Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics

M.F. Valstar and M. Pantic

Department of Computing, Imperial College London
180 Queen's gate, London SW7 2AZ, England
{Michel.Valstar, M.Pantic}@imperial.ac.uk

Abstract. The analysis of facial expression temporal dynamics is of great importance for many real-world applications. Being able to automatically analyse facial muscle actions (Action Units, AUs) in terms of recognising their neutral, onset, apex and offset phases would greatly benefit application areas as diverse as medicine, gaming and security. The base system in this paper uses Support Vector Machines (SVMs) and a set of simple geometrical features derived from automatically detected and tracked facial feature point data to segment a facial action into its temporal phases. We propose here two methods to improve on this base system in terms of classification accuracy. The first technique describes the original time-independent set of features over a period of time using polynomial parametrisation. The second technique replaces the SVM with a hybrid SVM/Hidden Markov Model (HMM) classifier to model time in the classifier. Our results show that both techniques contribute to an improved classification accuracy. Modeling the temporal dynamics by the hybrid SVM-HMM classifier attained a statistically significant increase of recall and precision by 4.5% and 7.0%, respectively.

1 Introduction

A system capable of analysing facial actions would have many applications in a wide range of disciplines. For instance, in medicine, it could be used to continuously monitor a patients pain level or anxiety, in gaming a virtual avatar could be directed to mimic the user's facial expressions and in security the analysis of facial expressions could be used to assert a person's credibility.

The method proposed in this paper is based on the analysis of atomic facial actions called Action Units (AUs), which are defined by the Facial Action Coding System (FACS) [1]. FACS is the best known and the most commonly used system developed for human observers to objectively describe facial activity in terms of visually observable facial muscle actions (AUs). FACS defines 9 upper face AUs and 18 lower face AUs. There are 5 additional AUs that occur irregularly in interpersonal communication and are therefore often not mentioned in the literature[2].

Previous work on automatic AU detection from videos includes automatic detection of 16 AUs from face image sequences using lip tracking, template matching and neural networks [3], detecting 20 AUs occurring alone or in combination by using temporal templates generated from input face video [4] and detection of 18 AUs using wavelets, AdaBoost and Support Vector Machines [5]. For an overview of the work done on AU and emotion detection from still images or face video the reader is referred to [6, 7].

Many of the aforementioned applications of automatic facial expression analysis require the explicit analysis of the temporal dynamics of AU activation. The body of research in cognitive sciences which suggests that the temporal dynamics of human facial behaviour (e.g. the timing and duration of facial actions) are a critical factor for interpretation of the observed behaviour, is large and growing [8–10]. Facial expression temporal dynamics are essential for categorisation of complex psychological states such as various types of pain and mood [11]. They are also the key parameter in differentiation between posed and spontaneous facial expressions [12, 13, 10]. For instance, it has been shown that spontaneous smiles, in contrast to posed smiles (such as a polite smile), are slow in onset, can have multiple AU12 apices (multiple peaks in the mouth corner movement), and are accompanied by other facial actions that appear either simultaneously with AU12 or follow AU12 within 1s [8]. Recently the automatic facial expression recognition community has published work that emphasises the importance of facial expression temporal dynamics for deception detection [14].

In light of the aforementioned, it is striking that very few research groups focus on the analysis of temporal dynamics. Almost all existing facial expression recognition systems are binary classification systems that are only capable of recognising the presence of a facial action, regardless whether that action has just begun and is getting stronger, is at its peak or is returning to its neutral state. Also, while many systems are in essence capable to compute the total duration of a facial action based on this activation detection, none do this explicitly [7]. What's more, this total activation duration information alone would be insufficient for the complex tasks described above.

A facial action, in our case an AU activation, can be in any one of four possible phases: (i) the onset phase, where the muscles are contracting and the appearance of the face changes as the facial action grows stronger, (ii) the apex phase, where the facial action is at its peak, (iii) the offset phase, where the muscles are relaxing and the face returns to its neutral appearance and (iv) the neutral phase, where there are no signs of this particular facial action. Often the order of these phases is neutral-onset-apex-offset-neutral, but other combinations such as multiple-apex facial actions are possible as well.

Only recently has the first system been proposed that is capable to explicitly model the temporal dynamics of facial actions, in terms of the phases neutral, onset, apex or offset [4]. The authors proposed to classify each frame of a video into one of the four temporal phases using features computed from tracked facial points and a multiclass Support Vector Machine (SVM). The authors of that work have successfully used this phase detection to define a set of high-level

parameters, such as the duration of each phase or the number of apices in the video. Using this high-level representation discrimination between posed and spontaneous brow actions [14] was possible.

Strangely though, the multiclass-SVM classification strategy adopted in their work does not incorporate any model of time. The dynamics are completely modeled by mid-level parameters such as the current speed of a facial point. We believe that the expressive power of the mid-level parameters proposed in [4] can be improved upon to better describe the characteristics of AU temporal dynamics.

The method we propose is fully automatic, operating on videos of subjects recorded from a near-frontal view. It uses data from 20 tracked facial points to analyse facial expressions. We propose two ways to improve on the work by Valstar and Pantic [4]. First we define a set of mid-level parameters that better encodes the dynamics of facial expressions. Instead of defining the features for each frame, we describe the evolution of the feature values over a period in time using polynomial parametrisation. This way, we capture how a feature related to a specific facial action behaves dynamically instead of observing its value at a single moment.

The second improvement we propose is to explicitly incorporate a sense of time in the classification procedure. We do so by combining a Hidden Markov Model (HMM) with a SVM. While the evolution of a facial action in time can be efficiently represented using HMMs, the distinction between the temporal phases at a single point in time is usually made using Gaussian mixture models, which do not offer a high discrimination. SVMs on the other hand are large-margin classifiers known for their excellent discrimination performance in binary decision problems but they do not incorporate a model of time. We will show that the combination of the high-margin SVMs with the excellent temporal modeling properties of HMMs increases the recognition performance significantly.

2 Automatic feature extraction

The features used in our study are extracted by a fully automatic method consisting of, consecutively, face detection, facial point detection, facial point tracking and the calculation of geometry based features. These features are used to train and test the classifier combination described in section 3. We will now describe each subsystem in some detail.

To detect the face in a scene we make use of a real-time face detection scheme proposed in [15], which represents an adapted version of the original Viola-Jones face detector [16]. The Viola-Jones face detector consists of a cascade of classifiers trained by AdaBoost. Each classifier employs integral image filters, which remind of Haar Basis functions and can be computed very fast at any location and scale. This is essential to the speed of the detector. For each stage in the cascade, a subset of features is chosen using a feature selection procedure based on AdaBoost.

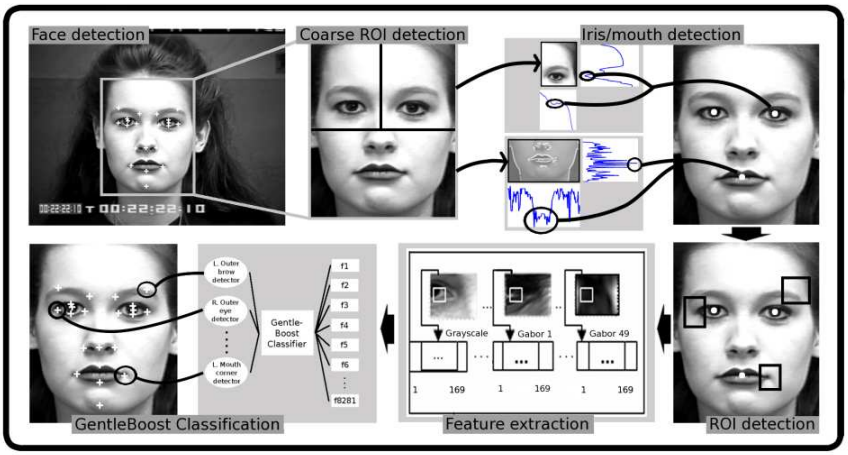

Fig. 1. Outline of the facial point detection system.

The adapted version of the Viola-Jones face detector that we employ uses GentleBoost instead of AdaBoost. GentleBoost has been shown to be more accurate and converges faster than AdaBoost [17]. At each feature selection step (i.e., for every feature selected by AdaBoost), the proposed algorithm refines the feature originally proposed by AdaBoost. The algorithm creates a new set of filters generated by placing the original filter and slightly modified versions of the filter at a two pixels distance in each direction.

The method that we use for fully automatic detection of 20 facial feature points plus the irises and the centre of the mouth in a face image, uses Gabor-feature-based boosted classifiers as proposed in [18]. The method, outlined in Fig. 1, assumes that the input image is a face region, such as the output of the face detection algorithm explained above. In this face region, the irises and the medial point of the mouth are detected first. A combination of heuristic techniques based on the analysis of the vertical and horizontal histograms of the upper and the lower half of the face-region image achieves this. Based on these three points and anthropomorphic rules, the input face region is divided into 20 regions of interest (ROIs), each corresponding to a facial point to be detected.

For each pixel in the ROI, a feature vector is computed that consists of the grey values of the 13x13 patch surrounding the pixel and the responses to 48 Gabor filters (8 orientations and 6 spatial frequencies, 2:12 pixels/cycle at 1/2 octave steps). This feature vector is used to learn the pertinent point's patch template and, in the testing stage, to predict whether the current point represents a certain facial point or not.

To capture all facial motion, we track the detected points through frames of the input video. The algorithm we used to track these facial points is Particle Filtering with Factorised Likelihoods (PFFL) [19]. We used the observation model proposed in [20], which is both insensitive to variations in lighting and able to cope with small deformations in the template. This polymorphic aspect is necessary as many areas around facial points change their appearance when a facial action occurs (e.g. the mouth corner in a smile). The facial point tracking scheme results for every image sequence with n frames in a set of points P with dimensions $20 * 2 * n$.

For all points $\boldsymbol{p_i} \in P$, where $i = [1 : 20]$ denotes the facial point, we compute first two features for every frame j to encode the y and the x coordinate deviation of a point relative to their position in the first frame of an input image sequence:

$$f_1\left(\boldsymbol{p_i}, t\right) = p_{i,y,t} - p_{i,y,1} \tag{1}$$
$$f_2\left(\boldsymbol{p_i}, t\right) = p_{i,x,t} - p_{i,x,1} \tag{2}$$

For all pairs of points $\boldsymbol{p_i}$, $\boldsymbol{p_j}$,$i \neq j$ we compute in each frame three features:

$$f_3(\boldsymbol{p_i}, \boldsymbol{p_j}, t) = \|p_{i,t} - p_{j,t}\| \tag{3}$$
$$f_4(\boldsymbol{p_i}, \boldsymbol{p_j}, t) = f_3(\boldsymbol{p_i}, \boldsymbol{p_j}, t) - \|p_{i,1} - p_{j,1}\| \tag{4}$$
$$f_5(\boldsymbol{p_i}, \boldsymbol{p_j}, t) = \arctan\left(\frac{\|p_{i,y,t} - p_{j,y,t}\|}{\|p_{i,x,t} - p_{j,x,t}\|}\right) \tag{5}$$

These features correspond to the distance between two points, the distance between two points relative to their distance in the first frame and the angle made by the line connecting two points and the horizontal axis. Also, to capture some of the temporal dynamics of a facial expression, we compute the first temporal derivative df/dt of all above defined features. This results in a set F of 840 features per frame.

The features f_1-f_5 are computed using the tracking information of at most two frames. As such, their temporal scope is limited. Not only does this mean that it is hard to model continuous behaviour of facial actions, it also makes the system very sensitive to inaccuracies of the tracker. A single tracking error will likely result in an outlier in our data representation. This, in turn, will lead to lower classification accuracy. To increase the temporal scope, we add a second set of features. Within a temporal window of duration T we describe each of the original mid-level parameters f_1-f_5 with a p^{th} order polynomial. We choose T and p such that with a given framerate we can accurately describe the feature shape in the fastest facial segment (the onset of AU45, a blink). For our data, recorded at 25 frames per second, this results in $T = 7$ and $p = 2$. Thus, the mid-level parameters f_{11}-f_{25} are found to be the values that fit the polynomial best:

$$f_k = f_{2k+1}t^2 + f_{2k+2}t + f_{2k+3}\ ,\ k \in [1 \ldots 5] \tag{6}$$

The addition of the polynomial description of mid-level parameters adds another 1260 features per frame, bringing the total feature dimensionality to 2100.

3 Hybrid classification

While the temporal dynamics of a facial action can be represented very efficiently by HMMs, the multiclass classification of the features on a frame-by-frame basis is normally done using Gaussian mixture models as the emission probabilities. These Gaussian mixtures are trained by likelihood maximisation, which assumes correctness of the models and thus suffers from poor discrimination [21]. It results in mixtures trained to model each class and not to discriminate one class from the other.

SVMs on the other hand discriminate extremely well. Using them as emission probabilities might very well result in an improved recognition. We therefore train a set of SVMs, one for every combination of classes (i.e., temporal phases neutral, onset, apex, and offset) and use their output to compute emission probabilities. This way we effectively have a hybrid SVM-HMM system. This approach has been previously applied with success to speech recognition [22].

Unfortunately we cannot use the output of a SVM directly as a probability measure. The output $h(\boldsymbol{x})$ of a SVM is a distance measure between a test pattern and the separating hyper plane defined by the support vectors. There is no clear relationship with the posterior class probability $p(y = +1|\boldsymbol{x})$ that the pattern $\boldsymbol{x}$ belongs to the class $y = +1$. Fortunately, Platt proposed an estimate for this probability by fitting the SVM output $f(\boldsymbol{x})$ with a sigmoid function [23]:

$$p(y = +1|\boldsymbol{x}) = g(h(\boldsymbol{x}), A, B) \equiv \frac{1}{1 + \exp(Ah(\boldsymbol{x}) + B)} \tag{7}$$

The parameters A and B of eq.(7) are found using maximum likelihood estimation from a training set p_i, t_i with $p_i = g(f(\boldsymbol{x}_i, A, B)$ and target probabilities $t_i = (y_i + 1)/2$. The training set can but does not have to be the same set as used for training the SVM.

Since SVMs are *binary* classifiers we use a one-versus-one approach to come to a multiclass classifier. This approach is to be preferred over the one-versus-rest approach as it aims to learn the solution to a more specific problem, namely, distinguishing between one class from one other class at a time. For this pairwise classification we need to train $K(K-1)/2$ SVMs, where in our case $K = 4$ is the number of temporal phases.

Our HMM consists of four states, one for each temporal phase. For each SVM we get, using Platt's method, pairwise class probabilities $\mu_{ij} \equiv p(q_i|\text{or}q_j, \boldsymbol{x})$ of the class (HMM state) q_i given the feature vector $\boldsymbol{x}$ and that $\boldsymbol{x}$ belongs to either q_i or q_j. These pairwise probabilities are transformed into posterior probabilities $p(q_i|\boldsymbol{x})$ by

$$p(q_i|\boldsymbol{x}) = 1/\left[\sum_{j=1, j\neq i}^{K} \frac{1}{\mu_{ij}} - (K-2)\right] \tag{8}$$

Finally, the posteriors $p(q|\boldsymbol{x})$ have to be transformed into *emission probabilities* by using Bayes' rule

$$p(\boldsymbol{x}|q) \propto \frac{p(q|\boldsymbol{x})}{p(q)} \tag{9}$$

where the a-priori probability $p(q)$ of class q is estimated by the relative frequency of the class in the training data.

4 Evaluation

We have evaluated our proposed methods on 196 videos selected from the MMI-Facial Expression Database [24], containing videos of 23 different AUs. We have chosen this database, instead of for example the Cohn-Kanade DFAT-504 dataset [25], because the videos in the MMI-Facial Expression Database display the full neutral-expressive-neutral pattern. This is essential, as it is this temporal pattern of facial actions that we are interested in. The videos were chosen so that we selected at least 10 videos of every AU we want to analyse.

The tests investigate the effects of our two proposals: adding the polynomial features described in section 2, and using the hybrid SVM-HMM system described in section 3. For ease of reference, in this section we denote the system using only the mid-level parameters $f_1 \dots f_{10}$ as the Traude system (for tracked action unit detector) and the system that uses all 25 mid-level parameters Traude-Plus. Thus we have compare four methods: Traude with SVM, Traude with SVM-HMM, Traude-Plus with SVM and Traude-Plus with SVM-HMM. A separate set of classifiers was trained for each AU. All methods perform feature selection using GentleBoost before training their respective classifiers (see [4] for details on feature selection using GentleBoost). Evaluation was done using 10-fold cross validation and measured the number of correctly classified frames, where the classes are neutral, onset, apex or offset. Table 1 shows the F1-value results per AU. The F1-value ϕ is a measure of performance that values recall as important as precision and is computed as follows:

$$\phi = \frac{2ab}{a+b} \tag{10}$$

where a is the recall and b the precision of a test. As we can see from table 1, the hybrid SVM-HMM method outperforms the SVM method for almost every AU. The effect of adding features $f_{11} \dots f_{25}$ is not that obvious. However, the mean of the results does suggest that the parametric features do benefit performance. We also see that the effect of the new features is more pronounced for the SVM-HMM classifier. This is another evidence that the new features indeed capture the temporal dynamics of the facial action, of which the SVM-HMM classifier makes full use.

Figure 4 shows the relative increase of performance of the temporal segments separately. In this figure we have averaged the results per temporal phase over all AUs. Again, from inspection we see that the SVM-HMM approach outperforms

Table 1. F1-value for the four tested systems, for all Action Units (AUs).

AU	Traude SVM	TraudePlus SVM	Traude SVM-HMM	TraudePlus SVM-HMM
1	0.614	0.612	0.703	0.714
2	0.590	0.657	0.663	0.734
4	0.584	0.570	0.623	0.607
5	0.447	0.480	0.482	0.560
6	0.579	0.496	0.672	0.614
7	0.275	0.289	0.362	0.327
9	0.699	0.638	0.763	0.790
10	0.620	0.656	0.715	0.678
12	0.719	0.780	0.797	0.786
13	0.817	0.804	0.836	0.839
15	0.491	0.523	0.583	0.612
16	0.615	0.620	0.581	0.637
18	0.719	0.723	0.742	0.726
20	0.732	0.731	0.801	0.804
22	0.570	0.611	0.677	0.658
24	0.443	0.380	0.393	0.379
25	0.694	0.715	0.705	0.735
26	0.582	0.607	0.636	0.607
27	0.743	0.686	0.840	0.815
30	0.527	0.484	0.564	0.595
43	0.662	0.710	0.716	0.778
45	0.685	0.718	0.714	0.787
46	0.378	0.382	0.332	0.397
Mean:	0.599	0.603	0.648	0.660

the SVM method. Shown are the recall, precision separately and finally the F1-measure.

We now turn our attention to the significance of our results. We first compare the results of the Traude with the TraudePlus methods. When we consider the averaged performance rate of all AUs and all temporal segments together, the recall of TraudePlus is 1.8% better and the precision is 0.5% better. However, these performance increases are not significant at a 95% confidence level. Comparing the SVM method with the SVM-HMM method, we find that the recall of SVM-HMM is 4.5% better and the precision even 7% better. Both performance increases *are* significant at a 95% confidence level.

When we take a closer look at the results of the detection of the temporal segments, we see that the detection of the offset phase has increased most from introducing the HMM, achieving an increase in recall of 6% and an increase in precision of 10.6%. The apex phase is a good second, with a recall increase of 10.1% and a precision increase of 5.1%. The neutral phase benefits least from the addition of the HMM. This is as expected, because by its very nature it is not a dynamic part of the facial action, apart from tracker noise the facial points are stationary during the neutral phase.

5 Conclusion

We have presented and evaluated two methods to improve the analysis of a facial action's temporal dynamics in terms of classifying the temporal phases neutral, onset, apex and offset. From our results we can conclude that replacing

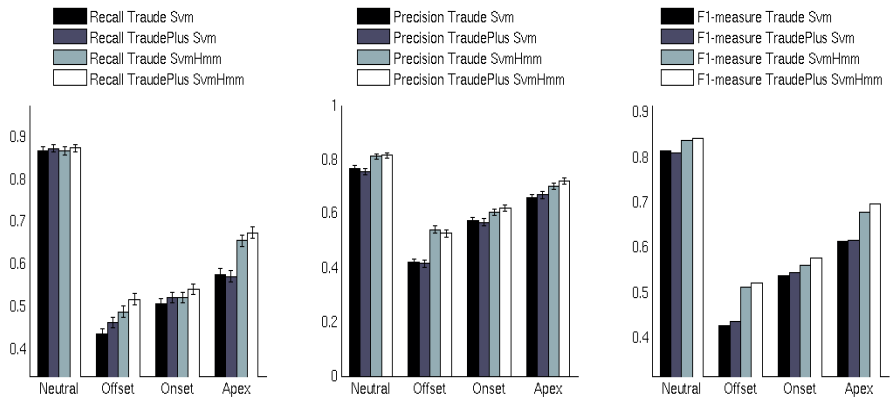

Fig. 2. Comparison of the classification results shown per temporal phase (onset, apex, offset and neutral). The results shown are the average over all 23 AUs, error bars depict the standard deviation.

the SVM classifier with a hybrid SVM-HMM classifier results in a significant classification accuracy improvement. It is harder to judge the effect of adding the polynomial representation of the original mid-level parameters. Using these features, there seems to be a slight but statistically insignificant improvement, which is more pronounced in combination with the SVM-HMM classifier than it is with the original SVM classifier. However, the performance increase due to the polynomial mid-level parameters is not significant and might not be high enough to make up for the fact that we have to compute 2.5 times as many features.

References

1. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System. A Human Face (2002) Salt Lake City.
2. Cohn, J.F.: Foundations of human computing: Facial expression and emotion. Proc. ACM Int'l Conf. Multimodal Interfaces **1** (2006) 610–616
3. Y. Tian, T.K., Cohn, J.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Analysis and Machine Intelligence **23**(2) (2001) 97–115
4. Valstar, M.F., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. (2006) 149
5. Bartlett, M., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.: Machine learning methods for fully automatic recognition of facial expressions and actions. Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics **1** (2004) 592–597
6. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human-computer interaction. Proc. IEEE **91**(9) (2003) 1370–1390
7. Tian, Y.L., Kanade, T., Cohn, J.F.: Handbook of Face Recognition. Springer (2005) New York.

8. Cohn, J.F., Schmidt, K.L.: The timing of facial motion in posed and spontaneous smiles. J. Wavelets, Multi-resolution and Information Processing **2**(2) (2004) 121–132
9. Bassili, J.N.: Facial motion in the perception of faces and of emotional expression. J. Experimental Psychology **4**(3) (1978) 373–379
10. Hess, U., Kleck, R.E.: Differentiating emotion elicited and deliberate emotional facial expressions. European J. of Social Psychology **20**(5) (1990) 369–385
11. de C. Williams, A.C.: Facial expression of pain: An evolutionary account. Behavioral and Brain Sciences **25**(4) (2006) 439–488
12. Ekman, P.: Darwin, deception, and facial expression. Annals of New York Ac. of sciences **1000** (2003) 105–221
13. Ekman, P., Rosenberg, E.L.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System. Oxford University Press, Oxford (2005) Oxford, UK.
14. Valstar, M.F., Pantic, M., Ambadar, Z., Cohn, J.F.: Spontaneous vs. posed facial behavior: automatic analysis of brow actions. Proc. ACM Intl. conf. on Multimodal Interfaces (2006) 162–170
15. Fasel, I., Fortenberry, B., Movellan, J.: A generative framework for real time object detection and classification. Comp. Vision, and Image Understanding **98**(1) (2005) 181–210
16. Viola, P., Jones, M.: Robust real-time object detection. Technical report CRL 200001/01 (2001)
17. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. The Annals of Statistics **28**(2) (2000) 337–374
18. Vukdadinovic, D., Pantic, M.: Fully automatic facial feature point detection using gabor feature based boosted features. Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics (2005) 1692–1698
19. Patras, I., Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features. Proc. Int'l Conf. Automatic Face & Gesture Recognition (2004) 97–102
20. Patras, I., Pantic, M.: Tracking deformable motion. Proc. Int'l Conf. Systems, Man and Cybernetics (2005) 1066–1071
21. Bourlard, H., Morgan, N.: Hybrid hmm/ann systemds for speech recognition: Overview and new research directions. Lecture Notes in Artificial Intelligence (1998) 389–417
22. Kruger, S., Schaffner, M., Katz, M., Andelic, E., Wendemuth, A.: Speech recognition with support vector machines in a hybrid system. In: Interspeech. (2005) 993–996
23. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. (2000) 61–74 Cambridge, MA.
24. Pantic, M., Valstar, M.F., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. Proc. Int'l Conf. Multimedia & Expo (2005) 317–321
25. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: IEEE Int'l Conf. on Automatic Face and Gesture Recognition. (2000) 46–53