Emotion Recognition Using a Cauchy Naive Bayes Classifier

Nicu Sebe, Michael S. Lew
Leiden Institute of Advanced Computer Science,
Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands
{nicu, mlew}@liacs.nl

Ira Cohen, Ashutosh Garg, Thomas S. Huang
Beckman Institute,
University of Illinois at Urbana Champaign
{iracohen, ashutosh, huang}@ifp.uiuc.edu

Abstract

Recognizing human facial expression and emotion by computer is an interesting and challenging problem. In this paper we propose a method for recognizing emotions through facial expressions displayed in video sequences. We introduce the Cauchy Naive Bayes classifier which uses the Cauchy distribution as the model distribution and we provide a framework for choosing the best model distribution assumption. Our person-dependent and person-independent experiments show that the Cauchy distribution assumption typically provides better results than the Gaussian distribution assumption.

1. Introduction

Faces are much more than keys to individual identity. Human beings possess and express emotions in day to day interactions with others. Emotions are reflected in voice, hand and body gestures, and mainly through facial expressions. While a precise, generally agreed upon definition of the emotion does not exist, it is undeniable that emotions are an integral part of our existence. The fact that we understand emotions and know how to react to other people's expressions greatly enriches the interaction. Computers today, on the other hand, are still "emotionally challenged." They neither recognize the user's emotions nor possess emotions of their own.

In recent years there has been a growing interest in improving all aspects of the interaction between humans and computers. Ekman and Friesen [4] developed the most comprehensive system for synthesizing facial expressions based on what they call Action Units (AU). In the early 1990s the engineering community started to use these results to construct automatic methods of recognizing emotions from facial expressions in images or video [6, 7, 11, 8, 1]. Work on recognition of emotions from voice and video has been recently suggested and shown to work by Chen, et al. [2], and De Silva, et al [3].

We propose a method for recognizing the emotions through facial expressions displayed in a video sequence. We consider a Naive Bayes classifier and we classify each frame of the video to a facial expression based on some set of features computed for that time frame. The novelty of this work is in proposing and showing the effectiveness of a Cauchy Naive Bayes classifier toward the problem of human emotion recognition. From a statistical perspective, we provide a framework to choose the model distribution for each emotion (class) according to the ground truth we have. Sebe, et al. [9] showed that the Gaussian assumption is often invalid and proposed the Cauchy distribution as an alternative assumption. Based on this, we propose a Naive Bayes classifier based on Cauchy model assumption and we provide an algorithm to test whether a Cauchy assumption is better than the Gaussian assumption.

The rest of the paper is organized as follows. Section 2 presents the features used for facial expression recognition. In Section 3 we present the Cauchy Naive Bayes classifier followed by the experimental setup and the framework for choosing the best model assumption (Section 4). In Section 5 we apply the theoretical results to determine the influence of the model assumption on the emotion classification results. Conclusions are given in Section 6.

2 Features for emotion recognition

The very basis of any recognition system is extracting the best features to describe the physical phenomenon. As such, categorization of the visual information revealed by facial expression is a fundamental step before any recognition of facial expressions can be achieved. First a model of the facial muscle motion corresponding to different expressions has to be found. This model has to be generic enough for most people if it is to be useful in any way.

The best known such model is given in the study by Ekman and Friesen [4], known as the Facial Action Coding System (FACS). Ekman has since argued that emotions are linked directly to the facial expressions and that there are six basic "universal facial expressions" corresponding to happiness, surprise, sadness, fear, anger, and disgust. The FACS codes the facial expressions as a combination of facial movements known as action units (AUs). The AUs have some relation to facial muscular motion and were defined based on anatomical knowledge and by studying videotapes of how the face changes its appearance. Ekman defined 46 such action units to correspond to each independent motion of the face. In our work, we consider a simplified model proposed by Tao and Huang [10] which uses an explicit 3D wireframe model of the face. The face model consists of 16 surface patches embedded in Bézier volumes. Figure 1 shows the wireframe model and the 12 facial motion measurements being measured for facial expression recognition, where the arrow represents the motion direction away from the neutral position of the face. The 12 features we use correspond to the magnitude of the 12 facial motion measurements defined in the face model and the combination of these features define the 7 basic classes of facial expression we want to classify (the Neutral class is also considered in classification).

3 Cauchy Naive Bayes classifier

Consider a classification problem with $y \in \{0, 1, \ldots, M\}$ (class label) and $X \in R^n$ (feature vector) the observed data. The

(a) Anger (b) Disgust (c) Fear (d) Happiness (e) Sadness (f) Surprise

Figure 2. Examples of images from the video sequences used in the experiment.

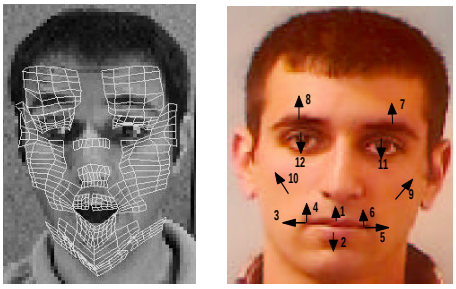

Figure 1. The wireframe model and the facial motion measurements

classification problem under the maximum likelihood framework (ML) can be formulated as:

$$\hat{y} = \underset{y}{argmax}\ P(X|y) \tag{1}$$

If the features in X are assumed to be independent of each other conditioned upon the class label (the Naive Bayes framework), equation (1) reduces to:

$$\hat{y} = \underset{y}{argmax} \prod_{i=1}^{N} P(x_i|y) \tag{2}$$

Now the problem is how to model the probability of features given the class label $P(x_i|y)$. In practice, the common assumption is that we have a Gaussian distribution and the ML can be used to obtain the estimate of the parameters (mean and variance). However, Sebe, et al. [9] have shown that the Gaussian assumption is often invalid and proposed the Cauchy distribution as an alternative model. Intuitively, this distribution can be thought of as being able to model the heavy tails observed in the empirical distribution. This model is referred to as *Cauchy Naive Bayes*.

The difficulty of this model is in estimating the parameters of the Cauchy distribution. For a sample of size n sampled from the Cauchy distribution the likelihood is given by:

$$L(x_i|y; a_i, b_i) = \prod_{d=1}^{n} \left[\frac{b_i}{\pi(b_i^2 + (x_i^d - a_i)^2)} \right] \tag{3}$$

where a_i is the location parameter, b_i is the scale parameter, and $i = 1, \ldots, N$. Note that similar with the Gaussian case we have to estimate only two parameters.

Let $\hat{a}_i$ and $\hat{b}_i$ be the maximum likelihood estimators for a_i and b_i. The maximum likelihood equations are

$$\sum_{d=1}^{n} \frac{x_i^d - \hat{a}_i}{\hat{b}_i^2 + (x_i^d - \hat{a}_i)^2} = 0 \tag{4}$$

$$\sum_{d=1}^{n} \frac{\hat{b}_i^2}{\hat{b}_i^2 + (x_i^d - \hat{a}_i)^2} = \frac{n}{2} \tag{5}$$

The equations (4) and (5) are high order polynomials and therefore a numerical procedure must be used in order to solve them for $\hat{a}$ and $\hat{b}$. For solving these equations we used a Newton-Raphson iterative method with the starting points given by the mean and the variance of the data. We were always able to find unique positive solutions for $\hat{a}$ and $\hat{b}$ which is in accordance with the conjecture stated by Hass, et al. [5]. In certain cases, however, the Newton-Raphson iteration diverged, in which cases we selected new starting points.

4 Experimental setup

We consider that representative ground truth is provided. We split the ground truth in two nonoverlapping sets: the training set and the test set. The estimation of the parameters is done using only the training set. The classification is performed using only the test set.

An interesting problem is determining when to use the Cauchy assumption versus the Gaussian assumption. One solution is to compute the distribution for the data and to match this distribution using a Chi-square or a Kolmogorov-Smirnov test with the model distributions (Cauchy or Gaussian) estimated using the ML approach described above. Another solution (considered here) is to extract a random subsample from the training set and to perform an initial classification. The model distribution which provides better results would be used further in the classification of the test set. The assumption behind this solution is that the training set and the test set have similar characteristics.

In summary, our algorithm can be described as follows:

Step 1. For each class consider the corresponding training set and estimate the parameters of the model (Gaussian and Cauchy) using the ML framework.

Step 2. Extract a random sample from the training set and perform classification. The model which provides the best results will be assigned for each individual class in the classification step.

Step 3. Perform classification using only the test set.

5 Experiments

The testing of the algorithm described in the previous section was performed on a database of five people who were not actors. They were instructed to display facial expressions corresponding to the six types of emotions. Each person displays six sequences of each one of the six emotions and always comes back to a neutral state between each emotion sequence. Figure 2 shows one frame of each emotion for one of the subjects.

The data was collected in an open recording scenario, where the person was asked to display the expression corresponding to a particular emotion. The ideal way of collecting emotion data would be using a hidden recording, inducing the emotion

through events in the normal environment of the subject, not in a studio. However, the authors are not aware of an international benchmark which was acquired through hidden recording. Also, hidden recording could bring up ethical issues.

5.1 Person dependent results

There are six sequences of each facial expression for each person. For each test, one sequence of each emotion is left out, and the rest are used as the training sequences. Table 1 shows the recognition rate for each person and the total recognition rate averaged over the five people when the Gaussian and Cauchy assumptions are used.

Person	Gauss	Cauchy
1	80.97%	81.69%
2	87.09%	84.54%
3	69.06%	71.74%
4	82.5%	83.05%
5	77.18%	79.25%
Average	79.36%	80.05%

Table 1. Person-dependent emotion recognition rates using different assumptions

The Cauchy assumption does not give a significant improvement in recognition rate mainly due to the fact that in this case there are fewer outliers in the data (each person was displaying the emotion sequences in the same environment). This may not be the case in a natural setting experiment.

Note that the third person has the lowest recognition rate. This fact can be attributed to the inaccurate tracking result (resulting in inaccurate features) and lack of sufficient variability in displaying emotions.

The confusion matrix for the Cauchy assumption is presented in Table 2. The analysis of the confusion between different emotions shows that Happy and Surprise are well recognized. The other more subtle emotions are confused with each other more frequently, with Sad being the most confused emotion. Note that in some cases Happy was confused with Surprise due to the fact that the subject smiled while displaying surprise. These observations suggest that we can see the facial expression recognition problem from a slightly different perspective. Suppose that now we only want to detect whether the person is in a good mood, bad mood, or is just surprised (this is separated since it can belong to both positive and negative facial expressions). This means that we consider now only 4 classes in the classification: Neutral, Positive, Negative, and Surprise. Anger, Disgust, Fear, and Sad will count for the Negative class while Happy will count for the Positive class.

The confusion matrix obtained in this case in presented in Table 3. The system can tell now with 88-89% accuracy if a person displays a negative or a positive facial expression.

5.2 Person independent results

From the previous experiments we noticed that the Cauchy assumption brings only a small improvement in the classification rate. A more challenging application is to create a system which is person-independent. In this case the variation of the data is more significant and we expect that using a Cauchy-based classifier we will obtain significantly better results.

Emotion	Neutral	Positive	Negative	Surprise
Neutral	74.52	0.48	20.79	4.18
Positive	2.77	87.16	4.97	4.08
Negative	7.83	0.61	89.11	2.43
Surprise	4.39	0	8.54	87.06

Table 3. Person-dependent average confusion matrix using the Cauchy assumption

For this test all of the sequences of one subject are used as the test sequences and the sequences of the remaining four subjects are used as training sequences. This test is repeated five times, each time leaving a different person out (leave one out cross validation). Table 4 shows the recognition rate of the test when the Gaussian and Cauchy assumptions were used. In this case the recognition rates are lower compared with the person-dependent results. This means that the confusions between subjects are larger than those within the same subject.

Set	Gauss	Cauchy
1	52.44%	58.02%
2	70.62%	75.00%
3	56.29%	60.41%
4	55.69%	63.04%
5	59.66%	61.41%
Average	58.94%	63.58%

Table 4. Person-independent emotion recognition rates using different assumptions

One of the reasons for the misclassifications is the fact that the subjects are very different from each other (three females, two males, and different ethnic backgrounds); hence, they display their emotion differently. In fact, the recognition rate of subject 2, an hispanic male, was the highest in this case (75% for Cauchy assumption). Although it appears to contradict the universality of the facial expressions as studied by Ekman and Friesen [4], the results show that for practical automatic emotion recognition, consideration of gender and race play a role in the training of the system.

Note that the Cauchy assumption is more appropriate in each case. The average gain in classification accuracy is almost 5%.

If we now consider the problem where only the person mood is important, the classification rates are significantly higher. The confusion matrix obtained in this case in presented in Table 6.

Emotion	Neutral	Positive	Negative	Surprise
Neutral	71.30	0.64	26.73	1.31
Positive	5.45	81.16	11.97	1.40
Negative	8.96	5.74	79.08	6.2
Surprise	10.81	8.79	12.23	68.15

Table 6. Person-independent average confusion matrix using the Cauchy assumption

Emotion	Neutral	Happy	Anger	Disgust	Fear	Sad	Surprise
Neutral	74.52	0.48	5.04	3.11	6.19	6.44	4.18
Happy	2.77	87.16	0.83	1.87	1.06	2.19	4.08
Anger	11.3	2.27	74.81	6.03	2.48	2.05	1.02
Disgust	0.92	0	2.73	86.39	2.66	4.03	3.23
Fear	5.51	0	2.96	8.36	77.09	2.43	3.61
Sad	13.59	0.19	2.18	5.61	2.10	74.45	1.84
Surprise	4.39	0	0	0.47	5.14	2.92	87.06

Table 2. Person-dependent confusion matrix using the Cauchy assumption

Emotion	Neutral	Happy	Anger	Disgust	Fear	Sad	Surprise
Neutral	71.30	0.64	3.75	4.06	8.29	10.62	1.31
Happy	5.45	81.16	1.41	8.13	0.15	2.27	1.40
Anger	11.19	2.64	59.27	14.87	0.86	11.14	0
Disgust	5.67	9.94	2.73	50.2	6.48	18.88	6.03
Fear	8.99	0	2.34	1.36	75.53	2.40	9.35
Sad	10.00	10.39	5.14	8.25	17.37	39.41	9.41
Surprise	10.81	8.79	0.98	2.35	4.49	4.40	68.15

Table 5. Person-independent average confusion matrix using the Cauchy assumption

Now the recognition rates are much higher. The system can tell now with about 80% accuracy if a person displays a negative or a positive facial expression.

6 Discussion

In this paper we presented a method for recognizing emotions through facial expressions displayed in video sequences using a Naive Bayes classifier. The common assumption is that the model distribution is Gaussian. However, we successfully used the Cauchy distribution assumption and we provided an algorithm to test whether the Cauchy assumption is better than the Gaussian assumption. We performed person-dependent and person-independent experiments and we showed that the Cauchy distribution assumption provides better results than the Gaussian distribution assumption. Moreover, we showed that when the emotion recognition problem is reduced to a mood recognition problem the classification results are significantly higher.

Are the recognition rates sufficient for real world use? We think that it depends upon the particular application. In the case of image and video retrieval from large databases, the current recognition rates could aid in finding the right image or video by giving additional options for the queries. For future research, the integration of multiple modalities such as voice analysis and context would be expected to improve the recognition rates and eventually improve the computer's understanding of human emotional states.

Acknowledgments. This work has been supported in part by the National Science Foundation Grants CDA-96-24396 and IIS-00-85980. The work of Ira Cohen and Asutosh Garg is supported by Hewlett Packard and IBM fellowships, respectively.

References

[1] L. S. Chen. *Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[2] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato, and R. Nakatsu. Emotion recognition from audiovisual information. In *Proc. IEEE Workshop on Multimedia Signal Processing*, pages 83–88, 1998.

[3] L. C. De Silva, T. Miyasato, and R. Natatsu. Facial emotion recognition using multimodal information. In *Proc. IEEE Int. Conf. on Information, Communications and Signal Processing (ICICS'97)*, pages 397–401, 1997.

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

[5] G. Haas, L. Bain, and C. Antle. Inferences for the Cauchy distribution based on maximum likelihood estimators. *Biometrika*, 57(2):403–408, 1970.

[6] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions*, E74(10):3474–3483, 1991.

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

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

[9] N. Sebe, M. Lew, and D. Huijsmans. Toward improved ranking metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1132–1143, 2000.

[10] H. Tao and T. S. Huang. Connected vibrations: A modal analysis approach to non-rigid motion tracking. In *CVPR*, pages 735–750, 1998.

[11] Y. Yacoob and L. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.