ECE423 Robotics and Automation

TITLE : K-Means Clustering-Based Kernel Canonical Correlation Analysis for Multimodal Emotion Recognition in Human-Robot Interaction

NAME : SOWMYA BARLA
ROLL NO : 2021BEC0027

Introduction:

- The paper presents a novel multimodal emotion recognition system that enhances human-robot interaction (HRI) by improving emotion detection accuracy.
- It focuses on using speech and facial expressions together to create a more accurate and robust emotion recognition model.
- The method is built upon machine learning and signal processing techniques, utilizing K-Means Clustering and Kernel Canonical Correlation Analysis (KCCA) to improve the fusion of different data sources.
- Unlike traditional unimodal methods, this approach ensures that complementary information is extracted from multiple modalities, leading to a more context-aware and adaptive emotion recognition system.
- The methodology aims to address limitations of existing approaches, making emotion recognition more reliable and accurate for real-world applications.

Introduction:

Human-Robot Interaction (HRI) is the study of how humans and robots communicate and work together.

Feature Fusion: Combining Multiple Data Sources for Improved Recognition

- Combining data from different sources (e.g., speech and facial expressions).
- Merges the strengths of each data type for more accurate emotion recognition.
- Speech tells us the tone of voice, while facial expressions show physical reactions, both contributing to a clearer emotional picture.

K-Means Clustering: Grouping Similar Data Points to Reduce Dimension

- A technique that groups similar data points together into clusters.
- Helps simplify data by grouping similar items, making it easier to analyze.
- Grouping facial expressions that show similar emotions, like happy or sad, to recognize patterns.

Introduction:

Kernel Canonical Correlation Analysis (KCCA): A technique to find relationships between two different datasets using kernel methods.

- A method to find connections between two datasets, even when the relationships are non-linear.
- Allows us to link data from different sources (like speech and facial expressions) in complex ways.
- Linking the tone of voice in speech with the movement of facial muscles to understand emotions better.

Multimodal Emotion Recognition: Identifying Emotions Using Multiple Data Types

- Recognizing emotions by combining different data sources, like speech and facial expressions.
- Provides a more complete picture of emotions, improving accuracy.
- Example: Using both tone of voice and facial expressions to detect if someone is angry or happy.

Why was this paper Proposed?

- Existing emotion recognition methods focus on single modalities (only speech or only facial expressions).
- Single-modality methods are less accurate due to noise, occlusions, or missing features.
- The paper proposes a better fusion approach to improve emotion recognition using both speech and facial expressions.
- **Need for Robust Emotion Recognition:** Current systems struggle with inconsistencies in real-world data, requiring a more adaptable and scalable method.

Limitations of Previous Studies:

Single-Modality Weaknesses:

- Speech-based methods fail in noisy environments.
- Facial expressions can be affected by occlusions or lighting.

Basic Fusion Techniques:

- Simple concatenation or PCA does not fully exploit the relationship between features.

No Proper Feature Selection:

- Unnecessary features introduce noise, reducing accuracy.

Limited Generalization:

- Existing methods often perform poorly on diverse datasets.

Proposed Method:

The main contributions include the following.

1. K-means clustering is introduced in fusion to strengthen the heterogeneity among emotions. It can extract features at the early fusion stage and remove the features with an adverse effect on recognition.
2. The in sufficient complementarity among different modalities is addressed for subsequent correlation analysis.
3. The kernel function is deployed to solve the nonlinear fusion problem.
4. The proposed multimodal fusion method can effectively improve the emotion recognition rate in HRI.

Block Diagram Explanation:

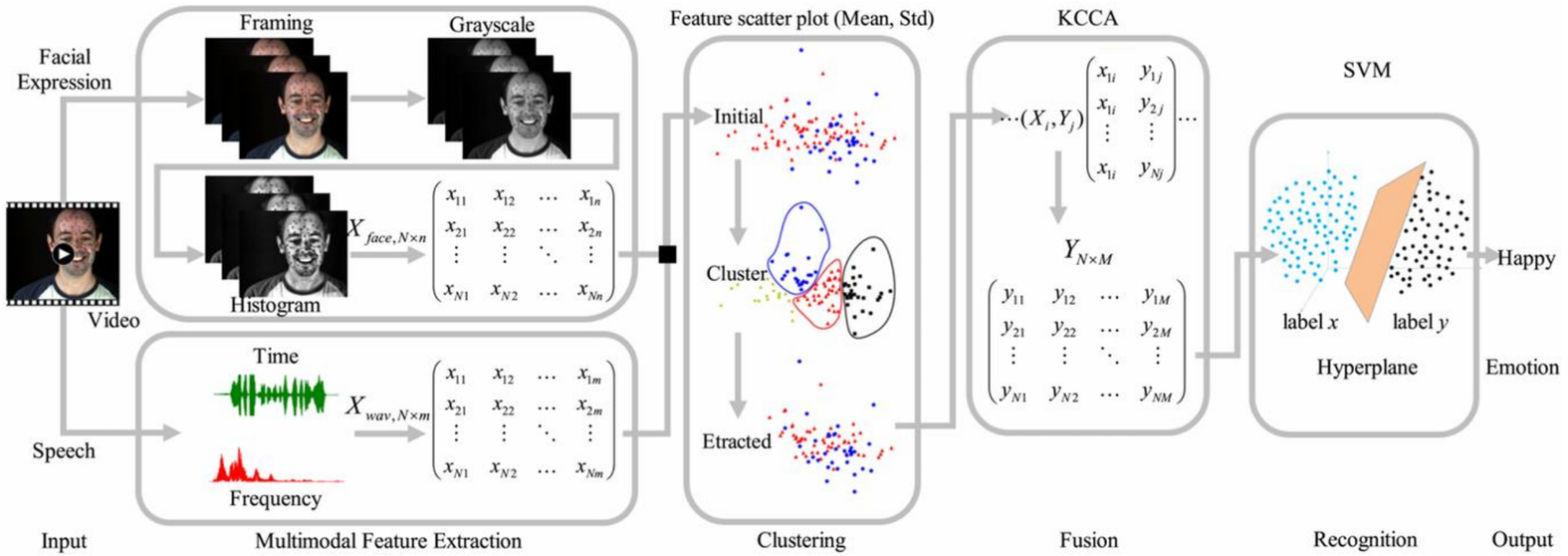

Fig. 1. Overview of our presented multimodal emotion recognition system based on feature fusion. The two-channel input of speech and facial expressions, respectively, clustered and extraction of multimodal features is to eliminate features that interfere the fusion between modalities. After the fusion of the feature matrix based on KCCA, SVM is used to identify the fused emotion labels.

- Data Input: Speech & Facial expressions.
- Feature Extraction: Extract relevant features from each modality.
- Feature Selection (K-Means): Remove unnecessary features.
- Feature Fusion (KCCA): Combine features while preserving nonlinear relationships.
- Classification (SVM): Predict the emotion based on the fused features.

K-MEANS KERNEL CANONICAL CORRELATION ANALYSIS:

A. Preprocessing

- For the facial expressions, first of all, equal frame distance of each video is extracted, and the face region is intercepted. Endpoint detection is carried out for each speech and divided into equidistant frames, with blank frames deleted.

B. Feature of Facial Expressions

- In pixel method, the feature dimensionality is high and affected by the pose.
- Adaptive histogram equalization suitable for producing more details improves image contrast. Histogram shearing is considered in this article

Histogram Shearing Equation:

$$n'_k = \begin{cases} s, n_k - s \geq 0 \\ n_k, n_k - s < 0 \end{cases} \quad k = 0, 1, \ldots, L-1 \ .$$

- nk → Number of pixels in the image that have gray level k(before shearing).
- nk′→ Number of pixels in the image with gray level k after applying shearing (adjusted version of nkn).
- s → Shear coefficient (a threshold value to clip excessive intensities).
- k → Gray level index, representing a particular intensity value in the grayscale range (from 0 to L−1).
- L→ Total number of gray levels in the image (usually 256 for an 8-bit grayscale image).
- If the number of pixels with intensity k exceeds s, it is clipped to s.
- This prevents over-saturation of any intensity level, helping in better contrast adjustment.

Histogram Equalization Equation:

$$s_k = T(r_k) = \sum_{j=0}^{k} P_r(r_j) = \sum_{j=0}^{k} \frac{n_j}{n}$$

- sk → New intensity value after applying histogram equalization.
- T(rk)→ Transformation function, which adjusts pixel intensity values.
- rkr→ Original intensity value before equalization (ranges from 0 to L−1).
- Pr(rj) → Probability distribution function (PDF) of intensity values, representing the fraction of pixels at each gray level j.
- nj → Number of pixels with intensity j in the original image.
- n → Total number of pixels in the image.
- L → Total number of intensity levels (usually 256 in an 8-bit grayscale image).
- This redistributes pixel intensities so that they are more evenly spread across all possible values.
- The transformation function T(rk) maps original pixel values to new values based on their cumulative probability.
- Enhances contrast, making darker regions brighter and lighter regions darker, improving the visibility of facial features.

C. K-Means Clustering of Features:

- Step 1: Select k initial samples as centers a1,a2,...,ak
- Step 2: For each sample x, calculate the distance to the k cluster centers and classify them into the cluster with the smallest distance.
- Step 3: For each category μ, recalculate its cluster centers

$$\mu_j = \frac{1}{|c_j|} \sum_{i \in c_j} a_i.$$

- μj → New cluster center for cluster j.
- Cj→ The set of all data points assigned to cluster j.
- |Cj| → Total number of points in cluster j.
- ai → Each individual data point in cluster j.

D. Normalization During Clustering:

- Normalization ensures all features have the same range, preventing large values from dominating the clustering process.

$$Y = \frac{X - X\,\min}{X\,\max - X\,\min}$$

- Y → Normalized feature value (scaled between 0 and 1).
- X → Original feature value.
- Xmin → Minimum value of the feature in the dataset.
- Xmax → Maximum value of the feature in the dataset.

E. Canonical Correlation Analysis:

- CCA projects the two feature sets (speech & facial expressions) into a new space where their correlation is maximized.

Correlation Formula:

$$\rho(M,N) = \frac{\text{cov}(M,N)}{\sqrt{D(M)}\sqrt{D(N)}}$$

- ρ(M,N) → Correlation coefficient (higher means stronger relationship).
- M, N → Feature sets of speech and facial expressions.
- cov(M,N)→ Covariance (measures how much the two sets vary together).
- D(M),D(N) → Variances of the feature sets.

Projection of Feature Matrices onto New Space:

$$M' = a^T M, N' = b^T N$$

- a, b → Projection vectors that transform the features.
- M’,N′ → Projected feature sets where correlation is maximized.

E. Canonical Correlation Analysis:

Optimization Problem for Maximum Correlation:

$$\underbrace{\arg\max}_{a,b} \frac{\operatorname{cov}(M', N')}{\sqrt{D(M')}\sqrt{D(N')}}.$$

- The goal is to find a and b that maximize the correlation between M′ and N′.
- This ensures that the best possible projections are selected to maximize correlation between the modalities.

$$\operatorname{cov}(M', N') = a^T E\left(MN^T\right) b$$
$$D(M') = D\left(a^T M\right) = a^T E\left(MM^T\right) a$$
$$D(N') = D\left(b^T N\right) = b^T E\left(NN^T\right) b.$$

- **Normalization Before Projection:**
- E(MNT) → Expectation of the product of feature matrices.
- E(M M^T), E(N N^T)→ Expectation of feature matrix self-products.

Meaning:

These equations standardize the data to ensure fair correlation computation.

E. Canonical Correlation Analysis:

Final Optimization Equation Using Singular Value Decomposition:

$$\underbrace{\arg\max}_{a,b} \frac{a^T S_{XY} b}{\sqrt{a^T S_{XX} a}\sqrt{b^T S_{YY} b}}$$
$$\text{s.t.} a^T S_{MM} a = 1, b^T S_{NN} b = 1.$$

- SXY → Cross-covariance matrix between M and N.
- SXX,SYY → Covariance matrices of and N.
- M → Feature matrix of speech features (e.g., MFCCs, prosodic features).
- N → Feature matrix of facial expression features (e.g., grayscale pixel intensities, shape features).
- CCA finds the best correlation between two feature sets (speech & facial expressions).
- This equation finds the best projection directions that maximize canonical correlation.
- The constraint ensures that the projections remain valid unit vectors.

F. Kernel Function:

- When speech and facial expression features have nonlinear relationships, simple linear methods like Canonical Correlation Analysis (CCA) fail to find meaningful patterns.
- Kernel functions help by mapping the features into a higher-dimensional space, where their relationships become linear.

Kernel Transformation:

- ϕ(M), ϕ(N)→ Nonlinear transformations of the original feature matrices M(speech) and NNN (facial expressions).
- α, β→ Weight vectors for projecting transformed features.
- KM and KN are kernel matrices, containing similarity measures between feature samples.

$$c\phi : \mathbf{x} = (\mathbf{x}_1, \ldots) \rightarrow \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots)$$

$$\psi(M) = \alpha^T_{\phi(M)} \phi(M)$$

$$\psi(N) = \beta^T_{\phi(N)} \phi(N)$$

$$K_M = \phi^T(M)\phi(M)$$

$$K_N = \phi^T(N)\phi(N).$$

- **Final Kernel CCA Optimization Equation:**
- This equation finds the best projection vectors a and b to maximize the correlation between transformed feature sets.

$$\underbrace{\arg\max}\ \frac{a^T K_M K_N b}{\sqrt{a^T K_M K_M a}\sqrt{b^T K_N K_N b}}.$$

G. Support Vector Machine

- SVM is used for emotion classification after feature fusion (KCCA).
- It finds the best boundary (hyperplane) to separate different emotions (e.g., Happy, Sad, Angry).
- If emotions are not linearly separable, SVM uses a kernel function to transform the data into a higher dimension.
- The polynomial kernel helps improve classification accuracy.
- SVM is chosen because it works well with small datasets and provides high recognition accuracy.

Fig. 5. Part of the sample frames in the SAVEE database.

Experimental Environment and Data Selection

Setup & Datasets:

- Tested on SAVEE and eNTERFACE‘05 datasets.
- SAVEE has 480 samples (speech + facial expressions).
- eNTERFACE‘05 has data from 44 people of different ethnicities

Feature Extraction:

- Speech features: MFCCs, prosody, spectral features.
- Facial features: Pixel grayscale values.
- PCA was used to reduce dimensionality.

Key Findings:

- Facial expressions had higher recognition rates in SAVEE (88.89%) than speech (59.72%).
- Speech had better recognition rates in eNTERFACE‘05 (77.42%) than facial expressions (52.80%)

Selection of Kernel :

- Polynomial Kernel (p = 2) gave the highest recognition rate for SAVEE (91.39%).
- If p increases, data becomes too scattered, reducing accuracy.
- For eNTERFACE‘05, Gaussian & low-order polynomial kernels worked better.

TABLE I
DIFFERENT OPTIMAL RECOGNITION RATES BASED ON DIFFERENT KERNELS BASED ON KCCA

Kernel	Recognition rate(%)
line	66.53
sigmoid	67.50
spline	70.28
RBF (gamma=0.01)	88.33
polynomial (p=2)	91.39

Feature Selection Based on K-Means Clustering:

- Some features do not contribute to emotion recognition.
- K-Means groups similar features together and removes irrelevant ones.
- **Feature Selection Process**:
- Step 1: Extract speech & facial expression features.
- Step 2: Normalize features using Min-Max scaling to [0,1].
- Step 3: Cluster features into K = 3 groups.
- Step 4: Remove one group with least variation, improving classification.
- **Key Findings (Table II):**
- KMKCCA (K=3) achieved 93.06% accuracy.
- KMKCCA (K=4) improved accuracy to 94.16% (best result).

TABLE III
RECOGNITION RATE (%) AND FEATURE DIMENSIONALITY OBTAINED UNDER DIFFERENT METHODS IN THE eNTERFACE'05 DATABASE

Method	Recognition rate	Dimension
Face	52.80	47
Speech	77.42	68
Series Fusion	81.21	115
PCA Fusion	81.67	94
KCCA Fusion	82.50	136
KMKCCA Fusion ($K = 3$)	85.68	102
KMKCCA Fusion ($K = 4$)	87.20	110

Feature Selection:

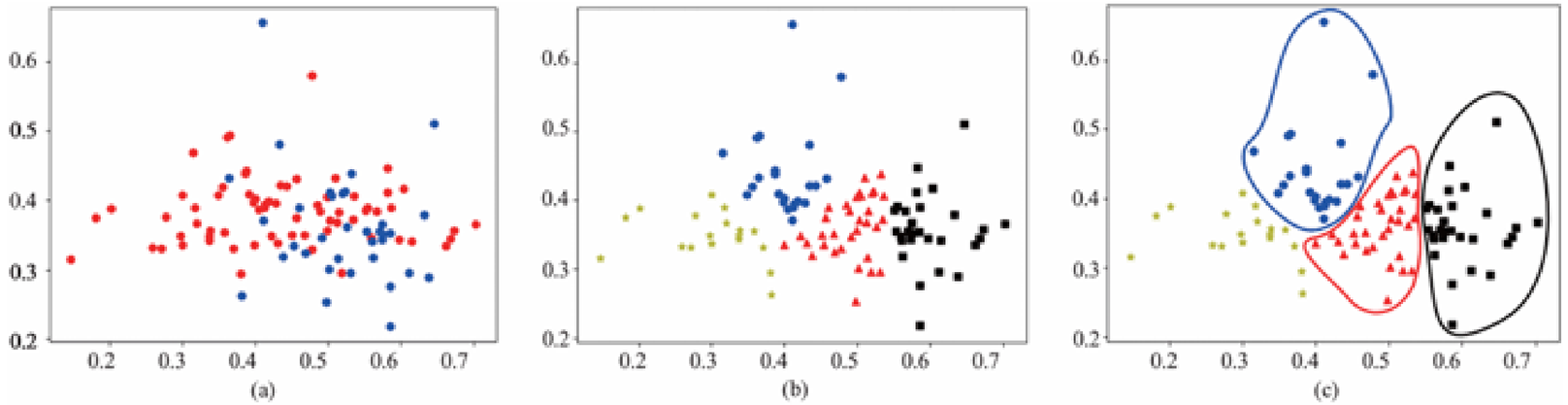

Fig. 6. All features are divided into four classes. After eliminating the class of features with the smallest mean and standard deviation, (a) is a scatter plot before clustering, (b) is after clustering without elimination, and (c) is with elimination of yellow star and its final rate is 94.16%.

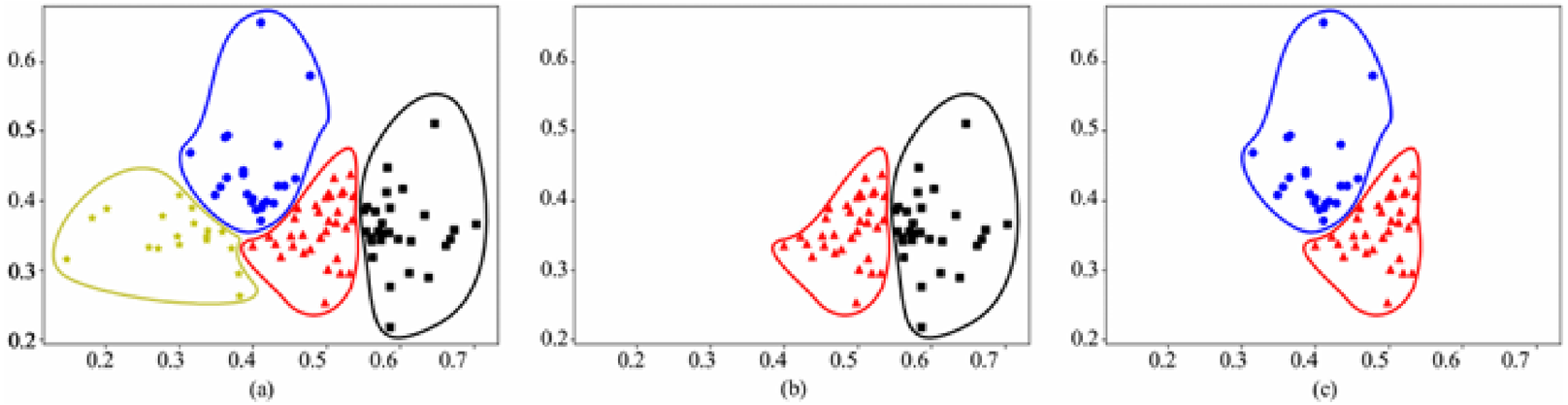

Fig. 7. When $K = 4$, (a) is a scattered plot after clustering and the result of 94.16%. After eliminating the set of features with the smallest mean and standard deviation, (b) and (c) are, respectively, the class with the highest mean and standard deviation. (b) is 91.38% and (c) is 88.61%.

Results:

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	.889	0	0	.022	0	.022	.067
DI.	0	.911	0	0	.044	.044	0
FE.	0	.089	.844	.022	0	0	.044
HA.	.022	0	.022	.911	0	0	.022
NE.	0	.011	0	0	.967	.022	0
SA.	0	.022	0	.022	.089	.867	0
SU.	.023	0	.133	.022	0	.067	.756

Fig. 8. Confusion matrix of facial expression in the SAVEE database.

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	.867	0	.067	.022	0	0	.044
DI.	0	.889	0	0	.022	.067	.022
FE.	.022	.044	.889	0	0	0	.044
HA.	0	0	.022	.956	0	0	.022
NE.	0	.011	0	0	.978	.011	0
SA.	0	.022	0	0	.044	.933	0
SU.	0	0	.044	0	0	0	.956

Fig. 9. Confusion matrix by KMKCCA ($K = 3$) in the SAVEE database.

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	.867	0	.067	.022	0	0	.044
DI.	0	.933	0	0	.022	.044	0
FE.	.022	.044	.889	0	0	0	.044
HA.	0	0	.022	.956	0	0	.022
NE.	0	.011	0	0	.989	0	0
SA.	0	.022	0	0	.044	.933	0
SU.	0	0	.022	0	0	0	.978

Fig. 10. Confusion matrix by KMKCCA ($K = 4$) in the SAVEE database.

- SAVEE Dataset (Figures 8–10)
- K=3 Results (Fig. 9):
- Surprise emotion recognition improved from 75.6% to 95.6%.
- Anger & Disgust slightly decreased due to similarity in speech samples.
- K=4 Results (Fig. 10):
- Improved heterogeneity among emotions by removing weak features.
- Higher overall accuracy (94.16%)

Results:

	AN.	DI.	FE.	HA.	SA.	SU.
AN.	.859	.023	.032	.023	.027	.036
DI.	.014	.841	.014	.027	.068	.036
FE.	.036	.014	.900	.018	.009	.023
HA.	.014	.027	.014	.895	.036	.014
SA.	.014	.086	.009	.023	.823	.045
SU.	.027	.032	.023	.045	.050	.823

Fig. 12. Confusion matrix by KMKCCA ($K = 3$) in the eNTERFACE'05 database.

	AN.	DI.	FE.	HA.	SA.	SU.
AN.	.868	.023	.032	.018	.027	.032
DI.	.014	.827	.014	.027	.073	.045
FE.	.032	.009	.914	.018	.009	.018
HA.	.014	.018	.009	.927	.023	.009
SA.	.014	.095	.005	.027	.818	.041
SU.	.027	.014	.014	.041	.027	.877

Fig. 13. Confusion matrix by KMKCCA ($K = 4$) in the eNTERFACE'05 database.

- K=3 Results (Fig. 12):
- Happiness increased from 74.1% to 89.5%.
- Sadness improved from 68.2% to 82.3%, showing better modality complementarity.
- K=4 Results (Fig. 13):
- Further increased heterogeneity among emotions.
- Higher accuracy overall due to better separation of emotions
- **KMKCCA (K=4) consistently outperformed previous models, proving its effectiveness.**

Comparision with old methods:

TABLE IV
COMPARISON WITH THE RECOGNITION RATE (%) OF EXISTING METHOD

Database	Recognition method	Fusion
SAVEE	ISLA [31]	86.01
	KMKCCA	94.16
eNTERFACE'05	HDM [32]	85.97
	RBML [33]	86.67
	KMKCCA	87.20

- ISLA (Informed Segmentation and Labeling Approach) → 86.01% (SAVEE)
- RBML (Rule-Based Machine Learning) → 85.97% (eNTERFACE'05)
- HDM (Hybrid Deep Model) → 86.67% (eNTERFACE'05)
- Our KMKCCA (K=4) method achieved the highest accuracy:
- 94.16% (SAVEE) → 8.15% better than ISLA.
- 87.20% (eNTERFACE'05) → 1.23% better than HDM.

How They Overcame Previous Limitations:

Enhanced Feature Selection:

- K-Means clustering removes redundant features, improving efficiency.

Better Fusion Technique:

- KCCA finds hidden patterns between speech and facial expressions.

Stronger Classifier:

- SVM ensures higher accuracy in emotion classification.

More Robust to Noise:

- Kernel-based correlation techniques handle nonlinear relationships effectively.

Key Contributions of the Paper:

1. Introduced K-Means Clustering for feature selection before fusion.
2. Used Kernel CCA for better multimodal fusion.
3. Achieved higher accuracy on standard emotion recognition datasets.
4. Provided insights on kernel selection for different datasets.
5. Enhanced robustness of emotion recognition models in real-world applications.

Limitations of Existing Methods:

- Challenges in real-time emotion recognition due to processing speed.
- Individual differences in emotion expression affect accuracy.
- Future work will focus on handling personalized emotion characteristics and improving multimodal feature integration for better human-robot interaction (HRI)

Overcoming the Limitations:

- Apply Advanced Deep Learning: CNNs or transformers could improve feature extraction.
- Improve Noise Handling: Use denoising techniques for speech and image data.
- Optimize Kernel Selection: Adaptive kernel selection can further improve KCCA results.
- Expand to Real-World Data: Test on larger, more diverse datasets.
- By integrating deep learning, advanced feature extraction, better fusion techniques, and optimized classification models, you can significantly improve emotion recognition accuracy and real-time performance!

Conclusion:

- The proposed KMKCCA algorithm effectively combines speech and facial expression features for emotion recognition.
- K-means clustering improves feature selection before fusion.
- Kernel CCA successfully captures nonlinear relationships between modalities.
- SVM classifier provides higher recognition rates compared to previous methods.
- Achieved 94.16% (SAVEE) and 87.20% (eNTERFACE’05), outperforming older models

Thank You