Local Inter-Session Variability Modelling for Object Classification

Kaneswaran Anantharajah
QUT, SAIVT and MILAB

ZongYuan Ge
QUT, CyPhy Lab.

Chris McCool
NICTA, Brisbane

Simon Denman
QUT, SAIVT

Clinton Fookes
QUT, SAIVT

Peter Corke
QUT, CyPhy Lab.

Dian Tjondronegoro
QUT, MILAB

Sridha Sridharan
QUT, SAIVT

Abstract

Object classification is plagued by the issue of session variation. Session variation describes any variation that makes one instance of an object look different to another, for instance due to pose or illumination variation. Recent work in the challenging task of face verification has shown that session variability modelling provides a mechanism to overcome some of these limitations. However, for computer vision purposes, it has only been applied in the limited setting of face verification.

In this paper we propose a local region based inter-session variability (ISV) modelling approach, and apply it to challenging real-world data. We propose a region based session variability modelling approach so that local session variations can be modelled, termed Local ISV. We then demonstrate the efficacy of this technique on a challenging real-world fish image database which includes images taken underwater, providing significant real-world session variations. This Local ISV approach provides a relative performance improvement of, on average, 23% *on the challenging MOBIO, Multi-PIE and SCface face databases. It also provides a relative performance improvement of* 35% *on our challenging fish image dataset.*

1. Introduction

Object classification is a challenging problem due to variations in the appearance of the objects and the environment in which they appear. One of the best known and most well investigated object classification problems is that of face recognition, where variations in subject pose and lighting present significant challenges [6]. A recent state-of-the-art face recognition approach uses session variability modelling [12] to provide a general model that describes the differences that occur between instances of the same class, whether that be from pose, illumination or expression variation. This session variability modelling approach is applied in the context of a free-parts model [16], which discards potentially useful spatial relationships.

The free-parts approach described in [16] divides the face into blocks and each block is considered to be a independent observation of the same object (the face). The distribution of these blocks is described by a Gaussian mixture model (GMM) and has been investigated by several researchers [16, 9, 10, 19]. Lucey and Chen [9] showed that a relevance adaptation approach, similar to the one used for speaker authentication [14], could be used to quickly obtain client (class) specific GMMs by using a universal background model (UBM). Furthermore, Lucey and Chen showed that adding spatial constraints to this free-parts approach could yield state-of-the-art face recognition performance on the BANCA dataset [13]. Sanderson et al. [15] proposed a multi-region probabilistic histogram (MRH) approach which used the free-parts approach as its basis but incorporates spatial constraints and also makes several simplifications for efficiency purposes. This efficient method provided state-of-the-art performance on the labeled faces in the wild (LFW) dataset [1].

Recently in [18, 12] the GMM free-parts (GMM-FP) model was extended to include an inter-session variability (ISV) modelling component. ISV learns a sub-space which models the differences in instances of the same object (the face). Such an approach was initially proposed to cope with similar problems in speaker authentication [17]. This model of session variability is used to estimate session variations in order to suppress, or account, for them. Using this model yielded state-of-the-art performance on several well known face datasets such as MOBIO [11] and Multi-PIE [6]. Despite this state-of-the-art performance, this approach has an obvious limitation as it does not enforce any spatial relationships between the blocks (observations), which discards spatial information which would help to disambiguate between the classes. Furthermore, its general applicability to vision problems has not been shown as it has only ever been applied to face recognition.

Contributions: In this paper we propose a local inter-

[1] http://itee.uq.edu.au/ conrad/lfwcrop/

session variability modelling approach that enforces local spatial relationships that were previously discarded. This approach is similar to [15] which adopts a multi-region probabilistic histogram approach. However, rather than using a probabilistic histogram that uses the zeroth order statistics of a GMM [15], we apply this to the GMM-FP and ISV approaches which, as has been shown in [12], uses the zeroth and first order statistics which provide a better approximation of the underlying data. We also apply, for the first time, the ISV model to the broader problem of object classification to examine the general applicability of this technique. To do this we use a large fish image dataset that contains challenging real-world images consisting of fish images captured in conditions ranging from controlled with a constant background and illumination, through to underwater imagery of fish in their natural habitat with significant illumination and pose variations.

We show that introducing spatial constraints leads to state-of-the-art performance for face and fish image classification. Spatial constraints are introduced by dividing the images into R regions and learning a model specific to each region. This allows us to locally model session variability and capture local identity information. For face recognition this Local ISV approach provides an average relative improvement of 23% for the MOBIO [11], Multi-PIE [6] and SCface [5] databases over the existing state-of-the-art. For fish classification, we show that using Local ISV provides a relative performance improvement of 35%.

Finally, we examine the sensitivity of the Local ISV approach to real-world problems such as errors in face localisation. Using the real-world MOBIO database, which consists of face images captured from a mobile phone, we introduce noise to the manually annotated landmarks to simulate misalignment, a problem often encountered in practical applications [7]. Empirically we show that the Local ISV approach is more sensitive to this misalignment, but still provides superior performance when the noise in the position of the landmarks is less than 20% of the inter-eye distance.

The remainder of the paper is organized as follows. An overview of existing work is presented in Section 2; the proposed region based GMM and ISV based face authentication frame works are explained in Section 3. Databases and protocols used in the experiments are presented in Section 4. In Section 5, we present the experimental results using our novel fish image database and three face databases. We conclude the paper in Section 6.

2. Prior work

2.1. GMM Free-Parts for Face Verification

Several researchers have examined the use of the GMM-FP framework to perform face verification [16, 9, 19]. Introduced in [16], this approach divides the image (the face) into N overlapping blocks which are considered to be independent observations of the same underling signal (the face), $\boldsymbol{O}$. From each block a 2D-DCT feature vector of dimension M is obtained to compactly represent each block, such that the n-th block yields the feature vector $\boldsymbol{o}_n$. Thus the j-th image of the i-th client yields the set of n observations $\boldsymbol{O}_{i,j} = [\boldsymbol{o}_{i,j,1}, \ldots, \boldsymbol{o}_{i,j,n}]$. The distribution of these feature vectors is then modelled using a GMM,

$$Pr\left(\boldsymbol{O} \mid \boldsymbol{\theta}\right) = \prod_{n=1}^{N} \sum_{c=1}^{C} \omega_c \mathcal{N}\left[\boldsymbol{o}_n \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right], \qquad (1)$$

where C is the number of components for the GMM, ω_c is the weight for component c, $\boldsymbol{\mu}_c$ is the mean for component c, and $\boldsymbol{\Sigma}_c$ is the covariance matrix (usually considered to be diagonal) for component c.

In order to overcome the limited number of samples per client, i, mean-only relevance MAP adaptation [9] is used to enrol the client (class). Originally proposed for speaker authentication [14], mean-only relevance MAP adaptation takes a prior model, usually referred to as a universal background model (UBM) GMM, and performs MAP adaptation on the means using the observations of the i-th client, $\boldsymbol{O}_i$, to obtain a model for the client. Since only the mean vectors change, it has been shown [17] that this can be written as,

$$\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_i, \qquad (2)$$

where $\boldsymbol{s}_i$ is the mean super-vector for the i-th client, $\boldsymbol{m}$ is the mean super-vector of the UBM GMM (the prior), $\boldsymbol{z}_i$ is a normally distributed latent variable, and $\boldsymbol{D}$ is a diagonal matrix that incorporates the relevance factor and the covariance matrix [17] and ensures the result is equivalent to mean-only relevance MAP adaptation.

To evaluate the likelihood that image t, described by a set of observations $\boldsymbol{O}_t$, was produced by client i a log-likelihood ratio is used. In this case the positive class is given by the claimed identity i and the negative class is represented by the UBM GMM. Thus, the log-likelihood ratio is,

$$h\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \log\left[p\left(\boldsymbol{O}_t \mid \boldsymbol{s}_i\right)\right] - \log\left[p\left(\boldsymbol{O}_t \mid \boldsymbol{m}\right)\right]. \qquad (3)$$

It was shown in [19] that this could be efficiently calculated using the linear scoring approximation [4] leading to,

$$h_{linear}\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \left(\boldsymbol{s}_i - \boldsymbol{m}\right)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{f}_{t|\boldsymbol{m}}, \qquad (4)$$

where the diagonal matrix $\boldsymbol{\Sigma}$ is formed by concatenating the diagonals of the UBM covariance matrices and $\boldsymbol{f}_{t|\boldsymbol{m}}$ is the super-vector of mean normalised first order statistics as given in [12]. A decision threshold, τ, is applied to this score to decide if the observations were generated by the model, $\boldsymbol{s}_i$. Image, $\boldsymbol{O}_t$, is classified as being of client i if and only if $h_{linear}\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) \geq \tau$.

Super-vector notation is a way of compactly representing data for a GMM. It is particularly useful when we consider mean-only relevance MAP adaptation as the only part of the model that changes is the means. Since the weights, $[\omega_1, \ldots, \omega_C]$, and variances, $[\Sigma_1, \ldots, \Sigma_C]$, are fixed each model can be described by the concatenation of their means to form a single super-vector $\boldsymbol{a} = [\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_C^T]^T$. More details for this notation can be found in [12].

2.2. Inter Session Variability Modelling

Inter-session variability modelling (ISV) has been applied successfully to speaker [17] and face verification [12]. ISV aims to model and suppress session variation, that is variation that makes one image of the same class look different to another image of the same class. For face recognition this is often considered to be illumination, pose or expression variation. At enrollment time session variation is suppressed by jointly estimating a latent session variable along with a latent identity variable, the latent session variable is then discarded. When scoring, an estimate of the latent session variable, $\boldsymbol{x}_t$, is obtained from the test samples, $\boldsymbol{O}_t$. This estimate, $\boldsymbol{x}_t$, is then used to offset the models so that the likelihood function now takes into account the session variation (noise), of the test samples; see [12] Section 3.5 for more details.

Enrolling a client for ISV consists of MAP adaptation, similar to mean-only relevance MAP adaptation. The difference is that a sub-space, $\boldsymbol{U}$, is introduced to model session variation and so restricts the movement for relevance adaptation such that the model for the j-th image of the i-th client (class) is,

$$\boldsymbol{u}_{i,j} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{i,j} + \boldsymbol{D}\boldsymbol{z}_i, \tag{5}$$

where $\boldsymbol{x}_{i,j}$ is the latent session variable and is assumed to be normally distributed. In this way each image is considered to have been produced with its own session variation; for instance due to pose or illumination variation. As previously mentioned when performing enrollment the session varying part ($\boldsymbol{U}\boldsymbol{x}_{i,j}$) is discarded and only those parts pertaining to identity are retained. Thus, the ISV client model is given by,

$$\boldsymbol{s}_{ISV,i} = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_i. \tag{6}$$

This should not be confused with mean-only relevance MAP adaptation (see Equation 2) as the latent variables $\boldsymbol{x}_{i,j}$ and $\boldsymbol{z}_i$ are jointly estimated for ISV.

Scoring with ISV is performed by first estimating the latent session variable, $\boldsymbol{x}_t$, for the test sample $\boldsymbol{O}_t$. This is then used to offset the client model ($\boldsymbol{s}_{ISV,i}$) and the UBM ($\boldsymbol{m}$) so that the log-likelihood is estimated in the session conditions of the test samples. This provides a mechanism to compensated for session variation. When used in the context of linear scoring, this leads to the following log-likelihood ratio (LLR),

$$h_{ISV}\left(\boldsymbol{O}_t, \boldsymbol{s}_{ISV,i}\right) = \left(\boldsymbol{s}_{ISV,i} - \boldsymbol{m}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{f}_{t|\boldsymbol{m}} - \boldsymbol{N}_t \boldsymbol{U} \boldsymbol{x}_{t|UBM}\right), \tag{7}$$

where $\boldsymbol{N}_t$ is the zeroth order statistics for the test sample in a block diagonal matrix as defined in Equation 11 of [12].

3. Proposed approach

We propose to overcome one of the major limitations of the ISV approach to image classification by dividing an image into local regions. Doing this allows us to re-enforce spatial constraints that were previously being discarded. To properly evaluate the local ISV approach we also have to evaluate the local GMM-FP approach to ensure that locally modelling session variability is not being boosted solely by being able to extract local class specific information.

The approach is similar to work conducted in [15] where a probabilistic histogram for local regions was formed using a GMM, termed a multi-region probabilistic histogram (MRH). This MRH approach collates the zeroth order statistics, the occupation probabilities, of a GMM to perform classification. By contrast, we propose to apply local region decomposition to the ISV approach due to their state-of-the-art performance when used globally in [12]. These techniques collate the zeroth and first order statistics of a GMM to perform classification, furthermore, ISV provides an additional constraint to the MAP equations to suppress session variations (noise).

3.1. Local GMM Free-Parts Approach

We propose an extension to the GMM-FP approach whereby the input images are divided into a set of R regions and each region is modelled independently. This approach, termed Local GMM-FP, allows us to derive local descriptions of the identity variation. Similar to the GMM-FP approach, the proposed Local GMM-FP technique divides each region into a set of overlapping blocks from which DCT features are extracted. A local GMM UBM is then learnt for each specific region M_r, $\boldsymbol{m}_r$, and local models of the identity are then obtained using region specific mean-only relevance MAP adaptation,

$$\boldsymbol{s}_{r,i} = \boldsymbol{m}_r + \boldsymbol{D}_r \boldsymbol{z}_{r,i}, \tag{8}$$

where $\boldsymbol{s}_{r,i}$ is the i-th client model corresponding to region r, $\boldsymbol{z}_{r,i}$ is a normally distributed latent variable for region r, and $\boldsymbol{D}_r$ is a diagonal matrix that incorporates the relevance factor and the covariance matrix [17] as per Section 2.1.

The t-th image is compared to the i-th client model in a region specific manner. Thus the observations from the r-th region of t-th image, $\boldsymbol{O}_{r,t}$, are compared to the i-th client's

model for the r-th region, $s_{r,i}$. Thus the LLR becomes region specific,

$$h_{linear}\left(\boldsymbol{O}_{r,t}, \boldsymbol{s}_{r,i}\right) = \left(\boldsymbol{s}_{r,i} - \boldsymbol{m}_r\right)^T \boldsymbol{\Sigma}_r^{-1} \boldsymbol{f}_{r,t|m_r}, \quad (9)$$

where $\boldsymbol{\Sigma}_r$ is the covariance matrix for the r-th region and $f_{r,t|m_r}$ is the mean normalised first order statistics for the r-th region. Subsequently, all region specific scores are summed and compared to the threshold, τ.

3.2. Local Inter-Session Variability Modelling

In this section we propose to apply ISV to local regions so that we can locally model session variability and capture local identity information. We apply a similar concept to Section 3.1 of dividing the image into R regions and again perform MAP adaptation for each region independently. Thus for the j-th image of the i-th client in the r-th region we obtain the model,

$$\boldsymbol{u}_{r,i,j} = \boldsymbol{m}_r + \boldsymbol{U}_r \boldsymbol{x}_{r,i,j} + \boldsymbol{D}_r \boldsymbol{z}_{r,i}. \quad (10)$$

A region specific ISV client model, $\boldsymbol{s}_{ISV,r,i}$, is formed by,

$$\boldsymbol{s}_{ISV,r,i} = \boldsymbol{m}_r + \boldsymbol{D}_r \boldsymbol{z}_{r,i}. \quad (11)$$

During the evaluation process, the region specific latent session variable $\boldsymbol{x}_{r,i}$ is estimated for $\boldsymbol{O}_{r,i}$ using the r-th region from the i-th client model. Then, session variation is compensated for by adding this estimated session offset to $\boldsymbol{s}_{ISV,r,i}$ prior to scoring.

4. Database and Evaluation Protocols

4.1. Fish Image Set

To evaluate the new ISV approach in the broader object classification domain we introduce a new, large fish image dataset consisting of $3,960$ images collected from 468 species. This data consists of real-world images of fish captured in conditions defined as "controlled", "out-of-the-water" and "in-situ". The "controlled" images consist of fish specimens, with their fins spread, taken against a constant background with controlled illumination, see Figure 2 (a) and (b). The "in-situ" images are underwater images of fish in their natural habitat and so there is no control over background or illumination, in addition there is the challenge of the unique underwater imaging environment, see Figure 2 (c) and (d). The "out-of-the-water" images consist of fish specimens, taken out of the water with a varying background and limited control over the illumination conditions, see Figure 2 (e) and (f).

There are two main difficulties when performing classification on the fish imagery. The first is that, in many cases, different species are visually similar, as shown Figure 1 (a)-(d) where it can be seen that four species are visually similar. The second is that there is a high degree of variability in the image quality and environmental conditions, see Figure 2 for example images [2] for some example images.

Approximately half of the images have been captured in the "controlled" condition, where the image of the fish has been captured out-of-the-water with a controlled background. The "in-situ" condition consists of images taken underwater with no control over the background and with significant pose and illumination variations. Approximately one third of the data was captured in this manner. Finally, the remaining images are captured "out-of-the-water", but without a controlled background and may contain some minor pose variation.

Evaluation Protocol: An evaluation protocol, similar to [11] and [3], has been developed for experiments on this dataset. We define three sets of data by splitting the data, based upon species (class), into a training set (*train*) to learn/derive models; a development set (*dev*) to determine the optimal parameters for our models; and an evaluation set (*eval*) to measure the final system performance.

Two protocols are defined to evaluate the system performance when high quality ("controlled") and low quality ("in-situ") data is used to enrol classes. Protocol 1a uses one enrollment image per species from the "controlled" data. Protocol 1b uses one enrollment image per species from the "in-situ" data. For both protocols, the same test imagery (a mix of "controlled", "in-situ" and "out-of-the-water" images) is used. The *train* set consists of $1,296$ images from 169 species, and can be used to learn or derive models for principal component analysis, probabilistic linear discriminant analysis, or for learning the UBM GMM [3]. The *dev* set consists of 958 images from 93 species, and the *eval* set consists of 963 images from 98 species. For these two protocols the *dev* and *eval* partitions consist of the sub-set of species for which we have at least three images, with at least one "controlled" and one "in-situ" image.

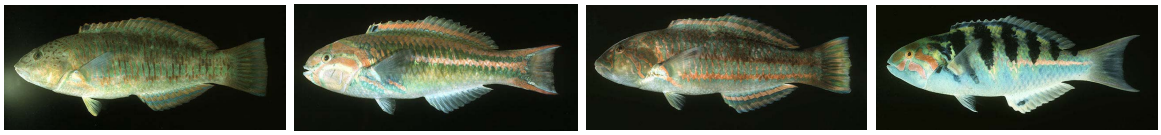
We evaluate system performance by measuring the Rank-n identification rate, using manually annotated bounding boxes.

Rank-n refers to the percentage of queries for which the correct result in within the top n matches. We measure performance at $n = 1$, $n = 5$ and $n = 10$. The bounding boxes were obtained by inscribing a region around the body of each fish, an extra 3% margin was added to avoid losing edge information, example bounding boxes are shown in Figure 2. The new fish database which has been presented will be made publicly available[4].

[2] images (a) and (c) in the Figure 2 are from Australian National Fish Collection CSIRO, (b) is taken by G. Edgar, and (d) is taken by Dennis King

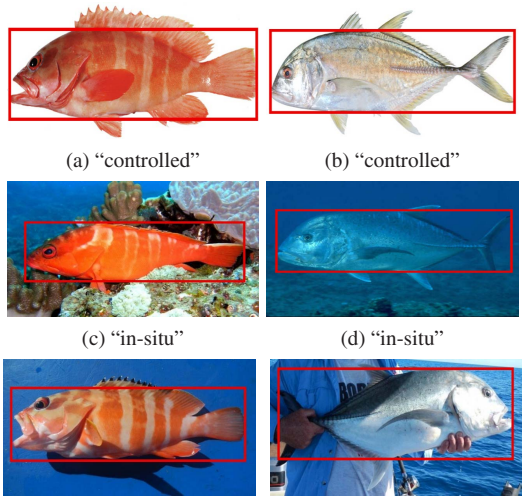
[3] to train ISV there we only use the 155 classes that have more than one image per species

[4] see http://tiny.cc/fishdataset for details



(a) Thalassoma Trilobatum (b) Thalassoma Quinquevittatum (c) Thalassoma Purporeum (d) Thalassoma Hardwicke

Figure 1: Example images of four different fish species, all which have similar visual appearance despite being distinct species. (Images taken by J.E. Randall)



(a) "controlled" (b) "controlled"

(c) "in-situ" (d) "in-situ"

(e) "out-of-the-water" (f) "out-of-the-water"

Figure 2: Example images of two different fish species captured under the three different capture conditions (from top to bottom): "controlled", "in-situ" and "out-of-the-water". Significant variation in appearance due to the changed imaging conditions (session variation) is evident. Ground truth bounding boxes are shown in red.

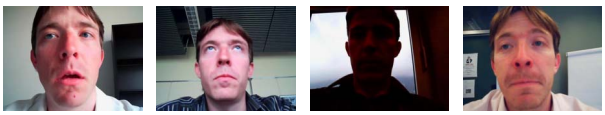


Figure 3: Example images from the MOBIO [11] database.

Figure 4: Example images from the Multi-PIE [6] database.

Figure 5: Example images from the SCface [5] database.

4.2. Face Databases

Three face databases are used to evaluate the proposed approach: MOBIO [11], Multi-PIE [6], and SCface [5]. Face verification is still a challenging classification problem and we want to compare the proposed approach to the current state-of-the-art. The MOBIO and Multi-PIE databases contain pose and illumination variations, while MOBIO and SCface contain images captured with different sensors. SCface also contains variations in the resolution of the captured images.

When performing evaluations on each database we use the well defined protocols that provide dedicated *train*, *dev* and *eval* sets. In each case approximately one third of the data is used for each set. The *train*, *dev* and *eval* datasets are used in the same manner as outlined in Section 4.1. For all three databases we use manually annotated eye locations and examples images are provided in Figures 3, 4 and 5 for the MOBIO, Multi-PIE and SCface databases respectively. More details on the protocols for the MOBIO and SCface databases are given in [18], and for the Multi-PIE database in [3].

System performance is presented in terms of equal error rate (EER) and half total error rate (HTER) [11]. EER is used for the development set and is the point at which the false alarm rate equals the false rejection rate (a smaller number is better). The threshold, τ, derived from the EER on the development set is then used on the evaluation set to obtain the HTER (the average of the false alarm rate and false rejection rate) to present the final system performance (a smaller number is better). Linear scoring and ZT-Normalisation are used for all evaluated systems, as it has previously been shown to be effective for face recognition [19].

4.3. Impact of Face Localisation Error

An issue for any real world face verification system is it's robustness to face mis-alignment; that is, the performance degradation when the face image is not extracted perfectly (based on the eye positions). Therefore, we evaluate the robustness of our proposed approach to errors in mis-

System	Protocol 1a		Protocol 1b	
	Dev	Eval	Dev	Eval
PCA+PLDA	23.8	23.8	16.4	17.9
RBF-SVM (HoG)	31.8	31.4	24.2	25.5
GMM-FP	29.5	32.6	25.2	28.0
Local GMM-FP	37.4	43.0	34.6	40.2
ISV	34.9	37.8	30.9	33.5
Local ISV	**43.1**	**49.3**	**40.8**	**46.7**

Table 1: Fish Identification Results. Rank-1 identification rate results are given, and the best performing system is shown in **bold**.

alignment by introducing noise into the manually annotated landmarks. We choose the MOBIO database for this evaluation, and add uniform random noise equal to 2%, 5%, 10% and 20% of the average inter-eye distance (119 pixels for the MOBIO database). The new landmark points which have been used in this experiment are publicly available [5].

5. Experiments

The proposed techniques have been implemented using the the freely available signal processing and machine learning tool box, BOB [1].

5.1. Evaluation on Fish Image Set

The images are cropped with an extra margin of 3% added to the ground truth bounding boxes. Images are then converted to gray-scale and resized to 160×64 pixels. DCT features are extracted exhaustively using a block size of 20×20 with $M = 65$. Mean and standard deviation is applied to each block, as such the zeroth DCT coefficient is discarded. GMM based approaches use 512 components, for the sub-space size is set to 64 for Protocol 1a and 32 for Protocol 1b. For the local approaches the optimal region size was found to be 4×4.

The fish image dataset is a new dataset and so in addition to the proposed approaches we also present several baseline systems. The baseline systems used in this work are probabilistic linear discriminant analysis (PLDA) which achieves state-of-the-art performance for face recognition [8], and a support vector machine (SVM) approach similar to that used for classifying pedestrians [2]. For both the PLDA and SVM approaches we used the gray-scale images which have been resized to 160×64 pixels. For PLDA we apply dimensionality using principal component analysis (PCA) as this showed improved performance, this is termed PCA+PLDA. For the SVM approach we use a histogram of oriented gradients as the feature and a radial basis function as this provides superior performance over a linear SVM, referred to as RBF-SVM.

[5] visit https://wiki.qut.edu.au/display/saivt/Noisy+MOBIO+Landmarks for details

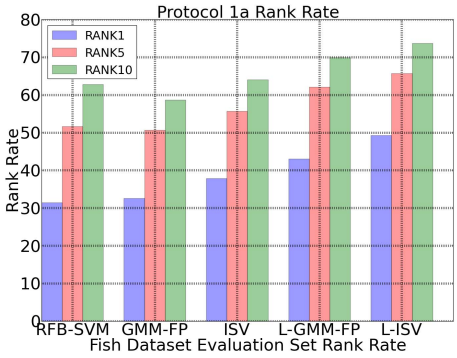


Figure 6: Rank-1, Rank-5 and Rank-10 identification rates for Protocol 1a on the evaluation set.

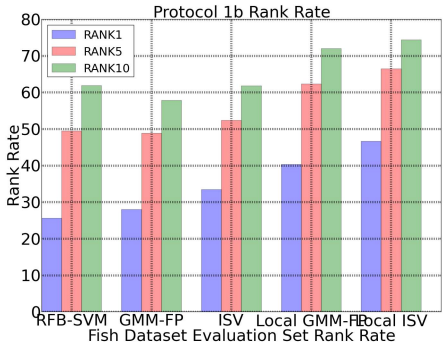


Figure 7: Rank-1, Rank-5 and Rank-10 identification rates for Protocol 1b on the evaluation set.

Results presented in Table 1 show that the Local ISV approach outperforms all other approaches. The standard ISV approach clearly outperforms the RBF-SVM and GMM-FP approaches, and the Local ISV approach provides a relative performance gain of 35% when compared to ISV. The next best system is the Local GMM-FP approach which provides a relative performance gain of 38% when compared to GMM-FP. The Rank-5 and Rank-10 identification results, in Figures 6 and 7, show that Local ISV and Local GMM-FP provide consistently improved performance.

A general trend for all of the classifiers is that Protocol 1a provides better performance than Protocol 1b. The average relative performance difference for all classifiers between Protocol 1a and Protocol 1b is 13%. This is likely due to the fact that for Protocol 1a the enrollment data consists of a "controlled" image, compared to Protocol 1b which uses an "in-situ" image. This demonstrates the importance of having high quality enrollment data with which to generate a model, even when session variability modelling is used.

5.2. Evaluation on Face Verification Databases

When extracting the DCT features we use a block size of 12×12 with $M = 44$ for the MOBIO and Multi-PIE databases. For the SCface database, we used a block size of 20×20 with $M = 65$. These optimal block and feature sizes were taken from [19].

We evaluated the proposed local face verification approach on three databases as outlined in Section 4.2. Our proposed technique is compared to three baseline techniques: MRH, GMM-FP and ISV. In this experiment UBMs are trained with 512 components for MOBIO and Multi-PIE and 256 components for SCface. In the ISV and Local ISV approaches a sub-space of 40 components is used for MOBIO and SCface, and 80 components is used for Multi-PIE. For the Local GMM-FP approach we use a region size of 4×4 for MOBIO and Multi-PIE, and 1×2 for SCface. For the Local ISV approach, we use region sizes of 4×4 for MOBIO, 2×2 for Multi-PIE and 2×2 for SCface.

Table 2 shows the performance of the proposed approaches and the baselines. It was found that the Local ISV approach performs best in all cases except for the SCface evaluation dataset, which obtains best performance using the ISV system. The Local ISV modelling technique marginally improves the verification performance in the *dev* set and marginally decreases the performance in the *eval*. This marginal performance degradation is likely due to the large block size used (20×20) in conjunction with many images in the SCface database being up-sampled to have an inter-eye distance of 33 pixels. The Local ISV system provides an average relative performance improvement of 32% for the MOBIO and Multi-PIE databases. We also note that the Local GMM-FP system consistently outperforms the GMM-FP system on all three databases, with an average relative improvement of 18%, further demonstrating the value of a region based approach. The Local ISV approach outperforms the Local GMM-FP system on all three databases, and demonstrates the value in modelling session variability and capturing identity information locally.

5.3. Evaluation of Face Verification Performance in the Presence of Localisation Error

The performance of face verification in the presence of localisation noise is evaluated as outlined in Section 4.3. Figures 8 and 9 show the half total error rate (HTER) of the Local GMM-FP and Local ISV face verification systems and their respective baselines (GMM-FP and ISV) in the presence of increasing levels of face localisation noise on the MOBIO database. The same systems configurations as those in Section 5.2 are used. We evaluate performance at five different noise levels: no noise; and with localisation error of up to 2%, 5%, 10% and 20% of the average inter-eye distance.

For both the proposed and baseline systems, system performance degrades as noise increases. At levels of noise up to 20% of the average inter-eye distance the proposed approaches outperform their baselines. However, as noise is increased above 10%, the proposed performance of all systems degrades considerably (see Figure 8).

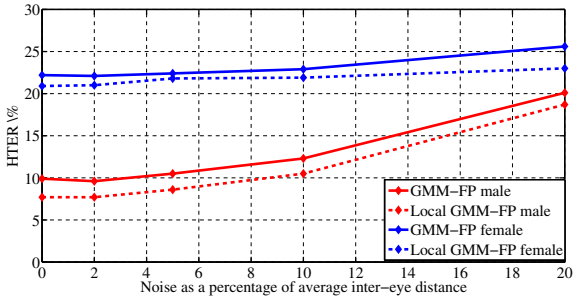


Figure 8: Performance of the Local GMM-FP and GMM-FP face verification systems in the presence of face localisation noise on MOBIO database evaluation set.

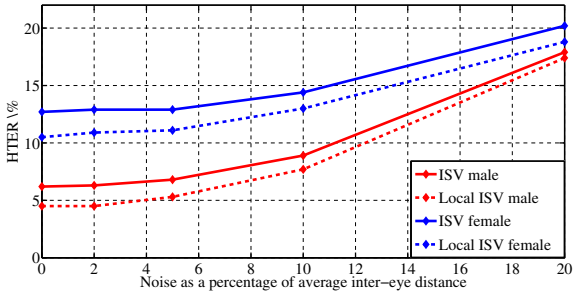


Figure 9: Performance of the Local ISV and ISV face verification systems in the presence of face localisation noise on MOBIO database evaluation set.

This increased degradation is likely caused by the nature of the region based systems. At high levels of noise and with small region sizes, the locations of the regions relative to the face changes significantly. Thus the assumption that corresponding regions between the client model and probe image are modelling the same portion of the face is increasingly likely to be violated as noise increases. However this effect could be mitigated by using fewer regions (i.e. 2×2 rather than 4×4), which would incur a small drop in performance under ideal conditions, but offer greater invariance to localisation errors.

6. Conclusions and Future Work

This works shows that state-of-the-art performance can be obtained for fish and face image classification through a region based, Local ISV modelling technique. This approach allows noise (in the form of session variation) to be modelled locally, while also capturing local identity information. For the first time, we have applied the ISV model

System	MOBIO (female)		MOBIO (male)		SCface		Multi-PIE	
	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval
MRH [12]	14.5	21.9	13.6	13.0	28.3	30.3	4.8	6.2
GMM-FP	11.5	22.2	7.5	9.9	16.7	16.3	3.1	3.8
Local GMM-FP	10.3	20.9	4.8	7.7	15.7	15.9	1.1	2.3
ISV	6.7	12.7	4.1	6.2	13.6	**12.8**	1.6	2.2
Local ISV	**5.2**	**10.5**	**2.5**	**4.5**	**12.0**	13.4	**0.6**	**1.1**

Table 2: Face Verification Results. The MRH results are taken from [12]. Results for the *Dev* data are equal error rates, while results for the *Eval* data are half total error rates. The best performing systems are shown in **bold**.

to challenging natural world images of fish to examine the broad applicability of this technique to the more general object classification domain, and have shown that the Local ISV approach outperforms the standard ISV by 35%. In the face verification task, the Local ISV technique outperforms the standard ISV technique by an average of 32% for the MOBIO database and Multi-PIE unmatched illumination data set. We have shown that the Local GMM-FP system also consistently outperforms the GMM-FP system on all three face databases with an average relative improvement of 18%, further demonstrating the value of a region based approach.

In addition to this, we have evaluated the real-world applicability of the Local ISV approach to face verification in the presence of face localisation error. It has been shown that Local ISV outperforms baseline systems at noise levels of up to 20% of the average inter-eye distance. Future work will consider the selection of weights for combining the region based models, and will investigate approaches to incorporate features such as colour into the models, which may be of particular use for classification of natural images.

Acknowledgements

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. This research was supported in part by a grant form Cooperative Research Centre for Smart Services - (CRC-SS) and by an Australian Research Council (ARC) Discovery grant DP110100827.

References

[1] A. Anjos, L. E. Shafey, R. Wallace, et al. Bob: a free signal processing and machine learning toolbox for researchers. In *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*. ACM Press, Oct. 2012.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, volume 1, pages 886–893 vol. 1, 2005.

[3] L. El-Shafey, C. McCool, and S. Marcel. A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[4] O. Glembek, L. Burget, N. Dehak, et al. Comparison of scoring methods used in speaker recognition with joint factor analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.*, pages 4057–4060, 2009.

[5] M. Grgic, K. Delac, and S. Grgic. Scface — surveillance cameras face database. *Multimedia Tools Appl.*, 51(3):863–879, Feb. 2011.

[6] R. Gross, I. Matthews, J. Cohn, et al. Multi-pie. *Image Vision Comput.*, 28(5):807–813, May 2010.

[7] G. B. Huang, M. Mattar, H. Lee, et al. Learning to align from scratch. In *Neural Information Processing Systems*, 2012.

[8] P. Li, Y. Fu, U. Mohammed, et al. Probabilistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):144–157, 2012.

[9] S. Lucey and T. Chen. A GMM parts based face representation for improved verification through relevance adaptation. In *Proc. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–855–II–861 Vol.2, 2004.

[10] C. McCool, V. Chandran, S. Sridharan, et al. 3D Face Verification using a Free-Parts Approach. *Pattern Recognition Letters*, 29:1190–1196, 2008.

[11] C. McCool, S. Marcel, A. Hadid, et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, 2012.

[12] C. McCool, R. Wallace, M. McLaren, et al. Session variability modelling for face authentication. *IET Biometrics*, 2:117–129(12), September 2013.

[13] K. Messer, J. Kittler, M. Sadeghi, et al. Face authentication test on the banca database. In *International Conference on Pattern Recognition*, pages 523–532, 2004.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:2000, 2000.

[15] C. Sanderson and B. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 199–208. Springer Berlin Heidelberg, 2009.

[16] C. Sanderson and K. K. Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, 24(14):2409 – 2419, 2003.

[17] R. Vogt and S. Sridharan. Explicit modelling of session variability for speaker verification. *Computer, Speech and Language*, 22:17–38, 2008.

[18] R. Wallace, M. McLaren, C. McCool, et al. Inter-session variability modelling and joint factor analysis for face authentication. In *International Joint Conference on Biometrics*, 2011.

[19] R. Wallace, M. McLaren, C. McCool, et al. Cross-pollination of normalisation techniques from speaker to face authentication using Gaussian mixture models. *IEEE Transactions on Information Forensics and Security*, 2012.