Exploring Automatic Translation between Affect Representation Schemes

Affective Image Content Analysis

Shuang Liu

Supervisor(s): Chirag Raman, Bernd Dudzik

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Shuang Liu
Final project course: CSE3000 Research Project
Thesis committee: Chirag Raman, Bernd Dudzik, Alan Hanjalic

Abstract

Images possess the ability to convey a wide range of emotions, and extracting affective information from images is crucial for affect prediction systems. This process can be achieved through the application of machine learning algorithms. Categorical Emotion States (CES) and Dimensional Emotion Space (DES) are two typical models used for representing emotions. Moreover, the development of a mapping schema between these representations can contribute to benefit the research in AI and psychology. Consequently, this study focuses on investigating the feasibility of translating emotions from DES to CES. To accomplish this goal, relevant databases are identified and combined as the training data, and machine learning models, namely Naive Bayes, K-nearest Neighbors, and Decision Tree are employed to perform the classification task. The results indicate the superior performance of the K-nearest Neighbors classifier, exhibiting higher mean accuracy (60%) and low standard deviation (0.004) among all implemented classifiers. Overall, the translation from DES to CES offers several advantages, including a simplified and interpretable representation of emotions, as well as the provision of a common language for discussing and expressing emotions.

1 Introduction

Nowadays, with the widespread popularity of social networks, people are increasingly expressing their opinions through multimedia data such as text, images, and video. Images as a medium is very powerful and can convey rich affect, which refers to the underlying experience of feeling, emotion, attachment, or mood[9]. Image analysis is the extraction of meaningful information from images[17]. If the information is about affect, then image content analysis can be seen as an affect prediction, which can be done by the use of machine learning. As what people feel may directly determine their decision making, affective image content analysis (AICA) is of great importance, which can enable wide applications[5] such as personalized recommendations, mental health monitoring, and user experience analysis.

The representation of affective states is a crucial component of affect prediction systems. It determines how the system understands and responds to affect. Psychologists mainly employ two typical models to represent emotions: categorical emotion states (CES), and dimensional emotion space (DES) [19]. CES is a framework that categorizes emotions into discrete and distinct states. It proposes a set of basic or primary emotions, such as happiness, sadness, anger, fear, disgust, and surprise, which are considered universal across cultures. On the other hand, DES is an alternative approach that represents emotions in a continuous dimensional framework. It suggests that emotions can be understood based on two or more underlying dimensions. The most commonly used dimensions are valence (ranging from positive to negative), arousal (ranging from low to high intensity), and dominance (ranging from controlled to in control). In practice, dominance is frequently omitted from the description of emotion space because it was shown to be the least informative measure of the elicited affect[1].

Only few studies deal with the translation between different emotion schemes. Moreover, most of these activities are only concerned with discrete representations of the Ekman model (see Table 5)[4]. However, having a robust, high-accuracy mapping schema for both representations may help further unify both lines of research (in AI, not limited to NLP, as well as in psychology)[18] and applications may benefit from being able to choose a specific scheme. Let us consider a scenario where a company wants to develop a sentiment analysis system for customer reviews. In this case, multiple scenarios are needed to adequately represent the various affects expressed in the reviews. Here is an example:

- Scenario 1: Categorical Emotion States

  Review: "I absolutely love this product! It was amazing."

  In this scenario, the sentiment analysis system needs to categorize the user's emotion as "joy" or "happiness" based on the positive nature of the statement.

- Scenario 2: Dimensional Emotion Space

  Review: "The product itself is great, but the delivery took forever. It was frustrating to wait so long to receive it."

  In this scenario, it is better for the system to represent the user's emotion in a dimensional emotion space, such as valence and arousal. The system would assign a negative valence (frustration) and moderate arousal to this particular message.

By considering both categorical emotion states and dimensional emotion space, the system can provide a more nuanced understanding of the user's affects. It allows for a broader range of emotional representation, capturing not only the basic emotions but also the intensity and valence of those emotions. Therefore, the feasibility of exploring automatic translation between affect representation schemes is important and useful for applications needed to have access to more than one representation scheme.

The exploration of translation between different schemes can be broken down into some research questions: What supervised machine learning model performs best in the translation and what are the relevant properties of datasets used for training that influence the capacity of translation models to generalize to unseen datasets? These questions will be answered in the following sections.

Section 2 provides a review of related work, highlighting prior research and relevant theoretical frameworks. Section 3 presents the methodology employed in the study and Section 4 offers a detailed account of the experimental process, presenting the obtained results and corresponding findings. Building upon the results, Section 5 engages in comprehensive discussions, encompassing the interpretation of the findings, identification of limitations inherent in the research, and proposing avenues for future investigation. Furthermore, Sec-

tion 6 addresses the ethical and responsible research considerations. Finally, Section 7 concludes the study, summarizing the key findings, discussing their implications, and offering final remarks on the significance of the research and potential directions for further inquiry.

2 Related Work

This section presents an overview of the existing research relevant to the current study. It begins by discussing various emotional models proposed by researchers in the field (Section 2.1). Subsequently, section 2.2 summarizes databases with multiple representation schemes. Finally, the mapping relationship between different emotional models is presented (Section 2.3).

2.1 Emotional Models

To facilitate the translation between various schemes, it is crucial to have a comprehensive understanding of the existing schemes employed in AICA. Some models that are used widely in AICA are listed in Table 5 [19].

2.2 Affective Databases

Subsequent to examining the representative models utilized in the field of AICA, it is pertinent to delve into the realm of relevant affective databases that incorporate multiple representation schemes. By thoroughly examining the characteristics of affective databases, researchers can make informed decisions regarding data selection and fusion strategies. Merging databases with the same representation scheme augments the volume of training data available for machine learning algorithms. The increased training set size contributes to improved model performance and generalization.

While exploring existing literature, it becomes evident that the majority of available databases predominantly adhere to a single representation scheme. This discrepancy underscores the scarcity of databases that encompass diverse representation schemes for emotions. Table 6 presents a compilation of datasets specifically associated with the investigation at hand.

The Geneva Affective Picture Database (**GAPED**) is a comprehensive collection of images designed to evoke and measure emotional responses. It consists of over 730 high-quality photographs in total, including 520 negative images, 121 positive images, and 89 neutral images. In addition, the database also includes ratings of valence and arousal for each image, collected from a large sample of participants.

The **Emotion6** consists of 1980 images which are classified into six basic emotions. For each image, the ground truth of VA scores and emotion distribution for evoked emotion are provided as well.

The Image-Emotion-Social-Net (**IESN**) database, with 1,012,901 images, contains a vast array of multimedia content, including images, along with associated emotion labels and social network information. Each image is annotated with Mikels' emotion category which indicates the corresponding emotional state and the average VAD values.

The International Affective Picture System (**IAPS**) is a standardized collection of emotionally evocative images developed for scientific research in the field of psychology. It encompasses a wide range of images. The original database only uses VAD as the representation scheme. However, there is another research[11] that extends the original annotations with the Ekman representation scheme. It should be noted that the original IAPS contains 716 standardized color photographs[3] but there are 703 slides collected because of oversight and slide duplication in [11].

2.3 Models Transformation

There is also research on the relationship between categorical and dimensional models. Figure 3 serves as a graphical representation to illustrate the placement of Ekman's basic emotions within the Valence-Arousal-Dominance (VAD) space[4]. The numerical ratings displayed in the figure are derived from the work of Russell and Mehrabian[15], who introduced the three-factor theory proposing that emotions can be characterized by valence, arousal, and dominance. To substantiate their claims regarding the three-factor theory, the authors conducted a series of experiments. These experiments aimed to provide average and variance values for various emotional categories, specifically valence, arousal, and dominance. The results of these experiments support the argument that these three factors represent crucial dimensions that capture the complexity and diversity of emotional phenomena.

The prior study conducted by Russell and Mehrabian provides a possibility for exploring the conversion of different representation schemes. However, while the study significantly contributed to the understanding of emotions by highlighting the significance of valence, arousal, and dominance as fundamental factors, it primarily focused on elucidating the importance of these dimensions in representing emotions rather than emphasizing the translation between different representation schemes. This research, however, aims to address this gap by leveraging the power of machine learning algorithms to facilitate the translation between distinct representation schemes.

3 Methodology

To address the research questions outlined in the introduction part, four primary tasks have been identified. The first task is to find databases possessing multiple representation schemes. The second task entails the merging of these selected datasets into a unified joint database, which will serve as the foundation for the subsequent supervised machine learning analyses. This amalgamation of datasets aims to leverage the combined information to explore the potential of machine learning models in effectively translating between different representation schemes. In this specific study, the translation process will be directed from the Dimensional Emotion Space (DES) representation to the Categorical Emotion Space (CES) representation, effectively framing the problem as a classification task. Furthermore, an important aspect of the research entails conducting a comprehensive analysis of what factors could potentially influence the performance of the employed machine learning models.

To execute the aforementioned tasks, the operations on the databases and the implementation of various supervised ma-

chine learning models will be carried out using Jupyter Notebook.

4 Experiment setup and Results

In this section, a comprehensive description of the experimental methodology is presented, outlining the integration of databases (section 4.1), the utilization of supervised machine learning models (section 4.2), along with their respective performance evaluations (section 4.3).

4.1 Databases Combination

Section 2.2 summarizes several databases containing multiple representation schemes that can be used in this research. However, due to limitations in time and database accessibility, two specific databases, namely **Emotion6** and **IAPS**, have been selected for the research investigation.

The Emotion6 dataset provides detailed information for each image, including valence and arousal values, as well as a probability distribution of seven emotion categories, including Ekman's basic emotions and neutral. The assignment of a dominant emotion is determined by selecting the emotion category with the highest probability value. Each record in the Emotion6 dataset after processing consists of three primary variables: the emotion category, valence, and arousal values.

On the other hand, the IAPS dataset has different mean values for the six basic emotion categories using the Ekman model for each image. Each image is labeled with the emotion category that has the highest mean value. Additionally, the IAPS dataset includes valence and arousal values for each image.

The valence and arousal scores for both databases are adopted by the Self-Assessment Manikin (SAM) 9-point scale[2]. The SAM scale represents valence on a continuum ranging from extremely negative (1) to neutral (5) to extremely positive (9). Arousal scores range from extremely calm (1) to neutral (5) to extremely excited (9). Because the measurement scale of these two databases is the same, there is no need to convert the range.

Consequently, both the Emotion6 and IAPS datasets after processing share a common structure in which each record is characterized by three variables: the emotion category, valence, and arousal values. It should be noted that before the combining operation, any record that includes NaN (not a number) values has been detected and dropped.

Figure 1 illustrates the distribution of various emotion categories within the valence and arousal space. The plot reveals that different emotions occupy distinct areas, indicating unique patterns of valence and arousal for each emotion. However, it is noteworthy that there exists a degree of overlap among the emotion categories. This observation highlights the importance of considering both VA dimensions and categorical emotion labels when studying and categorizing emotions.

Table 1 gives a more specific summary of statistics of each emotion the database contains. A closer examination of the table reveals that there are minimal differences in the mean values between certain emotions, namely "anger" and "disgust", as well as "neutral" and "surprise", across both the valence and arousal dimensions. The similarity in mean values

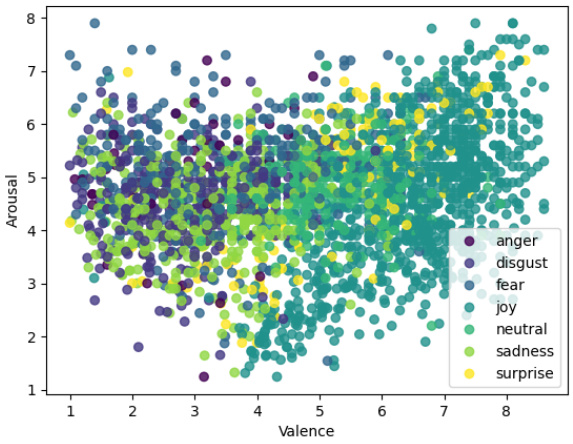

Figure 1: The distribution of VA scores of Emotion6 and IAPS

implies that these emotions share common underlying affective states, making them potentially difficult to differentiate using only the valence and arousal dimensions. This observation highlights the intricacies involved in accurately capturing and characterizing specific emotional states, particularly when relying solely on two dimensions of affective experience.

Table 1: Summary statistics of seven emotions (Ekman + neutral)

	Valence		Arousal	
Emotions	M	SD	M	SD
anger	2.953	1.055	5.032	1.089
disgust	2.864	0.957	4.536	0.840
fear	3.478	1.248	5.083	0.944
joy	6.421	1.169	4.618	1.341
neutral	5.145	0.950	4.886	0.669
sadness	3.275	1.108	4.378	0.848
surprise	5.230	1.398	4.941	1.210

Table 2: Number of each emotion category of Emotion6 and IAPS

Emotions	**Emotion6**	**IAPS**	**Total**
anger	31	22	53
disgust	245	68	313
fear	329	68	397
joy	638	373	1,011
neutral	325	0	325
sadness	308	107	415
surprise	104	64	168

The data presented in the Table 2 clearly indicates an imbalance in the sample distribution across different emotion categories. Notably, the "joy" emotion category has a significantly larger number of samples (1011) compared to the

"angry" emotion category, which only comprises 53 samples. This class imbalance can pose challenges such as accurately predicting the "angry" emotion for the models and may lead to biased or inaccurate results.

4.2 Models Implementation

In the field of machine learning, numerous classification algorithms have been developed and widely utilized for various tasks. In this study, three specific algorithms, namely Naive Bayes (NB), K-nearest Neighbors (KNN), and Decision Tree (DT) have been chosen for implementation and evaluation.

To ensure the identification of optimal hyperparameters for each model, the study employs the GridSearchCV function from the widely-used scikit-learn (sklearn) library. This function offers a systematic exploration of a predefined grid of hyperparameters for a given model, enabling the identification of the hyperparameter combination that results in the best performance. The types and ranges of hyperparameters of each model can be seen in Table 7.

Furthermore, in order to mitigate the issue of information leakage and overfitting, the study incorporates a *nested cross-validation* strategy in conjunction with GridSearchCV. This approach addresses the concern that model selection without nested cross-validation may inadvertently incorporate information from the evaluation data during the hyperparameter tuning process, leading to over-optimistic performance estimates. By employing nested cross-validation, the model's hyperparameters are tuned on the inner cross-validation loop, while the outer loop assesses the model's performance on an independent validation set. This nested approach helps to mitigate the risk of overfitting and provides a more reliable and unbiased evaluation of the model's performance.

In order to evaluate the models in a more robust and reliable manner, *5-fold Cross-Validation* is employed as both the cross-validation techniques for the inner and outer loops. This technique involves dividing the training set into five equally sized subsets, where each subset acts as a validation set once while the remaining subsets are used for training. This process is repeated five times, ensuring that each subset is utilized as the validation set exactly once. By averaging the performance metrics obtained from these iterations, a more accurate assessment of the model's capabilities can be obtained.

To mitigate the issue of imbalanced datasets, a technique known as Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE is a data augmentation approach that focuses on the minority class, generating synthetic samples to balance the class distribution. By augmenting the minority class, SMOTE helps address the bias caused by imbalanced data and improves the overall performance of classification models. After applying SMOTE, the number of each emotional state became 1,011.

In this study, the evaluation metric used to assess the performance of the translation models is accuracy, which measures the proportion of correctly classified instances. The accuracy values range between 0 and 1, with a higher value indicating better classification performance.

As a baseline for comparison, the majority classifier is employed in this study. The majority classifier, implemented using the "most_frequent" strategy in the sklearn library, always predicts the majority class. By comparing the performance of the translation models against the majority classifier, a valuable benchmark is established to determine whether a given model's performance surpasses random chance.

4.3 Results

The experimental design involved conducting a total of 30 trials to evaluate the performance of each model. Table 3 presents the mean and standard deviation of the accuracy values (Acc) obtained from these trials, comparing them with the baseline performance of the majority classifier (MC). In addition, a t-test was conducted to examine the statistical significance of the performance differences between the machine learning models, specifically in comparison to the majority classifier. This test aimed to determine whether the observed variations in performance scores were statistically significant. A graphical overview of performance can be seen in Figure 2.

In addition, multiple t-tests among classifiers were also performed and the result is shown in Table 4. It should be noted that all p-values reported in Table 3 and Table 4 have undergone Bonferroni correction. However, all p-values before and after correction are very small so the values in the tables are all written as "<0.01***".

Table 3: Summary performance of each machine learning model, compared with the majority classifier

	Acc		**vs.Majority**		
Models	M	SD	Δ M(Acc)	t	p
DT	0.586	0.005	+0.192	431	<.001***
KNN	0.610	0.004	+0.191	617	<.001***
NB	0.386	0.001	+0.201	581	<.001***
MC	0.131	0.002	0	0	1

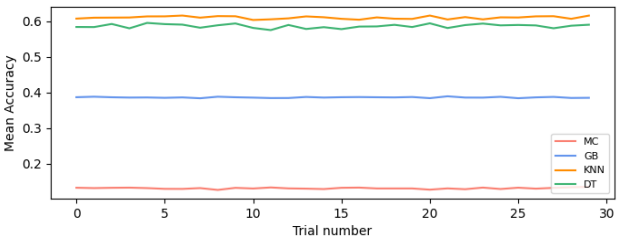

Figure 2: Mean Accuracy of each model across trials

Table 4: T-tests and p-values among classifiers

Models	**t**	**p**
KNN-DT	20	<.001***
DT-NB	196	<.001***
KNN-NB	306	<.001***

5 Discussions

Section 5.1 focuses on the implications and interpretations of the results. Section 5.2 aims to discuss the constraints, weaknesses, and areas for improvement in the study.

5.1 Empirical Findings

Table 3 provides an overview of the classification performance of the Decision Tree, K-nearest Neighbors, and Naive Bayes classifiers, as well as the majority classifier. The results of statistical analysis, as indicated by the p-values, suggest that all classifiers exhibit statistically significant performance differences compared to the majority classifier. Consequently, it can be inferred that the classifiers possess distinct predictive capabilities and are valuable for classification purposes.

In terms of performance comparability, Table 3 and Table 4 show that the K-nearest Neighbors classifier emerges as the most favorable option. This conclusion is drawn based on its superior performance metrics, specifically its highest mean value of accuracy and the small standard deviation. These findings suggest that the K-nearest Neighbors classifier demonstrates both high average accuracy and low variability, indicating its potential as an effective choice for accurate predictions.

The result that the Decision Tree classifier and K-nearest Neighbor perform better than Naive Bayes may indicate that non-linear classifiers (Decision Tree, K-nearest Neighbor) are more suitable for this affective prediction task than linear classifiers (Naive Bayes, Support Vector Machines, Logistic Regression). Linear Classification refers to categorizing a set of data points into a discrete class based on a linear combination of its explanatory variables. It is possible to classify data with a straight line or a hyperplane. On the other hand, Non-Linear Classification refers to separating those instances that are not linearly separable and it is not easy to classify data with a straight line[8]. This can also be seen from Figure 1 that it is almost impossible to classify points using lines.

It is pertinent to acknowledge that the inclusion of the Support Vector Machine (SVM) model was initially considered in this study. However, its implementation presented practical challenges, primarily about the considerable training time required for nested cross-validation. The SVM algorithm demonstrated substantially longer training durations compared to the other classifiers, thereby rendering it unfeasible to record its results within the specified time constraints. This predicament highlights the inherent disadvantages associated with the utilization of cross-validation techniques, including the increased training time and expensive computation cost.

Given the limitations imposed by time constraints, this research primarily focuses on presenting the results and analyses about the Decision Tree, K-nearest Neighbors, and Naive Bayes classifiers. In summary, the results from Table 3 and Figure 2 underscore the significant roles played by all the classifiers in classification tasks, as evidenced by their statistically significant performance differences compared to the majority classifier. Among all the classifiers, the K-nearest Neighbors classifier stands out as the optimal choice based on its higher mean accuracy and low standard deviation. In addition, future research endeavors could explore the inclusion of SVM and investigate its performance alongside the other classifiers in the context of this study. By allocating sufficient computational resources and time, a comprehensive evaluation of other classifiers' predictive capabilities could be conducted, enabling a more comprehensive comparison among different classifiers.

5.2 Limitations and Future Work

Database Shortage: Firstly, the utilization of only two databases may introduce a certain degree of bias and may not fully capture the variability and complexity of affective expressions across different populations or contexts. A broader sampling of databases, encompassing diverse demographic groups and cultural backgrounds, would contribute to a more comprehensive understanding of affective processing and facilitate the development of more generalized machine learning models.

One of the databases collected images from Flickr, a popular online image-sharing platform. All of the images are put on Amazon Mechanical Turk (AMT) to be labeled with emotional keywords and annotated with VA scores. As the images were sourced from a public platform, there may be variations in image quality, content, and relevance to the study's focus. This lack of control over the image selection process may introduce noise and potential biases into the dataset, affecting the generalizability and accuracy of the results.

The participants involved in the process of annotating images of the other dataset are 1,302 Midwestern university students who were 18 years old or older and included both males and females[11]. Therefore, another potential source of bias arises from the annotation process conducted by university students. The involvement of students from a specific region (Midwestern university students) may introduce regional and cultural biases in the annotations. Additionally, the age range and gender distribution of the participants may influence the interpretation and labeling of emotional keywords, potentially limiting the generalizability of the annotations to broader age groups and gender populations.

The joint database used for training machine learning algorithms is a combination of these two datasets. However, the distinct approaches employed for annotating the images in each dataset introduce certain constraints, weaknesses, and potential biases that should be carefully considered.

Additionally, exploring alternative annotation methods, such as involving domain experts or employing more advanced computational techniques (e.g., automated emotion recognition algorithms), could enhance the accuracy and reliability of the emotional labeling and VA scoring process. These approaches could minimize subjective biases and provide more objective and consistent annotations.

In summary, while the databases used in this study and the methodology of combining datasets have certain limitations and potential biases, there are several areas for improvement. By implementing stricter image selection criteria, diversifying the pool of annotators, conducting studies with larger and more diverse participant samples, and exploring alternative annotation methods, future research can enhance the quality,

reliability, and generalizability of affective datasets, leading to more robust findings in the field of affective computing.

Limited Dimensions: In this study, the classification of emotions is based on two dimensions, namely arousal, and valence. While these dimensions have proven to be valuable in capturing the affective states associated with different emotions, it is worth noting that emotions are complex and multifaceted phenomena that may involve additional dimensions such as dominance. By expanding the dimensions used for emotion classification and prediction, researchers can advance our understanding of emotions, and improve the accuracy of emotion prediction models.

Machine Learning Models: The time constraints imposed on the study may have restricted the scope and depth of the analysis. In-depth feature selection, optimization of hyperparameters, and exploration of various machine learning algorithms could provide further insights into the performance and effectiveness of the models. Additionally, the limited training time may have hindered the exploration of more advanced techniques such as deep learning, which could potentially yield improved accuracy and robustness in affective prediction tasks.

6 Responsible Research

In this study, it is important to note that datasets utilized in the present study have been collected and organized in previous research projects. No new data from external sources were incorporated into this research. This avoids duplicating data collection efforts. Moreover, this approach promotes transparency because the datasets used are publicly available or obtained with the necessary permissions, ensuring compliance with legal and ethical guidelines.

Furthermore, this study maintains transparency and objectivity. The findings are presented in a manner that accurately reflects the data and aligns with the established scientific process. No modifications or alterations have been made to manipulate the results or distort the meaning of the findings. In addition, the code and datasets utilized in this study have been made publicly available. They have been uploaded to a dedicated public repository on GitHub[1], ensuring accessibility and facilitating scrutiny by other researchers.

7 Conclusions

In conclusion, this research study aimed to investigate the feasibility of translating between different representation schemes, with a specific focus on the translation from dimensional emotion space (DES) to categorical emotion states (CES). To achieve this objective, two relevant affective databases, namely **Emotion6** and **IAPS**, were identified and combined to create a joint database for training machine learning models. Three commonly used models, namely Naive Bayes, K-nearest Neighbors, and Decision Tree were implemented and evaluated. The findings of this study demonstrate the superior performance of the K-nearest Neighbors classifier because of its high mean accuracy (60%) and low standard deviation (0.004) among all implemented classifiers. The translation from DES to CES provides a simplified and interpretable representation of emotions. By mapping the multidimensional space of emotions onto a categorical framework, it becomes easier to understand and communicate emotional experiences. Furthermore, CES offers a common language for discussing and expressing emotions, making it easier for people to connect and relate to each other's emotional experiences.

However, it is important to acknowledge the limitations of this study. The shortage of available databases, limited dimensions used for classification, and time constraints have impacted the generalizability and thoroughness of the findings. Addressing these limitations requires future research to focus on expanding the availability of affective databases that encompass multiple representation schemes and cover a wide range of emotional states. Furthermore, allocating adequate time and resources to explore advanced modeling techniques and optimization processes would contribute to the development of more accurate and reliable translation models.

[1] https://github.com/liushuang-118/ResearchProject.git

References

[1] Iris Bakker, Theo Van der Voordt, Jan Boon, and Peter Vink. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33:405–421, 10 2014.

[2] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[3] Margaret M. Bradley and Peter J. Lang. *International Affective Picture System*, pages 1–4. Springer International Publishing, Cham, 2017.

[4] Sven Buechel and Udo Hahn. Emotion analysis as a regression problem — dimensional models and their implications on emotion representation and metrical evaluation. 08 2016.

[5] Tao Chen, Felix Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and S. Chang. Object-based visual sentiment concept analysis and application. 11 2014.

[6] Elise Dan-Glauser and Klaus Scherer. The geneva affective picture database (gaped): A new 730-picture database focusing on valence and normative significance. *Behavior research methods*, 43:468–77, 03 2011.

[7] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6:169–200, 1992.

[8] Shabeg Singh Gill. Linear vs. non-linear classification. Accessed: 06 2023.

[9] M. A. Hogg and D. Abrams. *Psychology*. School of Psychology Publications. Pearson Education, 2004.

[10] Joonwhoan Lee and Eun-Jong Park. Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia*, 13:1031–1039, 10 2011.

[11] Terry M. Libkuman, Hájíme Otani, Rosalie Kern, Steven G. Viger, and Nicole Novak. Multidimensional normative ratings for the international affective picture

system. *Behavior Research Methods*, 39(2):326–334, May 2007.

[12] Joseph Mikels, Barbara Fredrickson, Gregory Samanez-Larkin, Casey Lindberg, Sam Maglio, and Patricia Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37:626–30, 12 2005.

[13] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. pages 860–868, 06 2015.

[14] R. Plutchik. *Emotion, a Psychoevolutionary Synthesis*. Harper & Row, 1980.

[15] James Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 09 1977.

[16] Harold Schlosberg. Three dimensions of emotion. *Psychological review*, 61 2:81–8, 1954.

[17] C.J. Solomon and T.P. Breckon. *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2010.

[18] Ryan Stevenson, Joseph Mikels, and Thomas James. Characterization of the affective norms for english words by discrete emotional categories. *Behavior research methods*, 39:1020–4, 12 2007.

[19] Sicheng Zhao, Guiguang Ding, Tat-Seng Chua, Björn Schuller, and Kurt Keutzer. Affective image content analysis: A comprehensive survey. pages 5534–5541, 07 2018.

[20] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Xie Wenlong, Xiaolei Jiang, and Tat-Seng Chua. Predicting personalized emotion perceptions of social images. pages 1385–1394, 10 2016.

A Models & Databases & Transformation

Table 5: Representative emotion models, reproduced from [19]

Models	References	Type
Ekman	[7]	CES
Mikels	[12]	CES
Plutchik	[14]	CES
Sentiment	-	CES
VA(D)	[16]	DES
ATW	[10]	DES

B Types & Ranges of Hyperparameters

Table 6: Databases with multiple representation schemes, reproduced from [19]

Datasets	References	Images	Emotion models
GAPED	[6]	730	Sentiment, VA
Emotion6	[13]	1,980	Ekman+neutral, VA
IESN	[20]	1,012,901	Mikels, VAD
IAPS	[3]	1,182	VAD

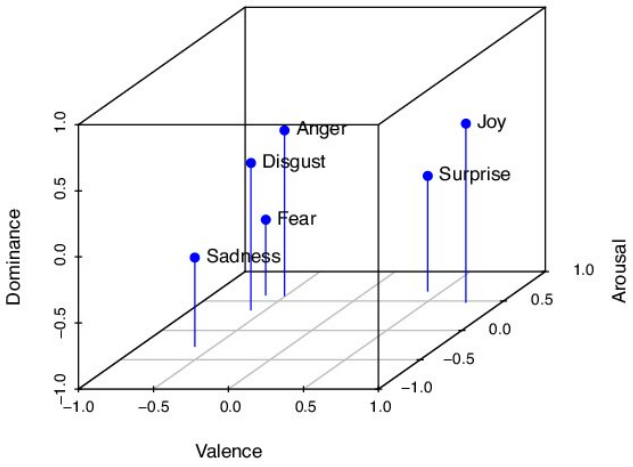

Figure 3: Positions of Ekman's basic emotions in VAD space [4]

Table 7: Hyperparameter types and ranges for each model

Models	**Types**	**Ranges**
DT	max_depth	2, 4, 8, 16, 32, 64
	min_samples_leaf	2, 4, 8, 16
KNN	n_neighbors	2, 5, 10, 15
	algorithm	ball_tree, kd_tree, brute, auto
	metric	minkowski, euclidean, manhattan, chebyshev
NB	var_smoothing	np.logspace(0,-9, num=100)