Multi-View Face Detection and Pose Estimation Using A Composite Support Vector Machine across the View Sphere

Jeffrey Ng and Shaogang Gong
Department of Computer Science, Queen Mary and Westfield College
London E1 4NS, UK
{jeffng,sgg}@dcs.qmw.ac.uk

Abstract

Support Vector Machines have shown great potential for learning classification functions that can be applied to object recognition. In this work, we extend SVMs to model the 2D appearance of human faces which undergo nonlinear change across the view sphere. The model enables simultaneous multi-view face detection and pose estimation at near-frame rate.

1 Introduction

Tracking people across a variety of views is becoming increasingly important in computer vision systems. Apart from traditional applications such as segmenting faces for identity recognition [2], multi-view face detection and tracking is also being used in smart-systems for visually mediated interaction [9], inferring user intentions in human-computer interaction [3] and incident monitoring [8]. The ability to extract visual cues such as gait or head orientation allows advanced vision systems to obtain better information about their contextual situation. A better perception of their prevailing operating conditions allows such systems to interact more intelligently with their environment.

Head pose, in particular, provides good cues about the general focus of attention of people. However, the appearance of the human head can change drastically across different viewing angles, mainly caused by nonlinear deformations during in-depth rotations of the head. Existing template-matching and neural network systems would be hard pressed to learn the whole gamut of multi-view face appearances. Systems based on similarity measures to prototypes, on the other hand, are to some extent still restricted to the available views for the prototypes selected in the training database [2]. Similarity measures can be noisy and sensitive to the choice of representation. In order to perform accurate and robust face tracking across views, its generalisation to novel views is aided by pose estimation.

In this work, we exploit the potential of Support Vector Machines (SVMs) [10] for generalising and transforming a generic 2D facial appearance model across the view sphere [1, 2]. In particular, we investigate plausible computational solutions for effective face detection at different views, tracking across views and pose estimation at no extra cost. We offer a viable solution for addressing the needs for both multi-view face detection and pose estimation at near-frame rate.

2 Structural Risk Minimisation with SVMs

SVMs are based on a generic learning framework that has shown unique potentials for object recognition [10, 7, 4, 5, 6]. Previous approaches to statistical learning have tended to be based on finding functions to map vector-encoded data to their respective classes. The conventional minimisation of the empirical risk over training data does not however imply good generalisation to novel test data. Indeed, there could be a number of different functions which all give a good approximation to a training data set. It is nevertheless difficult to determine a function which best captures the true underlying structure of the data distribution. Structural Risk Minimisation (SRM) aims to address this problem and provides a well defined quantitative measure for the *capacity* of a learned function to generalise over unknown test data. Due to its relative simplicity, Vapnik-Chervonenkis (VC) dimension [10] in particular has been adopted as one of the more popular measures for such a capacity. By choosing a function with a low VC dimension and minimising its empirical error to a training data set, SRM can offer a guaranteed minimal bound on the test error. For a two-class recognition problem, support vectors define a hyperplane and the decision boundaries of the two classes. However, a hyperplane classification function cannot be found for a two-class recognition problem when they are not linearly separable in the input space. To overcome

this problem, a high dimensional mapping such as

$$\phi \; : \; R^N \mapsto F$$

is used to build nonlinear support vector machines. As both the objective function and the decision function is expressed in terms of dot products of data vectors **x**, the potentially computationally intensive mapping $\phi(\cdot)$ does not need to be explicitly evaluated. A kernel function, $k(\mathbf{x}, \mathbf{z})$, satisfying Mercer's condition can be used as a substitute for $(\phi(\mathbf{x}) \cdot \phi(\mathbf{z}))$ which replaces $(\mathbf{x} \cdot \mathbf{z})$ [10].

For noisy data sets where a large overlap exists between data classes, error variables ε_i are introduced to allow the output of the outliers to be locally corrected, which constrains the range of the Lagrange multipliers α_i from 0 to C. C is a constant which acts as a penalty function, preventing outliers from affecting the optimal hyperplane. There are a number of kernel functions which have been found to provide good generalisation capabilities, e.g. polynomials. Here we explore the use of a Gaussian kernel function.

3 The Nature of Face Pose Distribution

Detecting human faces across views involves the recognition of a whole spectrum of very different face appearances. The pose of the head reveals some details about the 3-dimensional structure of the face while masking others. Head rotations introduce nonlinear deformations in captured face images while the rotation can occur in two axes outside the view plane of the camera. A face's main direction of light reflection can change and affect the illumination conditions of the captured image. For instance, ambient day-time lighting conditions in normal office environments are hardly symmetric for the top and bottom hemispheres of the face, while the bias towards the upper hemisphere is exacerbated by ceiling-fixed light sources during the night.

Understanding the face pose distribution across the view sphere can provide a basis for learning a generic face model based on multi-view support vector machines.

A face rotating across views forms a smooth trajectory as can be seen in Figure 1. In fact, faces form continuous manifolds across the view sphere in a Pose EigenSpace (PES) [1]. It is plausible to suggest that head rotations in depth subscribe to a continuous function in PES. This can be seen more clearly in Figure 2. In particular, a pattern appears for the vertical positioning (from the selected view angle) of groups of trajectories across the view sphere. The volume enclosed by the entire view sphere is more visible when the nodes of the sphere are plotted individually as in Figure 3. The distribution appears to be a convex hull.

Given the correlations of the lateral bands of the face sphere, we group the whole distribution into 19 different clusters according to their yaw orientation (0° to 180°). We

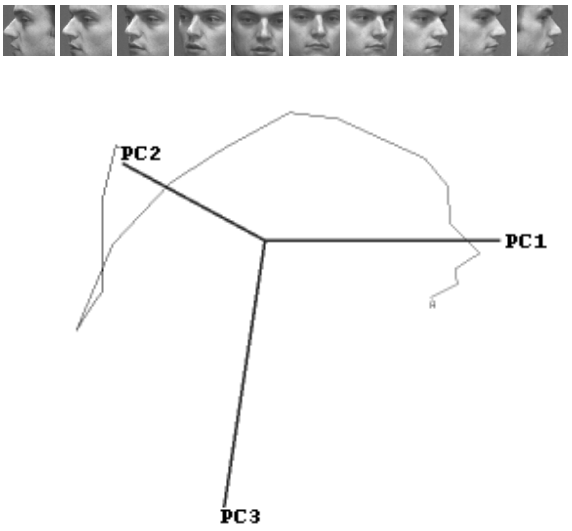

Figure 1. Face rotation in depth forms a smooth trajectory in a 3D pose eigenspace.

observed that the trajectory of the mean positions of the clusters, which are indeed their centroids in PES, structures the distribution across a main axis of variation. This notion is further supported by the tangentiality of the main axes of local variation inside the clusters across the mean trajectory as shown in the lower right picture in Figure 3. The above observations strongly suggest that the convex hull is more akin to a "tube", a volume function, through which data elements "flow" from one end to the other as their yaw angles increase from 0° to 180°.

4 A Multi-View Face Model using SVMs

Support Vector Machines perform automatic feature extraction and enable the construction of complex nonlinear decision boundaries for learning the distribution of a given data set. The learning process and the number of support vectors for a data set are determined in a principled way by only a few customisable parameters which defines the characteristics of the learned function. In our case, the parameters are limited to two: C, the penalty value for the Lagrange multipliers to distinguish between noisy data and, σ for determining the effective range of Gaussian Support Vectors. Effective values for the two parameters have already been reported for frontal view face detection [4].

We adopt a semi-iterative approach for obtaining good examples of negative training data. The ideal negative images chosen by SVM training algorithms for negative support vectors have been reported to be naturally occurring non-face patterns that possess a strong degree of similarity to a human face [4]. Given the highly complex distribution

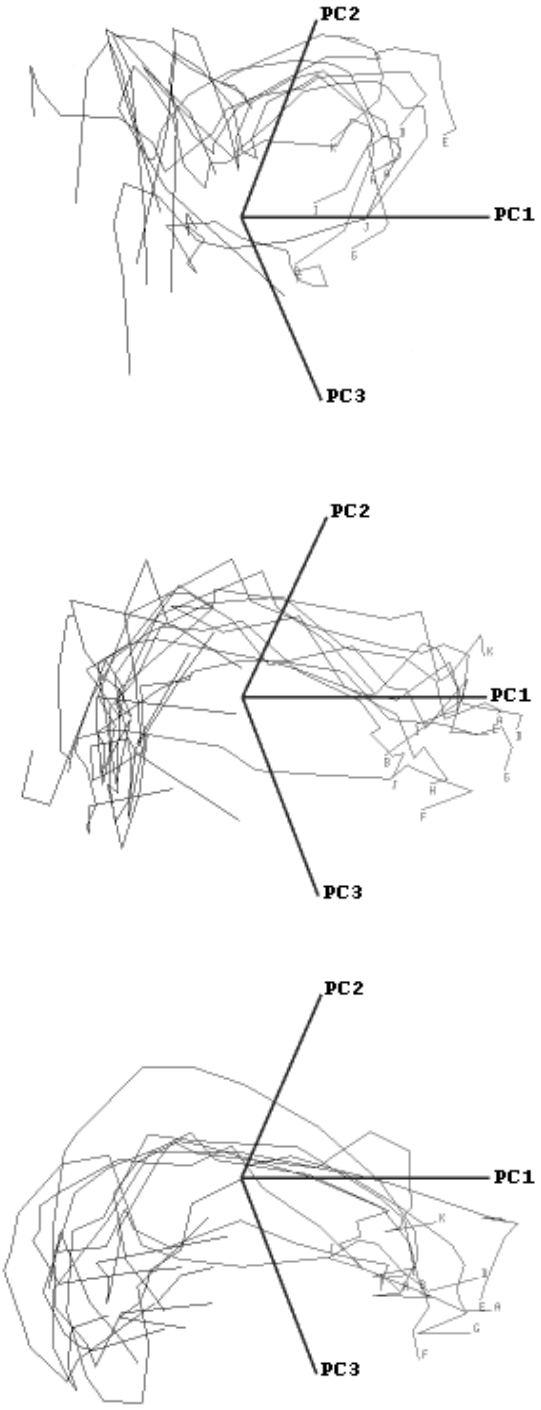

Figure 2. From top to bottom: The graphs show the PES trajectories for a set of 10 people rotating their heads from profile to profile, at 60° tilt, 90° tilt and 120° tilt respectively.

of the view sphere described in the previous section, it is crucial to find good examples of these to allow the training algorithm to construct accurate decision boundaries.

We first extend the training of a single frontal-view SVM face model to the use of face images across the view sphere. The process uses an iterative refinement methodology to find important negative training samples from a database of randomly selected scenery pictures. The resulting single SVM cannot cope with the degree of view generalisation required when applied to face detection across a significantly large range of views away from the frontal view. However, the model is very useful for iteratively collecting negative training samples beyond the near frontal view. Such negative samples are then used to train a multi-view face model based on a set of local component SVMs along the view sphere.

Given the face distribution in PES as shown in Figure 3, the view sphere can be divided into smaller, more localised yaw segments as in Table 1. The observed asymmetry of the view sphere distribution and the greater complexity of the left portion are reflected into the selection of smaller segments for that region.

Segments	1	2	3	4	5
Yaw angles	0-10°	20-40°	50-80°	90-130°	140-180°
Face data	140	210	280	350	350
Positive SVs	107	139	176	190	203

Table 1. The division of the view sphere for learning multi-view SVMs.

All the component SVMs were trained on the same global negative data set. The size of the negative training data is about 6000 images and of those, the SVMs selected 1666 as negative support vectors in total, with only 36 shared between two or more component SVMs. This shows that the negative support vectors are well localised to the sub-space of each yaw segment.

The modelling capabilities of the component SVMs and their tendency to overflow to the neighbouring segments corroborated with the previous observations of the structure of the distribution of the view sphere in pose eigenspace. In general, the component SVMs could detect faces at yaw angles of 10° on either side of their training ranges. In some cases, the overlap was as much as 30°. The observed phenomenon also shows that support vectors are localised in a composite distribution such as the view sphere. They can be used to detect either the whole distribution or smaller segments in that distribution.

For face detection across the view sphere, the component SVMs can be arranged into a linear array to form a compos-

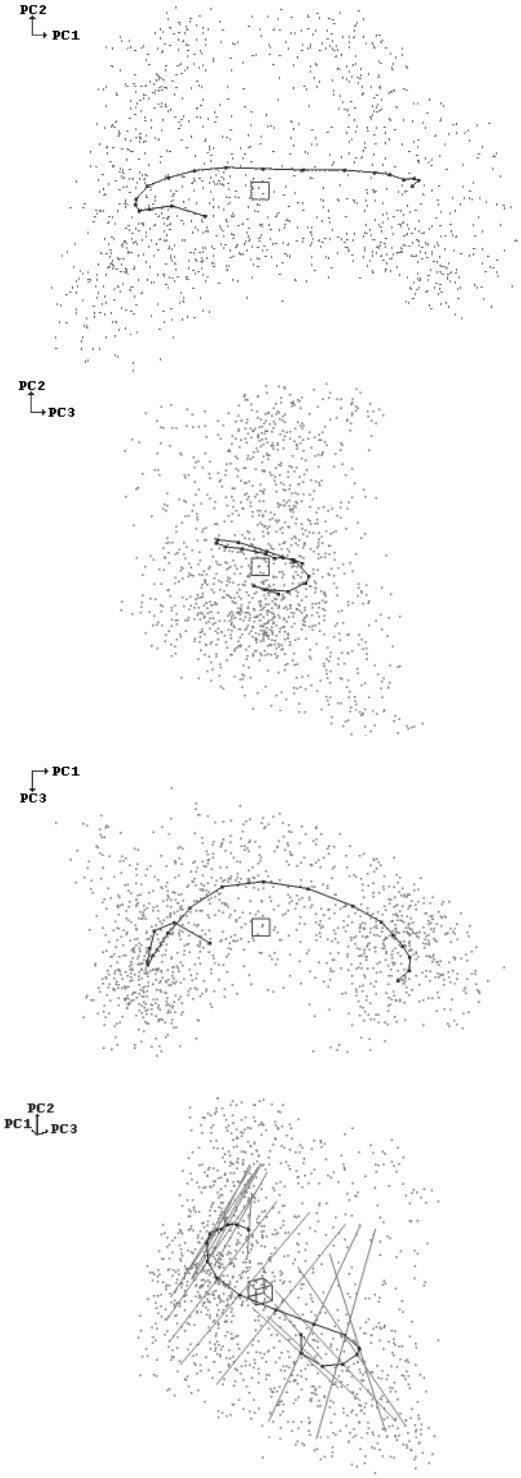

Figure 3. The side, frontal and top views of the face pose distribution in PES. The trajectories of the mean yaw clusters are also shown. The last plot shows the directions of the largest variance across yaw.

ite SVM classifier as follows:

$$\text{Composite SVM}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{5} SVM(i,\mathbf{x}) + 1\right) \quad (1)$$

where $SVM(i,\mathbf{x})$ is the decision function $f(\mathbf{x})$ for SVM number i.

This multi-view face model can also be applied to pose estimation across the view sphere. Figure 3 shows the correspondence of the yaw angles to the data elements' positions along the mean trajectory of the yaw clusters. A similar correspondence of the tilt angles to their "vertical position" from the selected viewing angles, with the variation lying approximately perpendicular to the mean yaw trajectory, can also be observed in Figure 2.

Since the support vectors define the boundaries of the face pose distribution, they lie on the "walls" of the "tube". Furthermore, they are also localised with regard to the pose sphere. Therefore, they can be effectively used to perform pose estimation by using nearest-neighbour matching, as illustrated in Figure 4. In fact, this process of pose estimation is retrieved at no extra computational cost to the calculation of the decision function.

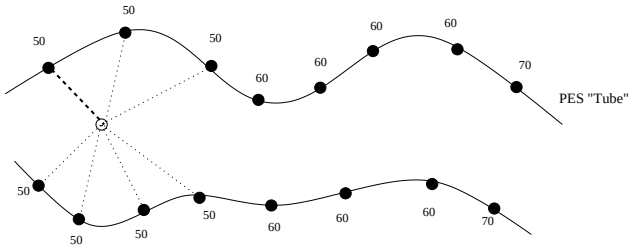

Figure 4. Top view of the face manifold across the pose eigenspace with pan angles labelled to each support vector (dark circles). The pose orientation of the classification image (white circle) is retrieved from that of the closest support vector.

5 Multi-View Face Detection and Pose Estimation

We applied the multi-view SVM-based face model to perform both multi-view face detection and pose estimation across views. First, we show the performance of a multi-view face detection system on training data in Table 2.

It is important to point out that the accuracy in the alignment of face images plays a crucial role in the learning process. Most of the misclassified elements of the view sphere were correctly recognised after multi-scale scanning of the

Training subsets	1	2	3	4
Full detection	100%	97.7%	94.7%	92.5%
multi-scale scan	100%	100%	100%	97.0%

Training subsets	5	6	7	8
Full detection	82.7	88.7%	94.7%	100%
multi-scale scan	85.7	99.2%	97.7%	100%

Table 2. Face detection on training data across the view sphere, grouped by human subject.

images. The multi-scale scanning is performed on the input images with a bias in each of the four directions to correct misalignment of the face images.

Our previous work reported that the variation of the view sphere distribution along the second principle component axis was highly related to the level of local lighting in the image [1]. Using an overhead light source can yield such an effect on the captured images. The lighting conditions must therefore help in the determination of the tilt orientation of the faces. However, it also makes down-facing poses very poorly illuminated and therefore, very difficult to detect by the system as shown in Figure 5.

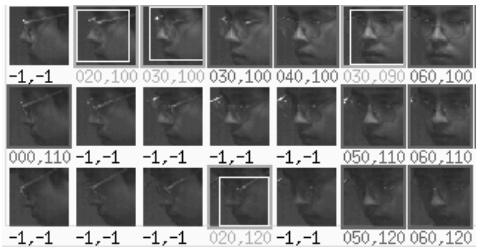

Figure 5. Misclassification in lower hemisphere of the view sphere (shown by -1,-1). Image multi-scale scanning is shown with white rectangles.

The multi-view system was tested over a number of test sequences of human subjects freely turning their heads in 3D space, with the ground-truths of the pose information measured for comparison. The system was also connected to an iso-tracking device, allowing face detection (alignment) and pose estimation to be independently evaluated. Experiments on three subjects are given here for illustration: the subject with the worst training detection results (test sequences A and B) and two novel subjects unknown to the training process (test sequences C and D). They were selected to test the generalisation capabilities of the system. Figures 6, 7 and 8 show example frames from different test sequences in which novel faces were detected in multi-views and tracked across views with their pose estimated simultaneously. Table 3 provides a summary of the results for the test sequences and it can be noticed that the detection rate and the average pose estimation error do not vary significantly between the sequences of known subjects and those of unknown subjects, namely sequences A and B against C and D.

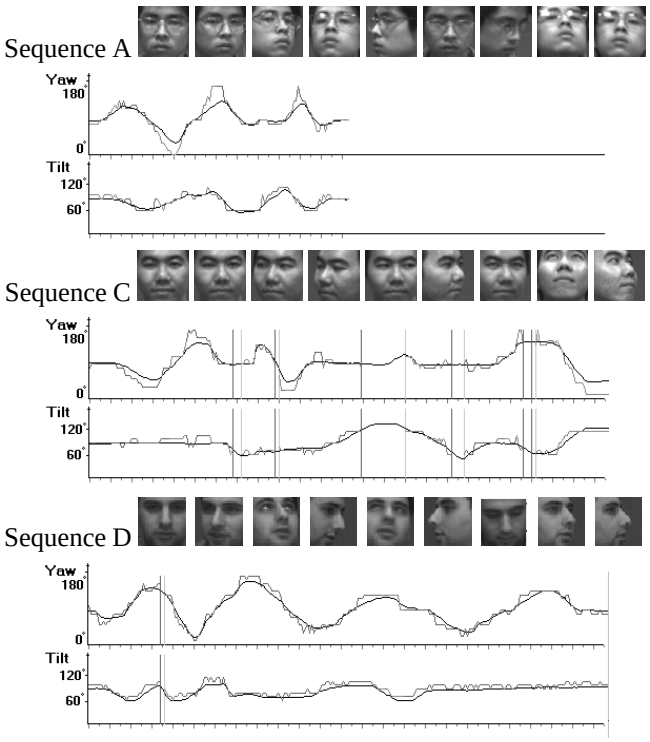

Figure 6. Examples of detected and tracked moving faces. The graphs also show the estimated face pose (in grey) over time and their corresponding ground-truths (in black), measured by electro-magnetic sensors. The vertical lines indicate moments in time where no face was detected.

Test Seq.	Det. Rate	M-Yaw Error	M-Tilt Error
A	100%	11.07°	6.62°
B	84.9%	11.467°	6.32°
C	82.9%	13.57°	7.29°
D	99.6%	8.73°	8.67°

Table 3. Test results of the multi-view face detector and pose estimator from a total of over 1000 images from the test-sequence set.

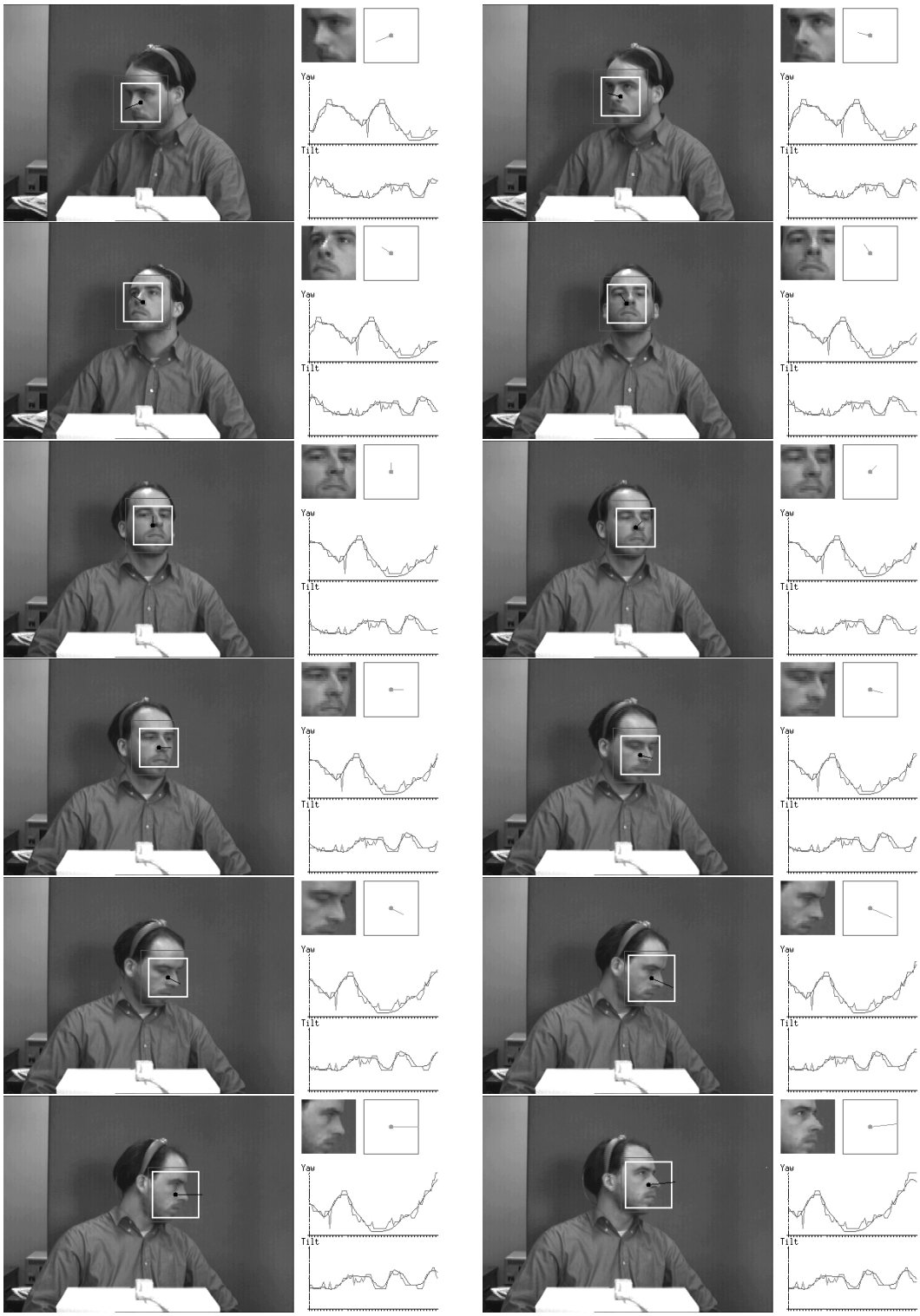

Figure 8. Examples of near-frame rate face detection, tracking and pose estimation using a multi-view composite SVM. On the right of each picture are detected faces in each image frame, its pose estimated in a dial, and the estimated pose versus the ground-truth from the Polhemus sensor over time.

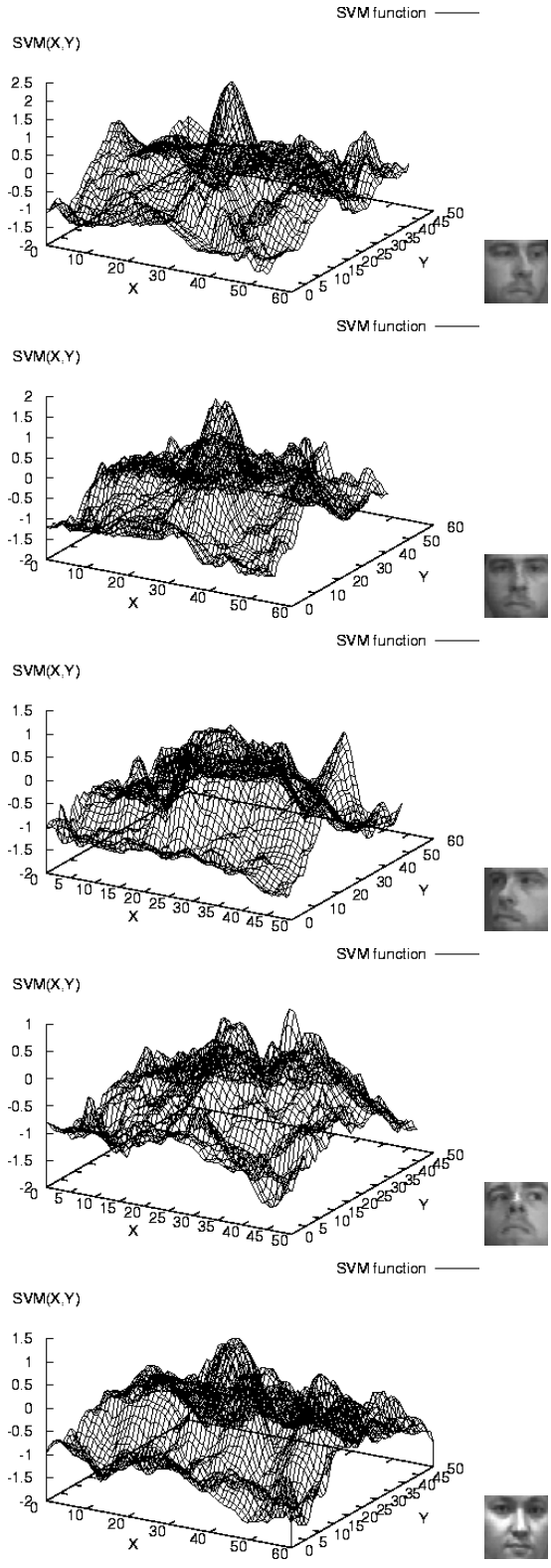

Figure 7. Examples of topographical outputs from a multi-view composite SVM classifier when applied to face detection at different views. The detected faces are also shown.

6 Real-Time Performance

Support vector machines use kernel functions to learn and classify nonlinearly separable data distributions. With typical support vector sets of SVMs ranging in the thousands and a kernel evaluation required for each support vector, classification can become computationally expensive. The problem can be further exacerbated by the type of hierarchical multi-resolution image scanning required for online face detection and tracking. SVM optimisation techniques such as the reduced or virtual support vector set methods cannot be easily implemented in the multi-view SVM system because of the importance of each original support vector for determining the pose of new images.

The performance of the SVM tracker was improved however by continuously tracking objects, restricting the range of resolution and image regions to be searched according to previous tracking results. The active set of component SVMs used for tracking can also be constrained to the local sub-spaces where objects were previously detected.

Furthermore, the output of the multi-view SVM classifier exhibits positive peaks at regions of faces in the image, as can be seen in Figure 7. The peaks are at their strongest at the centre of face regions where the best detection occurred. Threshold filtering was found to give best detection locations more quickly than uniformly scanning the image. However, noise exists due to translational misalignment in the training face images at some poses. This can be observed in some of the examples shown in Figure 7. It was again observed, however, that translational noise occurred mostly along the vertical axis of the image with each scanned line maintaining a perfectly distinguishable peak. This still enabled us to implement a peak detection mechanism for each scan line in order to avoid redundant scanning. By also performing temporal prediction, we achieved a multi-view face tracking and pose estimation at a frame rate greater than 1 Hz on a standard 330MHz Pentium PC running Linux without utilising any special hardware. An example of this multi-view tracking process is shown in Figure 8.

7 Conclusion

In this work, we have shown that the complex distribution of face poses can be modelled by a collection of view-based component support vector machines. Pose estimation can also be automatically performed by Gaussian kernel functions used in the multi-view SVMs, allowing both tasks to be performed by a single integrated process, thereby, greatly reducing computation. More accurate pose estimation can be achieved by using a better aligned training set. Future research into using the non-linear mapping learned by the SVM classifier can also provide an improve-

ment in pose estimation accuracy over the simple nearest-neighbour matching we adopted for retrieving pose information from support vectors.

On the matter of real-time performance, multi-view SVM classification is still not computationally attractive for real-time use. However, image-scanning using global optima search methods provides a promising future for faster tracking. Combined with motion prediction techniques, an example of which is CONDENSATION, multi-view SVM tracking and pose estimation has been shown to possess great potential for real-time systems.

References

[1] S. Gong, S. McKenna, and J. Collins. An investigation into face pose distributions. In *IEEE Int. Conf. on Face & Gesture Recognition*, pages 265–270, Vermont, 1996.

[2] S. Gong, E. Ong, and S. McKenna. Learning to associate faces across views in vector space of similarities to prototypes. In *BMVC*, volume 1, pages 54–64, UK, 1998.

[3] E. Ong, S. McKenna, and S. Gong. Tracking head pose for inferring intention. In *European Workshop on Perception of Human Action*, Freiburg, Germany, June 1998.

[4] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *CVPR*, 1997.

[5] J. C. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimisation*. Microsoft Research Technical Report MSR-TR-98-14, 1998.

[6] B. Schölkopf, C. Burges, and A. Smöla. *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[7] B. Schölkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In *International Conference on Artificial Neural Networks*, 1996.

[8] J. Sherrah and S. Gong. Exploiting context in gesture recognition. In *2nd Int. Interdisciplinary Conf. on Modelling and Using Context*, Trento, Italy, September 1999.

[9] J. Sherrah and S. Gong. Fusion of perceptual cues using covariance estimation. In *BMVC*, Nottingham, England, September 1999.

[10] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.