2009 IEEE Workshop on Advanced Robotics
and its Social Inpacts
Tokyo, Japan, November 23-25, 2009

A Hierarchal Approach to Facial Expression Recognition using Active Appearance Models for Service Robots

Abhishek Sharma
Dept. of Electrical and Electronics Engineering
Birla Institute of Technology, Mesra, India
abhishek.s@acm.org

Ronny Stricker
Neuroinformatics and Cognitive Robotics Lab
Ilmenau University of Technology, Germany
ronny.stricker@tu-ilmenau.de

***Abstract*— For better interaction between human beings and service robots we need to detect the emotional state of the human user. This can help the robot adapt to the user's needs and can be used to increase the naturalness of the interaction. Our work proposes a hierarchal approach to facial expression recognition based extracting some facial features with the help of an Active Appearance Model. We introduce a concept of dynamic feature parameters and then make use of both local static features and dynamic features for classification.**

Keywords- Active Appearance Model, Facial Expression Recognition, Service robots

I. Introduction

There has been a growing incidence of service robots and our work focuses on improved natural human-robot interaction vital for such robots. An important part of such an interaction would be improved systems to classify facial expressions of the human partner. Active appearance models, on the other hand, are a good way to characterize non rigid objects like human heads. The Active Appearance Model has been introduced by [1] , and some methods to classify facial expressions using AAM have already been proposed in [2], [3], [4] and [5]. In contrast to other methods our approach uses local features of the AAM fitted shape on an unknown face to classify the expression/emotion using a multi-class SVM classifier. Additionally, we introduce a concept of dynamic and static features set which is used for the hierarchal approach of classification that we present. The benefit of using local features and creating a dynamic feature set is that it is independent of the face contours and differences across persons in positions of the facial features.

This paper first discusses the techniques used and subsequently, we present the results on our tests on the Feedtum [6] database and at real-time on a service robot based on SCITOS [7]. We make fit the Active Appearance Model (AAM) onto an unknown face detected at real-time. This enables us to accurately get the points on the face which is used by our algorithm for expression recognition. Our algorithm classifies the expressions into 6 basic expression groups (Happy, Surprise, Sad, Neutral, Anger and Disgust).

Abhishek Sharma would like to thank Prof. Horst-Michael Großand all the members of the Neuroinformatics and Cognitive Robotics Lab at Ilmenau University of Technology (Technische Universität Ilmenau), Germany. His research internship stay at the lab in TU Ilmenau was funded by a scholarship from *Deutscher Akademischer Austausch Dienst(DAAD)* [German Academic Exchange Service]. This work was carried out under the supervision of Dipl. Inf. Ronny Stricker and Prof. H.-M. Groß.

If detected with fair accuracy, a service robot can make use of the emotional state of the human partner and decide to take the further course of action. For example, modify its action if the human is disgusted or angry, or proceed with its action if he/she is happy or neutral. The method proposed in [8], although a good attempt to classify using useful local facial features fails to achieve good results for classification into the target groups. Thus, we have proposed a hierarchal classification (Fig. 1 and Fig. 9) with separate models and the most suitable type of parameter set for each stage for classification.

Fig. 1. Multistage Expression Classification

A. Previous Works

AAMs have been used by many researchers previously to classify human emotions. In [9] AAMs are used to classify four basic emotions (happiness, sadness, anger and neutral) by means of a cascade of four Support Vector Machines. The classification rate was between 64% and 94%. Previous research in our lab by Wilhelm et al.[3], Gross et al. [10] and Martin et al. [5] have also proposed some methods for emotion classification.

In [10] used of shape and texture parameters along with SVM and MLP classifiers has been made. They have compared MLP and SVM classification rates and have achieved better results with use of an SVM classifier. A method to classify emotions using local features is in [8]. However, using static

local features over images is not very robust, and can lead to unpredictable results in real-time. This is because every person has different basic shape of the face and contours and thus using static [1] local features does not produce very good results with unknown faces.

II. Active Appearance Models (AAM)

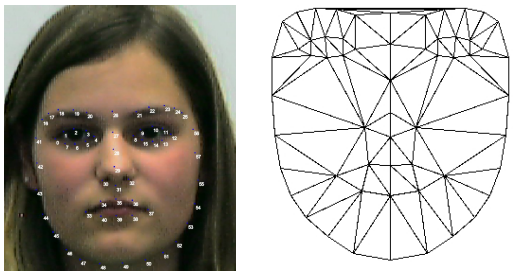
[2] In this section, we have given a brief introduction to AAM. For a detailed mathematical discussion, please refer [11] and [12]. Our Active Appearance Model framework is based upon work previously done at our lab by Martin et al. [5].

A. AAM Structure

AAM consists of a shape model, the texture model and the fitting algorithm. The shape and the texture are statistical models generally built using a PCA. The fitting algorithm used by us is Gradient Descend as introduced in [12].

B. Building the Shape model

In our work, we use a 2-dimensional shape model ***S*** consisting of v = 56 points. The points are placed in regions of the face, which typically have a lot of texture and shape information. Each instance **s** of the shape model can be described as a vector consisting of $2v$ elements, that is the x and y coordinates of each of the v points.



(a) A landmarked training set image (b) An example of a constructed Shape Model

Fig. 2. Shape Model Overview

$$\mathbf{s} = \begin{pmatrix} x_1, y_1, x_2, \cdots, x_v, y_v \end{pmatrix}^T$$

In a pre-processing step, all training shapes $\mathbf{t}_i$ are aligned by applying the *Generalized Orthogonal Procrustes Analysis* [13]. This algorithm removes all components from the data set, which are caused by scaling, translation and rotation. Furthermore, for a facial expression recognition system, it is very useful to generate additional shapes by mirroring the training shapes horizontally. This leads to a new training data set **t'** of $N' = 2N$ examples. Based on the training data set, the mean shape $\mathbf{s}_0$ can be computed as the mean of all N' training examples.

[1]The concept of static and dynamic features is introduced in Section III

[2]A part of the information in this section has been reproduced from previous work done in our lab by Martin et al. [5]

$$s_0 = \frac{1}{N'} \sum_{i=1}^{N'} t_i^{'} \tag{1}$$

The main components of all shapes in the training data set can be computed by a *Principle Component Analysis (PCA)*. The m components with the m biggest eigenvalues will be selected as the shape components $\mathbf{s}_1$ to $\mathbf{s}_m$. Now it is possible to reconstruct the shapes of the training data set and also to generate new shapes, which are not part of the data set by means of the basic shape $\mathbf{s}_0$ and a linear combination of the components $\mathbf{s}_i$.

$$s = s_0 + \sum_{i=1}^{m} p_i s_i \tag{2}$$

The quality of the reconstruction of the shapes from the training data set depends on the number m of used components. In order to make the model capable of generating new shapes not present in the training set, we need to have a diverse set of images for the training.

Depending on the training data set, the main components $\mathbf{s}_i$ typically represent global variations of the face (like pitch and yaw), which are mostly invariant to facial emotions and also local changes (like opening and closing of the eyes or mouth), which are involved in the emotional changes of the subjects. Figure 2(b) shows the basic shape $\mathbf{s}_0$.

C. Building the Appearance model

Besides the *Shape Model*, the other important part of an AAM is the *Appearance Model*, which uses a transformation of the high dimensional input images to a linear subspace of Eigenfaces. This results in a drastic reduction of the dimension of the parameter space. As a pre-processing step all input images **I(x)** are filtered by a Gauss Filter, to remove the image noise. By means of the piecewise affine transformation *W(x, p)* the input image will be transformed to the basic shape s_0 to **I**(W(x, p)). On this normalized images, a histogram equalization is applied to reduce the lighting influences. On the input images, normalized this way, a *PCA* is applied. The k components with the largest eigenvalues are selected as the appearance components $\mathbf{A}_1$**(x)** to $\mathbf{A}_k$**(x)**. The mean appearance components A_0**(x)** can be computed by a simple mean of all normalized input images:

$$A_0(x) = \frac{1}{N'} \sum_{i=1}^{N'} I(W(x,p))$$

The model is a standard *Gray Image Model* as introduced in [11], [12], [1]. Based on the appearance components $A_i(x)$, now it is possible to generate an image **A(x)** with the basic shape s_0, as follows:

$$A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) \tag{3}$$

Fig. 3. An example of a constructed Appearance Model

D. AAM instances

Combining the shape model and the appearance model leads to a model instance. The instance M(W(x, p)) describes a combination of an appearance model and its shape. For that, the appearance parameters $\lambda = (\lambda_1 \cdots \lambda_m)$ and the shape parameters $\mathbf{p} = (p_1, \cdots, p_n)$ are necessary. Using equation 3 the image **A(x)** in the form of the basic shape s_0 can be computed. After that, the image **A(x)** can be transformed to the shape s by using the warp W(x, p).

E. Model Adaptation

It is necessary to adapt the trained model to an unknown input image **I(x)** for the scope of our work. That means, that we have to find the optimal parameters **p** and λ.

$$\arg\min_{p,\lambda} \sum_{x \in s_0} [A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) - I(W(x,p))]^2 \quad (4)$$

For this optimization problem, the following error function E(x) can be defined:

$$E(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) - I(W(x,p)) \quad (5)$$

To solve this problem, in our work we use the a variant of the *Inverse Compositional Algorithm*, which was introduced in [11] and [12]. The problem of the original form of the adaptation algorithm is, that the appearance parameters λ_i are not part of the optimization. The Inverse Compositional Algorithm uses a projection algorithm, that was originally introduced in [14], which allows the optimization of the shape parameters **p** and the appearance parameters λ simultaneously. For more details, please refer to [11] and [12].

III. Feature Set for Expression Detection

In this section, we explain our proposed method for creating the feature set for classification of the expressions into 6 basic emotions of : *Happy, Sad, Neutral, Disgusted (Disapproval), Angry*. In the subsequent section we have discussed the classification method used. Once the face is detected as in Section V-A, the AAM is tried to be fit onto the face at real time. A good fitting is critical to our process since otherwise we will not be able to accurately get the points on the face required for feature extraction as described below. Once the AAM fitting is stable (which generally indicates a good fitting),we extract the feature set for classification. As proposed by Tang et al.[8] feature points extracted from certain areas on the face are good for classifying expressions. It is an easy way which has potential to be quite a powerful method. We investigated that, and found that the results are not very encouraging for classification of all the 6 desired emotions. Thus, we present a improved technique to come up with more discriminating features and construct two sets of features parameter set named the Static and Dynamic Set. Now, we describe the various feature parameters extracted, the method of extraction, and the concept of static and dynamic parameters in detail.

A. Static Parameters:

These parameters are extracted for each input image frame. We found that when the face is emoting, the changes in certain regions of the face are more pronounced than the others. Primarily, the region of the eye and mouth and distances between other organs of the face changes and can be used as emotion discriminating features. Based on experimentation and analysis of images in the Feedtum [6] database, we chose a set of ratios and normalized distances for the feature set. Our parameter set contains the ratios as in figures 4 to 7.

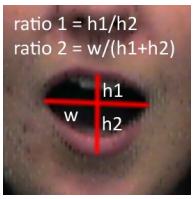


Fig. 4. Defining the origin as the point of intersection of the horizontal and vertical lines,
h_1 is defined as the distance between the upper lip tip and the origin.
h_2 is defined as the distance between the origin and the lower lip tip.
w_1 is defined as the width of the mouth.
Ratio 1 is defined as h_1/h_2.
Ratio 2 is defined as $w_1/(h_1 + h_2)$ or the ratio of the width to the height of the mouth.

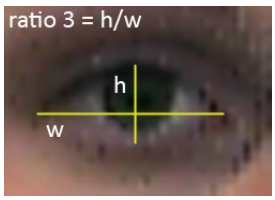


Fig. 5. Defining the h as the height of the eye and w as the width as the eye
Ratio 3 is defined as h/w. i.e. the ratio between the height and width of the eye.

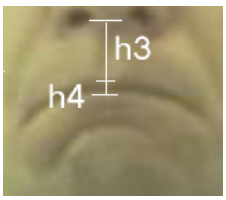
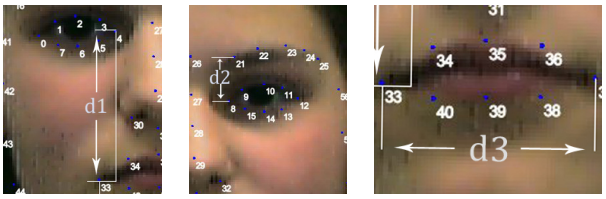


Fig. 6. Defining the h_3 as the distance between the nose tip and the upper lip tip and h_4 as the height of the mouth
Ratio 4 is defined as h_3/h_4.



(a) (b) (c)

Fig. 7. The normalized distance parameters used

B. Dynamic Parameters:

This subsection introduces the concept of dynamic feature parameter set. Based on our tests we came to the conclusion that while the static parameters are robust for classification of *happy* and *surprised* faces, it performs poorly when classifying *neutral*, *sad*, *disgusted* and *angry* faces. This may be because of the variations in face contours across people. Thus, in order to improve the recognition rate, we propose a three step process as shown in Figure 1. The dynamic feature set is an extension of the static feature set. In this, instead of extracting the parameters directly from the current image, we create a difference set between the current set of parameters and a base set. This base set is used to adapt the measures to a specific person. The base set is created when the AAM first converges on a given face. This is used as the base set until the instant when the fitting of the AAM fails and the convergence value is higher than the pre-defined threshold. Let the parameters in the base set be called, $\mathbf{B}_i$ (i=1$\cdots$6).

Algorithm 1 Steps to create Parameter Set for classification

Require: Convergence value lower than threshold
  if AAM fitting changes from unstable $\rightarrow$ stable **then**
    Create Base Static Parameter Set **B**
  end if
  while Model is Valid **do**
    Create Current Static Parameter Set **C**
    Create Dynamic Parameter Set **D** = **C** - **B**
    Classify
  end while

The features parameter set extracted from the current image be called, $\mathbf{C}_i$ (i=1$\cdots$6). Then, the dynamic parameter set created for classification is $\mathbf{D}_i = \mathbf{C}_i - \mathbf{B}_i$.This set based on our experiments performs up to 8-10 % better in the classification of neutral, sad, angry and disgusted faces. The improvement in performance can be attributed to the fact that the difference set would be comparatively freer from dependance on the individual's features compared to a static set.

IV. Classification

There are various classifiers available that can be used to classify the emotions. Based on the work of Gross et al. [10] and Simon et al.[15] which compares the results obtained by MLP, NN and SVM classifiers and shows that the SVM classifier performs marginally better for such classification tasks.Thus, we chose to use SVM for our work. The framework was based on LibSVM [16].

Support Vector Machine (SVM) : SVMs have been demonstrated to be extremely useful in a number of pattern recognition tasks including face and facial action recognition. This type of classifier attempts to find the hyper-plane that maximizes the margin between positive and negative observations for a specified class. SVMs offer additional appeal as they allow for the employment of non-linear combination functions through the use of kernel functions such as the *radial basis function (RBF), polynomial, sigmoid kernels*.

Since SVMs are intrinsically binary classifiers, special steps must be taken to extend them to the multi-class scenario required for facial action recognition. In our work, we adhered to the one-against-one approach [17] in which **K**(**K** - 1)/2 classifiers are constructed, where **K** are the number of the expression groups (**K** = 6 in our work), and each one trains data from two different classes. In classification we use a voting strategy, where each binary classification is considered to be a single vote. A classification decision is achieved by choosing the class with the maximum number of votes.

V. Experiment Procedure

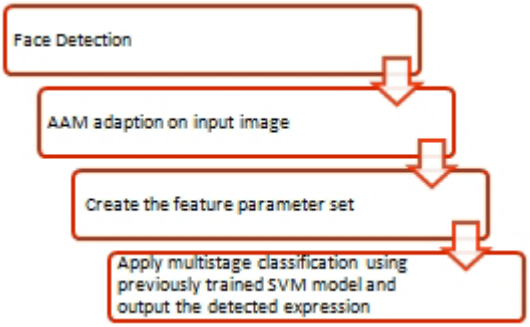


Fig. 8. System Architecture

A. Face Detection

For face detection our module uses the algorithm by Viola and Jones [18]. The OpenCV library comes with a detector based on this and is a cascaded detector. It is one of the best detector available in terms of speed and accuracy and hence widely use in similar works. The face detection takes place by scanning the image through the image window to find faces. The detector has a cascade of classifier layers. Each layer contains an *Adaboost* classifier that takes haar-like feature values as input. Once the face is detected, we proceed to fitting of the AAM.

B. AAM Adaption

The AAM is created offline as described in Section II. For the testing process the AAM fitting process is initiated for each input image. Since, we would like to address the results of the emotion estimation with our work, a good fitting is essential for accurate feature extraction and thus, we discard

the instances where the fitting is invalid or the convergence is above a certain threshold. If the fitting is good we proceed to construct the feature set.

C. Feature Set

The feature set was constructed as described in Section III. We used 324 images from the Feedtum [6] for our offline experiments.

D. SVM Parameter Selection

We made use of LibSVM[16] for our SVM requirements. Experiments were conducted to achieve best possible cross validation accuracy rates on the training set of images using a fixed recognition algorithm but varied parameters for the SVM model.

Based upon our experiments we found that, because the number of feature parameters was not very high a *Radial Base Function (RBF)* kernel was best suited for our problem. This is also supported by [17].

First, the feature parameters (the dynamic and static feature parameter set) were scaled to a range of [-1,1] before being used for classification by the SVM. This is done both during the training and test phase. It is important to scale the data to derive best possible performance of the SVM classifier. [17]

There are two parameters for the SVM RBF kernel - **C** and γ [17]. To achieve the highest possible level of accuracy the appropriate parameter set should be found out. It is pointless to use the classifier to predict training data because we may get very good results which may not be corroborated on an unknown data set. Thus, we used a 10-fold cross validation accuracy rate to determine the best set of SVM parameters. Please find more on feature scaling and SVM parameter selection in [17].

We performed a grid search to find out the parameters **C** and γ best suited for our classification problem. The parameter set achieving the highest rate of cross validation (10 fold) on the training set was used.

Once a satisfactory SVM parameter set has been determined, we proceed to save the model for classifying unknown images from another image set or at realtime.

E. Facial Expression Classification

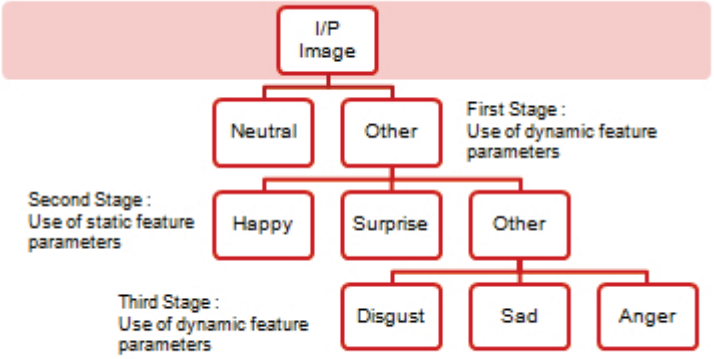
Using the SVM model created, we apply our hierarchal 3 stage approach as in figure 9 for facial expression classification. We present below the description of the stages 1-3.

I In the first stage, neutral was separated from the other expressions by means of the dynamic features parameter set. We achieved a 10 fold cross validation accuracy of around 88% on the data.

II In the second stage, a non-neutral expression is classified either as Happy , Surprise or Other. We achieved cross validation accuracy of around 91.8% on the database, with accuracy upto 98.9% for some test sets.

III In the third and final stage, the expressions were classified as Sad, Disgust or Anger. This we found is a difficult problem for the SVM because of the lack of highly discriminating features between the Disgust and Anger expressions. However, we still achieved cross validation accuracy of 82% in this stage.

Fig. 9. The 3 stage classification outline of facial expressions



Presented in figure 13 are test results on Feedtum[6] database images. We selected equal number of images for each of the 6 emotions for creating our training database. It was composed of 324 images with 7 males and 7 female subjects. Hence, a overall classification accuracy of 90.7% was achieved on our test set, while a cross validation accuracy of 86.7% was obtained on the training set. The same algorithm

(a) Realtime result screenshot from the robot SCITOS

(b) Neutral Expression

(c) Angry Expression

Fig. 10. A screenshot of the realtime implementation of facial expression recognition module detecting expressions of a human partner

was used for realtime detection of facial expressions on the robot SCITOS [7]. At realtime, in order to make the detection robust and add temporal smoothing, we implemented a historigram approach of choosing the expression with the highest number of votes in the last 20 iterations as the detected expression. In the absence of proper AAM fitting on the input image frame, we displayed the classification as 'unsure'.

VI. Social Impacts

The method was tested and implemented on a socially assitive robot used by our lab. This system will help robots like this to have greater understanding of the human partner.

	Neutral	Other
Neutral	12	5
Other	0	55
Accuracy rate = 93 % on test data set		
Cross validation accuracy of 88% on training set		

(a)

	Happy	Surprise	Other
Happy	11	0	0
Surprise	0	12	0
Other	1	0	36
Other includes Anger, Sad and Disgust			
Accuracy rate = 98.2% on test data set			
Cross validation accuracy of 92% on training set			

(b)

	Anger	Disgust	Sad
Anger	9	3	0
Disgust	1	8	0
Sad	2	1	12
Accuracy rate = 81 % on test data set			
Cross validation accuracy of 80% on training set			

(c)

Fig. 11. Results of the 3 stages of facial expression classification on the Feedtum Database images. The result is presented in the form of a confusion matrix. For instance, the first row of figure (a) suggests that 12 'Neutral' images were classified correctly. However 5 'Other' images were also classified as 'Neutral'. The matrix can be read similarly for other cases.

The capability of the system to classify on video inputs of 25 frames in a second, proves its suitability for realtime applications. The impact of such a system that if robust classification of human emotions is possible it can contribute to the learning of social assistant robots. Coupled with other systems they can help them to learn what actions make the users happy and what is confusing or overwhelming them. For instance, a robot teacher can learn what material is interesting its' pupils and whether it is confusing them or robots assisting the elderly or children can take specific actions on detecting the emotional state of the subjects being helped.

VII. Summary

In this paper we proposed a hierarchal approach to facial expression recognition within the AAM framework. We tested our method on a publicly available database and also on a socially interactive robot SCITOS [7] at realtime. Our results were quite encouraging showing the robustness of the classifier for the task.

A. Future Work

We intend to test our methods on more databases in the future. The classification using SVM is sensitive to the quality of images used and the level of emotions displayed by the subject. Thus, we would like to create more robust models and carry out tests on them, and use the same for realtime tests. In the future we would also like to work on decreasing the false positive rates especially for the negative emotions.

References

[1] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Proc. ECCV*, vol. 2, 1998, pp. 484–498.

[2] X. Feng, B. Lv, Z. Li, and J. Zhang, "Automatic facial expression recognition with AAM-Based feature extraction and svm classifier," in *MICAI*, 2006, pp. 726–733.

[3] T. Wilhelm, H.-J. Böhme, H.-M. Gross, and A. Backhaus, "Statistical and neural methods for vision-based analysis of facial expressions and gender," in *SMC (3)*, 2004, pp. 2203–2208.

[4] M. Ratliff and E. Patterson, "Emotion recognition using facial expressions with active appearance models," in *From Proceeding (611) Human Computer Interaction*, 2008.

[5] C. Martin, U. Werner, and H.-M. Gross, "A real-time facial expression recognition system based on active appearance models using gray images and edge images," in *Proc. 8th IEEE Int. Conf. on Face and Gesture Recognition (FG'08), Amsterdam, paper no. 299, 6 pages, IEEE*, 2008.

[6] F. Wallhoff, "Facial expressions and emotion database," *http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html, Technische Universität München*, 2006.

[7] Metralabs GmbH and Ilmenau University of Technology - IUT, "The robot SCITOS is a result of a close RnD collaboration between Neuroinformatics and Cognitive Robotics Lab, IUT and MetraLabs within the project SERROKON," 2004-2007.

[8] F. Tang and B. Deng, "Facial expression recognition using AAM and local facial features," 2007.

[9] Y. Saatci and C. Town, "Cascaded classification of gender and facial expression using active appearance models," 2006, pp. 393–400.

[10] T. Wilhelm, H.-J. Böhme, and H.-M. Gross, "Classification of face images for gender, age, facial expression, and identity," vol. I. Springer Verlag, 2005, pp. 569–574.

[11] S. Baker and I. Matthews, "Equivalence and efficiency of image alignment algorithms." in *Proc. of the 2001 IEEE Conf. on CVPR, volume 1, pages 1090-1097,*, 2001.

[12] ——, "Lucas-kanade 20 years on: A unifying framework," *Int. Journal of Computer Vision,56(3)*, pp. 221–255, 2004.

[13] A. Ross, "Procrustes analysis," *http://www.cse.sc.edu*, 2005.

[14] G. D. Hager and P. N. Belhumeur., "Efficient region tracking with parametric models of geometry and illumination." vol. 20(10), 1998, pp. 1025–1039.

[15] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J. Cohn, "AAM derived face representations for robust facial action recognition," *Proc. of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006.

[16] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification*, 2009.

[18] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," 2001, pp. 511–518.