International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

Identification of Facial Gestures and Audio Visual Interactions Using Ensemble Matrix of Multi Classifiers

Deepa S[1], Mamatha C R[2], Rashmi R[3]

[1]U G Scholar ,[2]Assistant Professor, [3]Assistant Professor, Department of Computer Science And Engineering, Vemana Institute of Technology.

[1]Deepasrinivas7@gmail.com,[2]mamathapradeep1983@gmail.com, [3]rashmiramreddy@gmail.com

***Abstract*— Human emotions are expressed using the facial expressions, the tone of voice, hands and body gestures .An approach of interaction between the computer-human must be accurate and robust. We use the concept of multi classifier systems to study the human emotions, audio-visual detection system. Automate recognition of emotional state, machines must be taught expressions to understand facial gestures. Here a proposal to identify the person's emotion state such as happy, anger, disgust etc. is stated. Audio visual detection is performed by fusing the results of separate audio and video classifiers on the decision level. We also study about the interactive visualization using Ensemble Matrix. It provides an approach for users to directly interact with the visualization in order to explore and build combination models. The efficiency of the system and approach in a user study is done. Multiple classifiers can be combined on multiple feature set to produce an ensemble classifier with accuracy that will provide a best-reported performance.**

***Keywords*—Emotions, Audio-Visual, Ensemble Matrix, SVM classifier, facial gestures**

I. INTRODUCTION

Human beings express their emotions in everyday interactions with others. Emotions are frequently reflected on the face, in hand and body gestures, in the voice, to express our feelings or liking. Emotions are feeling or response to particular situation or environment. Emotions are an integral part of our existence, as one smiles to show greeting, frowns when confused, or raises one's voice when enraged. It is because we understand other emotions and react based on that expression only enriches the interactions. Computers are "emotionally challenged". They neither recognize other emotions nor possess its own emotion. To enrich human-computer interface from point-and-click to sense-and-feel, to develop non-intrusive sensors, to develop life like software agents such as devices, this can express and understand emotion.

Since computer systems with this capability have a wide range of applications in different research arrears, including security, law enforcement, clinic, education, psychiatry and Telecommunications [7]. There has been much research on recognizing emotion through facial expressions Audio communication contains a wealth of information in addition to spoken words. Specifically, laughter provides clues regarding the emotional state of the speaker [1], topic changes in the conversation [2], and the speaker's identity. Accurate laughter detection could be useful in a variety of applications. A laughter detector incorporated with a digital camera could be used to identify an opportune time to take a picture [3]. Laughter could be useful in a video search of humorous clips [4]. In speech recognition, identifying laughter could decrease word error rate by identifying non speech sounds [2]. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. It has been employed in a wide range of real world problems such as text categorization, hand-written digit recognition, tone recognition, image classification and object detection, micro-array gene expression data analysis, data classification. SVMs are set of related supervised learning methods used for classification and regression [8]. They belong to a family of generalized linear classification. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin.

Machine learning is an increasingly used computational tool within human-computer interaction research. While most researchers currently utilize an iterative approach to ensemble classification techniques may be a viable and even preferable alternative. In ensemble learning, algorithms combine multiple classifiers to build one that is superior to its components.

Ensemble Matrix, an interactive visualization system that presents a graphical view of confusion matrices to help users understand relative merits of various classifiers. Ensemble Matrix allows users to directly interact with the visualizations in order to explore and build combination models. We evaluate the efficacy of the system and the approach in a user study. Results show that users are able to quickly combine multiple classifiers operating on multiple feature sets to produce an ensemble classifier with best accuracy.

II. Facial Detection

The design and implementation of the emotion classification using facial expression system can be subdivided into three main parts: Image Detection, Recognition technique which Includes Training of the images, Testing and then result of classification of images.

A. Image detection

Most systems detect face under controlled conditions, such as without facial hair, glasses, any rigid head movement. Locating a face in a generic image is not an easy task, which continues to challenge researchers. Once detected, the image region containing the face is extracted and geometrically normalized. References to detection methods using neural networks and statistical approaches can be found. Usage of Radial basis function network which is capable of handling noisy images. Its gives better result than back propagation neural network. For Face Data, The face slides are part of the facial emotion database assembled by Ekman and Friesen [24]. In Proposed method, the picture were tested on different age group of people. They are expressing one of the six emotions - happy, sad, fear, anger, surprise, disgust. We used three basic emotions such as happy, anger, disgust. The figures below will show the six emotions.

B. Recognition technique

It aims at modelling the face using some mathematical representation in such a way the feature vector can be fed into a classifier. The overall performance of the system mainly depends on the correct identification of face or certain facial features such as eyes, eyebrows and mouth. After the face is detected, there are two ways to extract the features: Holistic Approach, Analytic Approach.

In Holistic, raw facial image is subjected for feature extraction. While in analytic, some important facial features are detected.

C. Statistical method

After reading the image, the image must be analyzed for duplication. So that correlation of the matrix will be find. Since the correlation matrix for each image is square, we can calculate Eigen vector and Eigen value for each matrix. These are very important so that it gives useful information about the data.

a) Eigen Vectors and Eigen Value: In Eigen vectors any vector change in magnitude but not in direction is called as Eigen vector. In Eigen values, the magnitude that the vector is changed is called an Eigen value.

$$Ax=\lambda x \qquad \text{eq. (1)}$$

Where A is n x n matrix. X is the length of n column vector. λ is a scalar. It's an Eigen value and x is the Eigen vector.

b) Fisher's Linear Discriminant: It can reduce the number of variable in the input by projecting data onto a possibly uncorrelated and low dimensional space. It reduces the number of features in the input to a manageable level.

Fisher Linear Discriminant Analysis considers maximizing the following objective.

$$J\ (w) = w^T S_B w / w^T S_w w \qquad \text{eq. (2)}$$

Where SB is the "between class scatter matrix" and SW is the "within class scatter matrix"

c) Singular Value Decomposition: The Singular Value Decomposition (SVD) is one of the most important tools of numerical signal processing. It plays an interesting, fundamental role in many different applications. SVD in digital applications provides a robust method of storing large images as smaller, more manageable square ones. This is accomplished by reproducing the original image with each succeeding nonzero singular value. Computationally efficient and it is robust under noise conditions.

$$[U, S, V]=svd\ (sw) \qquad \text{eq. (3)}$$

$$Sw = U*S*V \qquad \text{eq. (4)}$$

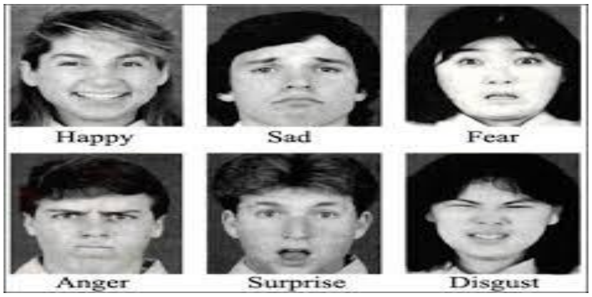


Figure 2: Human emotions

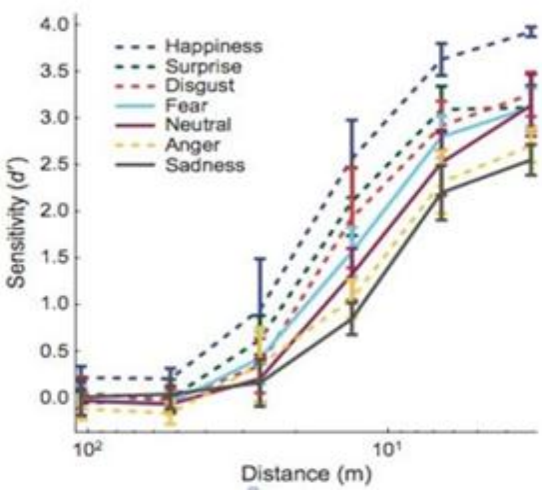


Figure 3: Graph showing the distance and sensitivity of emotions

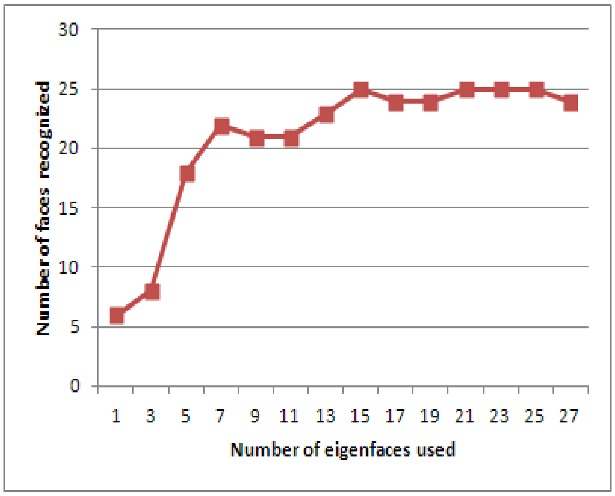


Figure1: Face recognition using Eigen values

III. AUDIO-VISUAL FUSION

The table 1 contains the performance using different modalities and fusion techniques; A indicates audio, V indicates video, FF indicates feature-level fusion, and DF indicates decision-level fusion. The performance is measured in classification accuracy, except for [12, 17, 18] for which we present the F1 measure instead of recall - precision pairs.

Table 1:
Fusion techniques studied so far

Study	Dataset	Performance
Petridis and Pantic	AMI, spontaneous, laughter	A:F$_1$=0.69,V:F$_1$=0.80, DF:F$_1$=0.88
Zeng et al.	AAI,spontanous,2 emotions	A:70%,V:86%,DF:90%
Hoch et al.	Posed ,3 emotions	A:82%,V:67%,DF:87%
Ito et al. Wang and Guan	Spontaneous, laughter Posed,6 emotions	A:F1=0.72,V:F$_1$=0.60,DF:F$_1$=0.72 A:66%,V:49%,FF:82%

The fusion of the audio and video modality boosts the classification performance generally with a few percent. However, most work does not report the significance of this gain in performance. The fusion of audio and video modalities seems to work best when the individual modalities both have a low performance, for example due to noise in the audio-visual speech recognition of Dupont [6]. When single classifiers have a high performance, the performance gain obtained by fusion of the modalities is low, and sometimes fusion even degrades the performance, as observed in the work of Gunes and Piccardi [9].

In order to overcome this we use RASTA-PLP features to encode the audio-signal.

RASTA-PLP adds filtering capabilities for channel distortions to PLP features, and yields significantly better results for speech recognition tasks in noisy environments than PLP [6]. We used the same settings as were used by Truong and Van Leeuwen for PLP features [21]. The 13 cepstral coefficients (12 model order, 1 gain) are calculated over a window of 32 ms with a step-size of 16 ms. Combined with the temporal derivative (calculated by convolving with a simple linear-slope filter over 5 audio frames) this resulted in a 26 dimensional feature vector per audio frame. We normalized these 26-dimensional feature vectors to a meanμ = 0 and a standard deviation σ = 1 using z-normalization.

A. Video features

The video channel is transformed into sequences of 20 two dimensional facial points located on key features of the human face. These point sequences are subsequently transformed into orthogonal features using a Principal Component Analysis (PCA). The points are tracked as follows. The points are manually assigned at the first frame of an instance movie and tracked using a tracking scheme based on particle filtering with factorized likelihoods [16]. We track the brows (2 points each), the eyes (4 points each), and the nose (3 points), the mouth (4 points) and chin (1 point). This results in a compact representation of the facial movement in a movie using 20 (x,y)-tuples per frame. This tracking configuration has been used successfully for the detection of the atomic action units of the Facial Action Coding System (FACS) [22]. After tracking, we performed a PCA on the 20 points per video-frame without reducing the number of dimensions; the principal components now serve as a parametric model, similar to the Active Shape Model of Cootes et al. [5]. No label information was used to create this model. An analysis of the eigenvectors revealed that the first five principal components encode the head pose, including translation, rotation and scale. In order to capture temporal aspects of this model, the first order derivative for each component is added to each frame. The derivative is calculated with Δt = 4 frames on a moving average of the principal components with a window length of 2 frames. Again, we normalized this 80-dimensional feature vector to a meanμ = 0 and a standard deviation σ = 1 using z-normalization.

B. Classification

We evaluate Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) for classification. GMMs and HMMs model the distribution for both classes and classify by estimating the probability that an instance was produced by the model for a specific class. GMMs and HMMs are frequently used in speech recognition and speaker identification, and have been used before for laughter recognition [3, 12, 14, 21] SVMs are discriminatory classifiers, and have been used for laughter detection in [13, 21].We used HMMs and GMMs for the audio-modality and SVMs for the video-modality as this resulted in the best performance [19].

The HMMs we use model the generated output using a mixture of Gaussian distributions. We used two different topologies; the left-right HMMs that are frequently used in speech recognition, and ergodic HMMs that allow transitions from all states to all states. For the SVMs we use a sliding window of 1.20 seconds to create fixed-length features from the video segments. During classification, a probability estimate for the different windows of an instance is calculated. The final prediction of an instance is the mean of its window-predictions. We use Radial Basis Function (RBF) kernel SVMs,

C. Classification Results

For audio, the GMM classifiers performed better than the HMM classifiers, resulting in a mean AUC-ROC of 0.825. On average 16.9 Gaussian mixtures were used to model laughter, on-laughter was modelled using35.6Gaussian mixtures. The HMM performed slightly worse with an AUC-ROC of 0.822. The HMMs used 11.6 fully connected states to model laughter, and 21.3 fully connected states to model non-laughter

Table 2. The performance of the audio and video classifiers. The standard deviation of the AUC-ROCs is displayed between parentheses.

Classifier	Params	Auc-roc	Eer
RASTA-GMM	16.9(3.2)pos.mix,35.6(5.9)neg.mix	0.825(0.143)	0.258
RASTA-HMM	11.6(1.9)pos.states,21.3(1.9)neg.states	0.822(0.135)	0.242
Video-SVM	C=2.46,γ=3.8*10^{-6}	0.916(0.114)	0.133

Table 3. Results of the decision-level fusion. The t-test is a paired samples t-test on the AUC-ROCs of the video-SVM (V-SVM) classifier and the specified fusion classifiers. The mean value of the AUC-ROCs is displayed with the standard deviation displayed between parentheses.

Fusion	Features	T-test	Auc-roc	Eer
RBF-SVM	V-SVM+R-GMM	t(29)=2.24,p<0.05	0.928(0.107)	0.142
RBF-SVM	V-SVM+R-HMM	t(29)=1.93,p=0.06	0.928(0.104)	0.142
$W_{sum,}\alpha=0.57$	V-SVM+R-GMM	t(29)=2.69,p<0.05	0.928(0.107)	0.142
$W_{sum,}\alpha=0.55$	V-SVM+R-HMM	t(29)=2.38,p<0.05	0.930(0.101)	0.142

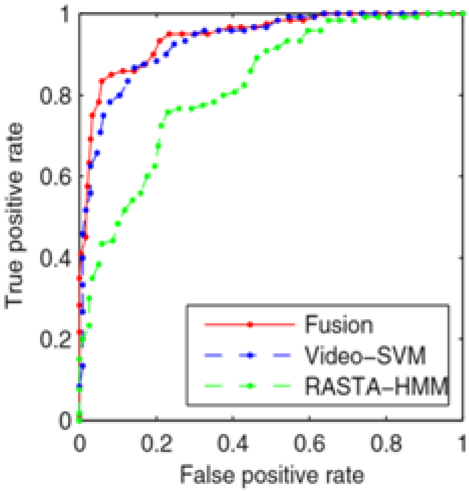


Figure 4: ROCs for the video-SVM, RASTA-HMM and weighted sum

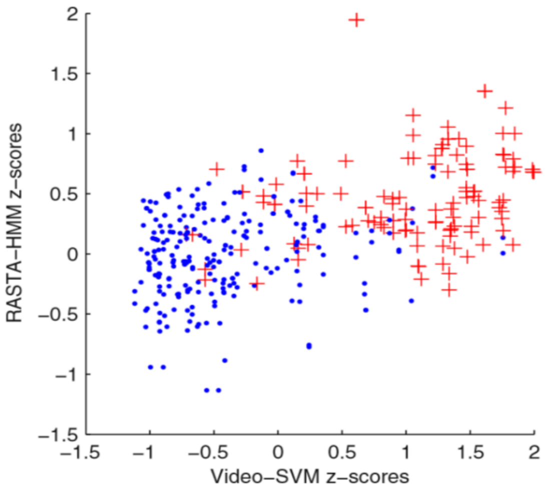


Figure 5: The normalized output of the audio and video classifiers on the test-sets.

IV. Ensemble Matrix

Machine learning, in particular classification, has become an increasingly important tool in HCI research, and more generally in the development of modern software. One problem with these fixed rules is that, it is difficult to predict which rule would perform best. On the other spectrum lie critic-driven approaches, such as layered HMMs [26], where the goal is to "learn" a good combination scheme using a hierarchy of classifiers. The disadvantage with these methods is that they require a large amount of labelled training data, often prohibitive for HCI work.

A. Visualizing the confusion matrix

The Ensemble Matrix interface consists of three basic sections: the Component Classifier view on the lower right, which contains an entry for each classifier that the user has imported to explore, the Linear Combination widget on the upper right, and the main Ensemble Classifier view on the left.

Both the Component Classifier and the Ensemble Classifier views visually represent a classified as a graphical heat-map of its confusion matrix. A confusion matrix represents classification results by plotting data instances in a grid where the column is an instance's predicted class and the row an instance's true class. Confusion matrices can reveal trends and patterns in the classification results and by extension reveal behavior of the classifier itself. We selected the confusion matrix as the core visual element of our system since it can represent results from all classification algorithms and interpretation is relatively algorithm agnostic. This nicely complements previous work that has focused on visualization properties of specific algorithms.

As with other matrix visualizations, the ordering of the matrix can greatly influence the patterns visible. Ensemble Matrix orders each of the Component Classifier matrices independently to highlight sets of classes which are frequently confused by that particular classifier. This corresponds to grouping clusters in an adjacency matrix. Additionally, as users update the Ensemble Classifier view, the main matrix is reordered interactively. This necessitates a fast reordering algorithm, so we chose to use the bary center heuristic [23], borrowed from the layout of bipartite graphs.

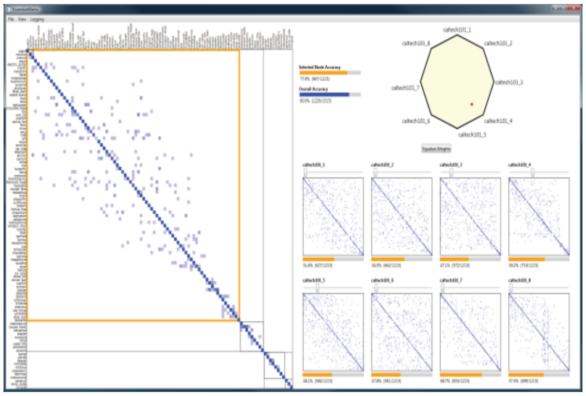
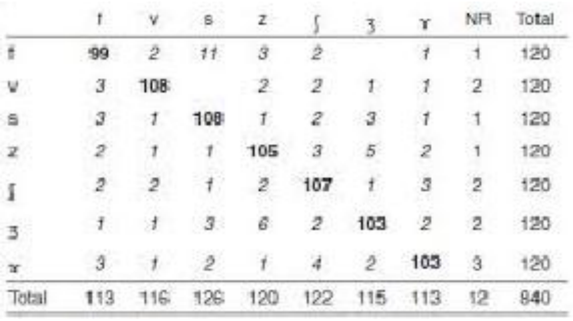


Figure 6: Primary view in Ensemble Matrix. Confusion matrices of component classifiers are shown in thumbnails on the right. The matrix on the left shows the confusion matrix of the current ensemble classifier built by the user.



	f	v	s	z	ʃ	ʒ	ɣ	NR	Total
f	99	2	11	3	2		1	1	120
v	3	108		2	2	1	1	2	120
s	3	1	108	1	2	3	1	1	120
z	2	1	1	105	3	5	2	1	120
ʃ	2	2	1	2	107	1	3	2	120
ʒ	1	1	3	6	2	103	2	2	120
ɣ	3	1	2	1	4	2	103	3	120
Total	113	116	126	120	122	115	113	12	840

Figure 7: Standard confusion matrix

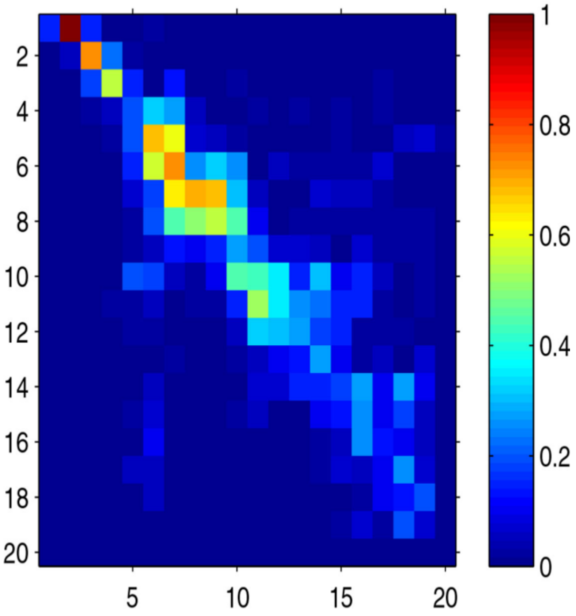


Figure 8: Heat-map confusion matrix

We conducted a formative user study to examine the usability of Ensemble Matrix and the efficacy with which users could use the system to explore component classifiers and create ensemble ones.

B. *Observations of confusion matrix*

Making sense of the individual classifiers by performing the ensemble classifier construction task in Ensemble Matrix a large number of strategies.

To understand the current state confusion the utilization of the tool was made specifically. Noticed confusion of class clusters under some of the classifiers and tried to determine why these classes were confused. By trying to intuit how the features used by a classifier led to particular class confusions or by looking across classifiers to find semantic relationships discovered by the classifiers. We would also examine individual classes trying to understand the behaviour of data instances across multiple classifiers to gain insights.

To build the combination classifier, the weights were frequently changed to see the accuracy. A relatively simple strategy to improve accuracy was seen. We used the matrix reordering to find strong confusion clusters in individual classifiers and recursively isolate the clusters. We then used linear combination to optimize the individual clusters in the leaf nodes. This algorithm led to high accuracy and good generalization. Relatively high classification accuracy was gained.

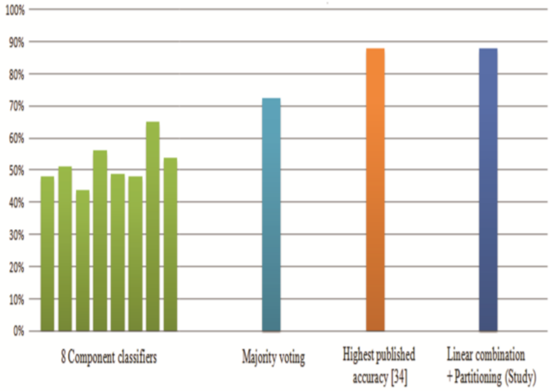


Figure 9: Average accuracy achieved in ensemble matrix

Each of the matrices is reordered independently in order to highlight structure in each matrix. At times this was confusing to users who tried to visually compare two matrices and mistakenly assumed that the ordering was the same. This was an especially large issue for the one user had split the matrices down to small sub matrices and tried to perform visual comparison of individual rows or columns. The reordering also caused the main matrix to visually jump around as the user changed the weighting. This was distracting and made visual comparisons between matrices difficult. We believe that it also made it hard to derive semantic insight into the classes since they were ordered differently in each classifier.

V. Conclusion

In this paper we presented Ensemble Matrix, an interactive visualization system for exploring the space of combinations of classifiers. We have described how we can improve the techniques in machine learning where we have used advanced techniques like SVM classifier for recognizing facial expression. We have also performed automatic laughter detection by fusing audio and video signals on the decision level. We have also used audio visual interaction using GMM classifiers for audio and HMM classifiers for video. We have also improved the performance gain obtained by fusion using RASTA-PLP features. While fusion on the decision-level improves the performance of the laughter-classifier significantly, fusion seems only beneficial for classification with unequal false-negative and false-positive rates. Finally we have used the higher end combination of audio visual fusion and reorganization of facial expressions and also describes about the Ensemble Matrix using multi classifier systems.

REFERENCES

[1] Truong, K.P. and Van Leeuwen, D.A., "Automatic detection of laughter", In Proceedings of Interspeech, Lisbon, Portugal, 2005.

[2] Kennedy, L. and Ellis, D., "Laughter detection in meetings", NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, 2004.

[3] Carter, A., "Automatic acoustic laughter detection", Master's Thesis, Keele University, 2000.

[4] Cai, R., Lu, L., Zhange, H.-J., Cai, L.-H., "Highlight sound effects detection in audio stream", in Proc. Intern. Confer. On Multimedia and Expo, Baltimore, MD, 2003.

[5] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models - their training and application. Computer Vision and Image Understanding (CVIU) 61(1), 38–59 (1995)

[6] Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Transactions on Multimedia 2(3), 141–151 (2000)

[7] G. Little Wort, I. Fasel. M. Stewart Bartlett, J. Movellan "Fully automatic coding of basic expressions from video", University of California, San Diego.

[8] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.

[9] Gunes, H., Piccardi, M.: Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 102–111. Springer, Heidelberg (2005)

[10] Nicu Sebe, Michael S, Lew, Ira Cohen, Ashutosh Garg, Thomas S. Huang, "Emotion recognition using a Cauchy naïve bayes classifier", ICPR, 2002.

[11] Mandeep Kaur, Rajeev Vashisht, Nirvair Neeru, "Recognition of Facial Expressions with principal component analysis and singular value decomposition", International Journal of computer Applications(009758887), volume 9-No.12, November 2010.

[12] Ito, A., Wang, X., Suzuki, M., Makino, S.: Smile and laughter recognition using speech processing and face recognition from conversation video. In: Proceedings of the International Conference on Cyberworlds (CW 2005), Singapore, November 2005, pp. 437–444 (2005)

[13] Kennedy, L.S., Ellis, D.P.W.: Laughter detection in meetings. In: Proceedings of the NIST Meeting Recognition Workshop at the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP 2004), Montreal, Canada (May 2004)

[14] Lockerd, A., Mueller, F.L.: Leveraging affective feedback camcorder. In: Extended abstracts of the Conference on Human Factors in Computing Systems (CHI 2002), Minneapolis, MN, April 2002, pp. 574–575 (2002)

[15] Dai, J. and Cheng, J. HMMEditor: a visual editing tool for profile hidden Markov models. BMC Genomics 2008, 9 (2008).

[16] Patras, I., Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features. In: Proceedings of theIEEE International Conference on Automatic Face and Gesture Recognition (FG 2004), Seoul, Korea, pp. 97–102 (2004)

[17] Becker, B., Kohavi, R. and Sommerfield, D. Visualizing the Simple Bayesian Classifier. [ed.] Fayyad, U., Grinstein, G. and Wierse, A. (2001), 237-249. 3

[18] Maja Pantic, Leon J.M. Rothkrantz, "Automatic analysis of facial expressions: the state of art", IEEE Transactions on pattern Recognition and Machine Intelligence, Dec. 2000, pp. 1424-1444.

[19] Reuderink, B.: Fusion for audio-visual laughter detection. Technical report, University of Twente (2007)

[20] Bosch, A., Zisserman, A. and Munoz, X. Representing shape with a spatial pyramid kernel. Proc. CIVR 2007, (2007), 401-408.

[21] Truong, K.P., van Leeuwen, D.A.: Automatic discrimination between laughter and speech. Speech Communication 49(2), 144–158 (2007)

[22] Valstar, M.F., Pantic, M., Ambadar, Z., Cohn, J.F.: Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In: Proceedings of the International Conference on Multimodal Interfaces (ICME 2006), Banff, Canada, November 2006, pp. 162–170 (2006)

[23] Mäkinen, E. and Siirtola, H. The Barycenter Heuristic and the Reorderable Matrix. Informatica (Slovenia), 29, 3 (2005), 357-364.

[24] P. Ekman and W. Friesen. Pictures of facial affect. 1976.

[25] Grimes, D., Tan, D.S., Hudson, S.E., Shenoy, P. and Rao, R.P. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. Proc. CHI 2008, (2008), 835-844.

[26] Oliver, N., Garg, A. and Horvitz, E. Layered representations for learning and inferring office activity from multiple sensory channels. Computer Vision and Image Understanding, 96, 2 (2004) 163-180.