Shared Gaussian Process Latent Variable Models for Handling Ambiguous Facial Expressions

Carl Henrik Ek * **Peter Jaeckel** ** **Neill Campbell** ***
Neil D. Lawrence **** **Chris Melhuish** †

** Oxford Brookes University, Oxford, OX3 0BP, UK (e-mail: cek@brookes.ac.uk)*
*** Bristol Robotics Laboratory, University of the West of England, Bristol, BS16 1QY, UK (e-mail: peter.jaeckel@brl.ac.uk)*
**** University of Bristol, Merchant Venturers Building, Woodland Road, Bristol, BS8 1UB, UK (e-mail: neill.campbell@bris.ac.uk)*
***** University of Manchester, School of Computer Science, Kilburn Building, Manchester M13 9PL, UK (e-mail: neill.@cs.man.ac.uk)*
† Bristol Robotics Laboratory, Bristol BS16 1QY, UK (e-mail: chris.melhuish.@brl.ac.uk)

Abstract: Despite the fact, that, in reality facial expressions occur as a result of muscle actions, facial expression models assume an inverse functional relationship, which makes muscles action be the result of facial expressions. Clearly, facial expression should be expressed as a function of muscle action, the other way around as previously suggested. Furthermore, a human facial expression space and the robots actuator space have common features. However, there are also features that the one or the other does not have. This suggests modelling shared and non-shared feature variance separately. To this end we propose Shared Gaussian Process Latent Variable Models (Shared GP-LVM) for models of facial expressions, which assume shared and private features between an input and output space. In this work, we are focusing on the detection of ambiguities within data sets of facial behaviour. We suggest ways of modelling and mapping of facial motion from a representation of human facial expressions to a robot's actuator space. We aim to compensate for ambiguities caused by interference of global with local head motion and the constrained nature of Active Appearance Models, used for tracking.

Keywords: Robot, Face, Ambiguity, Modelling, Gaussian Process

1. INTRODUCTION

Humans can mimic facial expressions they observe effortlessly, may it be unconsciously or deliberately. Facial mimicry is an essential mechanism often found during social and face-to-face interaction. Newborns and infants use mimicry as part of exploring their muscle spaces (motor babbling) for the acquisition of motor skills and culture specific behaviour patterns Meltzoff and Moore [1997]. Mapping observed facial expression into an intermodal presentation is an essential part of understanding one-another's emotional states and intentions Baron-Cohen [1991]. It also plays an important role in shared attention and engaging others. Clearly, there is a desire and motivation for socially and emotionally intelligent human-like robots to have similar ways of learning and processing perceived facial behaviour. An affective or empathetic agent must be capable of recognising, as well as mirroring emotional content Burleson et al. [2004].

Facial expressions are generated by muscles (or motors in robots) and act upon points or areas beneath the skin. Facial expressions are essentially skin deformation as a result of muscle contraction. Therefore, mapping facial expressions to a representation of muscle commands should use a model which is based on an explicit mapping from muscle to facial expressions, not the other way around. Previous work, however, proposes models which map from facial expressions to a robots muscle or servo space using explicit functional relationship such as a Gaussian Process Regression Jaeckel et al. [2008a], or Linear Partial Least Squares Jaeckel et al. [2008b]. Others have learned functional relationships, such as feed-forward neural networks to map individual areas of the face to a robot's facial motor coordinates Breazeal et al. [2005].

Facial expressions and facial motion contain the basic, fundamental parts, called action units Ekman et al. [2002]. Additionally, person specific characteristics are superimposed. These characteristics are due to individual face and muscle shapes. Similarly, one can consider robot specific characteristic as a result of the physical properties and constraints of a robotic face. It may or may not lack certain muscular features that are present in human faces. On the other hand, robotic faces may possess capabilities that are improbable or do not exist in ordinary human faces.

Hence, we need to acknowledge that along the shared facial features there are features that belong exclusively to a person or robot specific facial expression or muscle space. This suggests modelling of both observation spaces according to their shared and specific properties.

2. BACKGROUND AND MOTIVATION

There have been attempts to map and synthesise human behaviour and movements: The work reported in Littlewort et al. [2006] is based on recognition of stereotypical facial expressions of emotion. Input images are projected into a seven-dimensional emotion space, where each emotion is one of the six basic ones plus neutral. In a second step, the recognised emotional state (location in 7D emotion space) is fed into a character animation engine which produces facial expressions of emotion based on the 7D emotion code. Hence, in terms of mirroring capabilities, this system is limited to stereotypical and exaggerated facial expressions of emotion.

In previous work we have proposed a number of linear Jaeckel et al. [2007] and non-linear, probablistic Jaeckel et al. [2008a] approaches for modelling and mapping between an observed human's facial expressions and a humanoid robot's motor space. Active Appearance Models (AAM) Cootes et al. [1998], Stegmann [2003] have been used for obtaining a compact and low dimensional representation of facial expressions. AAMs are parametric, statistical generative models of facial appearance (texture) contained by a triangular mesh (shape). They are built by learning the combined texture and shape variation. Training data has been generated by hand-labelling (defining facial landmarks) training footage to best cover the range of facial expression and a small amount of global head motion. Minimising the error between a model instance and an input image, results in a model shape which represents the facial expression.

We have suggested mappings using Partial Least Squares Jaeckel et al. [2008b], modelling facial expressions and their relationship between a facial landmark space and a corresponding robot actuator space. The input consists of 25 2D-facial landmarks tracked in a video sequence showing a human subjects face. The landmark locations were scaled and normalised and projected into a low dimensional latent space, which was formed according to the correction between input and output space. A set of regression weight matrices is then responsible for generating the corresponding output pose.

The linear mapping worked well for mapping from footage that contained a small range of facial expressions. However, due to the model extrapolating linearly beyond the model borders, it cannot handle certain, previously unseen inputs very well. This occurs when the input faces go beyond model range. The linear model requires cross-validation and subsequent manual modification. The latter is essentially omitting latent variables. This is necessary because for a increasing number of latent variables PLS converges to multiple linear regression (MLR). It is well known that MLR cannot cope with multi-colinearities which causes large prediction errors.

GP-Regression Jaeckel et al. [2008a] provided a way to create models which do not need subsequent modification, handle noisy observation data and the aforementioned extrapolation-related problems. Furthermore, the non-linear functional relationships have been employed which extend the range of facial motion and enabled the inclusion of a small range of head motion.

3. DISCUSSION

The application of Active Appearance Models has been very useful for obtaining a compact representation of facial expressions. However, a lot of information is omitted by only considering $2D$ facial landmark locations, hence some facial motion may not be detected. The constrained nature of AAMs prevents facial landmarks from drifting and provides an efficient way of robust tracking. However, the constrained nature also causes undesired correlations between facial landmarks. This can have side-effects, such that global head motion can sometimes affect local facial features. Previously investigated GP-Regression cannot correctly map some facial features such as eyebrows and smiles when the head orientation differs greatly from neutral. Problems occurred when the input face was pointing down or up as well as left and right. Due to effects of projection and perspective, the movement can hardly be observed since it occurs orthogonally to the viewing and tracking plane. This suggests that model has to have some knowledge about the sometimes undetectable but existing facial motion.

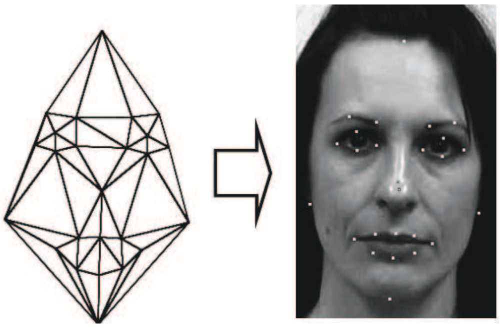


Fig. 1. The figure shows Active Appearance Model (AAM) shape (left) which, when fitted to an input image, gives 25 landmarks in $2D$ (right) which represent the input expression which is scaled and normalised prior to be fed into the model.

Facial expressions are essentially skin deformation as a result of muscle contraction. Therefore, mapping facial expressions to a representation of muscle commands should use a model which is based on a mapping from muscle to facial expressions, not the other way around, like in previous work Jaeckel et al. [2007, 2008b,a]. To overcome the problem of undetected facial motion and coupled facial landmarks, we suggest the application of Shared Gaussian Process Latent Variable Models (SGP-LVM). SGP-LVMs assume that observations are generated via a mapping from a low dimensional, latent manifold. Hence the manifold can be seen as a low dimensional, generic muscle representation, from which facial expressions and robot pose are generated.

The final point is that the representation of human facial expressions and robot pose do have shared and shared features. This suggests an appropriate low dimensional latent representation divided in to subspaces according to shared on non-shared variance.

In this work, we are focusing on the detection of ambiguities within training sets. We suggest ways of modelling and mapping of facial motion from a representation of human facial expressions to a robot's actuator space. We aim to compensate for ambiguities caused by global head motion and the constrained nature of Active Appearance Models used for tracking.

4. DATA ACQUISITION

Our input data consisted of 25 tracked feature points coordinates X (landmarks) in 2D. The feature point locations were

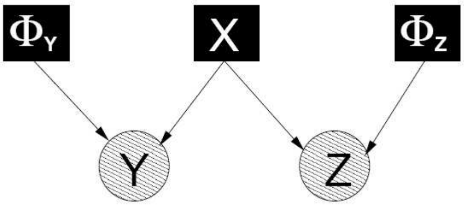


Fig. 2. Shared GP-LVM model. The shaded nodes indicates the observed variables while the black nodes shows variables found through Maximum Likelihood. Φ_Y and Φ_Z hold the parameters for the mappings from latent space $\mathbf{X}$ to the observation spaces $\mathbf{Y}$ and $\mathbf{Z}$, respectively.

obtained from fitting person specific Active Appearance Models to input video frames (320×240 pixels). AAMs are parameterised models which contain information on shape and texture. Fitting an AAM into an image is essentially finding the model parameters which minimise the error between input image and the instance of the active appearance model Cootes et al. [1998]. The search for an optimum fit is initialised with the model mean and the size of the face and its coordinates supplied by a face detector Viola and Jones [2001]. Initialisation is repeated each time the error threshold between input image and AAM instance is exceeded. Figure 1 shows the AAM shape (left) which is then fitted to frames which were supplied by video or camera, yielding 25 landmarks (right). Subsequently, the landmarks are centred and normalised.

5. THE MODEL

A Gaussian Process (GP) is the infinite dimensional generalisation of the Gaussian Distribution. Being defined over infinite domains GPs can be used to model functions and functional relationships. Lawrence Lawrence [2005] presented a generative dimensionality reduction technique based on GPs called The Gaussian Process Latent Variable Model (GP-LVM). The GP-LVM models the observed data as generated from a low-dimensional latent variable where a GP specifies the mapping to the observation. Minimising the GP-LVM objective is a non-convex optimisation and runs the risk of getting stuck in a local minima. As a solution, Lawrence Lawrence [2005] suggested the optimisation to be initialised from a analogous convex model such as spectral dimensionality reduction methods Tenenbaum et al. [2000], Belkin and Niyogi [2002], Weinberger et al. [2004].

In this paper we are interested in modelling both image features and robot actuator within the same model. Shon Shon et al. [2006] extended the GP-LVM to allow multiple observation spaces to be modeled using a single latent variable. The variable structure of a shared GP-LVM with two observation spaces is show in Figure 2. However, the proposed model assumes a one-to-one mapping between the observations which cannot be assumed for our facial behaviour data. The displacement of facial landmark cannot always be observed due to head orientation and effects of projection from $3D$-objects to a $2D$-plane where the tracking occurs. This, primarily, is the case for the eyebrow positions in the input (landmark) space. For some head orientations one eyebrow pose in the landmark space can cause several possible eyebrow poses in the robot. A latent variable model separately modelling variance shared between the observations from the non-shared is the NCCA model Ek et al. [2008]. Its graphical model is shown in Figure 3. The NCCA models uses a latent space divided into three orthogonal subspaces, one representing the shared variance and one space per observation space representing the non-shared or private variance. In Ek et al. [2008] the model is used for a multi-modal regression task where the private space represents variance in corresponding observation space that is ambiguous to the other.

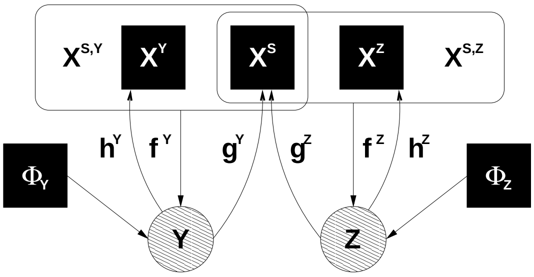


Fig. 3. The NCCA model. The central subspace $\mathbf{X}^S$ models variance that is shared between the observation spaces $\mathbf{Y}$ and $\mathbf{Z}$. The remaining subspaces $\mathbf{X}^Y$ and $\mathbf{X}^Z$ model variance that is private to its corresponding subspace. Φ_Y and Φ_Z hold the parameters for the mappings f^Y and f^Z from latent space $\mathbf{X}$ to the observation spaces $\mathbf{Y}$ and $\mathbf{Z}$, respectively. Note that additional mappings h^Y, g^Y and h^Z, g^Z may be used to provide back-constraints Lawrence and Quinonero-Candela [2006].

Fig. 4. The employed robot head 'Jules' has 34 servo motors emulating the majority of the human facial muscles in face and neck.

6. THE ROBOTIC FACE

For the experiments the robot head called 'Jules' (supplied by David Hanson [1]) shown in Figure 4, has been employed. It has 34 servo-actuators motors which emulate the majority of the muscle groups of the human face and neck. The actuators pull and/or push control points attached underneath the skin. It can move upper and lower eye lids and the eyes move vertically and horizontally. There are three motors manipulating each eyebrow. 'Jules' can nod, turn and tilt its head and is able to perform frowns and has a rich repertoire of sneers and smiles. The servo motors are controlled by servo controllers that receive desired positions for each servo motor via an RS232 interface at up to 30fps (frames per second).

7. TRAINING SETS

Initially, a GP-Model had been trained using around 65 training samples to best cover the range of facial expressions. It also

---

[1] web site: www.hansonrobotics.com

contains global head motion in the input space, but not in the output space. That means that all the global head motion is compensated and will not be mapped through to the output. The next step was to map a set of test sequences using this Gaussian Process Regression mapping Jaeckel et al. [2008a]. This gave a set of preliminary results which motivate further research: GP-Regression applied to a stream of test inputs gave a sequence which contains pairs of face-input and corresponding robot-pose-output. Subsets of this sequence serve as training sets for the Shared GP-LVM. There are two training sets containing one sequence each:

Training Set 1: The first set contains 425 frames of global head motion (up, down, left, right), while showing a neutral face. Additionally it contains smiles and jaw movement whilst the head is in neutral position. Subjective evaluation and preliminary results show that the GP-Regression Model cannot correctly map the eyebrows during head up or down. The output space contains lowered eyebrows when the input face is pointing down.

Training Set 2: This set shows 320 frames of combinations of global head motion, up, down, with eyebrows up and down. Unlike in the first sequence there are frames that contain simultaneous head and 'eyebrows-down' movements (see Table 1 for more detail). These simultaneous movements cannot be achieved by GP Regression but are given by manually editing the training data to give correct input - output pairs.

8. RESULTS

Modelling of Ambiguities: The first experiment shows that a Shared GP-LVM can be used to model ambiguities within data sets. To train the model we used Training Set 1 which contains faulty mapping of eyebrow motion, caused by head rotation. Results are shown for two frames, A and B.

The input face, in Figure 5 (a) in frame A of the test sequence shows 'head up' and a slight turn to the left and a neutral facial expression. Inspecting the resulting private latent space XZ in Figure 5 (b) shows two modes (bright patches). Brighter areas are those latent locations which the model assumes to be more likely to have generated the robot pose. The difference in servo outputs between the two modes is illustrated in Figure 5 (c).

The servo differences indicate an activation of servos '5 smile', '6 LipUpperCtr', '7 Brow Outer', '24 Brow Outer R' and negative '7 Brow Scowl L' and '25 Brow Scowl R'. The weak mode, the left patch in the latent space, tends to give a neutral robot pose, whereas the mode on the right, yields raised eyebrows and activation of smile and upper lip muscles. Hence the global head movement create a false facial expression. However, the weak mode correctly predicts a neutral facial pose.

Frame B of the test sequence shows a nod and the input face (top right) is pointing slightly downwards. The facial expressions are neutral. However, there are two modes (see mid right portion), the first one (top) gives a neutral facial expression, with mouth closed and a neutral eyebrow position. The second mode however, shows a strong activation of inner and outer eyebrows. Again, the mapping predicts an activation of facial features, due to their perspective changing with the head orientation. A second mode however, correctly suggests a neutral pose.

Model	Frame	Head Pose	Eyebrow Pose	Ambiguous Features
I	A	up/left	normal	EB, SM, L,
	B	down	normal	EB
II	C	down	down	EB, LL, BC
	D	up	normal	EB
	E	up	down	EB
	F	normal	up	- none -

EB = Inner/Outer Eyebrows (7,8,24,25))
SM = Smile Muscles(5,16), Lips (6,11)
LL = Lower Lids (22)
BC = Brow Center (26)
L = Lips (6,11)

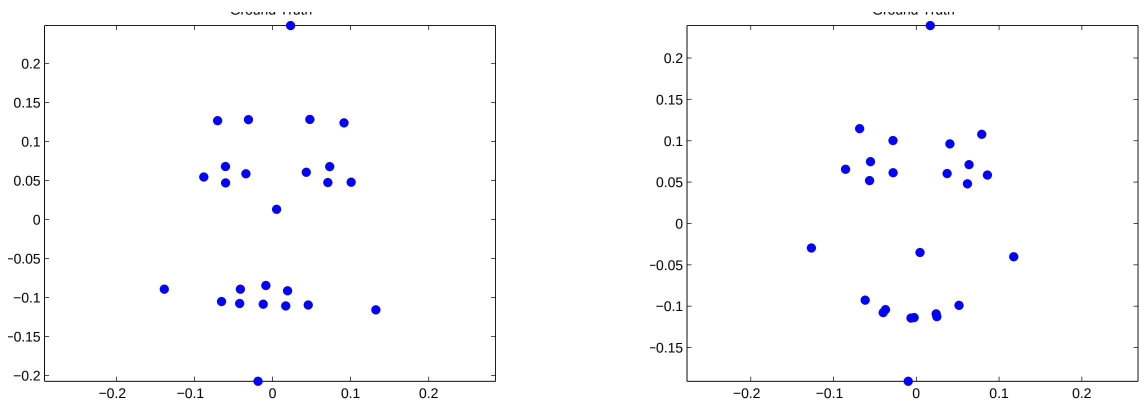
Table 1. Overview of head- and eyebrow poses and resulting ambiguous robot features in two different models. Model I is trained from regression data and is used for the detection of ambiguities within the data set. Model II has been primarily trained using combinations of eyebrow and head poses.

Disambiguating Eyebrow Pose: Training Set 2 has been used for this experiment: The sequence contains head poses (nod, up/down) accompanied by either neutral facial expression or by inner and outer eyebrows movements. Unlike in the previous experiment, the training set now does contain all possible combinations of head up/down and eyebrow poses. The training set obtained from GP-Regression has subsequently been edited to give a correct mapping. In normal regression the model cannot correctly map eyebrow movement when the head is pointing down. Eyebrow movement tend to be coupled to head movement - which is not desirable. To test the model, four selected test frames (Frames C,D,E,F) have been selected from the training set to suit a number of combinations of head and eyebrow poses (see Table 1). For each frame C - F the input face (see Figure 7) results in a corresponding likelihood map in the robot pose specific latent space shown in Figure 6. The servo differences for those frames that display two modes essentially show similar results to the one shown in Figure 5 (f) where there are large changes in servo values that manipulate inner and outer eyebrows.

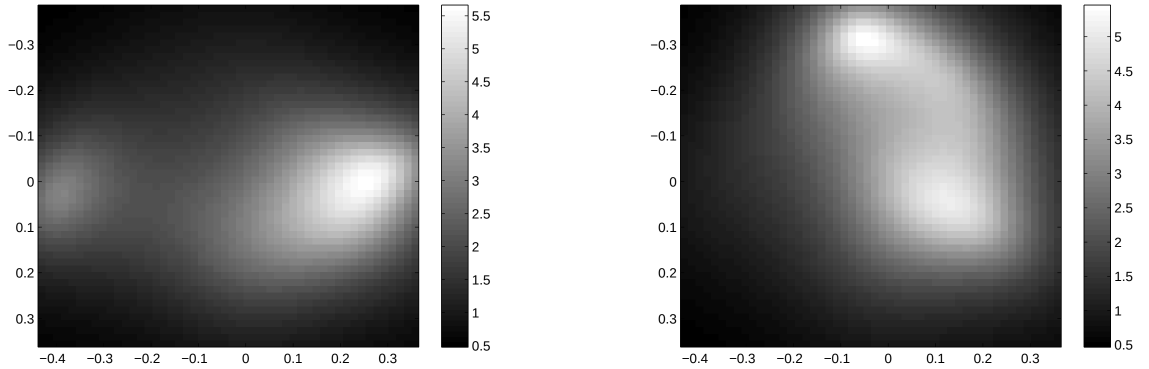
9. SUMMARY AND CONCLUSIONS

The work presented here has introduced Shared Gaussian Process Latent Variable Models for the detection and modelling of ambiguities in data sets of human and robotic facial behaviour. We have extracted facial motion from video by fitting Active Appearance Models into video frames. The shape of the fitted AAM gives a sequence of facial expressions in form of a set of 25 2D facial feature landmarks, which we suggest as representation for facial expressions.

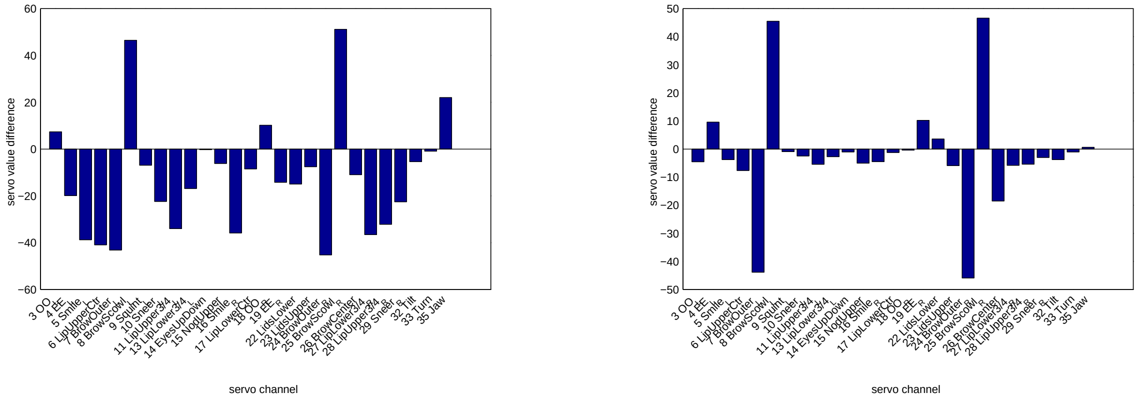
We used a Gaussian Process mapping from the facial landmarks location space to an robot actuator space to obtain the training data. The training set consists of a sequence of input faces and corresponding robot poses obtained from a previously investigated Gaussian Process (GP) Regression. A regression is not able to correctly map local facial features as global head motion tends to interfere. Hence the training data had to be created by manually correcting given sequences used for training in order to correctly map and model eyebrow motion, regardless of head motion. The second training set contains multiple input-output combinations for the same input face.



a) Input Face A (head up/left, eyebrows neutral) d) Input Face B (head down, eyebrows neutral)

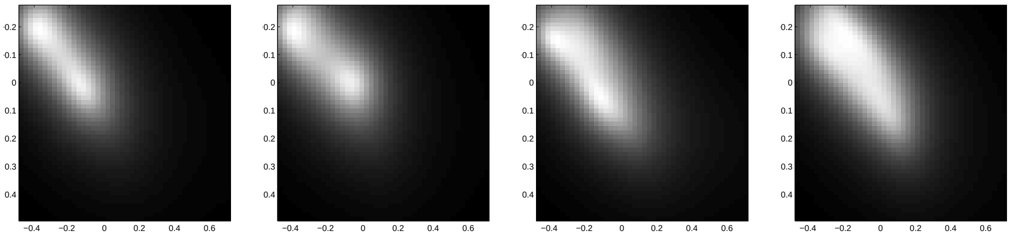


b) Two Modes - Frame A (bright patch left/right) e) Two Modes - Frame B (bright patch top/bottom)



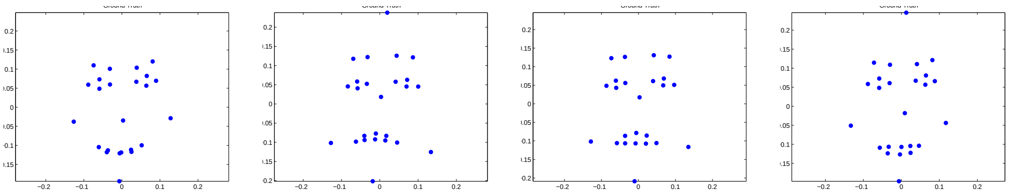
c) Servo Positions Differences between Modes - Frame A f) Servo Positions Differences between Modes - Frame B

Fig. 5. Frames A (left column) and B (right column) - Shows two frames where the input faces (top row) causes multiple modes in the private, servo space specific latent space. (mid row). The differences between two modes is expressed in terms servo output values. (bottom row)



Frame C Frame D Frame E Frame F

Fig. 6. Private, Robot Pose Specific Latent Spaces - Frame C, D and E have two modes, which indicates ambigiuties. Whereas Frame F has one mode only, hence the model output is certain.



Frame C Frame D Frame E Frame F

Fig. 7. Input Faces - Frame C, D and E involve head down, up and up, respectively. Frame F has neutral head position. The eyebrows are lowered in Frame C and E. Frame D shows a neutral eyebrow pose. Frame F shows raised Eyebrows.

The results suggest that Shared GP-LVMs can be used to model and detect ambiguities in data sets representing facial expressions and multiple solutions. Unlike in regression techniques, where this is not possible, because for each input exists only a single output. Ambiguities exist for smile muscles and eyebrows. This means that there are multiple robot poses which are most likely to be the solution to a facial expression in the facial expression space. Hence part of the features' variation are described by the robot pose specific subspace. This is because some of the variation only exists in the pose specific observation space. Conversely, jaw pose, for example is always confidently predicted, it is encoded in the shared latent subspace. Notably, the alternative neutral poses suggested by the model are not present in the training set. For example, the training set only contains inputs that have a 'head-down' pose and neutral eyebrows, but none which show a 'head-down' and 'eyebrows down' pose as suggested by the model. The corresponding robot output pose always shows lowered eyebrows associated with an input of 'head-down'. The model however, suggests a alternative solution which is a neutral eyebrow pose.

We have shown that ambiguities can occur due to projection of tracked facial landmarks onto a $2D$ plane. This leads to the inability to infer certain parts of robot pose from a set of facial landmarks with absolute certainty, since more than one solution are given. In particular this applies to the mapping of eyebrows and mouth corners under the influence of head motion. Head 'up' or 'down' causes ambiguities in eyebrow-positions. In general this means that for some head positions, multiple likely robot poses are suggested by the model. Since the model had been trained with multiple robot-eyebrow poses for a certain input eyebrow-pose, it suggests two solutions if the head is pointing down. Both of the solutions, suggest eyebrows down or a neutral pose, respectively.

Finally Shared GP-LVMs which model input and output observation separately, according to shared and space-specific feature variation, can be used to detect ambiguities in facial expression data. There are multiple solutions for the same input face, which shows that regression is not adequate for modelling shared facial behaviour structures of both, human face to robotic faces.

Future research should focus on disambiguating model outputs which requires the challenge of the modelling of dynamics. This can help to infer the right solution from previous model states. Once the model generates continuous robot behaviour one should attempt to investigate the possibility of finding objective measures of the robot outputs. Another, more feasible evaluation strategy is to survey human subjects observing the robot in psychological experiments.

REFERENCES

S. Baron-Cohen. *Precursors to a Theory of Mind:Understanding Attention in Others*. Blackwell Press, 1991.

M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 14:585–591, 2002.

C.L. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg. Learning From and About Others: Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots. *Artificial Life*, 11(1-2):31–62, 2005.

W. Burleson, R. Picard, K. Perlin, and J. Lippincott. A Platform for Affective Agent Research. In *Workshop on Empathetic Agents, International Conference on Autonomous Agents and Multiagent Systems, Columbia University, New York,NY*, 2004.

T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. In *Proceedings European Conference on Computer Vision 1998*, volume 2, pages 484–498, June 1998.

Carl Henrik Ek, Philip H.S. Torr, and Neil D. Lawrence. Ambiguity modeling in latent spaces. In *5th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2008)*, 2008.

P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System*. Research Nexus, Salt Lake City, Utah, 2002.

P. Jaeckel, N. Campbell, and C. Melhuish. Towards realistic facial behaviour in humanoids - mapping from video footage to a robot head. In *10th International Conference on Rehabilitation Robotics (ICORR 2007)*, pages 833–839, 2007.

Peter Jaeckel, Neill Campbell, and Chris Melhuish. Gaussian process regression for facial behaviour mapping. In *In Proceedings Taros 2008*, 2008a.

Peter Jaeckel, Neill Campbell, and Chris Melhuish. Facial behaviour mapping - from video footage to a robot head. *Robotics and Autonomous Systems*, 2008b.

N. D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, (6):1783–1816, 2005.

Neil D. Lawrence and Joaquin Quinonero-Candela. Local distance preservation in the gp-lvm through back constraints. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 513–520, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: http://doi.acm.org/10.1145/1143844.1143909.

G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. An Automatic System for Measuring Facial Expression in Video. *Image and Vision Computing*, 24(6):615–625, 2006.

N. Meltzoff and M. K. Moore. Explaining facial imitation: a theoretical model. *Early Development and Parenting*, pages 179–192, 1997.

A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and robotic imitation. *Advances in Neural Information Processing Systems*, 2006.

M. B. Stegmann. The AAM-API: An open source active appearance model implementation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003, 6th Int. Conference, Montréal, Canada*, LNCS 2879, pages 951–952. Springer, nov 2003. URL `http://www2.imm.dtu.dk/pubdb/p.php?2538`.

J.B. Tenenbaum, V. Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features, 2001. URL `citeseer.ist.psu.edu/viola01rapid.html`.

Kilian Q. Weinberger, Fei Sha, and Lawrence K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. volume 21, pages 839–846. Omnipress, 2004.

[!h]