Building Temporal Models for Gesture Recognition

Richard Bowden, Mansoor Sarhadi
Vision and VR Group
Dept of Systems Engineering, Brunel University
Uxbridge, Middlesex, UB8 3PH, UK
richard.bowden@brunel.ac.uk

Abstract

This work presents a piecewise linear approximation to non-linear Point Distribution Models for modelling the human hand. The work utilises the natural segmentation of shape space, inherent to the technique, to apply temporal constraints which can be used with CONDENSATION to support multiple hypotheses and quantum leaps through shape space. This paper presents a novel method by which the one-state transitions of the English Language are projected into shape space for tracking and model prediction using a HMM like approach.

1 Introduction

Previous work by the author and other researchers have investigated statistical models of deformation [1-8]. These deformable models have been used to learn *a priori* shape and deformation from a training set of examples which, represent the shape and deformation of an object or a class of objects. Models are typically constructed that know **what** is valid deformation but not **when** deformation is valid. This important temporal constraint is beneficial in disambiguating models.

A large body of work has been performed on the temporal mechanics of tracking. Many researchers have attempted to use predictive methods such as those based within a Kalman filter framework [1]. Hill *et al* proposed using genetic algorithms to model the discontinuous changes in shape space/model parameters [6][7]. Of particular interest to the work presented in this paper is the CONDENSATION algorithm [1][8] which is a method for stochastic tracking, where a population of model hypotheses are generated at each iteration. These populations are generated from pre-learnt Probability Density Functions (PDFs) generated over the model parameter space to provide a hypothosis-and-test approach to model prediction and tracking.

The key to this approach is an *a priori* model of motion from which populations are generated. Where motion is relatively uniform, such as the motion of an object within an image, the learning stage can be bootstrapped to the tracking process [8]. However, for the movement of the model within shape space (the deformation parameter space) this is not possible [4]. Instead, motion models must be pre-learnt in much the same way as deformable models; the temporal model merely augments that of deformation. Unfortunately, although a relatively small sample of training data can be used to

BMVC 2000 doi:10.5244/C.14.4

construct a model of deformation, considerably more examples are required to achieve an accurate representation of motion.

This paper addresses the problem of constructing a non-linear deformable model of the human hand for gesture recognition. It is demonstrated that the temporal model cannot be constructed from training data alone and a method which, allows temporal information about the English language to be projected down into shape space is presented. This generates a 1st order temporal model which, incorporates both information about shape space and the English Language.

Section 2 discusses the construction of a non-linear Constrained Shape Space Point Distribution Model (CSSPDM [4]) using a piecewise linear approximation. Section 3 demonstrates how the CSSPDM naturally lends itself to a CONDENSATION like approach to tracking. Section 4 presents a method by which the 1st order transition of the English Language are propagated into shape space. Finally the approaches are compared and conclusions drawn.

2 Constructing a CSSPDM for Gesture Recognition

3.1 American Sign Language

American Sign Language or ASL has a finger spelt alphabet similar to other national sign languages. These simple gesture alphabets are used to spell names or words (letter by letter), for which there is no sign either known or present in the vocabulary. ASL provides a more suitable problem domain over British Sign Language, as the BSL finger spelt alphabet is a two-handed system. This presents added difficulty for computer vision approaches due to the problems associated with occlusion. Figure 1 shows the ASL alphabet with the corresponding hand pose for each letter of the alphabet.

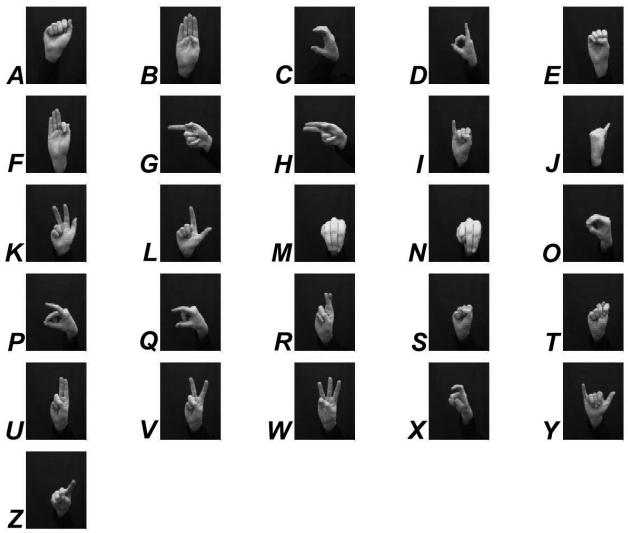

Figure 1 - The American Sign Language Finger Spelling Alphabet

3.1 The linear ASL PDM

Several image sequences were recorded which encapsulated numerous occurrences of each of the letters of the alphabet. These sequences included three 'runs' through the alphabet, along with a small selection of simple sentences and words. Once these sequences had been extracted, the hand was segmented to produce a binary image, and a contour-tracing algorithm initiated to extract the external contour of the hand for each image frame. After a standard alignment and resampling of the contour to 200 points (as described in [4]) a training set of 7441 examples was produced where each example $\mathbf{x}_i \in \Re^{400}$.

The Linear ASL model is generated by performing principal component analysis upon the training set [5]. Figure 2 shows the primary modes of the linear ASL PDM and how these modes deform the model from the mean shape.

By analysing the eigenvalues of the covariance matrix it can be determined that the first 30 eigenvectors corresponding to the 30 largest eigenvalues encompass 99.6% of the deformation within the model. However, due to the non-linearity of the model the linear PDM is insufficient for tracking as it encompasses too much deformation.

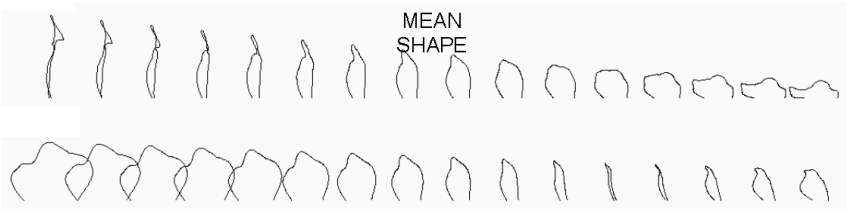

Figure 2 – First two primary modes of the ASL model

3.2 Applying non-linear Constraints to Shape Space

To further constrain the model the approach presented in [2][3][4] is applied. Non-linear constraints to the model are added by performing cluster analysis on the dimensionally reduced data set after it has been projected down into PCA space. From the linear model it has been determined that the 30 primary modes encompass 99.6% of the deformation, by projecting each of the training vectors down into this lower dimensional space, a dimensional reduction of 400 to 30 is achieved. Cluster analysis is now performed upon the dimensionally reduced data set.

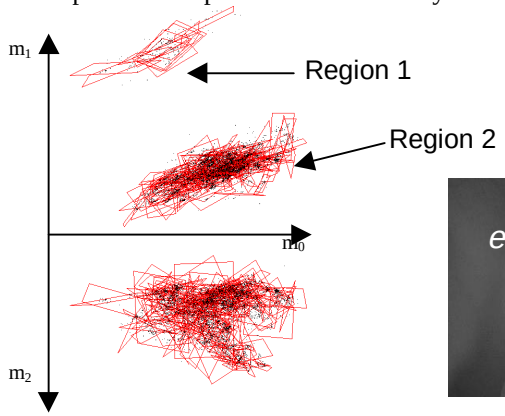

Figure 3 - Constrains on PCA space for the ASL Model

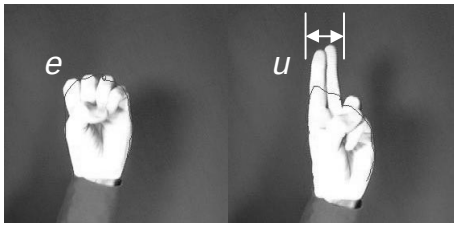

Figure 4 - ASL model Tracking an Image Sequence of the word 'gesture'

Figure 3 shows the PCA space for the model projected into 3 dimensions for visualisation purposes, with the constraints shown as the bounding boxes (first two primary modes) of the linear patches (clusters) extracted via PCA. By constraining the model to lie within a linear patch the non-linearity of the shape space is estimated and a robust model produced.

3 A hybrid PDF for CONDENSATION

3.1 Least Squares Gradient Descent Tracking

From figure 3 it can be seen that shape space is segregated into **at least** two separate regions due to the movement of landmark points. Furthermore, connected patches of the model may not represent consistent movement of the model in the image frame. This leads to the model *jumping* between patches, even when within region 2. Under these circumstances it is not possible for the iterative refinement algorithm used for the classic PDM/ASM [5] to provide the *'jump'* between regions.

An image sequence was recorded of a hand signing the word 'gesture' which consisted of 170 frames. Figure 4 shows the model attempting to track the image sequence for the letters 'e' and 'u'. The model successfully tracks the letter 'e' but when the image sequence reaches the letter 'u' and the fingers elongate, the model is unable to make the jump to the new cluster responsible for modeling this letter. This problem is fundamental to the operation of the least squares iterative refinement algorithm and is due to two reasons:

1. Only a small section of the contour (marked in frame 'u') is responsible for 'pulling' the contour up to follow the elongated fingers. As this section is relatively small, compared to the remainder of the contour, it has less influence over the overall movement.
2. The maximum movement of the contour per iteration is governed by the length of the normal used to search around the contour. Hence this factor limits the distance the model can move through shape space at each iteration.

An obvious solution to these problems is to increase the search length along normals. However, larger normal searches allow the contour to affix to incorrect features in the image and hence results in degradation and additional computational complexity.

3.2 Finding the Optimal Ground Truth for Tracking

To locate the optimum solution (i.e. the closest allowable shape from the Constrained Shape Space PDM, CSSPDM) for each iteration of the model, the space was exhaustively searched. If the assumption is made that any local patch of the CSSPDM can indeed be treated as a linear model, then the iterative refinement procedure can be used to move locally within that patch to the closest possible shape. Therefore, if the best match within each patch (cluster) is located for each frame, the resulting lowest cost solution must be the (near) optimum.

By analysing the optimum path through shape space and comparing this with the path taken by the least squares approach, the notion of discontinuity within shape shape can be confirmed.

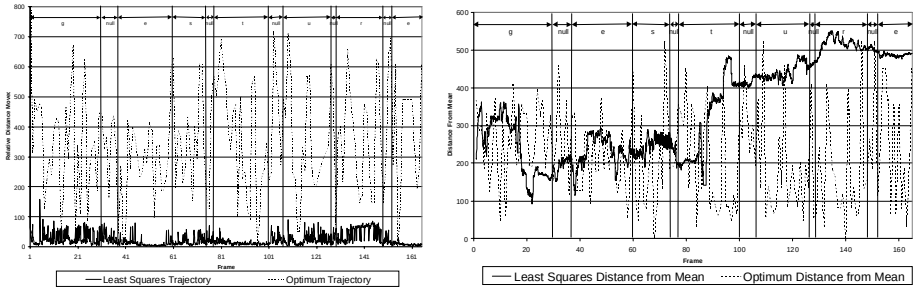

Figure 5 – (a) Graph of Distance Moved at each iteration for Least Squares Solution and Optimum Solution, (b) Graph of Distance from Mean of Shape Space at each frame for Least Squares Solution and Optimum Solution

Figure 5a shows the distance moved through shape space at each iteration for both the optimum trajectory and the iterative refinement algorithm. From this it can clearly be seen that the least squares iterative refinement algorithm makes small incremental movements at each iteration, whereas the optimum trajectory makes large 'jumps' at every frame. During the letters 'e' and 't' the least squares approach almost stops moving, which demonstrates that the model has converged upon a stable solution. However, the lack of such trends for other letters shows that the model is constantly struggling to better refine itself. Figure 5b shows distance from the centre of shape space for the two trajectories. Again this demonstrates that the optimum path jumps violently within the space whereas the least squares approach makes small movements.

The most interesting aspect of these figures is Figure 5b. The letter 'e' occurs twice during the sequence. However, during the first occurrence the least squares approach is at a distance of around 200 units from the mean whereas during the second occurrence it is at around 500. This demonstrates that there are at least two areas of shape space responsible for modelling the letter 'e' and these are distinctly separated in shape space. It also shows that the least squares approach can only use the local 'e' part of shape space and is incapable of jumping between them.

This confirms that not only is the non-linear shape space discontinuous but the least squares iterative refinement approach is incapable of providing a robust method for tracking. Instead a new method of applying CSSPDMs must be devised.

3.2 Supporting Multiple Hypotheses

Due to the discrete nature of the piecewise linear method of modelling non-linearity, the approach directly lends itself to a discrete PDF with the addition of a Markovian assumption. A 1st order model of temporal dynamics can be derived where the conditional probability $P\left(C_i^{t+1}\middle|C_j^t\right)$ provides the probability that the model will move to cluster C_i given it was at C_j at the last time step. This conditional probability can be calculated from the training sequence and produces a 2D PDF of motion within shape space.

Figure 7a shows the ASL PDF, which has a heavy diagonal dominance. This dominance is when $\mathbf{argmax}_i\left(P\left(C_i^{t+1}\middle|C_j^t\right)\right)$ and $i = j$ i.e. the highest probability is that the PDM will stay within the present cluster. The assumption can therefore be made that within any local patch the model can iterate to a local solution. This confirms the assumption used when calculating the optimum model trajectory. This assumption also provides two benefits:

1. The iteration to convergence of any global optimisation technique can be enhanced by allowing each hypothesis to iterate to a better solution within the present cluster.
2. A smaller population is required, as only global differences in hypotheses need to be supported.

From the 'learnt' probability density function, a sample population can be generated at each iteration of the model. Given a good initialisation of the model and the associated cluster $C^{t=0}$, this can then be used to predict the future movement.

However, this approach, unlike condensation, does not recover well from failures. As the new population is solely based upon the current best-fit cluster, the approach is highly sensitive to both an accurate PDF and a good fit to the current object pose. To help overcome this drawback less emphasis must be placed upon the current best-fit hypothesis being the optimum (and hence correct) solution, thus providing more robustness to failure. This can be addressed by creating a new population of hypotheses, not from the current best fit model, but from the weighted sum of the best *n* hypotheses as described thus:

Algorithm 1 - Weighted Condensation

- From the PDF $P\left(C_i^t \middle| C_j^{t-1}\right)$, extract the probability vector $P\left(C_i^{t=1}\right)$, which is the probability distribution of the first iteration, given $C_j^{t-1} = C^{t=0}$.
- Generate a randomly sampled distribution of *k* hypothoses $\mathbf{x}_\rho\,[\rho = 1,\ldots,k]$, where $\mathbf{x}_\rho$ is the mean shape of cluster C_i and $P(C_i) = P\left(C_i^{t=1}\right)$
- While still tracking,
  - Fit *k* hypotheses, applying CSSPDM constraints and assess fitness using error metric
  - Sort hypotheses into descending order according to error
  - Iteratively refine first *n* hypotheses and resort
  - Apply the CSSPDM constraints and determine the *n* clusters C_η^{t-1}, where $\eta = 1,\ldots,n$ which produce the lowest error
  - From the PDF $P\left(C_i^t \middle| C_j^{t-1}\right)$, extract the vector $P\left(C_i^t\right)_\eta$ using the *n* extracted clusters. Take the weighted sum using a Gaussian weighting distribution to form a new distribution $P'\left(C_i^t\right)$, where

$$P'(C_i^t) = \sum_{\eta=1}^{n} \omega_\eta P(C_i^t)_\eta \quad \text{and} \quad \omega_\eta = \exp\left[\frac{-9(1-\eta)^2}{2n^2}\right]$$

  - Normalise probability distribution $P'\left(C_i^t\right)$.
  - Generate a new random population of *k* hypotheses from the distribution $P'\left(C_i^t\right)$.

4. Extending Temporal Dynamics

It has been shown how, with the addition of a first order Markov chain to the CSSPDM, a hybrid approach to condensation can be used to provide robust tracking where either:

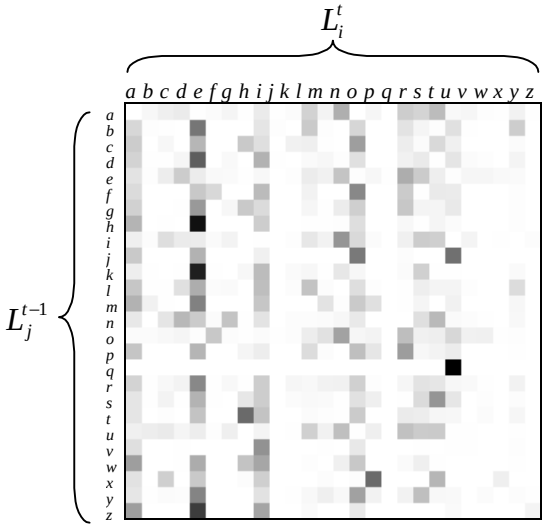

Figure 6 - Discrete Probability Density Function for the English Language

- The non-linearity of the PDM along with the discrete representation of the non-linear approximation leads to a discontinuous shape space.
- Rapid movement of the object produces large changes in the model parameters.

This Markovian model of dynamics can be used to explicitly constrain the movement of the model within shape space, or implicitly, using the hybrid condensation approach. However, the use of temporal constraints relies upon the assumption that the training set from which the model is built, contains a thorough representation of all-possible deformation and movement. For simple models this is often true. However, for ASL it is not, and it is important to ask the question, *'What exactly is the temporal model representing?'*

The ASL PDF represents two aspects of motion,

1. The non-linear representation of shape space, how the individual clusters relate and how the model moves throughout the space to form letters.
2. It also contains information about the English language and how letters relate to form words and sentences.

As the PDF encodes both of these attributes it must be constructed from a training set which has a good representation of how the model deforms and be representative of the English language. This is however infeasible. If the ASL image sequence used previously is considered, it took 165 frames to record the 7 letter word 'gesture'. Konheim reported a statistical study where the 1-state transition probabilities of the English Language were determined using 67,320 transitions between two successive letters [9]. As the 165 frames previously used produced an average of 20 frames per letter, this would constitute a training set in excess of 1.3 million frames not including transitional shapes between letters. As each frame produces a training shape this results in a training set which is of infeasible size.

The current ASL PDF (see Figure 7a) contains valuable information about how the model moves within shape space, but due to the deficiency in training it does not contain sufficient information to accurately model the transitions between the letters of the English language. Fortunately, it is relatively simple to gain a transition matrix for the English language by analyzing large samples of electronic text and calculating the

1-state transitions. What is required is a method of combining this knowledge of English into the ASL PDF, producing a more generic and accurate model for tracking and classification.

4.1 The Temporal Model

The ASL PDF $P\left(C_i^t \middle| C_j^{t-1}\right)$, constructed from the training set, provides the probability that the model will move to cluster C_i given it was at cluster C_j at the last time step. Similarly a 1st order Markov Chain can be constructed for the English language which provides a new PDF $P\left(L_i^t \middle| L_j^{t-1}\right)$. Figure 6 shows the PDF gained from this Markov Chain as taken from Konheim and shows the 1-state transitions calculated from a sample text of over 67 thousand letters [9].

Figure 6 does not demonstrate a diagonal dominance, unlike Figure 7a. This is because the English language has few occurrences of repetitive letters in words whereas the previous PDF resulted from operations involving a high degree of repetition. The main trend that can be seen are the vertical stripes that occur for many of the letters.

In order to incorporate this additional information learnt from sample text, a new ASL PDF must be constructed $P'\left(C_i^t \middle| C_j^{t-1}\right)$. To do this a mapping must be achieved which allows shape space to relate to gesture space.

4.2 Mapping Between Spaces

By labelling each training example with an associated letter a PDF can be generated which relates clusters in shape space to gestures. Here the conditional probability $P\left(L_i^t \middle| C_j^t\right)$ provides a probability of the occurrence of a letter L given the model is in cluster C in shape space at any time. This conditional probability provides a mechanism to relate shape space to the gesture space where the constraints of the English language (as learnt) can be applied. However, for this to be of use, a method that allows this information to be mapped back into the shape space must be provided. This can be done using the common form of Bayes theorem where

$$P\left(C_i^t \middle| L_j^t\right) = \frac{P\left(C_i^t\right)P\left(L_j^t \middle| C_i^t\right)}{P\left(L_j^t\right)}$$

However, where $P\left(C_i^t \middle| L_j^t\right)$ and $P\left(C_i^t\right)$ can both be gained from the training set, $P\left(L_j^t\right)$ (the probability of the occurrence of a letter) can only be gained from analyzing English text. As it is known that the training set does not fully represent the English Language this equation would lead to biasing of the final conditional probabilities. Instead, a variation of Bayes Theorem can be used, where

$$P\left(C_i^t \middle| L_j^t\right) = \frac{P\left(C_i^t\right)P\left(L_j^t \middle| C_i^t\right)}{\sum P\left(C_i^t\right)P\left(L_j^t \middle| C_i^t\right)}$$

Using this form, $\sum P\left(C_i^t\right)P\left(L_j^t \middle| C_i^t\right) \equiv P\left(L_j^t\right)$ but all probabilities are gained from the training set, and hence no bias occurs from mixing unrelated probabilities. This is possible as, although the training set does not contain a thorough representation of English, it does provide an accurate representation of the mapping between the two spaces.

4.3 The Hybrid ASL PDF

A new ASL PDF can now be constructed which incorporates the 1-State transitions of the English language by treating the system like a Hidden Markov Model and projecting the transitions of the observation layer down into the Hidden (parameter space). Taking the current cluster of the model the corresponding letter(s) associated with this cluster is determined and the 1-state transition matrix applied to extract the most likely next letter. The cluster(s) associated with this transition are then calculated.

Where,

$$P'\left(C_i^t \middle| C_j^{t-1}\right) = P\left(L_i^t \middle| C_j^t\right) P\left(L_i^t \middle| L_j^{t-1}\right) P\left(C_i^t \middle| L_j^t\right)$$

This produces a new ASL PDF which is shown in Figure 7b.

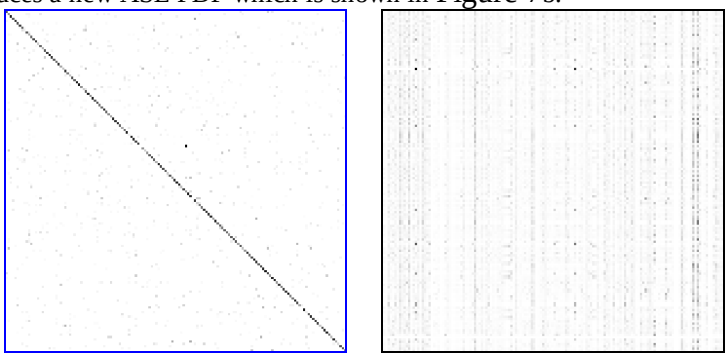

Figure 7 – (a) Discrete Probability Density Function for ASL Model, (b) Discrete Probability Density Function for hybrid ASL Model

Figure 7b demonstrates the same characteristic vertical strips seen from the English Language PDF, which it has inherited, and as such differs from the original ASL PDF in two ways.

1. Each cluster exhibits far more transitions to other clusters.
2. The diagonal dominance that is important to tracking, is missing.

Diagonal dominance can be forced upon the PDF by imposing diagonal dominance on either $P\left(L_i^t \middle| L_j^{t-1}\right)$ or $P'\left(C_i^t \middle| C_j^{t-1}\right)$. However, this is haphazard and risks over-biasing the hypothesis generated at each frame. An alternative is to simply ensure that the population generated at each step always includes at least one hypothesis from the current cluster.

Figure 8 shows the results of three of the techniques discussed, namely that of the least squares gradient descent (ASM, algorithm[5]) the optimal solution gained through an exhaustive search of shape space and that of the hybrid condensation approach. The error metric is the euclidean distance of a shape to the closest allowable point within shape space as gained from the different algorithms. It can clearly be seen that the optimum solution does indeed give the lowest results with the hybrid condensation producing only slightly higher error rates, both of which are significantly lower than those from the Least squares appraoch which fails catastophically.

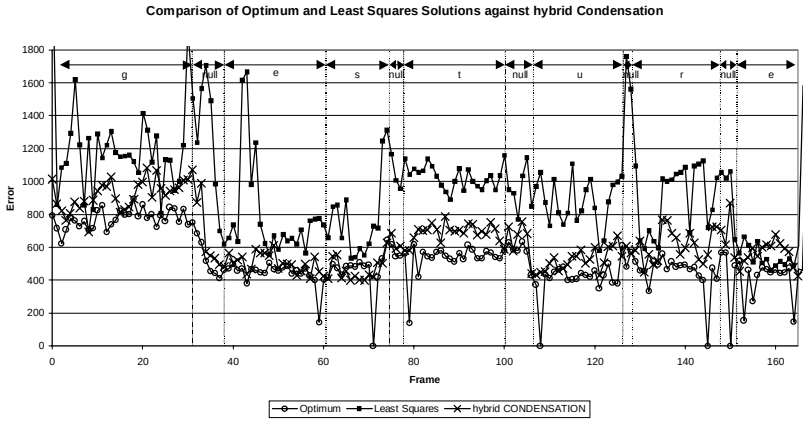

Figure 8 – Comparison of hybrid CONDENSATION against Optimum and Least Squares Approaches

5 Conclusions

This paper has presented the augmentation of statistical models with temporal dynamics gained through the probabilistic analysis of the training set and how this relates to movement within shape space. It has been shown how the discrete segregation of shape space used in the CSSPDM directly lends itself to a Markov chain approach to modelling temporal dynamics. It has been shown that the nature of shape space is often complex and discontinuous and how, using these additional learnt temporal constraints, tracking can be improved by supporting a population of multiple hypotheses. However, the key to this paper is the ability to project observation probabilities into a hidden shape space using an approach akin to a Hidden Markov Model where the simple acquisition of observation layer transitions can be propagated into the hidden parameter space to overcome the inadequacies of training. It has been shown how, using a hybrid CONDENSATION tracker, successful tracking can be achieved while maintaining a considerable lower population size to that of standard CONDENSATION.

6 References

[1] Blake, A. and Isard, M., Active Contours, Springer Verlag, 1998.

[2] Bowden, R., Mitchell, T. A., Sahardi, M., Cluster Based non-linear Principal Component Analysis, IEE Electronics Letters, 23rd Oct 1997, 33(22), pp1858-1858.

[3] Bowden, R., Mitchell, T. A., Sahardi, M., Non-linear Statistical Models for the 3D Reconstruction of Human Pose and Motion from Monocular Image Sequences. To appear in Image and Vision Computing.

[4] Bowden, R. Learning non-linear models of Deformation and Motion, PhD Thesis, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK, 2000.

[5] Cootes, T. F., Taylor, C. J., Cooper, D. H., Graham, J., Active Shape Models - Their Training and Application. Computer Vision and Image Understanding, 1995;61(1):38-59.

[6] Hill, A., Taylor, C. J., Model based image interpretation using genetic algorithms, In Proceedings British Machine Vision Conference, Springer-Verlag, 1991, pp 266-274.

[7] Hill, A., Taylor, C. J., Model based image interpretation using genetic algorithms, Image Vision Computing. 10, 1992, 295-300.

[8] Isard, M. and Blake, A., Condensation - conditional density propagation for visual tracking. International Journal of Computer Vision, 1998.

[9] Konheim, A., G., Cryptography: A Primer, John Wiley, New York, 1982.