Robust Registration of Dynamic Facial Sequences

Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro

***Abstract*—Accurate face registration is a key step for several image analysis applications. However, existing registration methods are prone to temporal drift errors or jitter among consecutive frames. In this paper we propose an iterative rigid registration framework that estimates the misalignment with trained regressors. The input of the regressors is a robust motion representation that encodes the motion between a misaligned frame and the reference frame(s), and enables reliable performance under non-uniform illumination variations. Drift errors are reduced when the motion representation is computed from multiple reference frames. Furthermore, we use the L_2 norm of the representation as a cue for performing coarse-to-fine registration efficiently. Importantly, the framework can identify registration failures and correct them. Experiments show that the proposed approach achieves significantly higher registration accuracy than state-of-the-art techniques in challenging sequences.**

I. Introduction

Face registration is the process of compensating for rigid transformations caused by head, body or camera movements in an image sequence. This is a fundamental pre-processing step for applications that interpret the non-rigid motions of facial features, such as facial action recognition [1], visual speech recognition [2], emotion recognition [3] and micro-expression recognition [4]. Rigid registration for facial analysis needs to address multiple challenges, namely *non-uniform illumination variations*, *occlusions* and *facial activity* itself, which generates non-rigid motions that become outliers for rigid registration. Moreover, significant *drift errors* may accumulate over time with online registration, even when individual registration errors remain under a tolerance threshold, thus leading to registration failures. Undetected *registration failures* then become false references for subsequent frames, thus generating additional registration errors.

Facial registration can be conducted considering the *whole* face or its *parts* [5]. Part-based registration refers to registering selected facial regions independently from one another (*e.g.,* each eye and the mouth are registered as three separate cropped sequences). Though part-based registration is useful to reduce the effect of out-of-plane head rotations [5], it is a challenging task as a large proportion of pixels undergo non-rigid motions.

Registration is often approached as an optimisation problem and solved with a gradient-descent method [6], [7], [8], [9], [10]. However, gradient descent may underperform with untextured regions, particularly when high-gradient regions are associated with outlier motions. An emerging approach to optimisation in computer vision is using statistical learning [11], [12], [13], [14]. The original idea of Cootes *et al.* [15] was to construct an algorithm by learning the relationship between the parameters to be optimised and the residual error caused by non-optimal parameters. We argue that optimisation based on learning is also promising for rigid facial registration, as invariance to non-rigid motions can be improved by training with sequences that contain facial activity. Moreover, robustness to non-uniform illumination variations can be improved with a robust feature extraction scheme without analytically modelling the relationship between features and misalignment parameters.

In this paper, we propose to use optimisation via statistical learning for rigid facial registration. The proposed iterative framework (Fig. 1) reduces drift errors by computing Gabor motion energy with respect to multiple reference frames, and can identify and correct registration failures via probabilistic learning. We show that, in iterative registration, misalignment can be estimated effectively with a pre-trained regressor of Gabor motion energy and that this regressor can generalise and perform accurately on data with illumination variations even when trained using controlled data. Moreover, we show that the L_2 norm of Gabor motion energy can be used to train multiple regressors with different granularities and also to efficiently perform coarse-to-fine registration with these regressors. We refer to the proposed framework as MUMIE (Multiple regressors for Misalignment Estimation), and evaluate it both for whole-face and part-based registration and obtain significantly higher accuracy than classical registration frameworks. Particularly notable is the part-based registration performance in the presence of large facial activity due to facial expressions, and its robustness to non-uniform illumination variations. The code of the method is made available for research purposes.

The paper is organised as follows. Section II reviews existing registration approaches. Section III presents the problem formulation. Section IV describes the registration process. Section V explains how registration failures are identified and corrected. Experimental results are discussed in Section VI. Section VII concludes the paper.

II. Related Work

We discuss existing registration approaches and focus on their ability to deal with illumination variations, outlier motions and drift errors. We first cover a method specific to faces, namely registration by localising fiducial points [5]. Then we cover generic registration techniques, which can be grouped in three main classes, namely keypoint, transformation-based and direct methods.

A popular approach to rigid facial registration is to localise and align fiducial landmark points (*e.g.,* eyes) [5]. There exist landmark localisers that are robust to illumination variations

E. Sariyanidi and A. Cavallaro are with the Centre for Intelligent Sensing in Queen Mary University of London, UK. H. Gunes is with the Computer Laboratory, University of Cambridge, UK.
E-mail: {e.sariyanidi, a.cavallaro}@qmul.ac.uk, hatice.gunes@cl.cam.ac.uk.

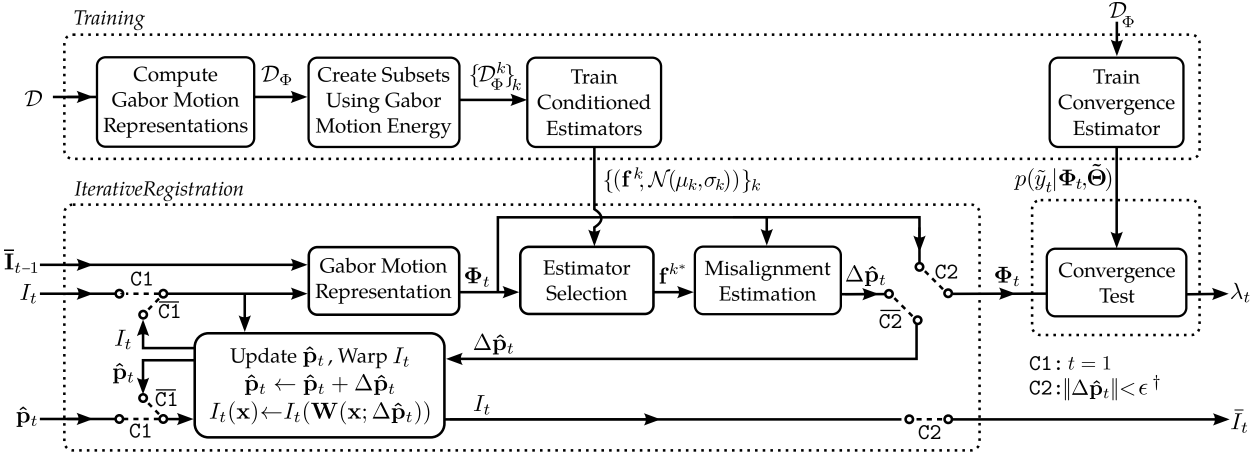
Training
$\mathcal{D}$
Compute Gabor Motion Representations
$\mathcal{D}_\Phi$
Create Subsets Using Gabor Motion Energy
$\{\mathcal{D}_\Phi^k\}_k$
Train Conditioned Estimators
$\tilde{\mathcal{D}}_\Phi$
Train Convergence Estimator
$\{(\mathbf{f}^k, \mathcal{N}(\mu_k, \sigma_k))\}_k$
$p(\tilde{y}_t|\mathbf{\Phi}_t, \tilde{\mathbf{\Theta}})$
IterativeRegistration
$\bar{\mathbf{I}}_{t-1}$
I_t
C1
$\overline{\texttt{C1}}$
Gabor Motion Representation
$\mathbf{\Phi}_t$
Estimator Selection
$\mathbf{f}^{k^*}$
Misalignment Estimation
$\Delta\hat{\mathbf{p}}_t$
C2
$\overline{\texttt{C2}}$
$\mathbf{\Phi}_t$
Convergence Test
λ_t
I_t
$\hat{\mathbf{p}}_t$
Update $\hat{\mathbf{p}}_t$, Warp I_t
$\hat{\mathbf{p}}_t \leftarrow \hat{\mathbf{p}}_t + \Delta\hat{\mathbf{p}}_t$
$I_t(\mathbf{x}) \leftarrow I_t(\mathbf{W}(\mathbf{x}; \Delta\hat{\mathbf{p}}_t))$
$\Delta\hat{\mathbf{p}}_t$
$\overline{\texttt{C1}}$
C1
$\hat{\mathbf{p}}_t$
I_t
C2
$\bar{I}_t$
$\texttt{C1}: t = 1$
$\texttt{C2}: \|\Delta\hat{\mathbf{p}}_t\| < \epsilon^\dagger$

Fig. 1: Overview of the proposed MUMIE framework. The top part represents the training of the misalignment estimators. The bottom part represents the iterative registration scheme, followed by a convergence test. The input to registration is an ordered set of reference frames $\bar{\mathbf{I}}_{t-1}$, the misaligned frame I_t and the initial misalignment estimation $\hat{\mathbf{p}}_t$. The dashed lines represent the conditional paths that are followed when the labelled conditions hold (C1/C2) or do not hold ($\overline{\texttt{C1}}/\overline{\texttt{C2}}$), and $||\cdot||$ is the L_2 norm. †The condition C2 is satisfied also if a maximal number of iterations, $K_{\max}$, is reached.

[16] (*e.g.,* [13], [11], [17]). Moreover, this approach can be made robust to outlier non-rigid motions by selecting landmarks that are less affected by facial activity (*e.g.,* eye corners). However, registration based on landmark localisation is prone to jitter among consecutive frames due to localisation errors [18] that occur even in relatively controlled conditions [2]. The detrimental effect of jittering caused by landmark localisation errors has already been observed in visual speech recognition [2]. Similar jittering errors are detrimental also to facial action analysis [1].

Keypoint methods perform registration using sparsely located image points that are centred on visually salient regions with rich texture [23]. While these methods are tolerant to large outlier motions thanks to the use of robust estimators such as RANSAC [27], keypoint methods may not perform reliably when outlier motions occur around visually salient regions (*i.e.,* regions with texture variations). This occurs with part-based registration or when illumination variations severely reduce the number of matched features [28].

Global transformation-based methods use the invariance properties of the Fourier transform [7], [22], Fourier-Mellin transform [29] or Radon transform [30], [23]. These methods are generally unsuitable for challenging real-life problems as they are sensitive to outlier motions and illumination variations [24]. Although a robust version of the fast Fourier transform (FFT) [24] is successful against these challenges, its accuracy in simpler conditions without illumination variations can be lower than that of keypoint-based methods [28].

Direct methods minimise an error function of a pair of misaligned frames. The Lucas-Kanade (LK) method minimises the sum of squared difference between two frames and can be rendered partially robust to outliers by dividing frames into blocks [6] or by employing robust estimators [31]. LK methods perform minimisation via gradient descent and may therefore not perform reliably if regions of outlier motions yield high gradient while the remaining regions are relatively flat, which is likely to happen in part-based registration. Extensions of LK differ in the error function that is optimised, the optimisation algorithm or the domain where the optimisation is performed [6], [7], [9], [25], [32], [33]. Methods that operate on the pixel domain are particularly sensitive to illumination variations [7]. Pre-processing with Gabor filters [25] helps improve robustness of LK methods against illumination variations [7]. One of the most robust methods against non-uniform illumination variations is based on the direct maximisation of the gradient correlation coefficient (GradCorr) [7]. GradCorr employs a cosine kernel, which improves robustness against outliers and illumination variations by eliminating local mismatches [7].

Keypoint, transformation-based and direct methods are prone to drift errors in long sequences as they register each frame with respect to a reference frame. This problem was highlighted for the LK framework [34] and addressed by a number of methods [10], [34], [26], [35], which were validated on data with limited illumination variations only.

Table I summarises the methods discussed in this section. Note that while methods exist that independently tackle drift errors, outlier motions or challenging non-uniform illumination variations, to the best of our knowledge no method addresses *all* these challenges within the same framework. We initially investigated the benefits of learning-based registration in our preliminary work [28]. With respect to our preliminary work the main novelties in this paper are (i) the misalignment estimation through a continuous model that is simpler to train than the discrete model; (ii) computing the motion representation using multiple reference frames to reduce drift errors; (iii) the strategy for correcting registration failures; (iv) using the magnitude of motion representation as a prior cue about the amount of misalignment, thereby improving computational efficiency.

TABLE I: Representative methods from various registration categories and how they address illumination variations, drift errors and outlier motions. Key: (K)eypoint, (T)ransformation-based, (D)irect, (S)tatistical Learning.

	Ref.	Approach	Illumination Variations	Drift Errors	Outlier Motions	Failure Identif.	Failure Correct.	Tested for part-based registration?
(K)	[19]	SURF feature matching	Robust features	—	RANSAC	—	—	—
	[20]	MSER feature matching	Robust features	—	RANSAC	—	—	—
	[21]	SIFT feature matching	Robust features	Drift correction	RANSAC	—	—	—
(T)	[22]	Multi-layer Fourier transf.	—	—	—	—	—	—
	[23]	Radon transf.	—	—	—	—	—	—
	[24]	Robust Fourier transf.	Gradient correlation	—	Cosine kernel	—	—	—
(D)	[6]	Lucas-Kanade (LK) matching	—	—	Robust estimator	—	—	—
	[25]	LK matching	Gabor Filtering	—	Robust estimator	—	—	—
	[26]	Robust LK matching	—	Drift correction	Robust estimator	—	—	—
	[10]	Extended LK matching	—	Backgr. modelling	Robust estimator	—	—	—
	[7]	Gradient correlation max.	Gradient correlation	—	Cosine kernel	—	—	—
(S)	This Work	Optimisastion via learning	3D Gabor representation	Multi-frame motion encoding	Pooling, training with noisy data	✓	✓	14 Sequences (294 Images)

III. Problem Formulation

Let $\mathbf{S} = (I_1, I_2, \ldots, I_t, \ldots, I_T)$ be a sequence of arbitrary length T with unregistered frames I_t. The goal is to generate a registered sequence $\bar{\mathbf{S}} = (\bar{I}_1, \bar{I}_2, \ldots, \bar{I}_T)$ with no rigid misalignment between any two frames $\bar{I}_j, \bar{I}_k$. When $\mathbf{S}$ is acquired via streaming, a frame I_t must be registered as soon as it is obtained (*online registration*). Let I_1 be the reference frame that subsequent frames will be registered to (*i.e.,* $\bar{I}_1 = I_1$).

Let $\mathbf{p}_t$ be the parameters of the rigid motion responsible for the misalignment in I_t. $\bar{I}_t$ can be obtained by transforming I_t with a warping operator $\mathbf{W}(\mathbf{x}; \mathbf{p}_t)$ that maps each pixel $\mathbf{x} = (x, y)^T$ based on $\mathbf{p}_t$ [6]:

$$\bar{I}_t(\mathbf{W}(\mathbf{x}; \mathbf{p}_t)) = I_t(\mathbf{x}). \tag{1}$$

The critical task is to obtain an accurate estimation of rigid motion, $\hat{\mathbf{p}}_t$. The rigid motion in I_t can be estimated with respect to a single frame (for example the most recently registered frame, $\bar{I}_{t-1}$); or by considering multiple past reference frames. For example, one can use an ordered set that contains the last T_R registered frames $\bar{\mathbf{I}}_{t-1} = (\bar{I}_\tau, \bar{I}_{\tau+1}, \ldots, \bar{I}_{t-1})$ where $\tau = \max\{1, t - T_R\}$. We refer to registration with $T_R = 1$ as *single-frame* registration and with $T_R > 1$ as *multi-frame* registration.

IV. Misalignment Estimation

Faces are non-planar objects and compensating for rigid motion with an affine transformation may distort facial geometry and undermine facial activity analysis. Therefore, we model rigid motion as a Euclidean transformation.

A. Optimisation via learning

Let $\mathbf{p}_t = (p_1, p_2, p_3, p_4)$ be a vector whose elements define the horizontal and vertical translation, scale and rotation, respectively. Registration via optimisation starts computing the rigid motion between two images $\bar{I}_{t-1}$ and I_t with an initial estimate $\hat{\mathbf{p}}_t$ that is then updated iteratively as:

$$\hat{\mathbf{p}}_t \leftarrow \hat{\mathbf{p}}_t + \Delta\hat{\mathbf{p}}_t, \tag{2}$$

until the norm of the increment, $||\Delta\hat{\mathbf{p}}_t||$, is smaller than a threshold ϵ. $\Delta\hat{\mathbf{p}}_t$ is generally computed with the LK algorithm [8], [10] that uses gradient descent for optimisation. Convergence is successful under constant illumination conditions and limited occlusions [6]. Extensions of LK can tackle illumination variations and occlusions using a robust estimator [31] or a cosine kernel that eliminates outliers caused by local texture mismatches [7]. However, as mentioned in Section I, algorithms based on gradient-descent may underperform when high-gradient image regions are related to outlier motions.

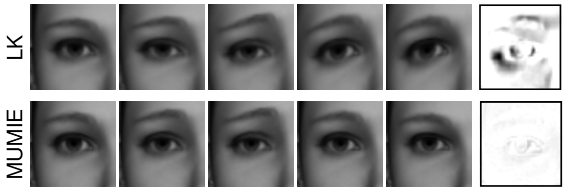

Fig. 2: Illustration of drift errors that can occur over time, through an exemplar sequence that starts and ends with the same eye expression. Registration output of a Lucas-Kanade (LK) method [7] (top) and MUMIE (bottom). LK is prone to drift errors, as seen by comparing the first and last frames of the registered sequences. Drift errors are highlighted in the last column where the difference between the first and last frames is depicted. (Dark values indicate registration errors.)

Registration for facial analysis needs to cope with the *non-rigid motions caused by facial activity*, which affect a large proportion of pixels and is problematic for part-based registration. Facial activity evolves slowly and may not be eliminated as a local mismatch, thus causing drift errors. Fig. 2 illustrates this problem: the first and last frame should be aligned as they depict the same eye expression at two different instants of a sequence; however, another expression appearing in the in-between frames causes drift errors for the LK-based algorithm [7].

An emerging approach to optimisation is to perform the updates with a pre-computed function [15], [13], [11]. The increment $\Delta\hat{\mathbf{p}}_t$ can be computed with a regressor that models the relationship between misalignment and the residuals

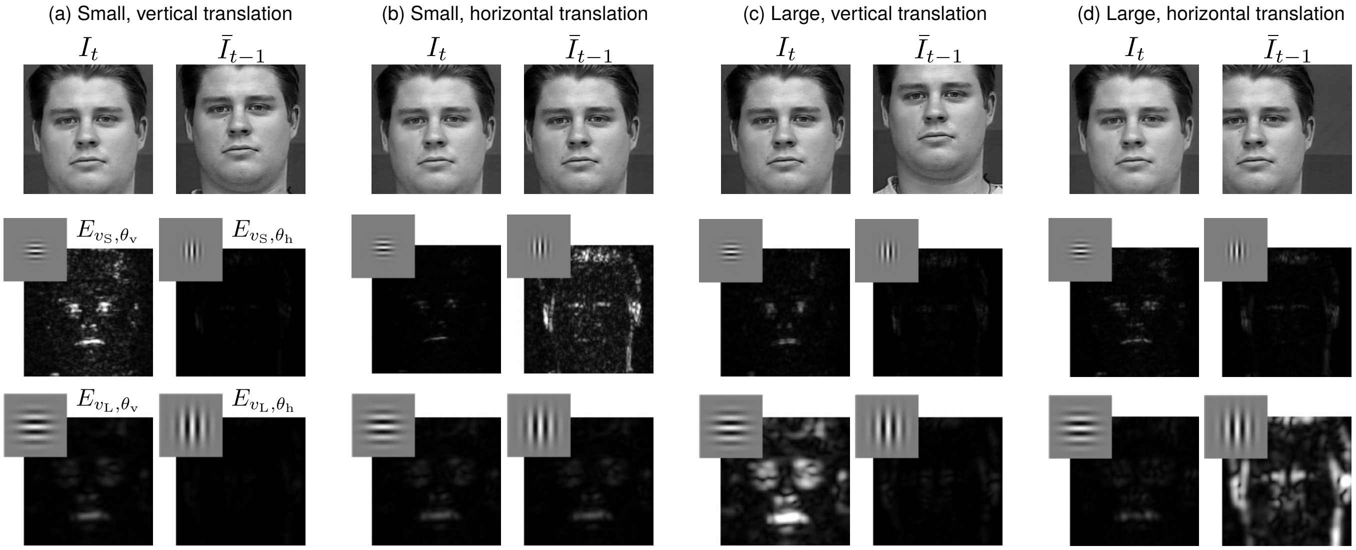
(a) Small, vertical translation
(b) Small, horizontal translation
(c) Large, vertical translation
(d) Large, horizontal translation
I_t
$\bar{I}_{t-1}$
$E_{v_\text{S},\theta_\text{v}}$
$E_{v_\text{S},\theta_\text{h}}$
$E_{v_\text{L},\theta_\text{v}}$
$E_{v_\text{L},\theta_\text{h}}$

Fig. 3: Illustration of the usefulness of the Gabor motion energy for registration via four example cases (a–d) that involve different types (horizontal/vertical translation) and amounts (small/large) of misalignment. For each case, the Gabor motion energy is computed with four different filter pairs tuned to a particular speed ($v_{\text{S(mall)}}$ or $v_{\text{L(arge)}}$) and orientation ($\theta_{\text{h(orizontal)}}$ or $\theta_{\text{v(ertical)}}$). The energy is maximal when the filters are in tune with the misalignment.

caused by misalignment. We use regression for rigid facial registration: at each iteration, we compute the rigid motion between $\bar{I}_{t-1}$ and I_t (or more generally, between $\bar{\mathbf{I}}_{t-1}$ and I_t) with a regressor $\mathbf{f}$ as:

$$\Delta\hat{\mathbf{p}}_t = \mathbf{f}(\boldsymbol{\Phi}(\bar{\mathbf{I}}_{t-1}, I_t); \boldsymbol{\Theta}), \tag{3}$$

where $\boldsymbol{\Theta}$ is the vector of input-independent regressor parameters, and $\boldsymbol{\Phi}(\cdot)$ is a feature extraction process that we discuss later in this section. $\boldsymbol{\Theta}$ is computed from a dataset $\mathcal{D} = \{(\bar{\mathbf{I}}^n, I^n, \mathbf{p}^n)\}_{n=1}^N$ that contains N misaligned samples and their misalignment labels. Invariance against an outlier can be encouraged by augmenting $\mathcal{D}$ with training samples that are affected by the outlier [36]. This strategy is particularly useful for dealing with outliers that are difficult to model analytically [36], such as the non-rigid motions caused by facial activity. Moreover, invariance against illumination variations can be encouraged with a robust feature extraction scheme. Unlike algorithms based on gradient descent (*e.g.,* LK), the computation in (3) does not require a differentiable expression to minimise. For this reason, we can employ feature extraction schemes that are difficult to differentiate or not differentiable. However, while optimisation with regression provides an efficient means for dealing with outliers, an important issue has to be addressed for a pre-computed regressor, namely the *generalisation to unseen faces and imaging conditions*.

Generalisation can be improved with a feature extraction scheme that is sensitive to rigid motion and insensitive to irrelevant factors, such as skin colour and illumination variations. To this end, we use a spatio-temporal Gabor representation, which encodes motion without computing motion vectors explicitly [37] and is robust against illumination variations [28]. The Gabor representation encodes the motion between two frames $\bar{I}_{t-1}$ and I_t by convolving this pair with speed- and orientation-selective Gabor filters that are defined as [38]

$$g^{\phi}_{v,\theta}(x,y,t') = \frac{\gamma}{\sqrt{8\pi^3}\sigma^2\tau} e^{-\frac{\bar{x}^2+\gamma\bar{y}^2}{2\sigma^2} - \frac{(t-\mu_t)^2}{2\tau^2}} \cos\frac{2\pi}{\lambda}(\bar{x} + vt' + \phi), \tag{4}$$

where $\bar{x} = x\cos\theta + y\sin\theta$ and $\bar{y} = -x\sin\theta + y\cos\theta$, and the phase offset ϕ can be set to $\phi = 0$ to obtain an even-phased (cosine) filter and to $\phi = \frac{\pi}{2}$ to obtain an odd-phased (sine) filter — the two filters together form a quadrature pair. The parameters θ and v define the orientation and speed of motion that the filter is tuned for (see [38] for the definition and details of the remaining parameters). An important property of the Gabor representation is direction selectivity (*e.g.,* distinguishing between leftwards and rightwards motion), which is acquired by computing the *Gabor motion energy* through a quadrature filter pair as:

$$E_{v,\theta} = ((\bar{I}_{t-1}, I_t) * g^{0}_{v,\theta})^2 + ((\bar{I}_{t-1}, I_t) * g^{\frac{\pi}{2}}_{v,\theta})^2, \tag{5}$$

where $*$ denotes convolution.

Fig. 3 illustrates why the Gabor representation is useful for registration. We plot four pairs of images along with the motion energies computed through four pairs of Gabor filters. The energy produced with filters tuned to small, horizontal motions gets maximal when the misalignment involves a small, horizontal translation, as illustrated in Fig. 3a. More generally, misalignment in different directions or magnitudes activate different Gabor filters. This property is critical for optimisation, as it guides which direction each optimisation step should take, and what the step size should be. The usage of such a motion representation enables generalisation; if we would replace the images $\bar{I}_{t-1}$ and I_t in Fig. 3 with the images of other subjects, the energy output would change. However, the essential relationship would not: each filter would still reach its maximal response only if the rigid motion (*i.e.,* the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2016.2639448, IEEE Transactions on Image Processing

misalignment) is in tune with the filter parameters. The overall representation, $\mathbf{\Phi}'(\bar{I}_{t-1}, I_t) = (\phi_{t,1}, \phi_{t,2}, \dots, \phi_{t,d}, \dots, \phi_{t,D})$, is computed by pooling the energy matrices $E_{v,\theta}$ after partitioning them into $M \times M$ non-overlapping subregions. An advantage of pooling is to facilitate generalisation in terms of image size. While the size of the energy matrices $E_{v,\theta}$ depends on the size of the images $\bar{I}_{t-1}$ and I_t, after pooling we have $M \times M = M^2$ coefficients per energy matrix independently of image size. The dimensionality of the overall representation is $D = M^2 \times K_G$, where K_G is the number of Gabor filter pairs or, equivalently, the number of energy matrices. The implementation details of the representation are provided in Section VI-C.

To reduce drift errors, we extend the above-described scheme to encode motion with respect to multiple reference frames, $\bar{\mathbf{I}}_{t-1}$. We denote this *multi-frame motion representation* between I_t and $\bar{\mathbf{I}}_{t-1}$ as $\mathbf{\Phi}_t = \mathbf{\Phi}(\bar{\mathbf{I}}_{t-1}, I_t)$. $\mathbf{\Phi}_t$ computes pair-wise motion representations between the misaligned frame and each of the reference frames in $\bar{\mathbf{I}}_{t-1}$, and then averages them over time:

$$\mathbf{\Phi}_t = \mathbf{\Phi}(\bar{\mathbf{I}}_{t-1}, I_t) = \frac{1}{t-\tau} \sum_{t'=\tau}^{t-1} \mathbf{\Phi}'(\bar{I}_{t'}, I_t), \tag{6}$$

where $\tau = \max\{1, t - T_R\}$. For brevity, we rewrite (3) as:

$$\Delta\hat{\mathbf{p}}_t = \mathbf{f}(\mathbf{\Phi}_t; \mathbf{\Theta}) \tag{7}$$

and denote the training set as $\mathcal{D}_\Phi = \{(\mathbf{\Phi}^n, \mathbf{p}^n) : \mathbf{\Phi}^n = \mathbf{\Phi}(\bar{\mathbf{I}}^n, I^n)\}_{n=1}^N$.

As Fig. 3 exemplifies, there is a non-linear relation between the Gabor representation (input) and rigid motion parameters (output). Therefore, it is reasonable to choose a non-linear regression function to model the intended input-output relationship. We choose $\mathbf{f}$ to be a single-hidden-layer neural network as it is a well-established non-linear regressor and one whose properties are well understood [39]. Then the parameter vector $\mathbf{\Theta}$ includes the hidden-layer weights, output layer weights and biases [36].

The optimal parameters $\mathbf{\Theta}^*$ are those that minimise the regularised mean squared error on $\mathcal{D}_\Phi$:

$$\mathbf{\Theta}^* = \underset{\mathbf{\Theta}}{\operatorname{argmin}} \sum_{n=1}^{N} ||\mathbf{p}^n - \Delta\hat{\mathbf{p}}^n||^2 + \alpha||\mathbf{\Theta}||^2, \tag{8}$$

where $\Delta\hat{\mathbf{p}}^n = \mathbf{f}(\mathbf{\Phi}^n; \mathbf{\Theta})$ and $\alpha \in (0, 1]$ is the regularisation parameter defined during training through cross-validation.

The iterative process in Fig. 1 can achieve accurate registration if the errors of the estimator $\mathbf{f}$ get smaller as the amount of rigid motion in I_t gets smaller. However, since the initial error in a given I_t may be high, $\mathcal{D}$ must contain samples with both large and small misalignments. In a dataset with such a broad range of input–output mapping, because of the bias/variance trade-off [40] the estimator may not be able to attain the desired level of accuracy. Although bias can be reduced by increasing model complexity, this would increase the variance of the estimator, thus increasing the risk of overfitting [40]. We address this problem with a coarse-to-fine misalignment estimation, as discussed next.

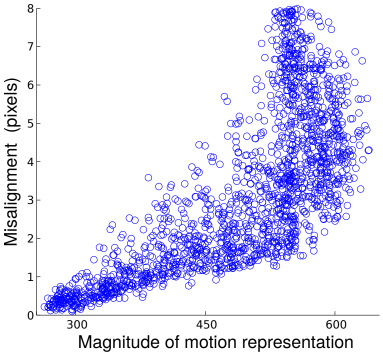

Fig. 4: Correlation between the magnitude of the Gabor representation and misalignment. This correlation suggests that the magnitude of the representation provides information about the amount of misalignment.

B. Coarse-to-fine Misalignment Estimation

To improve the bias/variance trade-off, one can employ a coarse-to-fine cascade of K estimators $\{\mathbf{f}^k\}_{k=1}^K$ with *coarse* estimators tuned to large amounts of misalignment and *fine* estimators tuned to small amounts of misalignment (*e.g.,* [41]). Such a cascade produces better bias/variance trade-offs as each estimator models an input-output mapping with a smaller range [40]. However, typical coarse-to-fine estimation schemes use all the estimators in the cascade, even when the initial registration is small and the finest estimator would suffice [42].

Coarse-to-fine estimation is more efficient when coarse estimators are used only if the initial registration error is large. However, this can be achieved only if we have a prior cue about the amount of misalignment in I_t. Our spatio-temporal Gabor representation provides this cue: Large-magnitude motion activates Gabor filters with large spatial support [38], as was exemplified in Fig. 3. For this reason, the L_2 norm (magnitude) of the representation,

$$\rho_t = ||\mathbf{\Phi}_t|| = \sum_{d=1}^{D} \phi_{t,d}^2, \tag{9}$$

generally gets larger as rigid motion gets larger. Fig. 4 illustrates this relationship, which allows us to use the L_2 norm of a representation as a prior on the amount of misalignment I_t.

We exploit magnitude while constructing the estimators of different granularities, $\{\mathbf{f}^k\}_{k=1}^K$. We choose all estimators $\mathbf{f}^k$ to have the same structure and therefore the estimators differ in their granularity due to the dataset they are trained with. Coarse estimators are trained with samples of larger magnitude and fine estimators with samples of smaller magnitude.

Let us denote the training dataset of each estimator as $\mathcal{D}_\Phi^k$, with $\bigcup_{k=1}^K \mathcal{D}_\Phi^k = \mathcal{D}_\Phi$. A simple way to create the sets $\{\mathcal{D}_\Phi^k\}$ is to first compute the magnitudes of all training samples, $\mathcal{D}_\rho = \{\rho^n : \rho^n = ||\mathbf{\Phi}^n||, \ \forall(\mathbf{\Phi}^n, \mathbf{p}^n) \in \mathcal{D}_\Phi\}$, and to partition the range of $[\min\{\mathcal{D}_\rho\}, \max\{\mathcal{D}_\rho\}]$ into K uniform intervals. However, this partitioning would be sensitive to the sample with maximal magnitude $\max\{\mathcal{D}_\rho\}$, as a large $\max\{\mathcal{D}_\rho\}$ value would affect all intervals. Instead, we allow for non-uniform lengths. To this end, we cluster the set $\mathcal{D}_\rho$ into K clusters by using a Gaussian Mixture Model. Each cluster is a distribution

$\mathcal{N}(\rho|\mu_k, \sigma_k^2)$ where the variance σ_k^2 controls distribution width and is learnt from data. We create a subset $\mathcal{D}_\Phi^k$ by picking the samples that are close to the k^{th} center. Specifically, we create $\mathcal{D}_\Phi^k$ as $\mathcal{D}_\Phi^k = \{(\mathbf{\Phi}^n, \mathbf{p}^n) : \mathcal{N}(\rho^n|\mu_k, \sigma_k) \leqslant 2\sigma_k\}$ (i.e. we cover approximately 95% of the distribution with the $2\sigma_k$ rule [43]). Then we train each $\mathbf{f}^k$ by applying the empirical risk minimisation in (8) using the dataset $\mathcal{D}_\Phi^k$.

We estimate misalignment at each iteration as:

$$\Delta\hat{\mathbf{p}}_t = \mathbf{f}^{k^*}(\mathbf{\Phi}_t), \tag{10}$$

where $k^* = \arg_k \max \mathcal{N}(\rho_t|\mu_k, \sigma_k)$. For clarity, we dropped the regressor parameters.

If no registration failure occurs, the procedure described in this section can register each I_t sequentially for $t = 2, \ldots, T$ (Fig. 1). However, when the registration of a frame fails, the corresponding frame must be identified and removed prior to registering subsequent frames, otherwise it becomes a false reference for subsequent frames. This problem is addressed in the next section.

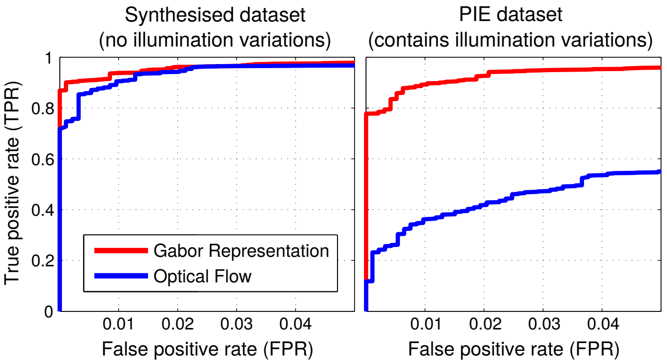

Fig. 5: Failure identification performance on the Synthesised dataset (left) and on the PIE dataset (right) illustrated via ROC curves. The FPR range is restricted to $[0, 0.05]$ for better interpretation. Each curve is computed from 500 positive and 500 negative samples for $\epsilon_y = 1$. Results suggest that the Gabor representation is more robust against illumination variations than the optical flow representation.

V. Failure handling

A. Probabilistic Failure Identification

To account for possible registration failures, it is desirable to generate a second output in addition to the registered sequence $\bar{\mathbf{S}}$. This second output, a vector $\boldsymbol{\lambda}$, should indicate whether the registration at each frame was successful: $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_T)$, where $\lambda_t = 1$ indicates that $\bar{I}_t$ was registered correctly and $\lambda_t = 0$ indicates that the registration failed.

Let $\langle\hat{\mathbf{p}}_t, \mathbf{p}_t\rangle_c$ define the average error in the estimation of canonical points [6] between two registered frames $\bar{I}_t$ and $\bar{I}_{t-1}$. We choose two canonical points[1] $\mathbf{x}_1$ and $\mathbf{x}_2$ as the leftmost and rightmost points in the vertical middle of the image plane. Then $\langle\hat{\mathbf{p}}_t, \mathbf{p}_t\rangle_c$ can be computed as:

$$\langle\hat{\mathbf{p}}_t, \mathbf{p}_t\rangle_c = \sum_{i=1}^{2} \sqrt{||\mathbf{W}(\mathbf{x}_i; \hat{\mathbf{p}}_t) - \mathbf{W}(\mathbf{x}_i; \mathbf{p}_t)||}. \tag{11}$$

When this error is smaller than a *convergence threshold*, ϵ_y, the registration is considered successful. We cast failure identification as a binary classification problem where the two classes are *converged* (*i.e.,* $\langle\hat{\mathbf{p}}_t, \mathbf{p}_t\rangle_c \leqslant \epsilon_y$) and *not converged* (*i.e.,* $\langle\hat{\mathbf{p}}_t, \mathbf{p}_t\rangle_c > \epsilon_y$). We denote those two classes with a binary variable $\tilde{y} \in \{0, 1\}$.

This problem could be solved with a classifier trained with a labelled dataset $\tilde{\mathcal{D}}_\Phi = \{(\mathbf{\Phi}_n, \tilde{y}_n)\}$. However, if we mislabel a frame $\bar{I}_t$ as *converged*, then the mislabelled frame will become a false reference to all subsequent frames; therefore, false positives are more costly than false negatives. A minimal false positive rate is therefore desirable, even if this causes a relatively higher rate of false negatives. False positives can be reduced if we have a confidence measure associated with each estimation, and we reject labelling a sample as *converged* unless the estimation confidence is above an *acceptance threshold* θ_{conv}. A probabilistic classifier can be used to this end, as the confidence value we seek is the probability assigned with the estimation.

We compute the convergence probability via Bayesian learning as [36]:

$$p(\tilde{y}_t = 1|\mathbf{\Phi}_t, \tilde{\mathcal{D}}_\Phi) = \int p(\tilde{y}_t = 1|\mathbf{\Phi}_t, \tilde{\mathbf{\Theta}}) p(\tilde{\mathbf{\Theta}}|\tilde{\mathcal{D}}_\Phi) \mathrm{d}\tilde{\mathbf{\Theta}}, \tag{12}$$

where $\tilde{\mathbf{\Theta}}$ is the vector that contains classifier parameters, $p(\tilde{\mathbf{\Theta}}|\tilde{\mathcal{D}}_\Phi)$ is the prior distribution over parameters $\tilde{\mathbf{\Theta}}$, and $p(\tilde{y}_t = 1|\mathbf{\Phi}_t, \tilde{\mathbf{\Theta}})$ is the probability of a segment with motion representation $\mathbf{\Phi}_t$ having converged when the parameters are $\tilde{\mathbf{\Theta}}$.

The closed-form expression of $p(\tilde{y}_t = 1|\mathbf{\Phi}_t, \tilde{\mathbf{\Theta}})$ depends on the classifier type, and the type of the distribution $p(\tilde{\mathbf{\Theta}}|\tilde{\mathcal{D}}_\Phi)$ is usually selected in a way that would allow (12) to have a closed-form approximation (*i.e.,* conjugate prior) [36]. Since the processes of failure identification and misalignment estimation share a common input space (*i.e.,* spatio-temporal Gabor representation), we choose statistical models with the same structure and use a single-hidden-layer neural network as a classifier. We implement Bayesian learning on this classifier through evidence approximation [44].

The decision on failure identification, λ_t, is defined as

$$\lambda_t = \lambda(\mathbf{\Phi}_t) = \begin{cases} 1 & \text{if } p(\tilde{y}_t = 1|\mathbf{\Phi}_t, \tilde{\mathcal{D}}_\Phi) > \theta_{\text{conv}} \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

We set the threshold θ_{conv} automatically as follows. We compute the ROC curve of the failure identification function of λ by evaluating the true positive rate (TPR) and false positive rate (FPR) on a validation set for a range of threshold values θ_{conv}, and select the θ_{conv} that produces a low false positive rate (*e.g.,* 0.01) on the ROC curve.

Fig. 5 illustrates the failure identification performance of the employed Bayesian neural network on two validation sets: one with constant illumination and one with illumination variations (the datasets are described in Section VI-B). Fig. 5a shows that

[1] Two points suffice to define Euclidean motion.

Algorithm 1 Procedure CORRECT

Input Failed frame, aligned frames: $(I_{t_f}, \bar{\mathcal{I}})$
Output Registered frame, convergence index: $(\bar{I}_{t_f}, \lambda_{t_f})$

```
for $\bar{I} \in \{\bar{I}_k\}$ do
    $\bar{I}_{t_f} \leftarrow IterativeRegistration(\bar{I}, I_{t_f})$
    $\lambda_{t_f} \leftarrow \lambda(\mathbf{\Phi}(\bar{I}, I_{t_f}))$
    if $\lambda_{t_f} = 1$ then
        break
    end if
end for
return $(\bar{I}_{t_f}, \lambda_{t_f})$
```

a Bayesian neural network enables reliable failure identification with a TPR larger than 0.90 for a FPR as low as 0.01 for both representations. To highlight the importance of a robust motion representation, we also compare the performance with an optical flow representation [45] that replaces the Gabor representation in the pipeline. Fig. 5b shows that the Gabor representation is significantly more robust against illumination variations than the optical flow representation. The robustness is due to the Gabor filters being localised in space and the Gabor response being normalised in time [28].

B. Failure Correction

Let t_f denote a time when a registration failure occurs. This failure may be corrected by registering with respect to temporally farther frames. To this end, we search for a reference within a set of previously registered frames

$$\bar{\mathcal{I}} = \{\bar{I}_\tau : \bar{I}_\tau = \bar{I}_{\tau_{\min}}, \bar{I}_{\tau_{\min}+1}, \ldots, \bar{I}_{t_f-1} \wedge \lambda_\tau = 1\}, \quad (14)$$

where $\tau_{\min} = \max\{t_f - T_D, 1\}$ and T_D is the length of the temporal window within which correction is attempted. If a reference frame is found, then the failure is corrected and the registration index is updated as $\lambda_{t_f} = 1$. The correction process is summarised in Algorithm 1. Note that *IterativeRegistration* refers to the set of operational blocks with the same label in the lower part of Fig. 1.

The likelihood of a correction can be increased by using frames after the failure time t_f, that is by constructing $\bar{\mathcal{I}}$ as

$$\bar{\mathcal{I}} = \{\bar{I}_\tau : \bar{I}_\tau = \bar{I}_{\tau_{\min}}, \bar{I}_{\tau_{\min}+1}, \ldots, \bar{I}_{\tau_{\max}} \wedge \lambda_\tau = 1\}, \quad (15)$$

where $\tau_{\max} = \min\{T, t_f + T_D\}$. In this case, the registration process will have a delay of T_D frames, which may become acceptable with small T_D values.

VI. EXPERIMENTAL VALIDATION

In this section we validate the ability of the proposed framework to prevent drift errors, to perform robustly in the presence of facial expressions and non-uniform illumination variations, to identify failures reliably and to generalise to unseen conditions. We first compare multi-frame and single-frame registration for MUMIE. Then we compare MUMIE with state-of-the-art methods on sequences with facial expression variations and on sequences with non-uniform illumination variations. The latter cause registration failures, enabling us to evaluate the failure identification and correction of the proposed framework. We validate generalisation by always conducting experiments in a cross-database manner, that is, by training only on one dataset and testing on different ones.

A. Evaluation Measures

We validate sequence registration performance by evaluating the ability of a method to reduce the overall registration error and its tendency to generate drift errors. To identify and compare drift errors, we also illustrate sequence registration performance by visualising the error variation over time.

We measure the *registration error*, $e_{s,t}$, of the t^{th} frame of the s^{th} sequence by measuring the error in the estimation of the canonical points (see Section V-A):

$$e_{s,t} = \frac{1}{2} \sum_{i=1}^{2} \sqrt{||\mathbf{x}_{i,t} - \mathbf{W}(\mathbf{x}'_{i,t}; \hat{\mathbf{p}}_t^s)||}, \quad (16)$$

where $\hat{\mathbf{p}}_t^s$ is the estimated transformation, $\mathbf{x}_{i,t}$ is a canonical point and $\mathbf{x}'_{i,t}$ is the canonical point after perturbation by a rigid motion $\mathbf{p}_t^s$. The *average error*, $\bar{e}_s$, over a sequence s is:

$$\bar{e}_s = \frac{1}{T-1} \sum_{t=2}^{T} e_{s,t}, \quad (17)$$

where T is the sequence length. (Note that the error is measured with respect to the initial frame.) The *overall average error*, $\bar{e}$, for a dataset is:

$$\bar{e} = \frac{1}{N_S} \sum_{s=1}^{N_s} \bar{e}_s, \quad (18)$$

where N_S is the number of sequences in the dataset. The *average drift error*, $\bar{e}_{drift}$, is defined as:

$$\bar{e}_{drift} = \frac{1}{N_S} \sum_{s=1}^{N_S} e_{s,T}. \quad (19)$$

Since drift error accumulates over time, the registration error between the first and last frames of the sequences serves as a useful measure of drift [46]. Finally, we use the *percentage of converged frames* measure, c, which is commonly used for registration algorithms [6], [7]:

$$c = 100 \times \frac{|\{e_{s,t} : e_{s,t} < 1, s \in \mathbb{N}_{[1,N_S]}, t \in \mathbb{N}_{[2,T]}\}|}{N_S(T-1)}, \quad (20)$$

where $|\cdot|$ denotes set cardinality. The measure c is a useful alternative to the overall average error when the average error is biased by a few frames with a high registration error.

Following [6], we introduce a registration error to frames by perturbing the canonical points with a random value drawn from a Gaussian white noise distribution with σ_{perturb} standard deviation. Since we focus on measuring registration accuracy and tendency to drift errors, we set $\sigma_{\text{perturb}} = 2$ when comparing with other methods, as LK methods may not converge for larger values [7]. However, we test our method with larger σ_{perturb} values when analysing its performance for coarse-to-fine registration in Section VI-F.

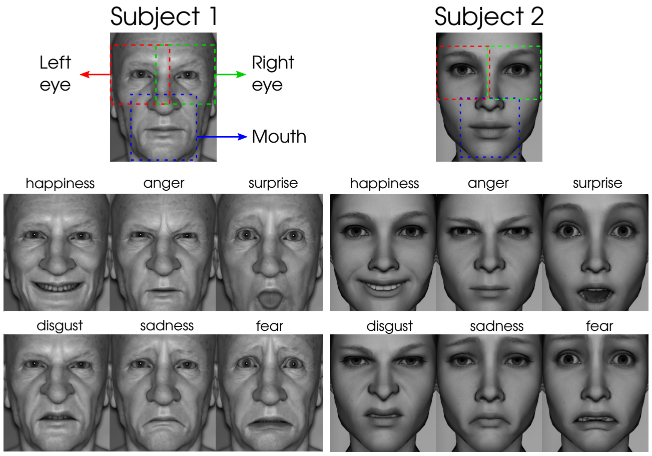

Fig. 6: The apex frame of the six-basic expressions in the Synthesised dataset. The top-left facial image shows the cropping regions for part-based registration.

B. Test Datasets

To validate performance with real sequences of facial expression variations, we perform registration on three facial datasets: CK+ [47], MMI [48] and AFEW [49]. CK+ and MMI contains sequences of posed facial expressions of frontal faces. AFEW comprises sequences cropped from movies; the challenges of this dataset include out-of-plane head pose variations, illumination variations and background motion. Registered videos from these sequences are available for qualitative analysis as supplementary material [2].

To quantify *robustness against illumination variations* we use the PIE dataset [50]. This dataset is collected from subjects that are sitting stably in front of a camera while the illumination conditions are changed rapidly in a controlled manner (see Fig. 7). We use 67 sequences (all the sequences that contain a frontal face). Each sequence is 21 frames long.

To quantify *robustness against non-rigid facial motions* we synthesised facial sequences with expression variations. We will refer to this as the *Synthesised* dataset. The need for such a test sequence arises from the goal of having only expression variations without head or body movements. People tend to move while displaying an expression even in controlled datasets such as MMI [48] and therefore a ground truth for rigid registration cannot be obtained. To produce realistic faces, we use Autodesk Maya and two publicly available facial rigs[3], Old Man (Subject 1) and Ilana (Subject 2). Subject 1 is an old male with a wrinkled face, whereas Subject 2 is a young female with a smooth skin (see Fig. 6). We created sequences that contain the six basic expressions by using the Action Units that are associated with those expressions. All sequences start with a neutral facial appearance, reach the apex, and then return to neutral appearance. We also include one sequence where there are no expression variations, thus yielding to a total of 14 sequences for the two subjects.

Prior to registration, we crop and resize the frames for all datasets. For whole-face registration, we first crop faces based on eye locations, and then resize the cropped frames to 200×200 pixels. For part-based registration, we first locate the centres of both eyes and mouth, and then crop each of these components so that the eye/mouth sits in the centre of frame after cropping. The cropped frames are then resized to 50×50 pixels. Fig. 6 illustrates the boundaries of the cropped components. For the Synthesised dataset we locate the centres of eyes and mouth manually. For the PIE and CK+ datasets we use the facial landmarks provided with the dataset. For the MMI and AFEW datasets we detect faces using OpenCV and locate landmarks using SDM [13].

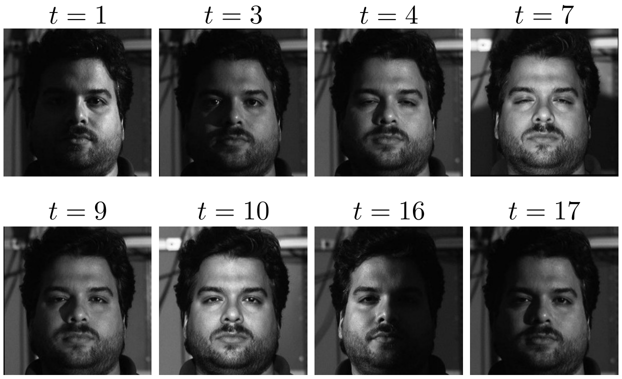

Fig. 7: Sample frames from the PIE dataset. All the sequences in this dataset undergo similar illumination variations.

C. Implementation details and parameter sensitivity

To compute the Gabor representation, we partition the energy matrices into $M \times M = 3 \times 3$ pooling subregions. We use standard deviation pooling (*i.e.,* compute the standard deviation of the values in each subregion), as it outperforms mean and max pooling [28]. We use Gabor filters across 8 orientations, $\{0°, 45°, \dots, 315°\}$ and 3 scales, $\{2^j\}_{j=0}^2$, yielding a filter bank with $K_G = 24$ filters, and an overall representation with $D = 9 \times 24 = 216$ features.

For optimisation we used the scaled conjugate gradients algorithm. We conducted the training on MATLAB using the NETLAB [44] library and the testing on a C++ implementation. We set the convergence threshold $\epsilon_y = 1$ pixel, which is the value used for the evaluation of the LK framework [6]. During correction, the width of the temporal window is set to $T_D = 5$ and we apply correction with a temporal window that considers also subsequent frames (*i.e.,* online-with-delay). We created the training samples from CK+ [47] by perturbing frames from 20 sequences where we perceived no head or body motion. We fix the number of training samples to $N = 15,000$. We set the maximum number of iterations to $K_{\text{max}} = 12$, which is sufficient for convergence for $\sigma_{\text{perturb}} = 2$ (see Section VI-A). Note that in Section VI-F we analyse performance against large registration errors with more iterations.

The number of estimators is $K = 5$, the number of hidden nodes in the neural network is $N_{\text{hidden}} = 10$, and the number of iterations is $N_{\text{iter}} = 500$. We will present below an experimental analysis that shows the effect of varying these parameters. For this purpose, we create testing data

[2] Supplementary material is on ftp://spit.eecs.qmul.ac.uk/pub/es/s.zip

[3] http://facewaretech.com/sdm_categories/rigs/

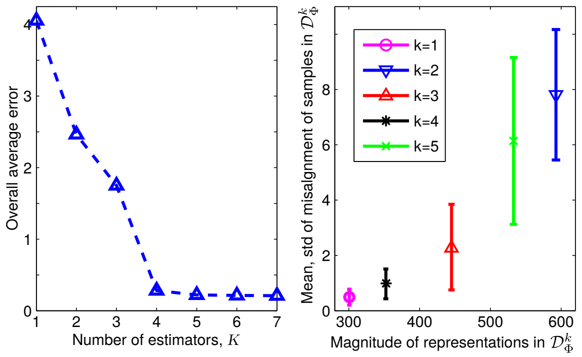

Fig. 8: (Left): average registration error against the number of estimators, K, suggests that $K < 4$ estimators are insufficient for accurate registration. (Right): The mean and standard deviation of misalignment of samples in each dataset $\mathcal{D}^k_\Phi$ against the average magnitude of representations in $\mathcal{D}^k_\Phi$, highlights the coarse-to-fine structure of the set of 5 estimators.

from the 6-basic expression sequences of the Synthesised dataset; specifically, we create misaligned image pairs (*i.e.,* two-frame sequences) by picking all consecutive image pairs and perturbing the second image with $\sigma_{\text{perturb}} = 2$, and thus obtain $N_S = 228$ two-frame sequences.

Fig. 8 (left) reports the performance in terms of overall average error, $\bar{e}$, when varying K. The error is particularly high for $K = 1$, which suggests that a single neural-network cannot model the entire range of input-output mapping efficiently (see Section IV-B). In fact, when $K = 1$, the optimisation algorithm stops after only 26 iterations, which is a symptom of inefficient learning. Limited improvement is obtained when K is increased up to 3 as, similarly, training stops early. When $K = 4$ (and beyond) there is a significant performance improvement. In Fig. 8 (right) we illustrate the coarse-to-fine structure of the $K = 5$ estimators through the statistics of their corresponding datasets, $\{\mathcal{D}^k_\Phi\}^5_{k=1}$. Specifically, we show the average and the standard deviation of the registration error of all the samples in a $\mathcal{D}^k_\Phi$ against the average magnitude of the representations in $\mathcal{D}^k_\Phi$. Some estimators have a coarse structure as their training samples have large misalignment (*e.g.,* $k = 2, 5$), and others have a fine structure as their training samples have smaller misalignment (*e.g.,* $k = 1, 4$).

To prevent overfitting, it is useful to add a random noise to the motion representations computed from the training images. This noise is drawn from an isotropic zero-mean Gaussian distribution with standard deviation $\sigma_{\text{noise}} = 0.5$. Other well-known strategies to prevent overfitting are reducing N_{hidden} (*i.e.,* using a simpler model) or reducing N_{iter} (*i.e.,* performing early termination) [36]. Fig. 9 compares the efficiency of these three approaches in preventing overfitting by providing the average convergence rate in the presence of three additional image variations. The first two variations are image blur with a Gaussian kernel of standard deviation 2 and additive white noise with standard deviation of 8 (similarly to [6]); these variations are applied to the second images of the test pairs created from the Synthesised dataset. The third variation is illumination, for which we created $N_S = 400$ two-frame test sequences from the PIE dataset. Fig. 9a shows that adjusting the N_{iter} parameter provides no performance improvement against image variations. The N_{hidden} parameter can be adjusted to improve performance against white noise. However, only limited improvement can be achieved against blur and illumination variations. On the other hand, adding noise to training samples, can improve performance significantly: With $\sigma_{\text{noise}} = 0.5$, $\sigma_{\text{noise}} = 0.6$, the performance in the presence of blur, white noise or illumination variations becomes similar to the performance without those variations.

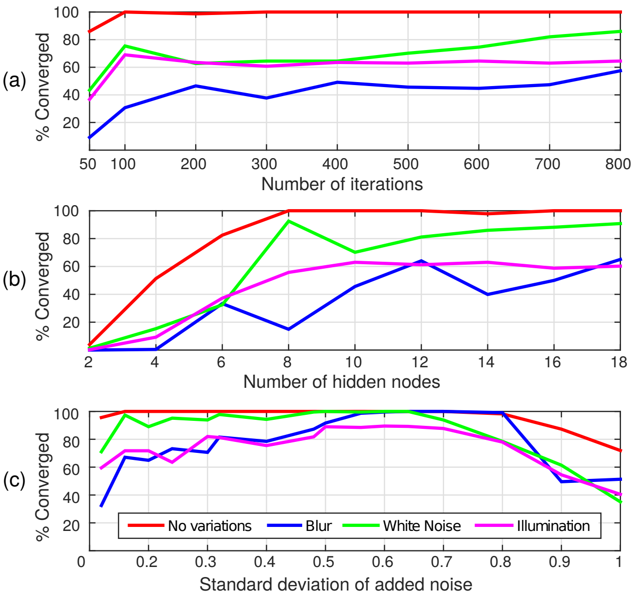

Fig. 9: Registration performance against the (a) number of iterations, (b) number of hidden nodes and (c) σ_{noise} of the training samples. Adding noise to training samples with a σ_{noise} of approximately 0.6 enables the best generalisation against image blur, white noise and illumination variations.

D. Methods under comparison

We compare the proposed framework, MUMIE, with a method from each of the categories listed in Table I. We selected recently proposed robust methods with available software. In categories where there are multiple methods, we select experimentally the best-performing ones for the comparison: (i) the SURF-based method as the keypoint-based method, which generally outperformed the MSER method; (ii) the Robust FFT (R-FFT) method [24] as the transformation-based method, which, to the best of our knowledge, is the only method that proved robust against illumination variations and other outliers; (iii) the GradCorr method [7] as the direct method, which outperformed a number of Lucas-Kanade (LK) variants, namely IC-LK [6], ECC-LK [9] and Fourier-LK [25]. We also compare with (iv) SDM [13], a registration based on landmark localisation. Specifically, we perform registration by computing an Euclidean transformation based on the eye corners, which are useful reference points for rigid registration. However, SDM (and recent landmark localisers [18]) requires the entire face for localising landmarks and therefore we perform only whole-face registration.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2016.2639448, IEEE Transactions on Image Processing

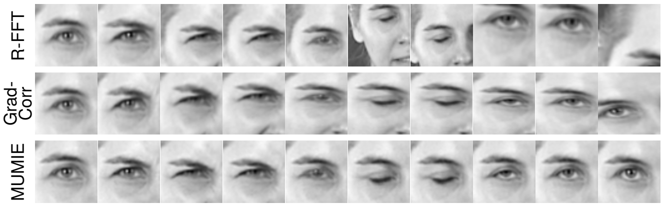

Fig. 10: Registration results for R-FFT, GradCorr our method, MUMIE, on a sequence with a disgust expression followed by blinking. MUMIE accumulates little drift error and is not affected by the sudden motions that occur during blinking.

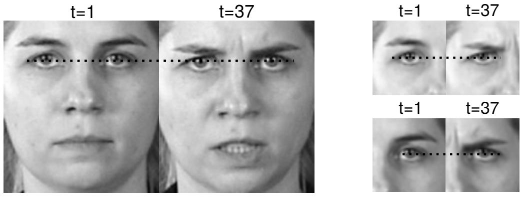

Fig. 11: Illustration that depicts the advantage of part-based registration for addressing out-of-plane rotations. The subject displays a small pitch rotation between the neutral phase ($t = 1$) and the apex phases ($t = 37$) of the expression. With whole-face registration (left), the effect of head-pose rotation is more evident, as the eye corners move visibly downwards in $t = 37$. The effect is less visible in part-based registration for left and right eye, as the eye corners are better aligned.

E. Results and discussion

The results produced by MUMIE to register real sequences from the CK+, MMI and AFEW datasets with various types of facial activity (*e.g.,* talking and facial expressions other than those of the six-basic emotions), out-of-plane head pose variations, occlusions and background motion are provided as supplementary material.

Fig. 10 shows registered frames from an 80-frame long MMI sequence that contains a disgust expression. The sequence contains also a blinking expression, which is a challenging quick facial action that may cause other algorithms to fail. MUMIE achieves accurate registration and a considerably smaller drift error.

Fig. 11 shows results from an anger sequence that contains a pitch rotation. Whole-face registration causes a downward motion around the eyes, which may be detrimental to the analysis of facial activity. When the eyes are registered independently, the effect of head rotation is reduced to a better extent. Sequences with head pose variation highlight the importance of doing part-based registration instead of whole-face registration.

We now quantify the benefits of using multiple reference frames and then compare MUMIE with other methods.

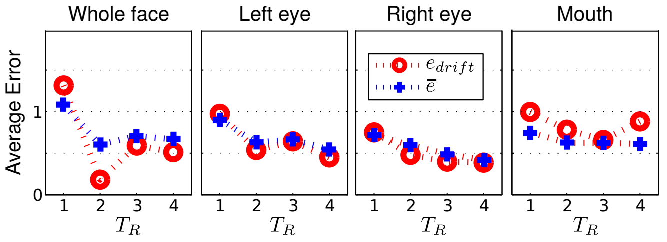

Fig. 12: Average drift error, e_{drift}, and overall average error, $\bar{e}$, on the Synthesised dataset for varying numbers of reference frames, T_R.

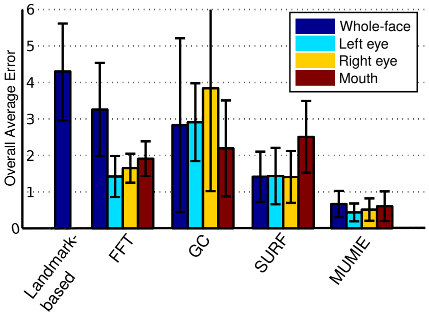

Fig. 13: Sequence registration performance in terms of average registration error over 14 sequences (Synthesised dataset).

Fig. 12 depicts the overall average error and average drift error on the Synthesised dataset when varying the number of reference frames, T_R. When $T_R = 2$ instead of $T_R = 1$ the error decreases consistently. The average registration error for the whole-face, left eye, right eye and mouth decrease respectively to 13%, 55%, 64%, 78% when T_R is set to 2 instead of 1. When T_R is larger than 2 the error decreases generally at a lower rate and sometimes increases. The fact that performance saturates with $T_R = 2$ is desirable from a computational complexity perspective, as computation time increases with T_R. In the remaining experiments, we therefore set $T_R = 2$ while obtaining the multi-frame registration results for our method. The averaging that takes place when we integrate information from multiple frames as in (6) may be responsible in providing little improvement when $T_R > 2$. While taking the average has the advantage of keeping the input representation at a limited length that is independent of T_R, it also reduces the weight of each individual frame as T_R increases, since the average is computed by dividing the contribution of each frame with T_R.

Fig. 13 compares the *average registration error* of MUMIE with other methods on the Synthesised dataset. Overall, Fig. 13 suggests that MUMIE outperforms other methods significantly on sequences with facial expression variations. The error variation for each sequence over time is plotted in Fig. 15 for landmark-based whole-face registration and in Fig. 14 for other methods. Small landmark localisation errors among consecutive frames cause jittering when using landmark-based registration. Fig. 16 shows the difference between a sample pair of consecutive frames from the neutral sequence of Subject 1. A similar jittering can also be observed when using the R-FFT method. On the other hand, registration using SURF, GradCorr and MUMIE produces only little jittering, also for non-neutral sequences (registered sequences are provided as supplementary material). Even though the registration error may increase with expression variations (see Fig. 14), this increase happens gradually without a jittering effect and the registration error at the end of the sequences becomes low,

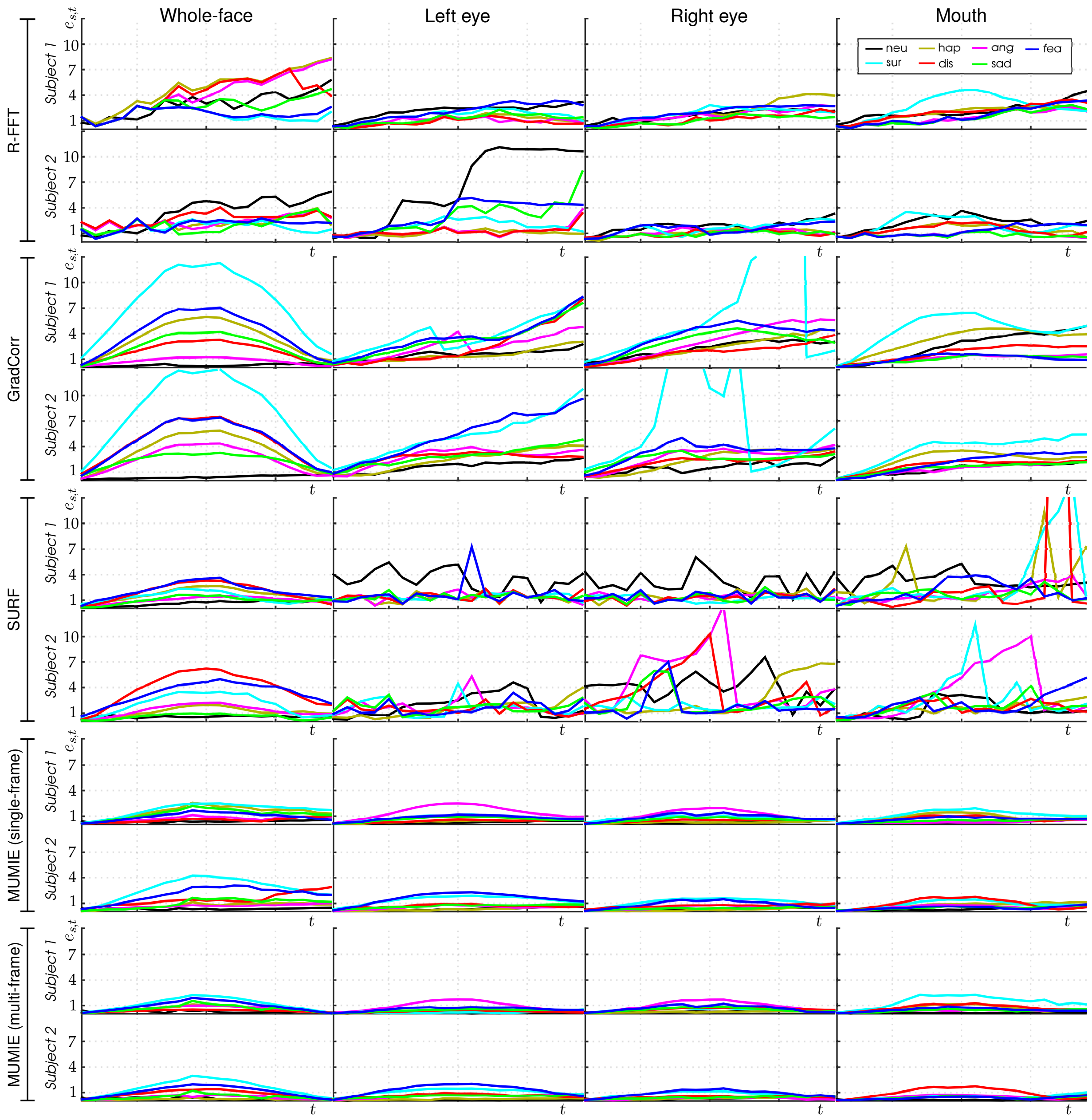

Fig. 14: Whole-face and part-based registration errors of compared methods on the Synthesised dataset. Each line represents the error over time, $e_{s,t}$, for one sequence (see legend on top right). MUMIE (multi-frame) results are obtained with $T_R = 2$. MUMIE outperforms other methods, with a notable difference in part-based registration.

which is indicative of low drift error. The best results are obtained with MUMIE (multi-frame) and, expectedly, MUMIE (single-frame) produces higher drift errors compared to its multi-frame variant.

However, the whole-face registration performance of SURF and GradCorr do not generalise to part-based registration. SURF keypoints are extracted from regions with rich texture. In part-based registration, frames contain less texture and relatively higher non-rigid motions. Therefore, finding keypoints for rigid registration becomes more challenging. The part-based registration error of GradCorr generally increases gradually over time. However, unlike whole-face registration, the error is not reversed in the offset of the expression; therefore, part-based registration with GradCorr yields visible drift errors. While the failure of a generic rigid registration method when the input has considerable non-rigid variations is not surprising, the large error for the neutral sequences is an unexpected result. This may suggest that GradCorr requires some texture variation to operate reliably, even for simple cases without outlier motions.

Part-based registration is problematic also for R-FFT: see for example the large performance difference between the left and right eye of Subject 2 in Fig. 14. While investigating this irregular outcome further, we noticed a difference between the unregistered versions of the left and the right eye sequences. The initial registration error in left eye sequences of Subject 2

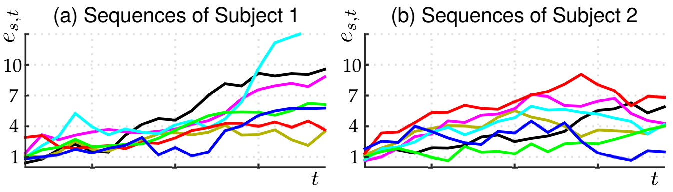

Fig. 15: Performance of landmark-based registration on sequences of the Synthesised dataset depicted separately for sequences of Subject 1 and Subject 2. See legend of Fig. 14 for the expression corresponding to each colour.

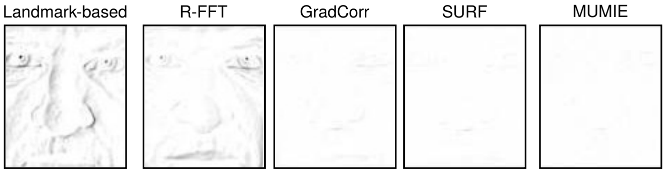

Fig. 16: Difference images computed from a consecutive pair of images from the neutral sequences of Subject 1 and Subject 2 of the Synthesised dataset. Grey levels visualise the registration errors. GradCorr, SURF and MUMIE produce little jittering error.

was causing the facial contour to appear in some frames and disappear in others. The R-FFT method operates on gradient images, and the contour of the face produces a high gradient which may be misleading the FFT-based algorithm.

The part-based registration errors of MUMIE are considerably smaller than those of other methods (Fig. 13 and Fig. 14). The error for whole-face registration *does* generalise to part-based registration; that is, even though the error grows as the expression evolves to apex, the error decreases during offset. MUMIE outperforms other methods, even when it is used with a single reference frame, *i.e.,* $T_R = 1$. The performance of MUMIE with $T_R = 2$ is particularly high, as the final error is less then 1 pixel for all sequences except the mouth sequences of Subject 1 and 2 for the surprise expression (see Fig. 14), which is the expression that involves arguably the largest non-rigid variation. The left and right eye performance of each subject is quite similar, which implies that symmetrical non-rigid motions of the same subject yield consistent results. The surprise and fear expressions cause higher errors when registering the eyes of Subject 2; this may be due to the eyebrows of Subject 2 being raised higher than those of Subject 1. (For both subjects, we raised eyebrows as much as possible when synthesising sequences, however, there are differences between facial rigs of the subjects.) Other inter-subject performance differences may be due to the skin texture; while Subject 1 has wrinkles, the skin of Subject 2 is smooth. Wrinkles are advantageous as additional texture if they are not moved by the expression, or they may be disadvantageous if they cause more non-rigid motion.

Non-uniform illumination variations typically cause registration failures on the PIE dataset, rendering the average sequence registration error of little use for quantitative evaluation. Therefore, we discuss the compared methods using the error visualised over time, where the performance of methods before and after failure is observed directly. Fig. 17 illustrates the performance of all compared methods for 5 randomly selected PIE sequences. (The results of all 67 sequences are shown as videos in the supplementary material.) Landmark-based registration deteriorates in the presence of illumination variations due to increased error in landmark localisation. SURF-based registration does not perform reliably in the PIE dataset as the number of matched keypoints falls significantly.

R-FFT is only slightly affected by illumination variations due to the robustness in the design of this method. R-FFT fails while registering the 17^{th} frame of almost all PIE sequences, due to the sudden illumination variation in this frame (see Fig. 7). GradCorr is also designed to be robust, and its performance deteriorates only slightly with illumination variations. Similarly to R-FFT, registration via GradCorr typically fails in the 17^{th} frames of the sequences. Compared to SURF or landmark-based registration, both R-FFT and GradCorr achieve significantly better performance in the presence of illumination variations. However, both methods accumulate drift errors over time.

Fig. 17 depicts the performance of MUMIE in after *failure identification and correction*. Uncorrected failures occurred only in two frames of Subject-34's sequence with MUMIE (single-frame). Overall, Fig. 17 suggests a considerable difference between MUMIE and other methods: the error is lower than that of other methods and, even though error does increase over time, the increase is lower than R-FFT or GradCorr. MUMIE (multi-frame) performs particularly well, as the error in the last frame is smaller than 1 pixel for all sequences.

Finally, Fig. 18 (bottom row) shows the average error, $\bar{e}_s$, for each sequence of MUMIE (multi-frame) with and without failure identification and correction. The former is computed by eliminating the frames where correction was not possible — Fig. 18 (top row) shows the ratio of those frames. The performance of our method is notably accurate after failures are automatically corrected, with an average error smaller than 1 pixel for 65 out of 67 sequences. Our method has successfully corrected most of failures of the PIE sequences (see Fig. 18 top). However, correcting a failure within a sequence may not be possible if a sudden appearance variation (*e.g.,* out-of-plane head rotation) makes a frame visually dissimilar from all preceding frames. This may cause subsequent registration failures, and a reasonable action to take after a number of failures is to restart the registration process by changing the reference frame to the one where the sudden appearance variation caused the registration failure.

F. Computation time and convergence rate

We report the computation time of the proposed framework and highlight the usefulness of employing the magnitude of the motion representation as prior information while performing coarse-to-fine estimation.

Fig. 19 (left) shows the computation time per frame with respect to the amount of initial registration error. The overall average computation takes 2.74 seconds when all estimators are applied in a cascaded manner, and 1.59 seconds when

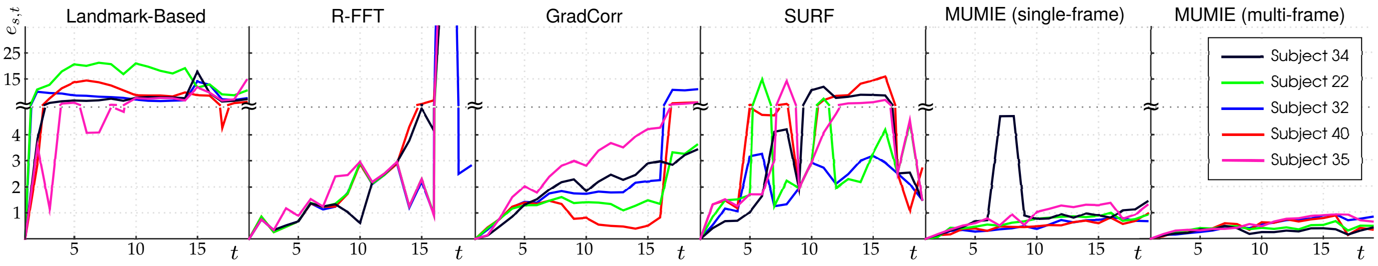

Fig. 17: The performance of compared methods on five randomly selected PIE sequences, illustrated as error per frame over time, $e_{s,t}$. Each sequence is represented with a different colour. Note that we depict error at two different scales by inserting a break into the vertical axis. Even the robust R-FFT and GradCorr methods accumulate significant drift errors over time, whereas MUMIE produces little drift error, particularly in a multi-frame setting (*i.e.,* with $T_R = 2$).

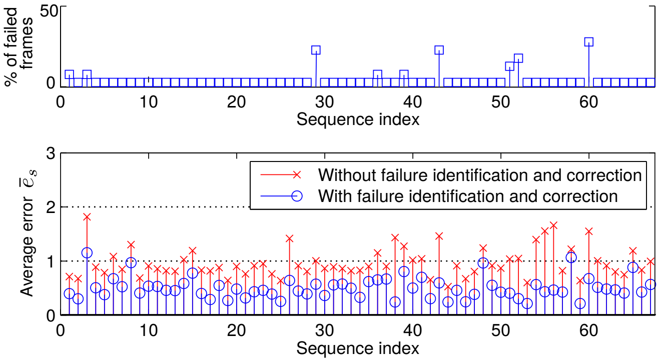

Fig. 18: Performance of MUMIE (multi-frame) for each sequence of the PIE dataset. (Top): The percentage of successfully registered frames. (Bottom): The average registration error with and without failure identification and correction.

TABLE II: Convergence rate against the amount of registration error. A representation computed from Gabor filters across 5 scales, $\{2^j\}_{j=0}^4$, can tackle larger registration errors than one computed from filters at 3 scales, $\{2^j\}_{j=0}^2$.

	Amount of registration error (pixels, ±1)							
	4	6	8	10	12	14	16	18
Conv. rate with 3 scales	100	100	100	66.7	38.9	16.7	16.7	0.0
Conv. rate with 5 scales	100	100	100	100	100	100	100	79.2

estimators are selected at each iteration based on the magnitude of the motion representation (*i.e.,* adaptively). For the cascaded approach, we allowed 6 iterations for the estimators except the finest one, as allowing for fewer iterations prevented convergence for some samples. The adaptive approach is on average faster, as coarse estimators are employed only when the initial registration error is large. Also, even when coarse estimators are used, they are generally used for less iterations, as we need not define a termination criterion for each estimator in the cascade.

Fig. 19 (right) highlights the advantage of the adaptive approach by showing the error against the number of iterations on two different test sets: one that includes samples with a registration error up to 4 pixels (*i.e.,* small misalignment) and one where the registration error of samples reaches up to 15 pixels (*i.e.,* large misalignment). As the registration error decreases, the magnitude of the representation also decreases, and therefore the adaptive approach proceeds registration with finer estimators, which results in a monotonic decrease in average error, and an earlier convergence compared to the cascaded approach when misalignment is small. With the cascaded approach, the error is not always reduced monotonically as in some cases the coarse estimators reach their granularity limit before they reach their limit of iterations, in which case they cannot reduce the registration error further. The cascaded and adaptive approaches reduce the error at a similar rate on the set with samples of large misalignment. However, the cascaded approach is still slower on average, as in some cases the convergence occurs before the last (*i.e.,* finest) estimator, yet, the cascaded approach needs to proceed with the subsequent estimators in the cascade at least for one iteration.

The maximal amount of registration error that can be tackled by our method depends on the scale of the Gabor filters that we use to compute the representation. We used filters at three scales, $\{2^j\}_{j=0}^2$, which may not converge if the registration error is 10 pixels or larger (see Table II). However, larger filters can tackle larger errors, as we report in Table II, where we trained a model that is based on a representation computed from filters at five scales, $\{2^j\}_{j=0}^4$.

VII. Conclusion

We proposed a novel rigid registration framework based on optimisation via statistical learning that can cope with outlier non-rigid facial motions, drift errors and registration failures. Extensive experiments showed that using multiple reference frames during registration reduces drift errors and the proposed framework performs accurate registration in the presence of facial expressions or non-uniform illumination variations. Overall, the proposed framework performs reliably and consistently across various scenarios, both for whole-face or part-based registration. The code of the proposed method and the synthesised facial expression sequences are available as supplementary material.

Future work includes investigating motion representations that are both robust and computationally efficient, and investigating the benefits of the proposed framework further in visual speech recognition and micro-expression detection.

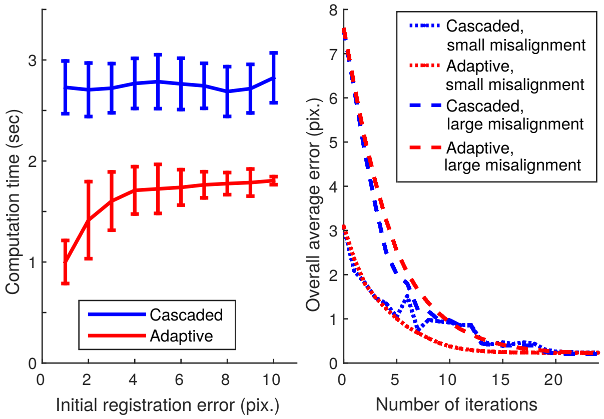

Fig. 19: The efficiency improvement achieved by choosing the estimators based on the motion representation's magnitude (*i.e.,* adaptively) instead of applying all estimators in a cascaded manner. (Left): Computation time against initial registration error, shows that registration takes less time with the adaptive approach as coarse estimators are used only when misalignment is large. (Right): Registration error against the number of iterations, depicted separately for samples of small misalignment and large misalignment. Note that the error decreases monotonically with the adaptive approach.

Acknowledgement

The research work of Evangelos Sariyanidi and Hatice Gunes has been partially supported by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood (Grant Ref.: EP/L00416X/1).

References

[1] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Int'l Conf. Affective Computing and Intelligent Interaction*, 2013, pp. 356–361.

[2] Z. Zhou, X. Hong, G. Zhao, and M. Pietikainen, "A compact representation of visual speech data using latent variables," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 1–1, 2014.

[3] Z. Ambadar, J. W. Schooler, and J. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions," *Psychological Science*, vol. 16, no. 5, pp. 403–410, 2005.

[4] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro facial expression database: Inducement, collection and baseline," in *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2013, pp. 1–6.

[5] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation and recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[6] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *Int'l J. Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[7] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Robust and efficient parametric face alignment," in *Proc. IEEE Int'l Conf. Computer Vision*, 2011, pp. 1847–1854.

[8] E. Muñoz, P. Márquez-Neila, and L. Baumela, "Rationalizing efficient compositional image alignment," *Int'l J. of Computer Vision*, vol. 112, no. 3, pp. 354–372, 2015.

[9] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, 2008.

[10] S. Oron, A. Bar-Hille, and S. Avidan, "Extended Lucas-Kanade tracking," in *Proc. European Conf. Computer Vision*, 2014, pp. 142–156.

[11] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3659–3667.

[12] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 1078–1085.

[13] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[14] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 824–832.

[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[16] N. Wang, X. Gao, D. Tao, and X. Li, "Facial feature point detection: A comprehensive survey," *arXiv preprint arXiv:1410.1037*, 2014.

[17] Q. Liu, J. Deng, and D. Tao, "Dual sparse constrained cascade regression for robust face alignment," *IEEE Trans. on Image Processing*, vol. 25, no. 2, pp. 700–712, 2016.

[18] O. Çeliktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP J. Image and Video Processing*, vol. 2013, no. 1, p. 13, 2013.

[19] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. European Conf. Computer Vision*, 2006, pp. 404–417.

[20] M. Okade and P. K. Biswas, "Video stabilization using maximally stable extremal region features," *Multimedia Tools and Applications*, vol. 68, no. 3, pp. 947–968, 2014.

[21] K.-p. Wang, Y. Gui, X.-h. Zhang, and Q.-f. Yu, "Robust tracking method with drift correction," in *Proc. Int'l Conf. on Image Analysis and Signal Processing*, 2010.

[22] W. Pan, K. Qin, and Y. Chen, "An adaptable-multilayer fractional Fourier transform approach for image registration," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 400–414, 2009.

[23] X. Xiong and K. Qin, "Linearly estimating all parameters of affine motion using radon transform," *IEEE Trans. on Image Processing*, vol. 23, no. 10, pp. 4311–4321, 2014.

[24] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki, "Robust FFT-based scale-invariant image registration with image gradients," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1899–1906, 2010.

[25] A. Ashraf, S. Lucey, and T. Chen, "Fast image alignment in the Fourier domain," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 2480–2487.

[26] D. Schreiber, "Robust template tracking with drift correction," *Pattern Recognition Letters*, vol. 28, no. 12, pp. 1483–1491, 2007.

[27] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[28] E. Sariyandi, H. Gunes, and A. Cavallaro, "Probabilistic temporal subpixel registration for facial expression analysis," in *Proc. Asian Conf. Computer Vision*, 2014, pp. 320–335.

[29] S. Kumar, H. Azartash, M. Biswas, and T. Nguyen, "Real-time affine global motion estimation using phase correlation and its application for digital image stabilization," *IEEE Trans. on Image Processing,*, vol. 20, no. 12, pp. 3406–3418, 2011.

[30] V. J. Traver and F. Pla, "Motion analysis with the radon transform on log-polar images," *J. of Mathematical Imaging and Vision*, vol. 30, no. 2, pp. 147–165, 2008.

[31] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework: Part 2," *Int'l J. Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[32] S. Lucey, R. Navarathna, A. B. Ashraf, and S. Sridharan, "Fourier Lucas-Kanade algorithm," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1383–1396, 2013.

[33] N. Dowson and R. Bowden, "Mutual information for lucas-Kanade tracking (MILK): An inverse compositional formulation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, no. 1, pp. 180–185, 2007.

[34] I. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, no. 6, pp. 810–815, 2004.

[35] N. Anantrasirichai, A. Achim, N. G. Kingsbury, and D. R. Bull, "Atmospheric turbulence mitigation using complex wavelet-based fusion," *IEEE Trans. on Image Processing*, vol. 22, no. 6, pp. 2398–2408, 2013.

[36] C. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006.

[37] E. H. Adelson and J. R. Bergen, "Spatio-temporal energy models for the perception of motion," *The J. of the Optical Society of America*, vol. 2, no. 2, pp. 284–299, 1985.

[38] N. Petkov and E. Subramanian, "Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal Gabor filters with surround inhibition," *Biological Cybernetics*, vol. 97, no. 5-6, pp. 423–439, 2007.

[39] S. Mallat, "Understanding deep convolutional networks," *arXiv preprint arXiv:1601.04920*, 2016.

[40] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[41] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 4998–5006.

[42] J. Le Moigne, N. S. Netanyahu, and R. D. Eastman, *Image registration for remote sensing*. Cambridge University Press, 2011.

[43] J. S. Oakland, *Statistical process control*. Routledge, 2007.

[44] I. Nabney, *NETLAB: algorithms for pattern recognition*. Springer Science & Business Media, 2002.

[45] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, 2003, pp. 363–370.

[46] A. Rav-Acha and S. Peleg, "Lucas-Kanade without iterative warping," in *Proc. IEEE Int'l Conf. Image Processing*, 2006, pp. 1097–1100.

[47] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2010, pp. 94 – 101.

[48] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int'l Conf. Multimedia and Expo*, 2005, p. 5 pp.

[49] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, 2012.

[50] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.

Evangelos Sariyanidi received his BS in 2009 and MS in 2012 from the Istanbul Technical University, Turkey. He is currently a Ph.D. candidate at the School of Electronic Engineering and Computer Science, Queen Mary, University of London, UK. His research interests include computer vision and machine learning, and current focus of interest is the automatic analysis of affective behaviour.

Hatice Gunes is an Associate Professor (Senior Lecturer) in the Computer Science Department at University of Cambridge, UK. Prior to that she led the Affective and Human Computing Lab at Queen Mary University of London. Her research expertise is in the areas of affective computing and social signal processing that lie at the crossroad of multiple disciplines including computer vision, signal processing, machine learning, multimodal interaction and human-robot interaction. She has published over 90 papers in these areas receiving awards for Outstanding Paper (IEEE FG11), Quality Reviewer (IEEE ICME11), Best Demo (IEEE ACII09) and Best Student Paper (VisHCI06). Dr Gunes is the Program Chair of IEEE FG 2017 and the President-Elect of the Association for the Advancement of Affective Computing (AAAC). She serves on the Executive Committee and the Management Board of AAAC and the Steering Committee of IEEE Transactions on Affective Computing. She is an Associate Editor of IEEE Transactions on Affective Computing, IEEE Transactions on Multimedia, and Image and Vision Computing Journal. She has edited Special Issues in International Journal of Synthetic Emotions, Image and Vision Computing, ACM Transactions on Interactive Intelligent Systems and Frontiers in Robotics and AI. Dr Gunes is a Senior Member of the IEEE.

Andrea Cavallaro is Professor of Multimedia Signal Processing and Director of the Centre for Intelligent Sensing at Queen Mary University of London, UK. He received his Ph.D. in Electrical Engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, in 2002. He was a Research Fellow with British Telecommunications (BT) in 2004/2005 and was awarded the Royal Academy of Engineering teaching Prize in 2007; three student paper awards on target tracking and perceptually sensitive coding at IEEE ICASSP in 2005, 2007 and 2009; and the best paper award at IEEE AVSS 2009. Prof. Cavallaro is Senior Area Editor for the IEEE Transactions on Image Processing; and Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technology and IEEE Multimedia. He is a past Area Editor for IEEE Signal Processing Magazine and a past Associate Editor for the IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Signal Processing, and IEEE Signal Processing Magazine. He has published over 170 journal and conference papers, one monograph on Video tracking (2011,Wiley) and three edited books: Multi-camera networks (2009, Elsevier); Analysis, retrieval and delivery of multimedia content (2012, Springer); and Intelligent multimedia surveillance (2013, Springer).