Eurographics Workshop on 3D Object Retrieval (2022)
S. Berretti, T. Thehoaris, M. Daoudi, C. Ferrari, and R. C. Veltkamp (Editors)

Short Paper

Reconstructing 3D Face of Infants in Social Interactions Using Morphable Models of Non-Infants

E. Sariyanidi[1], C. J. Zampella[1], M. N. Drye[1], M. L. Fecher[1], G. Megginson[1], L. Soskey Cubit[1], R. T. Schultz[1,2], W. Guthrie[1,2] & B. Tunc[1,2]

[1] Center for Autism Research, The Children's Hospital of Philadelphia, United States
[2] University of Pennsylvania, United States

Abstract

3D morphable models (3DMMs) simultaneously reconstruct facial morphology, expression and pose from 2D images, and thus could be an invaluable tool for capturing and characterizing the face and facial behavior in early childhood. However, 3DMM fitting on infants is a largely unexplored problem. All publicly available 3DMMs are developed for adults, and it is unclear if and to what extent they can be used on videos of infants. In this paper, we compare five state-of-the-art 3DMM fitting methods on data from naturalistic infant-caregiver interactions. Results suggest that it is possible to produce consistent and subject-specific reconstructions of 3D shape identity from multiple frames, but not from a single frame. Qualitative evaluation highlights that facial regions with high texture variation, such as eyes, brows and mouth, are captured with higher accuracy compared to the rest of the face. Thus, even though a 3DMM developed for adults has significant limitations when reconstructing the morphology of the entire facial region of infants, applications that involve analysis of facial behavior can be feasible. Our encouraging results, combined with the unique ability of 3DMMs to disentangle two major sources of noise for expression analysis (i.e., identity bias and pose variations), motivate future research on using 3DMMs to measure the facial behavior of infants.

CCS Concepts

• ***Human-centered computing*** → ***Empirical studies in HCI;*** • ***Computing methodologies*** → ***Shape representations;***

1. Introduction

Social interactions in early childhood play a critical role in development and learning, and studying behavior in the context of these interactions can provide rich information about a wide range of developmental skills and help to predict future outcomes [LAVS*16]. Facial expressions are one behavior employed by children from early infancy to communicate and bond with others during social interactions. Automated analysis via computer vision has the potential to significantly improve the accuracy and granularity with which the facial behavior of infants can be measured, and thus to become an invaluable tool for developmental and clinical researchers interested in early social-emotional development.

3D morphable model (3DMM) fitting is a topic that has regained significant momentum in recent years, due partly to the advent of deep learning, which led to the creation of methods that work robustly in uncontrolled imaging conditions. This reinvigorated interest is a promising development for studying facial behavior in infants, because 3DMM fitting has the potential to not only reconstruct the 3D face shape, but also to parametrize facial pose, identity and expression. However, despite a growing body of research, 3DMM fitting remains almost exclusively studied on the faces of adults. To our knowledge, there is no publicly available 3DMM constructed from infant data (Section 2.3), and learning-based 3DMM fitting approaches are trained predominantly by images of adults (Section 2.1). Faces of infants and adults have anatomical differences, therefore errors are likely to emerge when one uses a 3DMM from adults on videos of infants. Moreover, analyzing faces of infants has inherent difficulties, as infants tend to make large head movements and expressions, especially during naturalistic interactions. As a result, it is unclear if and to what degree state-of-the-art 3DMM fitting methods can work on infant faces.

This paper takes a critical step towards filling this gap in the research on 3DMM fitting methods by evaluating their performance on faces of infants during naturalistic interactions. Specifically, we conduct a quantitative comparison of five state-of-the-art approaches in terms of ability to produce consistent and person-specific estimates of neutral face shapes in infants. Furthermore, we provide qualitative results on measuring facial pose and expression in infants with 3DMMs. Our analysis allows us to identify 3DMM fitting methods and practices that lead to improved 3D reconstruction; and to discuss the applications for which 3DMM fitting on infants seem feasible even when a morphable model of non-infants is used. The unique contributions of this paper are as follows:

- To our knowledge, we conduct the first quantitative comparison of state-of-the-art 3DMM fitting methods on videos of infants from naturalistic dyadic interactions (Fig. 1).

DOI: 10.2312/3dor.20221178

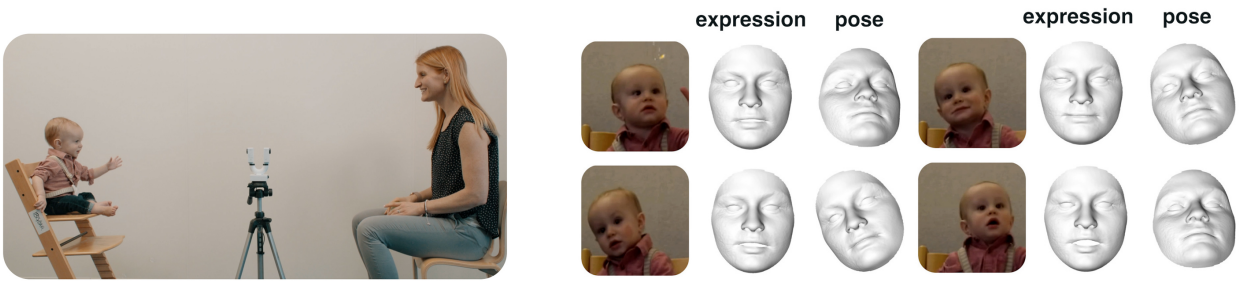

Figure 1: *Left: Illustration of the infant-caregiver interaction paradigm used in this study. Face videos of the infants were collected using the 2D camera device in the middle of the frame. Right: Several video frames of the infant during the interaction, along with the output of 3D face (expression and pose) reconstruction for the corresponding frames.*

- We show that producing estimates of (neutral) 3D face shapes that are consistent and person-specific is possible when estimation is done from multiple frames but not from a single frame.
- We show that the performance of methods that can *jointly* fit to multiple frames of a subject can be improved through repeated estimation on random combinations of frames.
- We provide qualitative evidence showing promise regarding 3DMMs' ability to quantify facial pose and expressions in infants, despite limitations in reconstructing the entire face.

The rest of this paper is organized as follows: Section 2 discusses recent progress in 3DMM fitting, further delineating the gaps addressed by this study. Section 3 provides the background and notation for 3DMM fitting and lists the ways in which multi-frame fitting can be performed, which turns out to be an impactful choice in this study. Section 4 presents the experiments, and Section 5 discusses the implications of this study, highlighting the applications of 3DMM fitting on infant faces that are feasible and those that seem less so. Finally, Section 6 concludes the paper.

2. Related work

In this section we first provide a brief summary of the methods for 3DMM fitting with an emphasis to their application on faces of infants (Section 2.1), then discuss metrics that can be used in our context (Section 2.2), and finally discuss recent progress in terms of 3DMM fitting on faces of infants (Section 2.3).

2.1. Methods for 3DMM fitting

3DMM fitting methods can be categorized as *optimization-based* and *learning-based*. Optimization-based methods estimate the 3DMM parameters in a given image by minimizing a cost function that measures the discrepancy between the given and the 3DMM-generated image. Most methods use unconstrained pseudo second-order optimization [RV05, SWTC14, BRV*18], but recently, inequality-constrained optimization [SZST20b] has also been proposed. Learning-based methods, on the other hand, do not optimize for every given image from scratch. Instead, they train a regressor a-priori (*i.e.,* offline) with large amounts of data, and use this regressor to estimate the 3DMM parameters from new images [EST*20]. It must be noted that the training of this regressor involves usage of face images of adults [DYX*19, GZY*20, BCLT21], which may be an extra limitation for 3DMM fitting on faces of infants, beyond the fact that the 3DMM itself is also from adults. Our experiments address this open issue, and suggest that some learning-based methods are significantly better than others in terms of producing consistent and subject-specific estimates of (neutral) face shapes (Section 4.3).

Most learning-based methods can only take a single frame as their input. Recently, multi-frame methods, such as INORig [BCLT21], have been proposed. The advantage of multi-frame methods is that, by *jointly* fitting a 3DMM to multiple frames of a person, they reduce the ill-posedness of the problem, since the reconstructed shape (identity) must be consistent across different frames. Note that optimization-based methods are readily capable of multi-frame fitting with rather straightforward modifications (*e.g.,* updating the Jacobian term [BRV*18]).

In this paper, we evaluate five state-of-the-art methods (Section 4.3), which include single- and multi-frame methods that are based on optimization or learning. To our knowledge, we present the first study to apply such a diverse set of techniques to data of infants within the context of naturalistic interactions. The results of our comparison and the implications of these results are discussed respectively in Section 4 and Section 5.

2.2. Evaluation metrics

The standard metric for validating 3DMM fitting methods is geometric error, which measures the discrepancy between the 3D ground truth shape and the estimated 3D shape. However, to our knowledge, there is no public dataset of infants in naturalistic interactions with 3D ground truth. Moreover, the geometric error is difficult to apply in the presence of strong facial expressions, as the correspondence problem becomes particularly difficult [FBPDB21].

A number of studies have recently used metrics that do not require 3D ground truth. One of these metrics is face recognition performance based on the 3DMM generated image [BCLT21]. This metric operates on the 2D image space, and uses the generated texture as well as the shape of the person. However, some 3DMM fitting methods do not even fit a texture model [GZY*20, SBFB19], but only a shape model. Moreover, the predominant interest in our study is analyzing facial shape (see also Section 3). Thus, we avoid using metrics that rely on estimated facial texture.

Recently, another metric that needs no 3D ground truth has been proposed: Within- and between-subject effect size (WBES) [SZST22]. This metric measures the ability of a method to estimate neutral face shapes (*i.e.,* shape identity; see Section 3) that are consistent (*i.e.,* similar for images of the same person) and person-specific. Since a 3D ground truth is not required, this metric can also be applied to infant videos collected only with 2D cameras (as will often be the case for naturalistic interaction data). The quantitative experiments of this study are based on WBES, which is defined more precisely in Section 4.2.

2.3. 3DMM fitting for faces of infants

Despite the significant research devoted to 3DMMs, their use on faces of infants has remained largely unexplored. A recent study presented a novel method to construct 3DMM from face scans of infants [MPT*20] and used it to construct BabyFM—a morphable model for infants. BabyFM was further applied to reconstruction of faces of infants from 2D data [MPL*22]. However, the test dataset of the latter study was from controlled face images, rendered from 2D scans. On the contrary, our study applies 3DMM fitting to infants in naturalistic social interactions, where the babies typically move their head freely and display large facial expressions. Moreover, to our knowledge, neither BabyFM nor any other infant-based 3DMM is available to third parties for research or industrial/commercial purposes. In contrast, many researchers have access to and use 3DMMs constructed from faces of non-infants, such as the Basel'09 [PKA*09], Basel'17 [GMB*18], FLAME [LBB*17] and LSFM [BRP*18] models. Therefore, it is an important open question if and to what degree one can use 3DMMs constructed from adults on infant data. While Morales *et al.* [MPT*20] partly answer the latter question by showing that 3DMMs of adults are limited in reconstructing the entire facial region (see also Section 4.5), 3DMMs are multi-purpose tools, and it remains to be seen whether adult 3DMMs can be applied to infant data for other applications, such as facial pose and expression quantification. Thus, our study provides further investigation to shed light on this question. The quantitative and qualitative experiments presented in this paper allow researchers to make more informed decisions about which applications are viable and which are less likely so when applying a 3DMM constructed from face scans of adults to images of infants (Section 5).

3. Background on 3DMM fitting

The main strength of 3DMM fitting, compared to alternative 3D face reconstruction approaches (*e.g.,* direct methods [JBAT17]), is that it can disentangle and parametrize the three predominant factors that underlie a facial image, namely, the facial pose, identity, and expression. Although 3DMM fitting can be done on a single frame, it may be unrealistic to expect the aforementioned 3-fold decomposition to be accurate in this case, particularly if the frame contains a spontaneous and unknown facial expression. For example, if the eyebrows of a subject appear to be higher compared to the average face, this may be either due to the subject's facial morphology, or due to the facial expression in the frame (*i.e.,* raised eyebrows). For this reason, it is critical to make the most of the available data and do *multi-frame* estimation when we have video data, particularly in the case of infants. Below we first introduce our notation for 3DMM fitting, and then describe three ways in which multi-frame estimation can be performed.

3.1. Notation

Suppose that the 3D shape of the face of a person in a world coordinate system is represented with a dense mesh of N points, denoted as $\mathbf{p} \in \mathbb{R}^{3N}$. It is reasonable to assume that $\mathbf{p}$ depends on three factors, namely, person-specific facial morphology, facial expression and facial pose (*i.e.,* rotation and translation). Fitting a 3DMM to a given image frame $\mathbf{I}$ amounts to estimating the parameters that govern these factors. That is, if we use a linear 3DMM, $\mathbf{p}$ can be assumed to be the result of the following equation

$$\mathbf{p} = \mathbf{R}(\bar{\mathbf{p}} + \mathbf{A}\boldsymbol{\alpha} + \mathbf{E}\boldsymbol{\varepsilon}) + \boldsymbol{\tau}, \tag{1}$$

where $\bar{\mathbf{p}}$ is the average face of the 3DMM; $\mathbf{A}$ and $\mathbf{E}$ are matrices that respectively represent the identity and expression basis of the 3DMM; $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ are the vectors that respectively represent the identity and expression of the subject; $\mathbf{R}$ is a matrix that applies a (common) rotation to all points in the mesh, and $\boldsymbol{\tau}$ is a vector represents the translation from origin.

3.2. Multi-frame fitting

When we have multiple frames per subject (*e.g.,* a video), it is natural to try and make the most of available data. In this case, our input is a set of T frames of a subject $\mathbf{I}_1, \ldots \mathbf{I}_T$, and we can define the dense mesh at the tth frame as $\mathbf{p}_t$, and the expression, rotation and translation parameters as $\boldsymbol{\varepsilon}_t$, $\mathbf{R}_t$ and $\boldsymbol{\tau}_t$ for $t = 1, \ldots, T$. Note that the identity parameter, $\boldsymbol{\alpha}$, does not depend on frame, since the identity-specific facial characteristics of a person do not change within the video. Thus, multi-frame fitting can improve the accuracy of the estimated identity parameter, $\hat{\boldsymbol{\alpha}}$, since we can observe the face of the same subject from different angles and with various expressions, thus rendering the problem less ill-posed. As explained below, multi-frame fitting can be done in at least three ways.

3.2.1. Naive averaging

One can fit a 3DMM to each of the frames in $\{\mathbf{I}_t\}_{t=1}^{T}$ independently, thus produce estimates $\{\hat{\boldsymbol{\alpha}}_t\}_{t=1}^{T}$, and then average them over to produce the final estimate as

$$\hat{\boldsymbol{\alpha}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\alpha}}_t. \tag{2}$$

Most 3DMM fitting methods are single-frame, but naive averaging allows them to be extended to multi-frame fitting. Despite its simplicity, naive averaging has been used widely [GCM*18, DYX*19] and can improve performance significantly [DYX*19].

3.2.2. Joint multi-frame estimation

Recently, a number of natively multi-frame 3DMM fitting methods have been proposed. The advantage of these approaches is that they can simultaneously fit a 3DMM to M frames by estimating separate expression and rigid transformation parameters per frame but keeping the estimated identity $\hat{\boldsymbol{\alpha}}$ common across the M frames. This approach better represents the problem by using the fact that

identity-related shape details do not change within a video, and typically outperforms naive averaging [DYX*19].

3.2.3. Averaging over multi-frame estimations

While multi-frame 3DMM fitting can improve performance over naive averaging, it may not be feasible if the number of frames is too large, since computational complexity can become prohibitive due to the added per-frame expression and pose parameters. In such cases, it can be more appropriate to split the T frames into K groups of M frames, $\mathcal{I}_1, \ldots, \mathcal{I}_K$, where $\mathcal{I}_k$ is a set that contains the indices of the frames corresponding to the kth group. Then, the identity $\hat{\boldsymbol{\alpha}}$ can be estimated by performing K multi-frame estimations on the K groups and averaging them as

$$\hat{\boldsymbol{\alpha}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\boldsymbol{\alpha}}^k \tag{3}$$

where $\hat{\boldsymbol{\alpha}}^k$ is the identity estimated from frames indexed in $\mathcal{I}_k$.

We can also consider the case where the different groups of frames, $\mathcal{I}_i$ and $\mathcal{I}_j$ with $i \neq j$, have some overlap. In other words, we can take random combinations of M frames from the entire set of T frames. In this case, we have a particularly large pool of subsets to select from, because there are $\binom{T}{M}$ combinations. Multi-frame estimation on partially overlapping groups of frames can be useful when there are not many frames at our disposal, as it allows us to produce many subsets of frames and hence many estimations to average over. It is reasonable to expect the accuracy of the average in Eq. (3) to increase as it is computed from more estimations, and our experimental results are consistent with this expectation.

4. Experiments

4.1. Dataset

We conducted our experiments on a sample of 47 infants collected to study facial behavior in the autism spectrum. 23 of the infants were female and 19 had an older sibling with autism. Infants were seen between 11 and 15 months of age (M = 12.60, SD = 0.72) and underwent concurrent developmental assessment. Overall, standard scores of infants' development ranged from extremely low to very high among cognitive (M = 102.93, SD = 8.14), language (M = 94.76, SD = 8.45), and motor (M = 104.15, SD = 9.86) domains. The videos were collected at The Children's Hospital of Philadelphia (CHOP) as part of a study approved by the IRB at CHOP, using a 2D camera with 1080p 60FPS resolution.

4.2. Evaluation metric

For quantitative analysis, we use the WBES metric, which allows us to infer the degree to which the reconstructed 3D face shapes are consistent and person-specific. As described below, this metric does not require a 3D ground truth.

Suppose that our dataset is comprised of videos of S subjects, and that we perform three reconstructions per subject, using three sets of (non-overlapping) frames per person. Ideally, all three reconstructions of the same person must be highly similar. Moreover, the reconstructions of faces of different subjects should be dissimilar. WBES measures to which degree a method succeeds in achieving these two criteria—the degree to which reconstructions are consistent (for the same-subject) and person-specific. This is achieved by constructing the *within-subject* distribution, which is comprised of all the pairwise (Euclidean) distances of the identity reconstructions of the same subjects; and the *between-subject* distribution, which is comprised of the pairwise distances of all reconstructions of different subjects. In accordance with the criteria described above, these two distributions should be ideally disjoint, and the within-subject distribution should be as narrow as possible. WBES measures these aspects via Cohen's effect size, which is computed as

$$\frac{\mu_B - \mu_W}{\sqrt{((N_W-1)\sigma_W + (N_B-1)\sigma_B)/(N_W+N_B-2)}}, \tag{4}$$

where μ_W and μ_B are respectively the means of the within- and between-subject distributions; σ_W and σ_B are the standard deviations of the two distributions; and N_W and N_B are the sizes of the distributions. Since WBES measures how disjoint the within- and between-subjects distributions are, the higher the WBES the better.

While measuring the WBES, one must compute the (neutral) face reconstructions of the same subject from frames with divergent poses. Otherwise, reconstructions from very similar poses are likely to be consistent even though they do not successfully capture the 3D shape, leading to misleadingly high WBES. For this purpose, we preprocessed all the videos as follows and ensured that the set of frames we used from each infant contains frames that are distributed approximately uniformly in terms of head pose. First, we detected the facial landmarks in each frame of a video using the 2D-FAN method [BT17], and used them to estimate the head pose per frame. Then, we split the frames into bins based on the angle of the head. Next, we discarded frames from the bins that were too populated, until all bins contained a similar number of frames.

4.3. Compared methods and parameters

We compare five 3DMM fitting methods: 3DDFA (v2) [GZY*20], 3DI [SZST20b], Deep3DFace [DYX*19], INORig [BCLT21] and RingNet [SBFB19]. These methods were selected based on the criteria of having an implementation that runs with reasonable speed (*e.g.,* not more than a few seconds per frame) and user effort on a CPU or an NVIDIA GPU with 8GB RAM; and having an expression shape model alongside identity model.

3DI and INORig are natively multi-frame. For these methods, the number of frames fit jointly, M (Section 3.2.2 and Section 3.2.3) was fixed to $M = 9$. The remaining methods were extended to multi-frame through naive averaging (Section 3.2.1). We investigated how performance improves with T, the total number of frames used per subject, by testing for $T = 9, 18, \ldots, 45$. Unless specified otherwise, we use multi-frame methods as described in Section 3.2.2 with non-overlapping groups of frames. We used 2D-FAN [BT17] to estimate 2D landmarks, which are required by some methods. All methods are based on the Basel Face Model'09 [PKA*09]. For 3DI, which can incorporate the camera model, we use the (perspective) camera transformation to ensure that expressions are encoded with maximal accuracy [SZST20a].

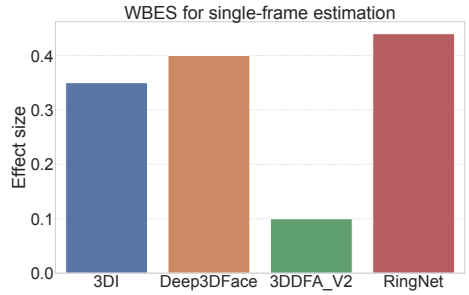

Figure 2: *Face reconstruction performance of four methods in terms of within- and between-subject effect size (WBES) from a single frame. Higher effect sizes indicate better performance.*

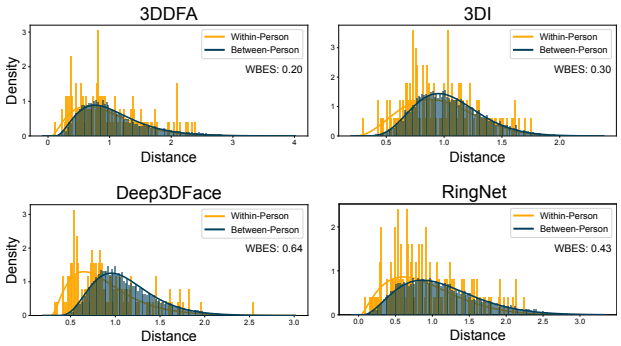

Figure 3: *Distribution of within- and between-subject face reconstruction distances (*i.e., *histograms and fit continuous densities), shown for the four single-frame methods, namely 3DDFA, 3DI, Deep3DFace and RingNet. Within- and between-subject distances are largely overlapping for all methods, highlighting the limitations of 3D (neutral) face reconstruction from a single frame.*

4.4. Quantitative results on identity estimation

Fig. 2 compares four single-frame 3DMM fitting methods in terms of WBES and shows that RingNet stands out as the best method in the single-frame scenario. However, the within- and between-subject distributions, which are computed while measuring WBES (Section 4.2), are highly overlapping for all methods (Fig. 3). In other words, reconstructions of the same infants across different frames are inconsistent, and no method is capable of reliably capturing person-specific shape characteristics from a single frame. This is not surprising, because 3D shape estimation from a single image is a highly ill-posed problem, and infants often show large expression and pose variations throughout the interaction (Section 3), rendering the problem even more difficult.

We next investigate whether performance improves when facial shape is estimated from multiple frames. Fig. 4 shows WBES against the number of frames used per reconstruction. Results show that the performance of all methods improves as they use more frames per infant while reconstructing the face; however, RingNet and 3DDFA visibly underperform compared to other methods. We observed that that the output of these methods is very conservative –it changes little for different subjects– which can explain why

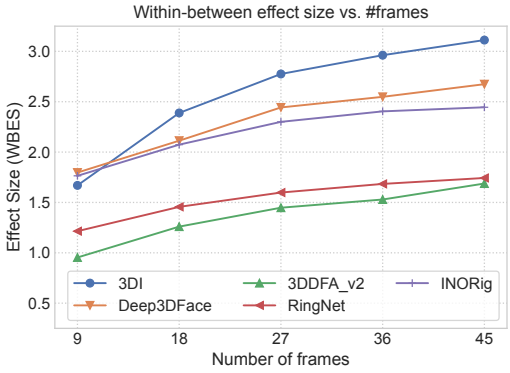

Figure 4: *Face reconstruction performance of five methods against the number frames used per reconstruction, measured in terms of within- and between-subject effect size (WBES). The performance of all methods improves as they use more frames per reconstruction.*

WBES is not increasing further with more frames per reconstruction. 3DI, Deep3DFace and INORig stand out as the best methods, and we observed that these methods indeed produce more diverse facial shapes from different subjects, which indicates higher capacity to capture person-specific shape cues. Fig. 5 shows the within- and between-subject distributions for all five methods for reconstruction from $T = 45$ frames, confirming the significant improvement achieved by using multiple frames as opposed to a single frame. One may think that 45 frames are too many, but our data suggests that a few minutes of infant social interaction is more than enough to produce 45 frames with large head pose diversity.

Our multi-frame analysis so far relied on naive-averaging (for single-frame methods) or averaging over multi-frame estimations from non-overlapping groups of frames (for multi-frame methods). As highlighted in Section 3.2.3, multi-frame methods have one more option to estimate neutral face shape: averaging over estimations from partially overlapping combinations of frames.

Fig. 6 shows WBES *w.r.t.* number of combinations for the two inherently multi-frame methods, namely, 3DI and INORig. Performance improves for both methods, but more so for 3DI. The latter reaches a WBES of 3.71 when it uses 30 combinations from the $T = 45$ frames, which is a nearly 20% improvement compared to the WBES of 3.11, which is achieved by 3DI when it uses 5 (non-overlapping) combinations of the same 45 frames. In sum, multi-frame reconstruction performance can be improved significantly by using the same frames but producing more estimations over random combinations of frames.

4.5. Qualitative results

Fig. 7 and Fig. 8 show qualitative face reconstruction results. Results in this section are obtained with the method that emerged as the most successful in quantitative analysis, namely 3DI.

Fig. 7 depicts how 3DI captures subject-specific shape cues. The face shown at top row of Fig. 7 has asymmetric eyebrows, which

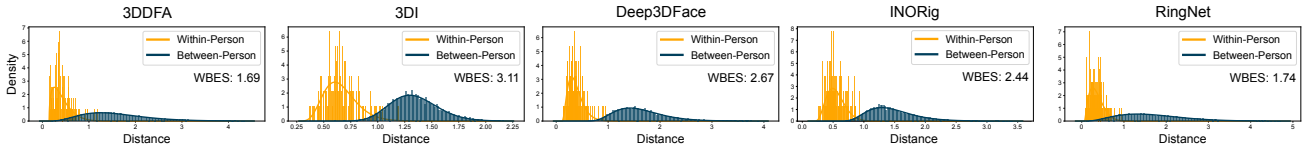

Figure 5: *Distribution of within- and between-subject face reconstruction distances from multiple frames ($T = 45$) for four methods, namely 3DDFA, 3DI, Deep3DFace, INORig and RingNet. Unlike the single-frame counterparts of these distributions (Fig. 3), distributions of within- and between-subject distances are highly disjoint overlapping, indicating that all methods yield significantly more consistent and person-specific results when they reconstruct (neutral) face shape from 45 frames.*

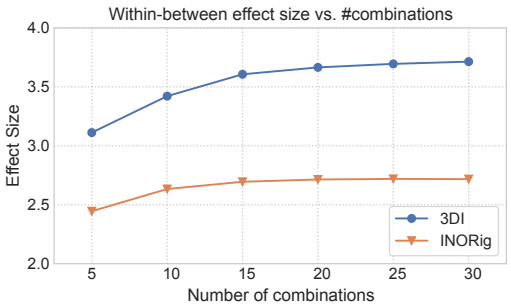

Figure 6: *Comparison of the two multi-frame methods, 3DI and INORig, against the number of combinations, K. All effect sizes were computed by performing 3D reconstruction using the same $T = 45$ frames, but averaging over estimations from an increasing number of (9-frame) combinations (Section 3.2.3).*

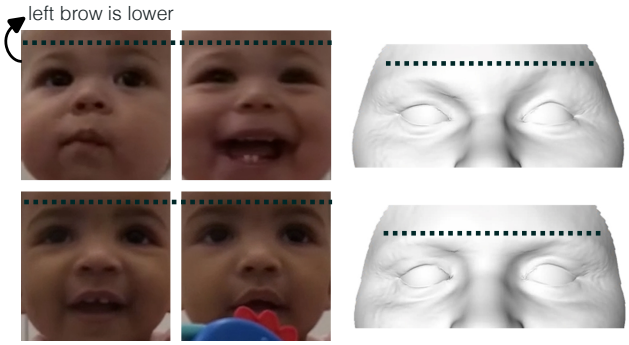

Figure 7: *Illustration of how the 3D reconstructions produced by the 3DI method capture subject-specific facial morphology. Top row: images of an infant along with the reconstruction of the eye region. The images show that one of the eyebrows is lower than the other, and the reconstruction seems to successfully capture this facial shape detail. Bottom row: Images and 3D reconstruction of another baby, both showing more symmetric eyebrows.*

is likely due to the infant's unique facial morphology. The 3D reconstruction for this infant, shown right next to the frames, also captures this asymmetric relationship between the eyebrows. The face on the second row, on the other hand, has more symmetric eyebrows, as the reconstruction also captures. The ability to capture such subject-specific characteristics is the likely explanation for high WBES. However, it must be highlighted that not all facial characteristics of infants are captured adequately. Results in Fig. 8 suggest that the reconstructed cheek regions are, unlike cheeks of infants, too rugged. Moreover, the nose and chin can be too pointy (see bottom left and bottom right reconstructions) for an infant. The lower success in these regions compared to eye regions is likely due to two reasons. First, the cheeks, nose and jaw of infants are in general very different from those of adults. Second, the brows, eyes and mouth are the regions with the highest texture variation and 3DI works more successfully on these regions, since it fits a 3DMM by maximizing gradient correlation [TZP11]. These results are consistent with the study of Morales *et al.* [MPT*20], which suggests that applying the Basel Face Model to data of infants is most problematic for regions such as chin, cheeks and forehead.

The reconstructed facial expressions of all three infants shown in Fig. 8 seem to be accurately capturing the true expression of the infants, which also supports the statement that brows, eyes and mouth are the most accurately captured regions of the face, as these regions are where the facial expressions are most clearly observed. Careful inspection of Fig. 8 suggests that using 3DMM fitting is a highly promising approach for facial expression quantification, as the diverse set of expressions displayed by the infants are generally well-captured in the 3D reconstruction, often despite pose variations, which is a significant strength of 3DMM fitting.

5. Discussion

Our experimental results and analysis suggest that fitting a 3DMM of adults to images of infants is not appropriate for applications that require the entire head or face to be accurately reconstructed, such as identifying craniofacial dysmorphology patterns [MPL*22]. However, regions of high texture variation (eyes, brows and mouth) are reconstructed with visibly higher accuracy than the remaining parts of the face, and, as a result, facial expressions and pose appear to be captured successfully. This shows that 3DMM fitting is a promising tool for measuring facial behavior of infants, even if the 3DMM that is used is constructed from data of adults. Clearly, one should use a 3DMM constructed from data of infants if that is possible, but to our knowledge, there is no publicly available 3DMM of infants, and most researchers are currently confined to using 3DMMs of adults. Given that 3D scans of infants are sensitive data, a public 3DMM of infants for research and commercial applications may not be available in the foreseeable future.

The promise of using 3DMMs to quantify the facial expressions

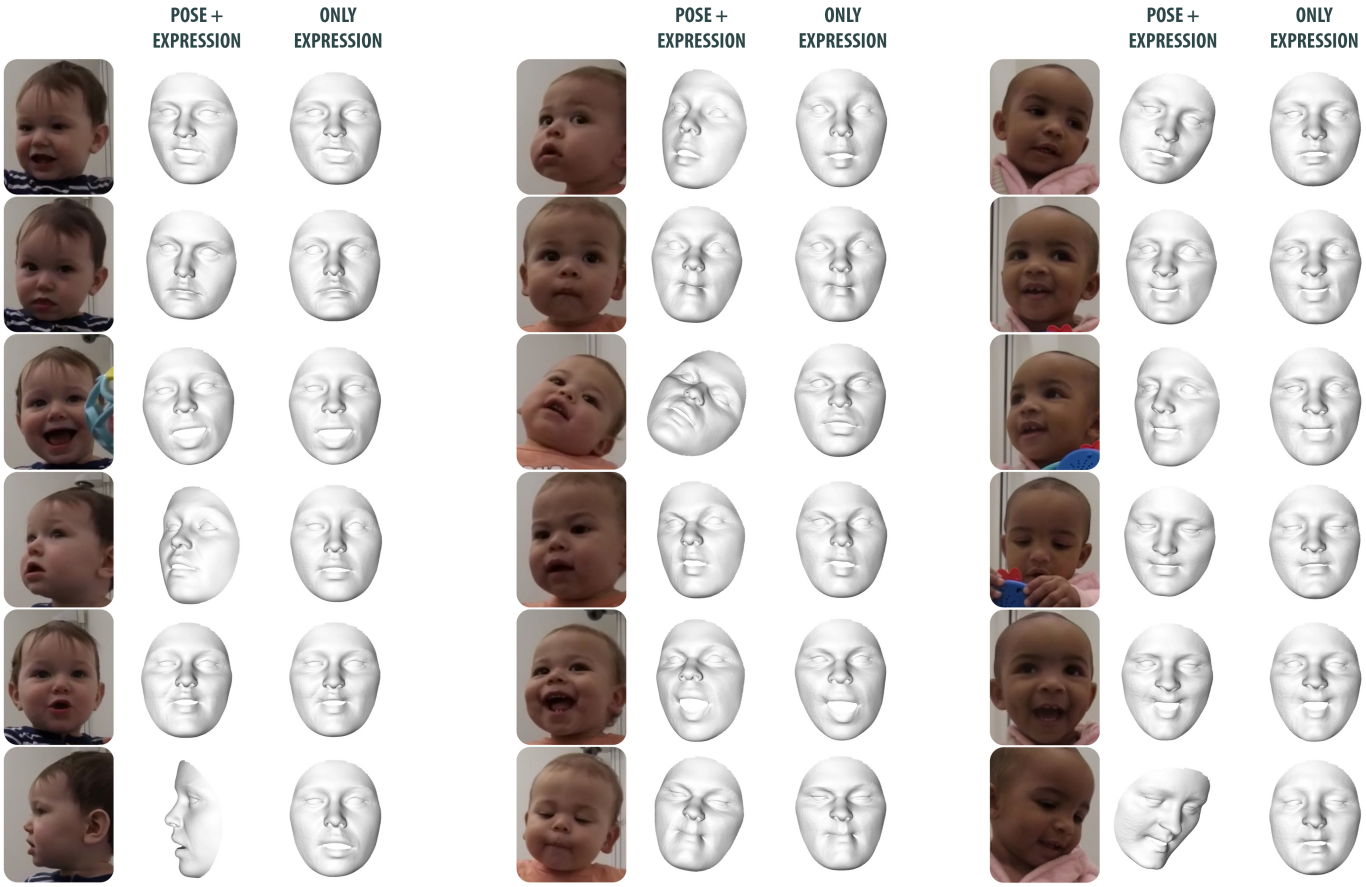

Figure 8: *Reconstruction of facial pose plus expression, as well as frontal expression, from frames of 3 babies, obtained using the 3DI method with the Basel face model. Results indicate that facial expression and pose are generally captured successfully, suggesting that 3DMM fitting is a promising tool for measuring infant facial behavior. However, the facial features that typically differ significantly between adults and infants* (e.g., *cheeks, nose) are generally not captured accurately (*e.g., *see nose in lower left or lower right reconstructions, or cheeks in the reconstructions in the second column), as the 3DMM that is used is constructed from face scans from adults.*

of infants is a positive outcome of this study, as there are very few methods for quantifying infant facial expressions [HCC*17, OEAB*22], and, to our knowledge, no method can decouple identity from facial expressions. On the contrary, our quantitative and qualitative results showed that 3DMMs *are* capable of identifying subject-specific facial morphology in infants, and thus correct for it while measuring facial expressions. For example, a 3DMM is capable of identifying that the eyebrow that appears raised in Fig. 7 (top) is due to facial morphology and not expression. In sum, our study suggests that 3DMMs merit further research due to their promise to accurately measure not only facial expression but also pose, which is of importance on its own [HCM*13, MHR*18, HC14, KAH*20].

Importantly, our study shows that single-frame (neutral) 3D identity reconstruction is likely not a feasible problem to solve. In other words, to fully benefit from the ability of 3DMMs to tease apart facial identity and expression cues, it is likely necessary to use multiple frames. Our results show not only that reconstructing 3D shape from multiple frames improves results by a large margin, but also that methods that can *jointly* fit a 3DMM to multiple frames have a significant advantage, by virtue of producing and averaging over a large number of combinations of frames.

6. Conclusions and Future Work

In this study, we compared five state-of-the-art 3D morphable model (3DMM) fitting methods in terms of their ability to reconstruct faces of infants from videos recorded in dyadic infant-caregiver interactions. Given the lack of publicly available 3DMMs of infants, we investigated the degree to which a 3DMM constructed from data of adults can be used in this context. Results suggest that it is possible to extract 3D representations that are consistent and subject-specific by producing reconstructions from multiple frames but not a single frame. Consistently with prior work [MPT*20], we observed that using infant faces reconstructed through a 3DMM for adults is likely inadequate for applications or studies that require an accurate reconstruction of the entire head or face, such as identifying craniofacial dysmorphology patterns [MPL*22]. However, results suggest that regions with high

texture variation, such as the eyes, brows and mouth, are identified with higher accuracy, making applications that involve studying facial expression or pose variation within reach. These promising results motivate immediate future work, namely, quantitatively evaluating the ability of publicly available 3DMMs and fitting methods in terms of their ability to quantify expression. Given the unique strength of 3DMMs to decouple expressions from two major sources of nuisance, *i.e.,* identity and pose, the use of 3DMMs could have significant implications for developmental and clinical research in infancy.

Acknowledgments

This work is partially funded by the National Institutes of Health (NIH), Office of the Director (OD), National Institute of Child Health and Human Development (NICHD), and National Institute of Mental Health (NIMH) of US, under grants R01MH118327, R01MH122599, 5P50HD105354-02 and R21HD102078.

References

[BCLT21] Bai Z., Cui Z., Liu X., Tan P.: Riggable 3d face reconstruction via in-network optimization. In *CVPR* (2021), pp. 6216–6225. 2, 4

[BRP*18] Booth J., Roussos A., Ponniah A., Dunaway D., Zafeiriou S.: Large scale 3d morphable models. *International Journal of Computer Vision 126*, 2 (2018), 233–254. 3

[BRV*18] Booth J., Roussos A., Ververas E., Antonakos E., Ploumpis S., Panagakis Y., Zafeiriou S.: 3D reconstruction of "in-the-wild" faces in images and videos. *IEEE TPAMI 40*, 11 (2018), 2638–2652. 2

[BT17] Bulat A., Tzimiropoulos G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV* (2017). 4

[DYX*19] Deng Y., Yang J., Xu S., Chen D., Jia Y., Tong X.: Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW* (2019). 2, 3, 4

[EST*20] Egger B., Smith W. A., Tewari A., Wuhrer S., Zollhoefer M., Beeler T., Bernard F., Bolkart T., Kortylewski A., Romdhani S., et al.: 3d morphable face models—past, present, and future. *ACM TOG 39*, 5 (2020), 1–38. 2

[FBPDB21] Ferrari C., Berretti S., Pala P., Del Bimbo A.: A sparse and locally coherent morphable face model for dense semantic correspondence across heterogeneous 3d faces. *IEEE TPAMI* (2021). 2

[GCM*18] Genova K., Cole F., Maschinot A., Sarna A., Vlasic D., Freeman W. T.: Unsupervised training for 3d morphable model regression. In *CVPR* (2018), pp. 8377–8386. 3

[GMB*18] Gerig T., Morel-Forster A., Blumer C., Egger B., Luthi M., Schoenborn S., Vetter T.: Morphable face models - an open framework. In *IEEE Int. Conf. Automatic Face and Gesture Recog.* (2018), pp. 75–82. 3

[GZY*20] Guo J., Zhu X., Yang Y., Yang F., Lei Z., Li S. Z.: Towards fast, accurate and stable 3d dense face alignment. In *ECCV* (2020), pp. 152–168. 2, 4

[HC14] Hammal Z., Cohn J. F.: Intra-and interpersonal functions of head motion in emotion communication. In *Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges* (2014), pp. 19–22. 7

[HCC*17] Hammal Z., Chu W.-S., Cohn J. F., Heike C., Speltz M. L.: Automatic action unit detection in infants using convolutional neural network. In *Affective Computing and Intelligent Interaction* (2017), IEEE, pp. 216–221. 7

[HCM*13] Hammal Z., Cohn J. F., Messinger D. S., Mattson W. I., Mahoor M. H.: Head movement dynamics during normal and perturbed parent-infant interaction. In *Affective Computing and Intelligent Interaction* (2013), IEEE, pp. 276–282. 7

[JBAT17] Jackson A. S., Bulat A., Argyriou V., Tzimiropoulos G.: Large pose 3D face reconstruction from a single image via direct volumetric cnn regression. *ICCV* (2017). 3

[KAH*20] Klein L., Ardulov V., Hu Y., Soleymani M., Gharib A., Thompson B., Levitt P., Matarić M. J.: Incorporating measures of intermodal coordination in automated analysis of infant-mother interaction. In *ACM Int. Conf. Multimodal Interaction* (2020), pp. 287–295. 7

[LAVS*16] Leclere C., Avril M., Viaux-Savelon S., Bodeau N., Achard C., Missonnier S., Keren M., Feldman R., Chetouani M., Cohen D.: Interaction and behaviour imaging: a novel method to measure mother–infant interaction using video 3d reconstruction. *Translational Psychiatry 6*, 5 (2016), e816–e816. 1

[LBB*17] Li T., Bolkart T., Black M. J., Li H., Romero J.: Learning a model of facial shape and expression from 4d scans. *ACM TOG 36*, 6 (2017), 194–1. 3

[MHR*18] Martin K. B., Hammal Z., Ren G., Cohn J. F., Cassell J., Ogihara M., Britton J. C., Gutierrez A., Messinger D. S.: Objective measurement of head movement differences in children with and without autism spectrum disorder. *Molecular autism 9*, 1 (2018), 1–10. 7

[MPL*22] Morales A., Porras A. R., Linguraru M. G., Piella G., Sukno F. M.: Babynet: Reconstructing 3d faces of babies from uncalibrated photographs. *arXiv preprint arXiv:2203.05908* (2022). 3, 6, 7

[MPT*20] Morales A., Porras A. R., Tu L., Linguraru M. G., Piella G., Sukno F. M.: Spectral correspondence framework for building a 3d baby face model. In *IEEE Int. Conf. Automatic Face and Gesture Recog.* (2020), IEEE, pp. 708–715. 3, 6, 7

[OEAB*22] Onal Ertugrul I., Ahn Y. A., Bilalpur M., Messinger D. S., Speltz M. L., Cohn J. F.: Infant afar: Automated facial action recognition in infants. *Behavior Research Methods* (2022), 1–12. 7

[PKA*09] Paysan P., Knothe R., Amberg B., Romdhani S., Vetter T.: A 3D face model for pose and illumination invariant face recognition. In *IEEE Conf. on Advanced Video and Signal based Surveillance* (2009), pp. 296–301. 3, 4

[RV05] Romdhani S., Vetter T.: Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR* (2005), vol. 2, pp. 986–993. 2

[SBFB19] Sanyal S., Bolkart T., Feng H., Black M. J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR* (2019), pp. 7763–7772. 2, 4

[SWTC14] Shi F., Wu H.-T., Tong X., Chai J.: Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM TOG 33*, 6 (2014), 1–13. 2

[SZST20a] Sariyanidi E., Zampella C. J., Schultz R. T., Tunc B.: Can facial pose and expression be separated with weak perspective camera? In *CVPR* (2020), pp. 7173–7182. 4

[SZST20b] Sariyanidi E., Zampella C. J., Schultz R. T., Tunc B.: Inequality-constrained and robust 3d face model fitting. In *ECCV* (2020), pp. 433–449. 2, 4

[SZST22] Sariyanidi E., Zampella C., Schultz R. T., Tunc B.: Inequality-constrained 3d morphable face model fitting. URL: https://figshare.com/s/7f39f034d5489b3d8476, doi:10.36227/techrxiv.20233860. 3

[TZP11] Tzimiropoulos G., Zafeiriou S., Pantic M.: Robust and efficient parametric face alignment. In *ICCV* (2011), pp. 1847–1854. 6