hz by estimating intermediate frames. In contrast to one-directional estimates, we obtain a fluid animation of a speaking mouth, which was confirmed by subjective tests.

One general drawback remains the manual evaluation of results, because there is no objective measurement available yet which serves the human sense of visual disagreement and realism in mouth sequences and their comparison, which can be seen as a possible future work task.

Additionally the different types of deformations for lips and teeth are not yet incorporated. It might be useful to incorporate a rigidity constraint for the teeth only and allow for more freedom in the surrounding.

References

1. Dedeoglu, G., Kanade, T., Baker, S.: The Asymmetry of Image Registration and Its Application to Face Tracking. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 807–823 (2007)
2. Liu, Kang, Weissenfeld, A., Ostermann, J., Luo, Xinghan: Robust AAM building for morphing in an image-based facial animation system. In: 2008 IEEE International Conference on Multimedia and Expo, pp. 933–936, (2008)
3. Modersitzki, J.: Numerical Methods for Image Registration. Oxford University Press (2003)
4. Modersitzki, J.: FAIR: Flexible Algorithms for Image Registration. SIAM (2009)
5. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer (2006)
6. Johnson, H.J., Christensen, G.E.: Consistent landmark and intensity-based image registration. In: IEEE Transactions on Medical Imaging, vol. 21, pp. 450–461 (2002)
7. Papenberg, N., Olesch, J., Lange, T., Schlag, P., Fischer, B.: Landmark Constrained Non-parametric Image Registration with Isotropic Tolerances. In: Bildverarbeitung für die Medizin 2009, pp. 122–126, Springer Berlin Heidelberg (2009)
8. Gay-Bellile, V. and Bartoli, A. and Sayd, P.: Direct Estimation of Non-Rigid Registrations with Image-Based Self-Occlusion Reasoning. In: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007, pp. 1–6 (2007)
9. Gay-Bellile, V., Bartoli, A., Sayd, P.: Direct Estimation of Nonrigid Registrations with Image-Based Self-Occlusion Reasoning, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 87–104 (2010)
10. Bacivarov, I., Ionita, M.C., Corcoran, P.: A combined approach to feature extraction for mouth characterization and tracking. In: Signals and Systems Conference, 208. (ISSC 2008). IET Irish, pp. 156–161 (2008)