Real Time Facial Expression Recognition in Video using Support Vector Machines

Philipp Michel & Rana El Kaliouby

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Introduction

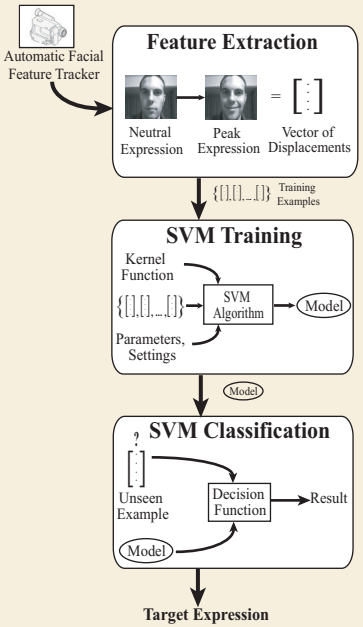
Human beings naturally and intuitively use facial expression as an important and powerful modality to communicate their emotions and to interact socially. There has been continued research interest in enabling computer systems to recognize expressions and to use the emotive and communicative information embedded in them in human-machine interfaces.

This poster presents an application of the machine learning system of support vector machines (SVMs) to the automatic recognition and classification of facial expressions in live video.

Our approach exhibits high performance for real time classification, is unobtrusive and requires no preprocessing, allowing for a variety of unconstrained interaction scenarios.

Implementation Overview

An automatic facial feature tracker performs face localization and feature extraction. It gathers a set of displacements from feature motion in a video stream. These are then used to train an SVM classifier to recognize previously unseen expressions.



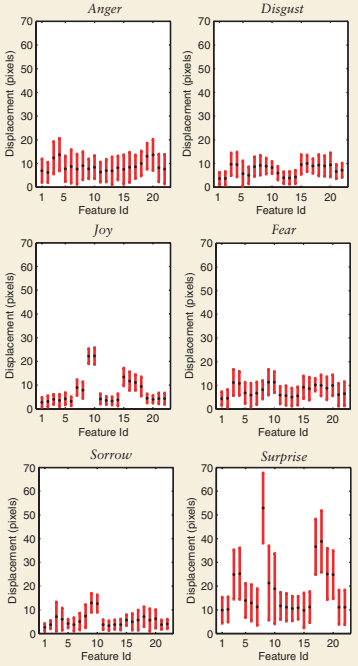
Feature Extraction

Our tracker continuously extracts the position of 22 facial features from the video stream.

For each spontaneous expression, a vector of feature displacements is calculated by taking the euclidean distance between feature locations in a neutral and a "peak" expressive frame. This location capture takes place automatically when the total amount of motion across all features is at a minimum.

Feature displacements between neutral and peak expressive frames

This allows characteristic motion patterns to be established for each expression.



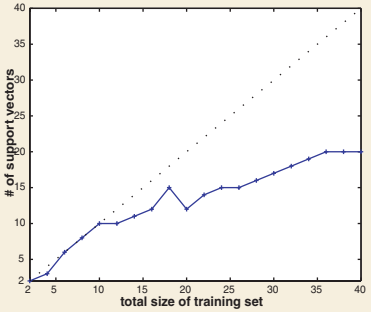
Mean and std. dev. for the characteristic motion of the 22 facial features for the six basic emotions

Training & Classification

The vector of displacements of each example expression together with a user-supplied label is used as input to the SVM training stage. The classifier is re-trained each time a new example expression is added.

The SVM subsequently assigns unseen expressions the label of the target expression that has the most similar displacement pattern.

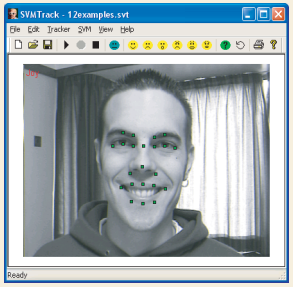
Training overhead is low, due in part to the sparseness of the SVM solution: the number of examples which actually define the SVM decision surface increases sublinearly with the size of the training set.



Number of support vectors defining decision surface vs. training set size

The overhead of evaluating the learned SVM decision function on unseen data is negligible.

Our approach thus allows classification to take place in real time, for every frame in the video stream. The current classification result is continuously reported back to the user.



Typical interactive expression recognition session

Evaluation

To ease comparison, we evaluated the classification performance of our system on the six basic emotions, even though the approach allows arbitrary, user-defined emotion categories to be used for training and classification.

We established an upper bound on the accuracy of a displacement / SVM approach by evaluating it on neutral-peak pairs of still images with manually defined facial features. A standard SVM algorithm with a linear kernel was used.

10 TRAINING EXAMPLES + 15 UNSEEN EXAMPLES PER EMOTION

Emotion	Percent correct
Anger	82.2%
Disgust	84.6%
Fear	71.7%
Joy	93.0%
Sorrow	85.4%
Surprise	99.3%
Average	**86.0%**

Baseline recognition accuracy on displacements extracted from still images

Use of a gaussian radial basis function kernel boosts accuracy to 87.9%.

For an expert user familiar with the system and the typical expression of the basic emotions, little penalty in accuracy is incurred when moving from images to video.

10 TRAINING EXAMPLES + 12 UNSEEN EXAMPLES PER EMOTION

Emotion	Anger	Disgust	Fear	Joy	Sorrow	Surprise	*Overall*
Anger	10	0	0	0	0	2	83.3%
Disgust	0	12	0	0	0	0	100.0%
Fear	2	0	10	0	0	0	83.3%
Joy	0	0	0	9	0	3	75.0%
Sorrow	0	0	2	0	10	0	83.3%
Surprise	0	0	0	0	0	12	100.0%
						Total accuracy:	87.5%

Video: Person-dependent confusion matrix for training & test data supplied by expert user

For wholly inexperienced users during very unconstrained, ad-hoc interaction in terms of pose, lighting and head motion, we established the worst case total accuracy at 60.7%.

Evaluation cont'd

To evaluate person-independent classification, we had our system recognize a user's expresions given only someone else's training data. The systems allows for training data to be re-used and augmented over multiple sessions.

1 TRAINING EXAMPLE + 12 UNSEEN EXAMPLES PER EMOTION

Emotion	Percent correct
Anger	66.7%
Disgust	64.3%
Fear	66.7%
Joy	91.7%
Sorrow	62.5%
Surprise	83.3%
Average	**71.8%**

Video: Person-independent recognition accuracy

Conclusion & Future Work

SVMs exhibit high classification accuracy for small training sets and good generalization performance on data that is highly variable and difficult to separate, making them particularly suitable for expression recognition in video.

We found SVMs to perform well in the presence of noise due to pose variation, lighting, etc. and to meet the temporal constraints imposed by a real-time video environment.

Our recognition system compares favorably to previous approaches and yields usable results even for very unconstrained interaction.

It runs on commodity hardware and is completely unobtrusive.

A typical video interaction session

Improvements being worked on:

1. More sophisticated normalization of feature displacements in the presence of head motion to include rotational movement such as tilting or nodding.

2. Use of automatic SVM model selection to determine optimal parameters of the classifier.

RAINBOW RESEARCH GROUP

William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
http://www.cl.cam.ac.uk/~re227/