Facial Motion Analysis using Clustered Shortest Path Tree Registration

David Cristinacce, Tim Cootes

Facial Motion Analysis using Clustered Shortest Path Tree Registration

David Cristinacce and Tim Cootes

Dept. Imaging Science and Biomedical Engineering
University of Manchester, Manchester, M13 9PT, U.K.
david.cristinacce@manchester.ac.uk

Abstract. We describe a method of automatically annotating video sequences, defining a set of corresponding points in every frame. This is an important pre-processing step for many motion analysis systems. Rather than tracking feature points through the sequence, we treat the problem as one of 'groupwise registration', in which we seek to find the correspondence between every image and an automatically computed model reference, ignoring the ordering of frames. The main contribution of this work is to demonstrate a method of clustering the frames and constructing a shortest path tree over the clusters. This tree defines the order in which frames will be registered with an evolving estimate of the mean. This technique is shown to lead to a more accurate final result than if all frames are registered simultaneously. We describe the method in detail, and demonstrate its application to face sequences used in an experiment to assess the degree of facial motion. The resulting ranking is found to correlate well with that produced by human subjects.

1 Introduction

Our aim is to achieve accurate registration between all frames of a video sequence, by finding corresponding control points in all frames. The corresponding points then define a dense correspondence between frames, which can serve as a basis for motion analysis.

A critical step in the registration process is the initialisation of control points in each frame. A previous approach described by Cootes *et al.* [1], initialises a grid of control points at the same location in all images. It then performs a series of translation searches and similarity transforms, followed by more complex warps, to register the whole set of images. Hence each image is treated equally and processed simultaneously.

In our approach we divide the video frames into smaller clusters of similar images, that are easier to register using the groupwise approach. The clustering method also provides an approximate starting point in each frame and the model generated from previous clusters is used to initialise the registration process in later clusters. A shortest path tree structure between the clusters enables an incremental approach to the registration which allows the initial processing to occur on easy data (in our case -frontal faces) and later matching to occur on

outlier frames. We find this incremental clustered tree approach to give superior results to the single groupwise method, when tested on the manually annotated FGNET data set [2].

Our tree based registration is applied separately to 59 video sequences of different individuals. The corresponding points are used to build a multi-person statistical shape model [3]. The paths of each individual sequence are analysed in the combined shape space and used to determine the relative amount of facial motion for different individuals.

2 Background

There are many approaches to vision based motion analysis of human faces.

One method is to track facial features approximately and use machine learning techniques to make deductions about the underlying image, treating small tracking errors as noise. This method is adopted by Pantic *et al.* [4] and successfully used to identify facial expressions. Similar machine learning approaches are adopted by face recognition systems [5] and could be applied to the task of estimating facial motion.

An alternative approach is to compute a dense motion correspondence explicitly from the image data and apply further analysis to tracked features, independent of the original images. To do this successfully we require accurate localisation of corresponding features on each frame of a video sequence. To achieve this we use an extension of the groupwise registration approach due to Cootes *et al.* [1], which is similar to the image encoding method described by Baker *et al.* [6]. An overview of general image registration methods is found in Zitova [7].

Image registration is computationally expensive compared to sequential tracking methods, e.g., Simultaneous Modelling and Tracking (SMAT) [8] or Constrained Local Models [9], however there are several advantages from processing the whole set of frames together instead of in chronological order. With tracking methods there is the risk of "losing track" and the danger of "drift" accumulating over the sequence. Tracking methods use the previous frame for initialisation, however when considering the whole sequence another frame may be more appropriate. As our processing is offline, we process the whole set of frames simultaneously to obtain a more accurate feature correspondence between frames.

Accurate feature correspondence is essential for motion analysis, therefore the main contribution of this paper is a method of improving image registration across a video sequence. We find that a simple detection and clustering approach combined with a shortest path tree data structure improves the registration significantly.

Given our improved registration and a set of corresponding features, we build a statistical shape model [3] of variation over a set of facial motion videos. We then deduce the average distance travelled across each frame of the sequence to

evaluate the magnitude of the motion of each face, which is shown to correlate with subjective scores provided by human subjects.

3 Methodology

3.1 Initialisation and Clustering

The aim of the initialisation step is to find a set of regions r_i which encompass the face in each frame F_i. Additionally a set of cluster templates $\{X_j\}$ are computed and each region r_i is associated with a particular template X_j. We therefore aim to simultaneously detect and cluster the face region across a set of video frames.

Our method requires an initial starting region r_0 in the first frame of the sequence F_0. A template is generated from this region X_0. This template is then used to search subsequent frames F_i and the best matching region r_i selected. If the best match of the template falls below a threshold T then a new template X_j is created and added to the current list of templates. The current list of templates are ordered according to the their similarity to the current best match and the number of templates used to search any individual frame F_i is limited to k_{max}.

Formally the detection and clustering algorithm is as follows:-

Algorithm 1 Initialisation and Clustering Algorithm

1. Start with initial region r_0
2. Create template X_0 from region r_0 from frame F_0
3. Set frame number $i = 0$ and number of templates $k = 1$
4. $i => i + 1$
5. For frame F_i:-
   (a) Set $j = 0$
   (b) For template X_j compute best match score m_j
   (c) If $m_j < T$ then $j => j + 1$ Go to step 5b until $j = \min(k, k_{max})$
6. Store best matching template $j*$ with region r_i and match score m_{j*}
7. If best match $m_{j*} < T$ then
   (a) Store new template X_{j+1} from best matching region r_i
   (b) $k => k + 1$
8. Re-order templates according to similarity to current template X_{j*} or the new template X_{j+1}
9. Go to Step 4 until $i = i_{max}$

Note in our experiments the matching score m_j was computed using normalised correlation and a sliding window search around the previous frame region r_{i-1}. The normalised correlation threshold limit to create a new template is set to $T = 0.90$. The maximum number of templates applied to each frame was $k_{max} = 10$. These settings were able to cluster the FGNET [2] talking face sequence (see Figure 2) into 53 separate clusters. The generated templates $\{X_j\}$ are shown in Figure 1 as nodes of the shortest path tree. The construction of the tree is described in Section 3.2.

3.2 Shortest Path Trees

The clustering method described in Section 3.1 produces a set of templates $\{X_j\}$. We wish to construct a tree from this set of clusters which arranges similar clusters together and also minimises the distance from a given root template.

A suitable structure is the "shortest path" tree which minimises the total distance from the root to each node, over all nodes. Given a selected root node and distance measure between each pair of templates this can be computed using Dijkstra's algorithm [10].

The distance measure we use is the sum of squared pixel errors between the cluster templates, which is found to produce a shortest path tree which split the video sequence into associated templates. For example templates where the head is orientated in a particular direction are grouped together (see Figure 1).

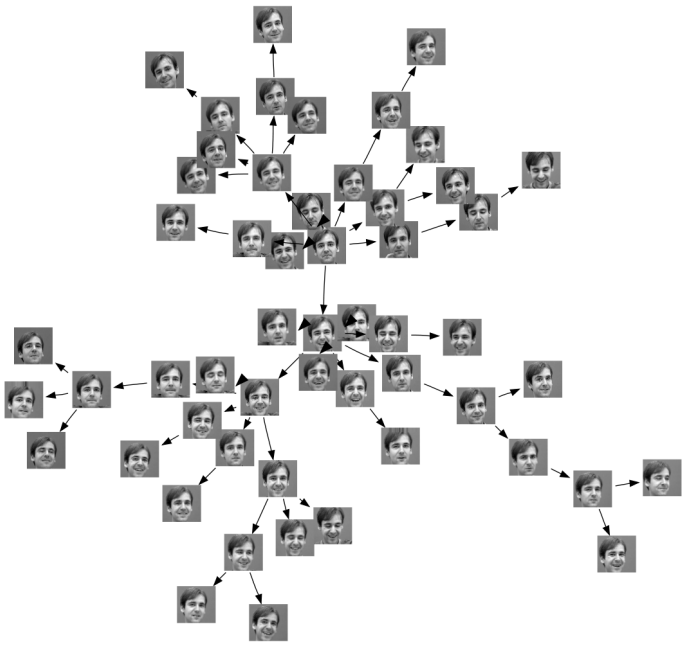


Fig. 1. FGNET talking face template cluster tree diagram

The root template selected to compute the shortest path tree is typically the template associated with the largest number of frames. One frame associated with this template will be labelled manually. Then the registration algorithm will construct a correspondence from this frame to all other frames. The shortest path tree minimises the distance from the labelled frame and therefore reduces the chances of the registration failing.

3.3 Groupwise Image Registration

Given a set of unlabelled images, we wish to automatically find corresponding points in all of the images. We use the method described by Cootes *et al.* [1] which optimises the total cost of encoding exact copies of the training set, the corresponding control points in each image being synonymous with a compact representation of the data.

Our groupwise approach uses a very simple mean image I_m and mean shape S_m, which has been shown to be effective in [1]. The training set images $\{F_i\}$ are then approximated by placing control points $\{c_i\}$ in each image, which represent a mapping from the mean shape S_m to each individual frame. The encoding cost of the training set C_{total} is as follows:-

$$C_{total} = C_{model} + C_{params} + C_{residuals} \tag{1}$$

Where C_{model} is the cost of encoding the model, C_{params} the cost of encoding the model and warp parameters and $C_{residuals} = \sum_i R_i$ is the cost of encoding the residual pixel errors over all frames. In our case C_{model} and C_{params} can be ignored, as the changes in encoding cost of the mean image I_m , mean shape S_m and warp parameters are negligible compared to the cost of encoding the residuals R_i for each frame.

Therefore the only variable to optimise is $\sum_i R_i$, which has two components for each image i.e. $R_i = S_i + T_i$. Where S_i is the cost of storing the residual shape variation between the mean shape S_m warped into each frame and the exact control points c_i. T_i is the cost of encoding the texture error over all pixels for each frame F_i given the warp defined by c_i. Since both S_i and T_i depend on the control points c_i, the principal task of registration is to optimise the control point location c_i in each frame of the video.

The groupwise algorithm proceeds by computing the mean image I_m and mean shape S_m from a set of starting control points $\{c_i\}$ and frames $\{F_i\}$. The control points c_i can then be independently optimised for each image F_i, to lower the total encoding cost C_{total}. When the control points c_i have been updated in each image, the mean image I_m and shape S_m are updated. We refer to the process of computing the mean model and adjusting the control points across all images as a registration "stage".

Typically a full registration consists of several such stages, which perform a image pyramid based coarse-to-fine registration of the whole data set with increasingly complex warps of the control points c_i in each image, eventually resulting in the individual control point co-ordinates being optimised at the maximum image resolution.

3.4 Incremental Tree Registration

The groupwise registration method described above can be applied incrementally given a clustered tree of templates (see Figure 1).

Each node of the tree actually represents a set of individual frames. Initially the groupwise registration algorithm can be applied to frames associated with

the root template in isolation. Then for each node of the tree the grid of control points can be initialised from the closest example in the parent cluster. Also the region r_i associated with frame F_i is used to warp the grid points from the closest frame. For each node, 16 registration "stages" are applied.

Each time a node registration is completed, the running mean texture and shape are updated, which maintains efficiency as each cluster node can then be processed individually, with the current mean and groupwise registration. The computational complexity of the incremental tree registration is the same as the single groupwise method in the sense that the number of times the control points c_i are optimised in each image is the same, whether the stages are applied to all images together or to separate image clusters.

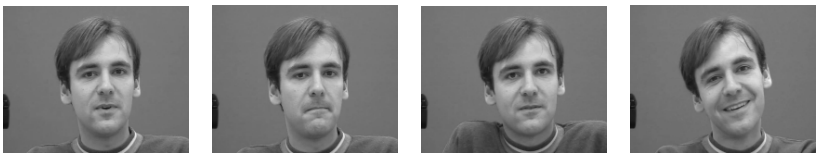
The rational behind the incremental approach is that similar frames will be processed together at the start, which is a relatively simple registration task. Later more difficult clusters will be registered and the mean updated, whilst the registration points in previous clusters remain unchanged.

If all frames are registered together in one single registration process (as described in Section 3.3) there is some risk that the outlier frames will distort the mean shape and texture. The incremental method is designed to reduce the affect of outliers by allowing the registration to concentrate on the largest template clusters in isolation, then attempt to fit to the more difficult images.

4 Experiments

4.1 Incremental Tree vs Single Groupwise Registration

The FGNET talking face data set [2] is used to test the performance of the tree based registration (see Section 3.4) and the simple groupwise registration (see Section 3.3). Example frames from this publicly available data set are shown in Figure 2.



(a) Frame 850 (b) Frame 1030 (c) Frame 2290 (d) Frame 4030

Fig. 2. Example of frames from the FGNET talking face sequence

The original FGNET data consists of 5000 frames of an individual talking in front of a camera. Each frame has been manually landmarked with 68 feature points. We select every 10^{th} frame to create a smaller data set of 500 images.

We applied the clustering algorithm described in Section 3.1 to deduce a set of regions r_i for each of the 500 frames F_i. On the first frame F_0 a regular grid of control points is generated over the region r_0.

We then applied the following two methods:-

1. Initialise the control grid in each frame using regions $\{r_i\}$ computed by the cluster method and run the single groupwise registration (see Section 3.3), for 56 stages.
2. Use the shortest path tree and regions $\{r_i\}$ to run the incremental groupwise registration (see Section 3.4), using 16 stages for each of the 53 nodes of the tree.

When applying single groupwise registration. The evolution of the mean image is shown in Figure 3.

(a) Stage 1 (b) Stage 7 (c) Stage 16 (d) Stage 36 (e) Stage 56

Fig. 3. Single groupwise mean texture on FGNET talking face data, after each stage of the registration

Figure 3 shows the mean image becoming sharper as the registration algorithm progresses. The mean image is initially blurred (see Figure 3(a)) but as the grid points in each image move to equivalent locations in each frame the image becomes sharper (see Figure 3(e))

When applying the incremental tree method, the progression of the mean images is different, see Figure 4.

(a) Node 1 (b) Node 5 (c) Node 16 (d) Node 43 (e) Node 53

Fig. 4. Incremental tree registration mean texture on FGNET talking face data, after registration of each tree node

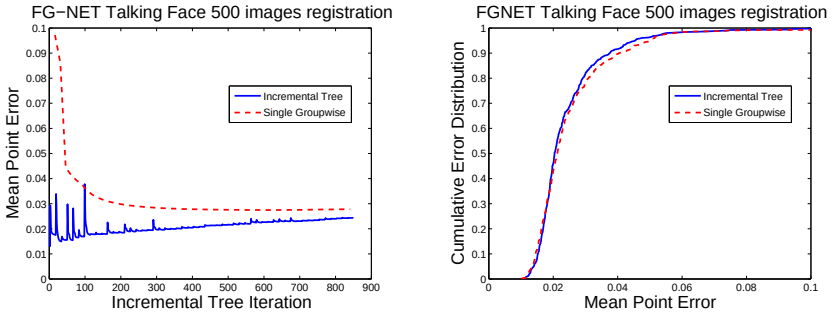
Figure 4(a) shows the mean image after completing the groupwise registration on the root template cluster, which consists of 16 stages on the 30 images in this

cluster. The mean image is quite sharp, which reflects the fact that all the images in the first cluster are similar. The later mean images, after completing the registration on the later clusters show similar sharpness, but the mean texture and mean shape are updated as more clusters are added. The final mean image of the incremental registration method (Figure 4(e)) looks similar to the single groupwise mean image (see Figure 3(e)), because at this point all the data has been registered.

Figure 5(a) shows the accuracy of the registration methods relative to the ground truth, for both the incremental and single groupwise approaches. The registration accuracy (mean point error on y axis) is determined at each stage using the following method:-

1. Warp all the manually labelled ground truth points from each image into the reference frame (using the computed control points c_i in each image)
2. Compute the mean x and y co-ordinates for each point in the reference frame
3. Warp the mean points in the reference frame back to the original images
4. Compute the average point to point error over all points and images

This determines the consistency of the registration relative to the ground truth. The average point to point error is normalised by the ground truth inter-ocular eye separation, to give an accuracy measure which is independent of image size.



(a) Errors after each iteration

(b) Cumulative error of final image registration

Fig. 5. Single Groupwise vs Incremental Tree Registration on FGNET data

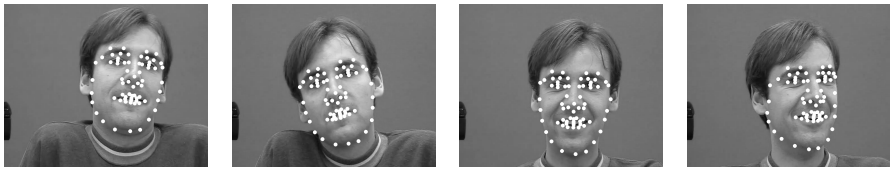
The dashed line on Figure 5(a) shows that when applying the single groupwise registration to the FGNET data the point to point error gradually decreases over time. Therefore with each iteration the registration improves, until the error reduces to 0.0275. Which represents an average point to point distance error of 2.75% (standard error 0.21%) of the eye separation, i.e. a reasonably accurate registration.

When applying the incremental tree registration (see solid line Figure 5(a)) the pattern of point to point errors over time is different. For each individual cluster (i.e. every 16 stages) the point to point error first rises as the new cluster is added, due to the new images being only approximately registered. Within each cluster the error decreases as the registration is applied to the new images. However as more clusters are added the registration task becomes more difficult, which steadily increases the registration error over time, resulting in the jagged pattern shown by the solid line Figure 5(a).

However the final incremental registration error on the whole data set is 2.43% (standard error 0.05%) of the eye pupil separation, which is a 12% improvement over the single groupwise approach. The small standard errors indicate that the incremental method achieves a statistically significant better result for the same computational effort compared to the single groupwise approach.

The cumulative error of the final registration for both methods is plotted in Figure 5(b), which shows that 90% of the images marked up using the incremental tree method (solid line) have $< 4\%$ mean point error relative to the ground truth, which is greater than the 86% success rate achieved by the single groupwise approach (dashed line) at the 4% threshold.

Example automatically found feature points from the incremental tree registration are shown in Figure 6.



(a) Frame 1250 (b) Frame 1820 (c) Frame 4260 (d) Frame 4510

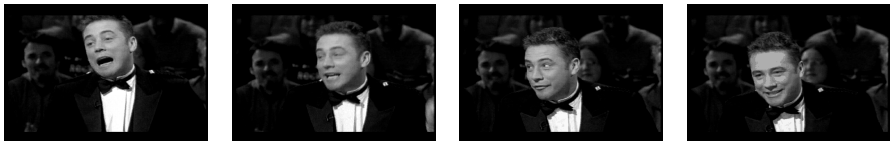
Fig. 6. Example of automatic labelling on FG-NET talking face sequence

This process simulates the user manually landmarking the mean image then generating the labelled points automatically from the registration. The whole data set (500 images) can be labelled accurately by only manually marking one image (see Figure 6).

4.2 Facial Motion Experiment

This experiment was designed to assess the ability of human volunteers to recognise celebrities under difficult conditions (the celebrities being famous people in the country where the experiment was performed, i.e. the UK). The videos are poor quality images collected from television broadcasts. See Figure 7 for two examples from the 59 sequences used.

The psychological experiment tested the hypothesis that human face recognition is more successful when observing celebrities who show a large amount



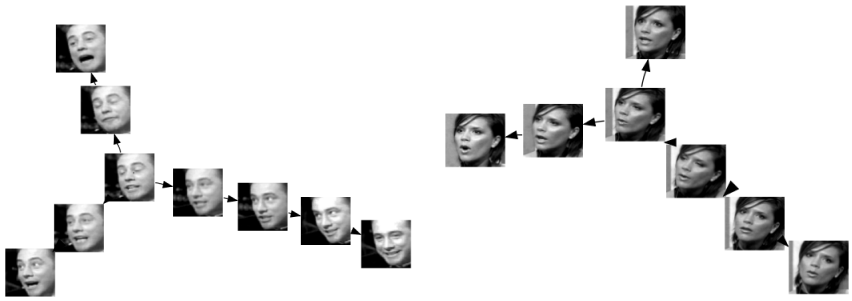
(a) McCoist 1 (b) McCoist 15 (c) McCoist 40 (d) McCoist 63

(e) VBeckham 1 (f) VBeckham 13 (g) VBeckham 55 (h) VBeckham 63

Fig. 7. Example frames from two video sequences of 'celebrities'

of facial movement. As part of this experiment the volunteer subjects rated the amount of facial motion present in each video (giving a subjective score from 0-9). To corroborate the manual motion ratings, we manually marked up one frame of each sequence and using incremental tree registration (Section 3) to automatically mark the whole set of frames for each video. We then compared the amount of shape variation for each individual with the human rankings.

Figure 8 shows the shortest path trees computed from the videos, by taking the most common template as the root node, for two individuals. In both cases the shortest path tree divides the templates extracted from the 64 frame video sequence in a visually sensible form.

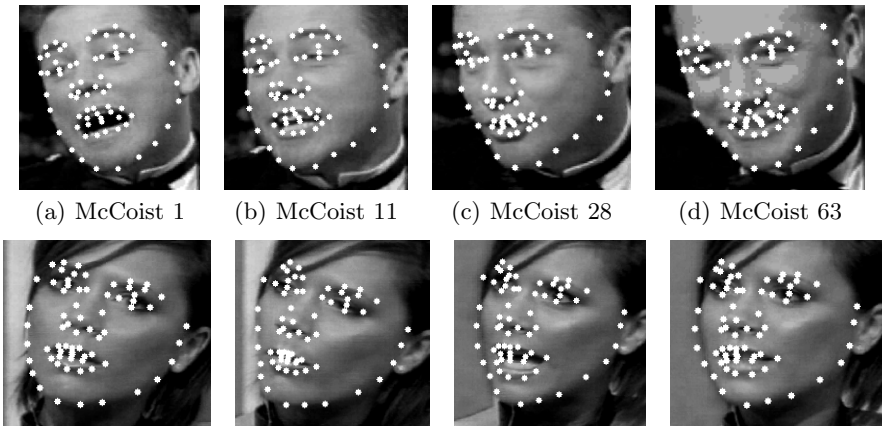


(a) Ali McCoist (b) Victoria Beckham

Fig. 8. Example shortest path trees built from 2 of the 59 celebrity videos sequences

Results of automatic annotation for the two sequences are shown in Figure 9, which show that the incremental registration was successful. The method is able

to cope with the opening and closing of the subject's mouth and head variation as the celebrity speaks, approximately facing the camera in each video.



(a) McCoist 1 (b) McCoist 11 (c) McCoist 28 (d) McCoist 63

(e) VBeckham 1 (f) VBeckham 13 (g) VBeckham 36 (h) VBeckham 62

Fig. 9. Example of automatic labelling for two video sequences of 'celebrities'

Given automatically labelled points for each of the 59 sequences. A shape model of the whole set of (64*59=3776) images was built and the euclidean distance travelled in shape space calculated for each of the 59 individuals. This automatic estimate of facial motion was then used to rank the individuals and the ranks compared with the human perceived motion ranks. The top ten ranks are shown in Figure 10.

Rank	Human Perceived Motion	Automatic (Shape Space Distance)
1	Graham Norton	Graham Norton
2	Victoria Beckham	Jermemy Clarkson
3	Jermemy Clarkson	Chris Tarrant
4	Matt Perry	Michael Parkinson
5	Bruce Forsyth	Bruce Forsyth
6	Angus Deayton	Robin Williams
7	Jack Nicholson	Sanjeev Bhaskar
8	Stephen Fry	Ali McCoist
9	Chris Tarrant	Stephen Fry
10	Jennifer Lopez	Victoria Beckham

Fig. 10. First 10 individual faces ranked by human perceived motion ratings (column 2) and automatic facial motion estimate using distance travelled in the shape space (column 3)

In Figure 10, 7 out of the top 10 individuals are the same. Indicating a strong correlation between the automatic computer estimate of facial motion and the motion perceived by humans. A spearman rank correlation test confirms this with a correlation value of 0.68 which is significant at the $p < 0.01$ level for a sample size of 59, against the null hypothesis of random permutations.

5 Conclusions

This paper has described an incremental registration method which is shown to outperform a single groupwise registration on a publicly available data set [2]. The method relies on a simple clustering algorithm and the building of a shortest path tree to cluster the data, see Figure 1.

The tree structure allows the incremental registration of the largest cluster of data first and leaves the registration of outlier data until later. The resulting registration is shown to be 12% more accurate than a simple groupwise approach. Due to the incremental computation of the mean image and shape, the tree registration method has the same complexity as the groupwise method.

The incremental method was successfully applied to low quality videos collected from television broadcasts and the resulting motion analysis was able to achieve a high correlation with human facial motion perception. The approach is general and could easily be applied to other types of video motion analysis.

References

1. Cootes, T., Twining, C., V.Petrović, R.Schestowitz, Taylor, C.: Groupwise construction of appearance models using piece-wise affine deformations. In: 16^{th} British Machine Vision Conference 2005, Oxford, England. Volume 2. (2005) 879–888
2. FGNET: Face and gesture network: Talking face database http://www.isbe.man.ac.uk/~bim/data/talking_face/talking_face.html.
3. Cootes, T.F., Taylor, C.J., Cooper, D., Graham, J.: Active shape models - their training and application. Computer Vision and Image Understanding **61**(1) (January 1995) 38–59
4. Valstar, M., Pantic, M.: Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In: Proceedings of IEEE Workshop on Human Computer Interaction. (2007)
5. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.: Face recognition: A literature survey. ACM Computing Surveys (2003) 399–458
6. Baker, S., Matthews, I., J.Schneider: Automatic construction of active appearance models as an image coding problem. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(10) (2004) 1380–84
7. Zitová, B., Flusser, J.: Image registration methods: A survey. Image and Vision Computing **21** (2003) 977–1000
8. Dowson, N., Bowden, R.: Simultaneus modeling and tracking (smat) of feature sets. In: 23^{rd} Computer Vision and Pattern Recognition Conference 2005, San Diego, USA. (2005) 99–105
9. Cristinacce, D., Cootes, T.: Detection and tracking with constrained local models. In: 17^{th} British Machine Vision Conference 2006, Edinburgh, Scotland. (2006) 929–938
10. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, Second Edition- Dijkstra's algorithm, pp.595601. MIT Press (2001)