Towards Authentic Emotion Recognition*

Nicu Sebe[1], Yafei Sun[2], Erwin Bakker[2], Michael S. Lew[2], Ira Cohen[3], Thomas S. Huang[4]

[1]Faculty of Science, University of Amsterdam, The Netherlands

[2]Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

[3]HP Labs, USA

[4]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, IL, USA

Abstract

In human computer interaction, the ultimate goal is to have effortless and natural communication. In the research literature significant effort has been directed toward understanding the functional aspects of the communication. However, it is well known that the functional aspect is insufficient for natural interactions. Indeed, the emotional or affective aspect has been shown in the psychology literature to be at least if not more important. As emotional beings, we interact most comfortably with other emotional beings. In this paper we give an overview of our current research toward automatic recognition of human emotions.

1 Introduction

While a precise, generally agreed definition of emotion does not exist, it is undeniable that emotions are an integral part of our existence. One smiles to show greeting, frowns when confused, or raises one's voice when enraged. The fact that we understand emotions and know how to react to other people's expressions greatly enriches the interaction. There is a growing amount of evidence showing that emotional skills are part of what is called "intelligence". Computers today, on the other hand, are still quite "emotionally challenged." They neither recognize the user's emotions nor possess emotions of their own [27].

Computer systems which have the ability to sense emotions, have a wide range of applications in different research areas, including security, law enforcement, clinic, education, psychiatry, and telecommunications. A new wave of interest in researching on emotion recognition has recently risen to improve all aspects of the interaction between humans and computers. This emerging field has been a research interest for scientists from several different scholastic tracks, i.e. computer science, engineering, psychology, and neuroscience [27]. In the past 20 years there has been much research on recognizing emotion through facial expressions. This research was pioneered by Paul Ekman [8] who started his work from the psychology perspective.

Ekman and his colleagues have performed extensive studies of human facial expressions. They found evidence to support universality in facial expressions [9]. These "universal facial expressions" are those representing happiness, sadness, anger, fear, surprise, and disgust. They studied facial expressions in different cultures, including preliterate cultures, and found much commonality in the expression and recognition of emotions on the face. However, they observed differences in expressions as well, and proposed that facial expressions are governed by "display rules" in different social contexts. For example, Japanese subjects and American subjects showed similar facial expressions while viewing the same stimulus film. However, in the presence of authorities, the Japanese viewers were more reluctant to show their real expressions. Babies seem to exhibit a wide range of facial expressions without being taught, thus suggesting that these expressions are innate [15].

Ekman's work [10] inspired many researchers to analyze facial expressions by means of image and video processing. By tracking facial features and measuring the amount of facial movement, they attempt to categorize different facial expressions. Recent work on facial expression analysis and recognition [6, 11, 18, 20, 21, 26] has used these "basic expressions" or a subset of them. These methods are similar in that they first extract some features from the images, then these features are used as inputs into a classification system, and the outcome is one of the preselected emotion categories. They differ mainly in the features extracted from the video images and in the classifiers used to distinguish between the different emotions. The recent surveys in the area [12, 22, 23] provide an in depth review of many of the research done in in automatic facial expression recognition in recent years.

Our goal is to perform real-time emotion classification using automatic machine learning algorithms. Our real-time system uses a model based non-rigid face tracking algorithm to extract motion features that serve as input to a classifier used for recognizing the different facial expressions and is discussed briefly in Section 2.

One current difficulty in evaluating automatic emotion

* **0-7803-8566-7/04/$20.00**

detection is that there are currently no international databases which are based on authentic emotions. The current facial expression databases contain facial expressions which are not naturally linked to the emotional state of the test subject. As a concequence, we decided to create an authentic facial expression database where the test subjects are showing the natural facial expressions based upon their emotional state. As far as we are aware, this is the first attempt to create such a database. We shall come back to this subject in Section 3.

We also evaluate several promising machine learning algorithms for emotion detection which include techniques such as Bayesian Networks, SVMs, and Decision trees (Section 4). We have concluding remarks in Section 5.

2 Expression Recognition System

There are three main challenges in designing a facial expression recognition system, namely face detection, facial feature extraction, and emotion classification. An ideal emotion analyzer should recognize the subject regardless of gender, age, or ethnic background. The system should be invariant to different lightning conditions and distraction as glasses, changes in hair style, facial hair, moustache, beard, etc. and also should be able to "fill in" missing parts of the face and construct a whole face. It should also perform robust facial expression analysis despite large changes in viewing condition, rigid movement, etc. A good reference system is the human visual system. The current systems are far from ideal and must still address many problems.

2.1 Face Detection and Feature Extraction

Most systems detect the face under controlled conditions, i.e. no facial hair, glasses, nor varying lighting, and thus more general face detection algorithms have drawn more attention [22]. Normally the face detection is done in two ways. In the holistic approach, the face is determined as a whole unit, while in a local feature-based approach only some important facial features are analyzed. After the face is detected, there are two ways to extract the features. In the holistic face model, a template-based method is used. In the local feature-based face model, featured-based methods will be used to track the facial features while people are showing the facial expression.

In our system, we mainly focus on the emotion classification part, not on face detection nor on facial feature extraction. For the extraction of the facial features we use the real time facial expression recognition system developed by Cohen et al [5]. This system is composed of a face tracking part, which outputs a vector of motion features of certain regions of the face. The features are used as inputs to a classifier (see Figure 1).

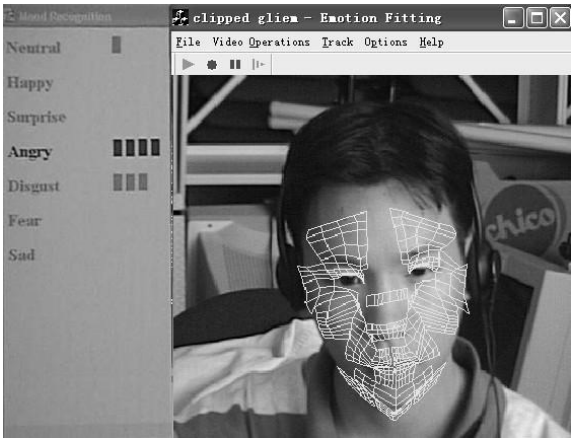


Figure 1: A snap shot of our facial expression recognition system. On the right side is a wireframe model overlayed on a face being tracked. On the left side the correct expression, Angry, is detected (the bars show the relative probability of Angry compared to the other expressions).

This face tracker uses a model-based approach where an explicit 3D-wireframe model of the face is constructed. In the first frame of the image sequence, landmark facial features such as the eye corners and mouth corners are selected interactively. The generic face model is then warped to fit the selected facial features. Once the model is constructed and fitted, head motion and local deformations of the facial features such as the eyebrows, eyelids, and mouth can be tracked.

The recovered motions are represented in terms of magnitudes of some predefined motion of various facial features. Each feature motion corresponds to a simple deformation on the face, defined in terms of the Bézier volume control parameters. We refer to these motions vectors as Motion-Units (MU's). Note that they are similar but not equivalent to Ekman's AU's and are numeric in nature, representing not only the activation of a facial region, but also the direction and intensity of the motion. The MU's are used as the basic features for the classifiers described in the next section.

2.2 Classifiers

Several classifiers and classification strategies from the machine learning literature were considered in our system and are listed below. We give a brief description for each of the classifiers point the reader to the original references.

Bayesian Networks classifiers. A Bayesian network is composed of a directed acyclic graph in which every node is associated with a variable X_i and with a conditional distribution $p(X_i|\Pi_i)$, where Π_i denotes the parents of X_i in the graph. The directed acyclic graph is the *structure*, and the distributions $p(X_i|\Pi_i)$ represent the *parameters* of

the network. We consider three examples of generative Bayesian Networks: (1) Naive-Bayes classifier [7] (**NB**) makes the assumption that all features are conditionally independent given the class label. Although this assumption is typically violated in practice, NB have been used successfully in many classification problems. Better results may be achieved by discretizing the continuous input features yielding the **NBd** classifier. (2) The Tree-Augmented Naive-Bayes classifier [14] (**TAN**) attempts to find a structure that captures the dependencies among the input features. In the structure of the TAN classifier, the class variable is the parent of all the features and each feature has at most one other feature as a parent, such that the resultant graph of the features forms a tree. (3) The Stochastic Structure Search classifier [4] (**SSS**) goes beyond the simplifying assumptions of NB and TAN and searches for the correct Bayesian network structure focusing on classification. The idea is to use a strategy that can efficiently search through the whole space of possible structures and to extract the ones that give the best classification results.

The Decision Tree Inducers. The decision tree represents a data structure which efficiently organizes descriptors. The purpose of the tree is to store an ordered series of descriptors. As one travels through the tree he is asked questions and the answers determine which further questions will be asked. At the end of the path is a classification. When viewed as a black box the decision tree represents a function of parameters (or descriptors) leading to a certain value of the classifier. We consider the following decision tree algorithms and use their $\mathcal{MLC}$++ implementation [17]: (1) **ID3** [24] is a very basic decision tree algorithm with no pruning. (2) **C4.5** is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, and pruning of decision trees [25]. (3) **MC4** is similar to C4.5 [25] with the exception that unknowns are regarded as a separate value.

Other inducers. (1) Support vector machines (**SVM**) were developed based on the Structural Risk Minimization principle from statistical learning theory [28]. They are one of the most popular classifiers and can be applied to regression, classification, and density estimation problems. (2) **kNN** is the instance-based learning algorithm (nearest-neighbor) by Aha [1]. This is a good, robust algorithm, but slow when there are many attributes.

Voting algorithms. Methods for voting classification, such as Bagging and Boosting (AdaBoost) have been shown to be very successful in improving the accuracy of certain classifiers for artificial and real-world datasets [2]. A voting algorithm takes an inducer and a training set as input and runs the inducer multiple times by changing the distribution of training set instances. The generated classifiers are then combined to create a final classifier that is used to classify the test set. The **bagging** algorithm (**B**ootstrap **agg**regat**ing**) by Breiman [3] votes classifiers generated by different bootstrap samples (replicates). Bagging works best on unstable inducers (e.g., decision trees), that is, inducers that suffer from high variance because of small perturbations in the data. However, bagging may slightly degrade performance of stable algorithms (e.g. kNN) because effectively smaller training sets are used for training each classifier. Like bagging **AdaBoost** (**Ada**ptive **Boost**ing) algorithm [13] generates a set of classifiers and votes them. The AdaBoost however, generates classifiers sequentially, while bagging can generate them in parallel. AdaBoost also changes the weights of the training instances provided as input to each inducer based on classifiers that were previously built. The goal is to force the inducer to minimize the expected error over different input distributions.

3 Authentic Expression Analysis

In many applications of human computer interaction, it is important to be able to detect the emotional state of the person in a natural situation. However, as any photographer can attest, getting a real smile can be challenging. Asking someone to smile often does not create the same picture as an authentic smile. The fundamental reason of course is that the subject often does not feel happy so his smile is artificial and in many subtle ways quite different than a genuine smile.

3.1 Posed versus Authentic Expressions

The issue of whether to use posed or spontaneous expressions in selecting facial stimuli, has been hotly debated. Experimentalists and most emotion theorists argue that spontaneous expressions are the only ”true” expressions of facial emotion and therefore such stimuli are the only ones of merit.

When recording authentic facial expressions several aspects should be considered. Not all people express emotion equally well; many individuals have idiosyncratic methods of expressing emotion as a result of personal, familial, or culturally learned display rules. Situations in which authentic facial expression are usually recorded (e.g., laboratory) are often unusual and artificial. If the subject is aware of being photographed or filmed, facial expressions may not be spontaneous anymore. Even if the subject is unaware of being filmed, the laboratory situation may not encourage natural or usual emotion response. In interacting with scientists or other authorities, subjects will attempt to act in appropriate ways so that emotion expression may be masked or controlled. Additionally, there are only a few universal emotions and only some of these can be ethically stimulated in the laboratory.

On the other hand, posed expressions may be regarded as an alternative, provided that certain safeguards are followed. Increased knowledge about the face, based in large part on observation of spontaneous, naturally occurring facial expressions, has made possible a number of methods of measuring the face. These measurement techniques can be used to ascertain whether or not emotional facial behavior has occurred and what emotion is shown in a given instance. Such facial scoring provides a kind of stimulus criterion validity that is important in this area. Additionally, posers can be instructed, not to act or pose a specific emotion, but rather to move certain muscles so as to effect the desired emotional expression. In this way, experimental control may be exerted on the stimuli and the relationship between the elements of the facial expression and the responses of observers may be analyzed and used as a guide in item selection.

It should be noted that the distinction between posed and spontaneous behavior is not directly parallel to the distinction between artificial and natural occurrences. Though posing is by definition artificial, spontaneous behavior may or may not be natural [8]. Spontaneous behavior is natural when some part of life itself leads to the behavior studied. Spontaneous behavior elicited in the laboratory may be representative of some naturally occurring spontaneous behavior, or conceivably it could be artificial if the eliciting circumstance is unique and not relevant to any known real life event.

From the above discussion, it is clear that the authentic facial expression analysis should be performed whenever is possible. Posed expression may be used as an alternative only in restricted cases and they can be mostly used for benchmarking the authentic expressions.

3.2 Authentic Expression Database

Construction and labeling of a good database of facial expressions requires expertise, time, and training of subjects. Only a few such databases are available, such as the Cohn-Kanade [16] and JAFFE [19] databases. Most (or perhaps all) of these existing facial expression data have been collected by asking the subjects to perform a series of expressions. The main problem with this approach is that these deliberate facial action tasks typically differ in appearance and timing from the authentic facial expressions induced through events in the normal environment of the subject. Kanade et al. [16] consider the distinction between the deliberate and spontaneous/authentic facial actions and show that deliberate facial behavior is mediated by separate motor pathways than spontaneous facial behaviors. As a consequence, for a representative test for detecting human emotions in spontaneous settings, we need a test set which captures facial expressions in spontaneous settings.

Our goal for the authentic expression database was to create ground truth where the facial expressions would correspond to the current emotional state of the subject. We consulted several members of the psychology department who recommended that the test be constrained as follows to minimize bias. First, the subjects could not know that they were being tested for their emotional state. Knowing that one is in a scientific test can invalidate or bias the results by influencing the emotional state. Second, we would need to interview each subject after the test to find out their true emotional state for each expression. Third, we were warned that even having a researcher in the same room with the subject could bias the results.

We decided to create a video kiosk with a hidden camera which would display segments from recent movie trailers. This method had the main advantages that it would naturally attract people to watch it and we could potentially elicit emotions through different genres of video footage - i.e. horror films for shock, comedy for joy, etc. Examples of facial expressions from the authentic database are shown in Figure 2. From over 60 people who used the video kiosk, we were able to get the agreement of 28 students within the computer science department for the database. After each subject had seen the video trailers, they were interviewed to find out their emotional state corresponding to the hidden camera video footage. We also secured agreement for the motion data from their video footage to be distributed to the scientific community which is one of the primary goals for this database.

(a) neutral (b) happy (c) disgust

Figure 2: Examples from the authentic database

In this kind of experiment, we can only capture the expressions corresponding to the naturally occurring emotions. This means that our range of emotions for the database was constrained to the ones genuinely felt by the subjects. For this database, the emotions found were either (1) Neutral; (2) Joy; (3) Surprise, or (4) Disgust. From having created the database, some items of note based purely on our experiences: (1) It is very difficult to get a wide range of emotions for all of the subjects. Having all of the subjects experience genuine sadness for example is difficult. (2) The facial expressions corresponding to the internal emotions is often misleading. Some of the subjects appeared to be sad when they were actually happy. (3) Students are usu-

Classifiers	Datasets	
	Authentic	Cohn-Kanade
NB	8.46 ± 0.93%	24.40 ±0.85%
NB bagging	8.35 ± 0.92%	24.33 ± 0.82%
NB boosting	8.25 ± 0.97%	24.45 ± 0.82%
NBd	8.46 ± 0.93%	24.40 ±0.85%
NBd bagging	9.26 ± 1.15%	21.08 ± 0.49%
NBd boosting	8.65 ± 1.03%	18.95 ± 0.53%
TAN	6.46 ± 0.34%	13.20 ±0.27%
SSS	5.89 ± 0.67%	11.40 ±0.65%
ID3	9.76 ± 1.00%	16.70 ±0.53%
ID3 bagging	7.45 ± 0.66%	11.82 ± 0.48%
ID3 boosting	6.96 ± 1.00%	10.70 ±0.53%
C4.5	8.45 ± 0.91%	16.10 ±0.69%
MC4	8.45 ± 0.94%	16.32 ±0.53%
MC4 bagging	7.35 ± 0.76%	12.08 ± 0.32%
MC4 boosting	5.84 ± 0.78%	8.28 ±0.43%
SVM	13.23 ± 0.93%	24.58 ±0.76%
kNN	4.43 ± 0.97%	6.96 ±0.40%
kNN bagging	4.53 ± 0.97%	6.94 ± 0.40%
kNN boosting	4.43 ± 0.97%	6.96 ±0.40%

Table 1: Classification errors for facial expression recognition together with their 95% confidence intervals.

ally open to having the data extracted from the video used for test sets. The older faculty members were generally not agreeable to being part of the database.

4 Emotion Recognition Experiments

In our experiments we use the authentic database described in Section 3 and the Cohn-Kanade AU code facial expression database [16].

The Cohn-Kanade database [16] consists of expression sequences of subjects, starting from a Neutral expression and ending in the peak of the facial expression. We selected 53 subjects, for which at least four of the sequences were available.

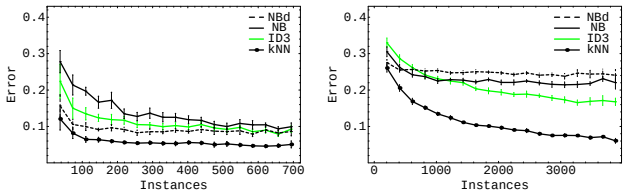
When performing the error estimation we used n-fold cross-validation (n=10 in our experiments) in which the dataset was randomly split into n mutually exclusive subsets (the folds) of approximately equal size. The inducer is trained and tested n times; each time tested on a fold and trained on the dataset minus the fold. The cross-validation estimate of error is the average of the estimated errors from the n folds. To show the statistical significance of our results we also present the 95% confidence intervals for the classification errors.

We show the results for all the classifiers in Table 1. Note that the results for the authentic database outperform the ones for the Cohn-Kanade database. One reason for this is that we have a simpler classification problem: only 4 classes are available. Surprisingly, the best classification results are obtained with the kNN classifier (k=3 in our experiments). This classifier is a distance-based classifier and does not assume any model. It seems that facial expression recognition is not a simple classification problem and all the models tried (e.g., NB, TAN, or SSS) were not able to entirely capture the complex decision boundary that separates the different expressions. This argumentation may also explain the surprisingly poor behavior of the SVM.

kNN may give the best classification results but it has its own disadvantages: it is computationally slow and needs to keep all the instances in the memory. The main advantage of the model-based classifiers is their ability to incorporate unlabel data[4]. This is very important since labeling data for emotion recognition is very expensive and requires expertise, time, and training of subjects. However, collecting unlabel data is not as difficult. Therefore, it is beneficial to be able to use classifiers that are learnt with a combination of some labeled data and a large amount of unlabeled data.

Another important aspect to notice is that the voting algorithms improve the classification results of the decision trees algorithms but do not significantly improve the results of the more stable algorithms such as NB and kNN.

We were also interested to investigate how the classification error behaves when more and more training instances are available. The corresponding learning curves are presented in Figure 3. As expected kNN improves significantly as more data are used for training.



(a) authentic database (b) Cohn-Kanade database

Figure 3: The learning curve for different classifiers. The vertical bars represent the 95% confidence intervals.

5 Conclusion

In this work we presented our efforts in creating an authentic facial expression database based on spontaneous emotions. We created a video kiosk with a hidden camera which displayed segments of movies and allowed filming of several subjects that showed spontaneous emotions. One of our main contribution in this work was to create a database in which the facial expressions correspond to the true emotional state of the subjects. As far as we are aware this is the first attempt to create such a database and our intention is to make it available to the scientific community.

Furthermore, we tested and compared a wide range of classifiers from the machine learning community includ-

ing Bayesian Networks, decision trees, SVM, kNN, etc. We also considered the use of voting classification schemes such as bagging and boosing to improve the classification results of the classifiers. We demonstrated the classifiers for facial expression recognition using our authentic database and the Cohn-Kanade database. Finally, we integrated the classifiers and a face tracking system to build a real time facial expression recognition system.

References

[1] D.W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *Int. Journal of Man-Machine Studies*, 36(1):267–287, 1992.

[2] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms:Bagging, boosting, and variants. *Machine Learning*, 36:105–142, 1999.

[3] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[4] I. Cohen, F. Cozman, N. Sebe, M. Cirello, and T. Huang. Semi-supervised learning of classifiers: Theory, algorithms for bayesian network classifiers and applications to human-computer interaction. *IEEE Trans. PAMI*, to appear in 2004.

[5] I. Cohen, N. Sebe, A. Garg, L. Chen, and T.S. Huang. Facial expression recognition from video sequences: Temporal and static modelling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003.

[6] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski. Classifying facial actions. *IEEE Trans. on PAMI*, 21(10):974–989, 1999.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[8] P. Ekman, editor. *Emotion In the Human Face*. Cambridge University Press, New York, NY, 2nd edition, 1982.

[9] P. Ekman. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, 115(2):268–287, 1994.

[10] P. Ekman and W.V. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, 1978.

[11] I.A. Essa and A.P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. on PAMI*, 19(7):757–763, 1997.

[12] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36:259–275, 2003.

[13] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Int. Conf. on Machine Learning*, pages 148–156, 1996.

[14] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2):131–163, 1997.

[15] C.E. Izard. Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115(2):288–299, 1994.

[16] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[17] R. Kohavi, D. Sommerfield, and J. Dougherty. Data mining using $\mathcal{MLC}$++: A machine learning library in C++. *Int. Journal on Artificial Intell. Tools*, 6(4):537–566, 1997.

[18] J. Lien. *Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity*. PhD thesis, Carnegie Mellon University, 1998.

[19] M. Lyons, A. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with Gabor wavelets. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 200–205, 1998.

[20] N. Oliver, A. Pentland, and F. Bérard. LAFTER: A real-time face and lips tracker with facial expression recognition. *Pattern Recognition*, 33:1369–1382, 2000.

[21] T. Otsuka and J. Ohya. Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences. In *Int. Conf. on Image Processing*, pages 546–549, 1997.

[22] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on PAMI*, 22(12):1424–1445, 2000.

[23] M. Pantic and L.J.M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[24] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[25] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1993.

[26] M. Rosenblum, Y. Yacoob, and L.S. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Trans. on Neural Network*, 7(5):1121–1138, 1996.

[27] N. Sebe and M. Lew. *Robust Computer Vision - Theory and Applications (Chapter 7)*. Springer, 2003.

[28] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.